KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

PRINCIPAL ANGLES IN SYSTEM THEORY, INFORMATION THEORY AND SIGNAL PROCESSING

Promotor:
Prof. dr. ir. B. De Moor

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Katrien DE COCK

Mei 2002

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

PRINCIPAL ANGLES IN SYSTEM THEORY, INFORMATION THEORY AND SIGNAL PROCESSING

Jury:
Prof. dr. ir. J. Berlamont, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. R. Boel (Universiteit Gent)
Prof. dr. A. Bultheel
Prof. dr. ir. M. Gevers (UC Louvain)
Prof. dr. ir. M. Moonen
Prof. dr. ir. G. Picci (Università di Padova)
Prof. dr. ir. J. Vandewalle
Prof. dr. ir. J. Willems

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Katrien DE COCK

U.D.C. 681.3*I54

Mei 2002

D/2002/7515/28

ISBN 90-5682-351-5

Voor mijn grootouders

Voorwoord

Vele mensen hebben er mee voor gezorgd dat mijn proefschrift er is gekomen. Eerst en vooral dank ik Professor Bart De Moor, die me overtuigd heeft om bij SISTA te beginnen werken en gezorgd heeft voor een boeiend onderwerp. Zijn ideeën en belezenheid vormen een belangrijke factor in de realisatie van dit werk. Met zijn aanstekelijk enthousiasme heeft hij me bovendien in kritieke momenten aangemoedigd om door te zetten. Zijn motiverende opmerkingen en stimulerende gesprekken heb ik enorm geapprecieerd. Ik ben hem ook dankbaar omdat hij me van bij het prille begin in contact heeft gebracht met de internationale wetenschappelijke wereld.

De leden van het leescomité, Professor Bultheel, Professor Moonen en Professor Vandewalle wil ik bedanken voor het zorgvuldig nalezen van de tekst. De rake opmerkingen en vragen van Professor Moonen hebben me weer aan het rekenen gezet. Professor Willems heeft een voorlopige versie van dit werk grondig nagelezen. Ik dank hem voor zijn vele waardevolle suggesties.

Tijdens mijn beginjaren bij SISTA kon ik altijd terecht bij Bart De Schutter. Zijn hulp en begeleiding heb ik erg op prijs gesteld. Ook dank ik Professor Boel voor zijn interesse en opmerkingen tijdens de jaarlijkse bijeenkomsten.

Mijn collega's in binnen- en buitenland zorgden voor de vriendschappelijke sfeer die het wetenschappelijk onderzoek zo aangenaam maakt. Vooral de vele Alma-uren en de jaarlijkse SISTA-weekends zullen mooie herinneringen achterlaten. Internationale conferenties waren een leerrijke, maar ook prettige ervaring, onder andere dankzij het gezelschap van Wouter en Tony.
I would also like to thank the members of the European TMR network on system identification (ERNSI) and especially Professor van Schuppen, the network coordinator, and Professor Gevers, the coordinator of the Belgian team. I enjoyed the yearly workshops very much. Moreover, it was always very pleasant to meet ERNSI-people at other conferences.

I would like to express my gratitude to Professor Picci for inviting me to LADSEB in Padova. The kind care of Nicole, Lorenzo, Per and the other colleagues made me feel at home immediately. I also want to thank Professor Picci for his immediate willingness to be a member of the jury. I am grateful to Alessandro Chiuso for his interest and for drawing my attention to a discrepancy in my work.

Ik ben het I.W.T. erkentelijk voor de financiële ondersteuning. Ook dank ik alle SISTA-leden die op een of andere manier meewerken aan projecten. Zij zorgen ervoor dat we niets te kort komen. Ida ben ik dankbaar voor alle hulp tijdens de voorbije jaren.

De altvioollessen van Marc Claes vormden een ideale afwisseling en uitlaatklep. Ik wil Marc bedanken voor zijn begrip. Als er iets geleden heeft onder de afwerking van dit proefschrift, is het wel mijn altvioolspel.

Petrus en Lief, An, Peter, Maarten en Helena, en ook Hanna en Helena, dank ik uit heel mijn hart voor de warme thuis. En natuurlijk vergeet ik Pieter-Jan niet, die meer heeft bijgedragen dan hij beseft. Bedankt, schat van mij, voor al je liefdevolle zorgen.

Abstract

We establish new connections between system theory, information theory and signal processing by means of the geometric notion of principal angles between linear subspaces.
We compute the principal angles between the different combinations of past and future input and output spaces of a linear stochastic model. By using the properties of linear stochastic models, we can express these angles in terms of the model parameters. The asymptotic values are equal to the canonical correlations of the respective processes. Hence, we have expressions for the canonical correlations in terms of the model parameters, from which we derive relations between the different canonical correlations.
Since the canonical correlations are related to the mutual information, we obtain formulas for the mutual information of the different pairs of processes.
We show how a recently defined cepstral norm for single input single output linear stochastic models can be interpreted in terms of principal angles between subspaces that are derived from the model parameters. Via the geometric insights that we have of linear stochastic models, we show that the cepstral norm is also directly related to the information theoretic notion of mutual information. The squared cepstral norm of a single input single output linear stochastic model is proportional to the the mutual information of the past and the future of the output process obtained by sending white Gaussian noise through the model.

Korte inhoud

In dit proefschrift worden nieuwe verbanden gelegd tussen de volgende disciplines: systeemtheorie, informatietheorie en signaalverwerking. De meetkundige techniek die we daarvoor gebruiken is het berekenen van canonieke hoeken tussen lineaire deelruimten.
We berekenen de canonieke hoeken tussen de verschillende combinaties van verleden en toekomstige ingangs- en uitgangsruimten van een lineair stochastisch model. Via de parametrische beschrijving van het model, kunnen deze hoeken uitgedrukt worden in functie van de modelparameters. De limietwaarden van de parametrische uitdrukkingen van de cosinussen zijn gelijk aan de canonieke correlaties van de overeenkomstige processen. Via deze uitdrukkingen voor de canonieke correlaties in functie van de modelparameters bewijzen we verbanden tussen de verschillende correlaties.
De wederzijdse informatie van twee Gaussische stochastische processen kan beschreven worden in termen van hun canonieke correlaties. Omdat we uitdrukkingen hebben voor de canonieke correlaties van de in- en uitgangsprocessen van een lineair model, kennen we dus ook de wederzijdse informatie van die processen in functie van de modelparameters.
We tonen aan dat een recent gedefinieerde afstandsmaat voor modellen van scalaire stochastische processen met een rationeel spectrum kan uitgedrukt worden in functie van de hoeken tussen deelruimten die afgeleid worden uit de modelparameters. Dezelfde afstandsmaat kan ook geïnterpreteerd worden als de wederzijdse informatie van twee processen. Deze interpretatie vinden we door de canonieke hoeken tussen de modeldeelruimten te relateren aan de canonieke correlaties van de ingangs- en uitgangsprocessen en de meetkundige verbanden tussen de verschillende hoeken tussen de verleden en toekomstige ingangs- en uitgangsruimten toe te passen.
Onze resultaten werden reeds toegepast op problemen in het domein van de computervisie. We geven aan dat ze mogelijk ook nuttig kunnen zijn bij foutdetectie, het clusteren van signalen en biomedische signaalverwerking.

Glossary

Symbols

The following table gives an alphabetical enumeration of the symbols used in this text. We give a short explanation of each symbol or mathematical representation and refer to the page where it is first encountered or defined.

symbol	**explanation**	**page**
$*$	convolution	190
$\#(S)$	cardinal number, i.e., number of elements, of finite set S	5
$0_{m,n}$	$m \times n$ matrix with zero elements	131
$\|\|a\|\|$	Euclidean norm (or l_2 norm) of vector a	38
$a \perp b$	vector a is orthogonal to vector b	39
A^T	transpose of matrix A	27
A^H	complex conjugate transpose of matrix A	28
$A^\dagger$	Moore–Penrose inverse of matrix A	33
$A(i,j)$	element of matrix A on row i and column j	128
$A(\delta)$	the area enclosed by the curve $\delta(\theta)$	228
$[a \sphericalangle b]$	angle between vectors a and b	38
$[A \sphericalangle B]$	ordered set (in non-descending order) of the principal angles between $\mathrm{row}(A)$ and $\mathrm{row}(B)$	40
(A,K,C)	state space matrices of linear stochastic model in forward innovation form	69
$(\bar{A},\bar{K},\bar{C})$	state space matrices of linear stochastic model in backward innovation form ($\bar{A} = A^T,\ \bar{C} = G^T$)	73
(A_z,K_z,C_z)	state space matrices of inverse model $((A_z,K_z,C_z) = (A-KC,K,-C))$	80
$\arg(S(z))$	phase of $S(z)$	201
B/A	the matrix $B\Pi_{\mathrm{row}(A)}$	37
$\binom{n}{m}$	binomial coefficient $\left(\binom{n}{m} = \frac{n!}{m!(n-m)!}\right)$	87
$\bar{c}$	complex conjugate of complex number c	129

symbol	explanation	page
$c_y(k)$	power cepstrum of $\{y(k)\}$	195
$\bar{c}_h(k)$	complex cepstrum of $h(k)$	200
$\mathcal{C}_i$	$n \times mi$ controllability matrix of (A, K)	74
$\mathcal{C}_\infty(H(z))$	$n \times \infty$ controllability matrix of SISO model with transfer function $H(z)$	129
$\mathcal{C}_j^*$	controllability matrix of SISO model in special state space basis	86
$\mathrm{cc}(X, Y)$	canonical correlations of random variables X and Y	56
$\mathrm{cc}(x, y)$	canonical correlations of $\{x(k)\}$ and $\{y(k)\}$	60
$C_k(A)$	k-th compound matrix of A	261
$C_{xy}(\tau)$	cross-covariance function of $\{x(k)\}$ and $\{y(k)\}$ $(C_{xy}(\tau) = E\left\{X(k+\tau)Y(k)^T\right\})$	58
$\mathrm{codom}(A)$	codomain of matrix A	32
$\cos[A \sphericalangle B]$	cosines of the principal angles between $\mathrm{row}(A)$ and $\mathrm{row}(B)$	42
$d(\log H, \log G)$	cepstral distance between SISO models with transfer function $H(z)$ and $G(z)$	216
$\delta(\tau)$	Kronecker delta function (identically zero except $\delta(0) = 1$)	64
$\delta(\theta)$	polar plot of $\log H(e^{i\theta})$ where θ is running from 0 to 2π	228
Δ_i	$n \times mi$ reversed controllability matrix of (A, K)	77
$\bar{\Delta}_i$	$n \times mi$ controllability matrix of backward innovation model $(\bar{A}, \bar{K})$	79
$\mathcal{D}_i$	$= \mathcal{C}_i Q_{u_i} \mathcal{C}_i^T = \Delta_i Q_{u_i} \Delta_i^T$	106
$\det A$	determinant of matrix A	157
$\mathrm{diag}(\sigma_i)$	diagonal matrix with σ_i on diagonal	28
$\mathrm{diag}(A, B, C)$	block diagonal matrix with matrices A, B, C on diagonal	75
$\dim(S)$	dimension of subspace S	27
$\mathrm{dom}(A)$	domain of matrix A	32
$E\{\cdot\}$	expected value operator	52
$\mathcal{F}(\cdot)$	Fourier transform	191
$\Phi_y(e^{i\theta})$	power spectral density of $\{y(k)\}$	81
G	covariance matrix of $X(k+1)$ and $Y(k)$	71
$g(A)$	Grassmann vector of A	263
$\gamma(\theta)$	polar plot of $H(e^{i\theta})$ where θ is running from 0 to 2π	229
Γ_i	$mi \times n$ observability matrix of (A, C)	74
$\bar{\Gamma}_i$	reversed observability matrix of backward innovation model $(\bar{A}, \bar{C})$	79

symbol	explanation	page
Γ_{z_i}	$mi \times n$ observability matrix of the inverse model (A_z, K_z, C_z)	80
$\mathcal{G}_{z_i}$	$= \Gamma_{z_i}^T Q_{u_i}^{-1} \Gamma_{z_i}$	99
$\mathcal{G}_z$	$= \lim_{i\to\infty} \mathcal{G}_{z_i}$	101
$H(z)$	transfer function	69
$H_z(z)$	transfer function of inverse model	80
$H(k)$	kth Markov parameter of (A, K, C) $(H(0) = I_m, H(k) = CA^{k-1}K$ for $k > 0)$	69
$H_z(k)$	kth Markov parameter of inverse model	80
H_i	$mi \times mi$ lower block triangular Toeplitz matrix with Markov parameters of (A, K, C)	77
$\bar{H}_i$	$mi \times mi$ upper block triangular Toeplitz matrix with Markov parameters of $(\bar{A}, \bar{K}, \bar{C})$	79
$\mathcal{H}_i$	$mi \times mi$ block Hankel matrix with Markov parameters $H(1), \ldots, H(2i-1)$	89
$\\|H(z)\\|_{HSH}$	Hilbert–Schmidt–Hankel norm of $H(z)$	229
$\mathcal{H}$	double infinite Hankel matrix with complex cepstral coefficients $\bar{c}(1), \bar{c}(2), \ldots$	228
$\\|\mathcal{H}\\|_{HS}$	Hilbert–Schmidt norm of $\mathcal{H}$	228
$H(X)$	entropy of random variable X	149
$H(X, Y)$	joint entropy of X and Y	149
$H(X\\|Y)$	conditional entropy of X, given Y	149
$h(X)$	differential entropy of X	155
$h(x)$	entropy rate of $\{x(k)\}$	161
I_n	identity matrix of dimension $n \times n$	28
$I(X;Y)$	mutual information of X and Y	164
$I_r(x;y)$	mutual information rate of $\{x(k)\}$ and $\{y(k)\}$	173
$I(y_p;y_f)$	mutual information of past and future of $\{y(k)\}$	177
K	Kalman gain	69
$\lambda(A)$	eigenvalues of matrix A (in non-ascending order when real)	47
$\\|\log H\\|$	cepstral norm of SISO model with transfer function $H(z)$	216
$M_{(A,B)}$	$= (AA^T)^{-1}AB^T(BB^T)^{-1}BA^T$	98
$\max(m, n)$	maximum of real numbers m and n	85
$\min(m, n)$	minimum of real numbers m and n	26
$\mathrm{null}(A)$	null space of matrix A	31
P	controllability Gramian of (A, K) $(P = \mathcal{C}_\infty \mathcal{C}_\infty^T)$	74
Π	(orthogonal) projector	33
Π_S	orthogonal projector onto subspace S	35
Q	observability Gramian of (A, C) $(Q = \Gamma_\infty^T \Gamma_\infty)$	75
Q_z	observability Gramian of inverse model $(Q_z = \Gamma_{z_\infty}^T \Gamma_{z_\infty})$	81

symbol	explanation	page
Q_x	covariance matrix of random variable X ($Q_x = E\left\{XX^T\right\}$)	54
Q_{xy}	cross-covariance matrix of random variables X and Y ($Q_{xy} = E\left\{XY^T\right\}$)	54
$R_x(\tau)$	autocovariance function of $\{x(k)\}$ ($R_x(\tau) = E\left\{X(k+\tau)X(k)^T\right\}$)	58
$\mathcal{R}_i$	$mi \times mi$ block Hankel matrix with autocovariance samples $R(1), R(2), \ldots, R(2i-1)$	65
$r_{(A,B)}$	number of orthogonal principal directions between row(A) and row(B)	98
R_i	$= \Delta_i \Gamma_{zi}$	111
range(A)	column space of matrix A, i.e., span of the columns of A	31
rank(A)	rank of matrix A	26
res(f,g)	resultant of the polynomials f and g	232
row(A)	row space of matrix A, i.e., span of the rows of A (row(A) = range(A^T))	31
row(B/A)	orthogonal projection of row(B) onto row(A)	37
ρ_{xy}	correlation coefficient of X and Y ($\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$)	53
$S^\perp$	orthogonal complement of subspace S	32
S_u	input (innovations) covariance matrix ($S_u = E\left\{U(k)U(k)^T\right\}$)	70
$\sigma(A)$	singular spectrum of matrix A, i.e. singular values of A (in non-ascending order)	30
(u_i, σ_i, v_i)	ith singular triplet, consisting of ith left singular vector, singular value and right singular vector	30
Σ_f	state covariance matrix of forward model	68
Σ	state covariance matrix of forward innovation model	70
$\bar{\Sigma}_b$	state covariance matrix of backward model	72
$\bar{\Sigma}$	state covariance matrix of backward innovation model	73
σ_x^2	variance of random variable X ($\sigma_x^2 = E\left\{X^2\right\}$)	52
σ_{xy}	cross-covariance of X and Y ($\sigma_{xy} = E\{XY\}$)	52
$t_{(A,B)}$	dimension of intersection of row(A) and row(B)	97
T_i	$= \left(\Sigma^{-1} + \mathcal{G}_{z_i}\right)^{-1}$	111
u_p	past of $\{u(k)\}$	92
u_f	future of $\{u(k)\}$	92
U_p	$mi \times j$ past block Hankel matrix of input process $\{u(k)\}$ ($U_p = U^{(j)}_{0\|i-1}$)	76
U_f	$mi \times j$ future block Hankel matrix of input process $\{u(k)\}$ ($U_f = U^{(j)}_{i\|2i-1}$)	76

symbol	explanation	page
$\mathrm{vol}(A)$	volume of A	265
$x^{(j)}$	vector with j samples of scalar random variable X normalized by factor $\frac{1}{\sqrt{j}}$	52
$X^{(j)}$	matrix with j samples of random variable X normalized by factor $\frac{1}{\sqrt{j}}$	55
$\{x(k)\}$	stochastic process	58
$X_{0\|i-1}$	compound random variable of i subsequent random variables $X(0), \ldots, X(i-1)$ of $\{x(k)\}$	58
$X_{0\|i-1}^{(j)}$	$mi \times j$ block Hankel matrix with subsequent samples $x(0), \ldots, x(i+j-2)$ of $\{x(k)\}$, normalized by $\frac{1}{\sqrt{j}}$	59
$\mathcal{X}$	alphabet of discrete random variable X	149
X_i	$n \times j$ state sequence matrix of samples $x(i), x(i+1), \ldots, x(i+j-1)$, normalized by $\frac{1}{\sqrt{j}}$	76
X_p	$n \times j$ past state sequence matrix ($X_p = X_0$)	77
X_f	$n \times j$ future state sequence matrix ($X_f = X_i$)	77
y_p	past of $\{y(k)\}$	92
y_f	future of $\{y(k)\}$	92
Y_p	$mi \times j$ past block Hankel matrix of output process $\{y(k)\}$ ($Y_p = Y_{0\|i-1}^{(j)}$)	76
Y_f	$mi \times j$ future block Hankel matrix of output process $\{y(k)\}$ ($Y_f = Y_{i\|2i-1}^{(j)}$)	76
$\mathcal{Z}(\cdot)$	Z-transform	199

Remark 0.1. Some symbols or mathematical representations have more than one meaning, which should be clear from the context. For example, $\bar{c} \in \mathbb{C}$ is the complex conjugate of the complex number c, whereas $\bar{c}(k) \in \mathbb{R}$ denotes a complex cepstrum. Also, Q_z is the controllability Gramian of the inverse model, while Q_x is the covariance matrix of the random variable X. ◯

End Symbols

As usual, the end of a proof is indicated by □. We introduce two other end symbols: ◊ for the end of an example and ◯ for the end of a remark.

Acronyms

AR	Autoregressive	83
ARMA	Autoregressive Moving Average	84
CCA	Canonical Correlation Analysis	51
DFT	Discrete Fourier Transform	197
FFT	Fast Fourier Transform	82
IFFT	Inverse Fast Fourier Transform	197
MA	Moving Average	82
MIMO	Multiple Input Multiple Output	75
SISO	Single Input Single Output	81
SVD	Singular Value Decomposition	26

Contents

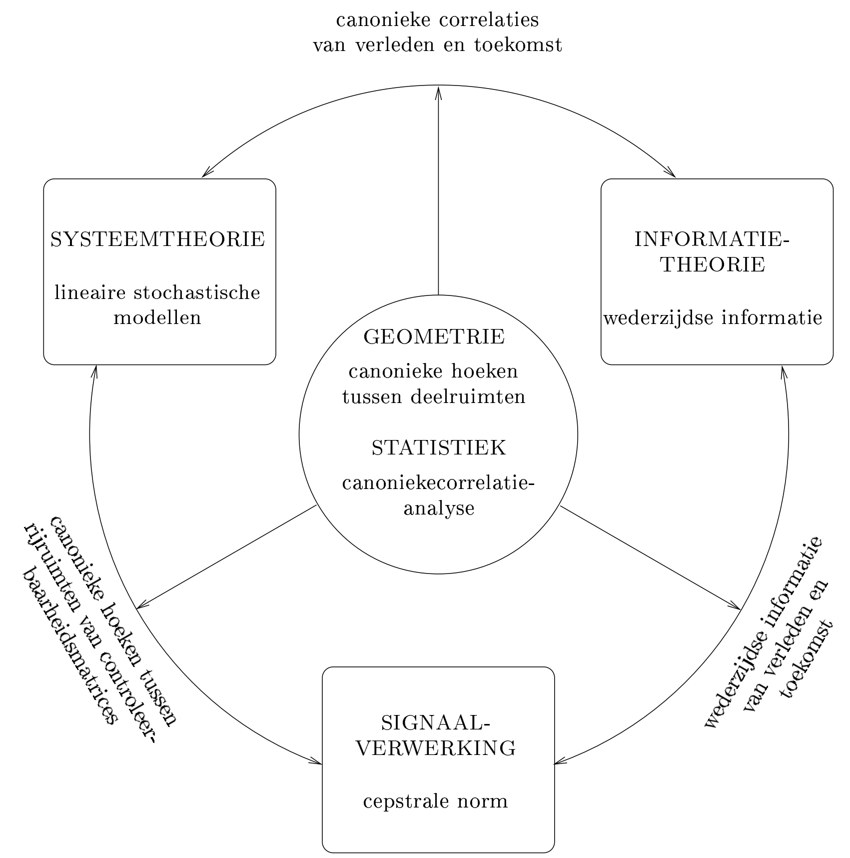
Canonieke hoeken in systeemtheorie, informatietheorie en signaalverwerking

Nederlandse samenvatting

Hoofdstuk 1 Inleiding

In dit proefschrift worden verbanden gelegd tussen de volgende disciplines: systeemtheorie, informatietheorie en signaalverwerking. De meetkundige techniek die we daarvoor gebruiken is het berekenen van canonieke hoeken tussen lineaire deelruimten. In figuur 1 geven we onze aanpak schematisch weer. Centraal staan de canonieke hoeken tussen deelruimten en hun statistische tegenhangers, de canonieke correlaties tussen random variabelen. Rondom vinden we de drie disciplines, systeemtheorie, informatietheorie en signaalverwerking terug. Uit systeemtheorie gebruiken we de beschrijving van lineaire stochastische modellen. Het informatietheoretische concept waarmee we vooral werken, is de wederzijdse informatie van twee stochastische processen. Uit signaalverwerking, tenslotte, komt vooral het cepstrum en een cepstrale norm voor lineaire modellen aan bod.
Door de canonieke hoeken te berekenen tussen welgekozen deelruimten, leggen we de volgende verbanden, zoals weergegeven in figuur 1.

(1) **Een link tussen systeemtheorie en informatietheorie: de canonieke correlaties van de verleden en toekomstige ingangs- en uitgangsprocessen.**
We berekenen de canonieke hoeken tussen de verschillende combinaties



Figuur 1: Een schematische voorstelling van de verbanden tussen systeemtheorie, informatietheorie en signaalverwerking die in dit proefschrift behandeld worden.

van verleden en toekomstige ingangs- en uitgangsruimten van een lineair stochastisch model. Via de parametrische beschrijving van het model, kunnen deze hoeken uitgedrukt worden in functie van de modelparameters. De cosinussen van deze hoeken zijn een benadering voor de canonieke correlaties van de overeenkomstige processen. De limietwaarden van de parametrische uitdrukkingen van de cosinussen zijn gelijk aan de canonieke correlaties van de overeenkomstige processen. Via deze uitdrukkingen voor de canonieke correlaties in functie van de modelparameters bewijzen we verbanden tussen de verschillende correlaties. Die verbanden worden onder andere gebruikt om de cepstrale afstandsmaat een informatietheoretische interpretatie te geven.

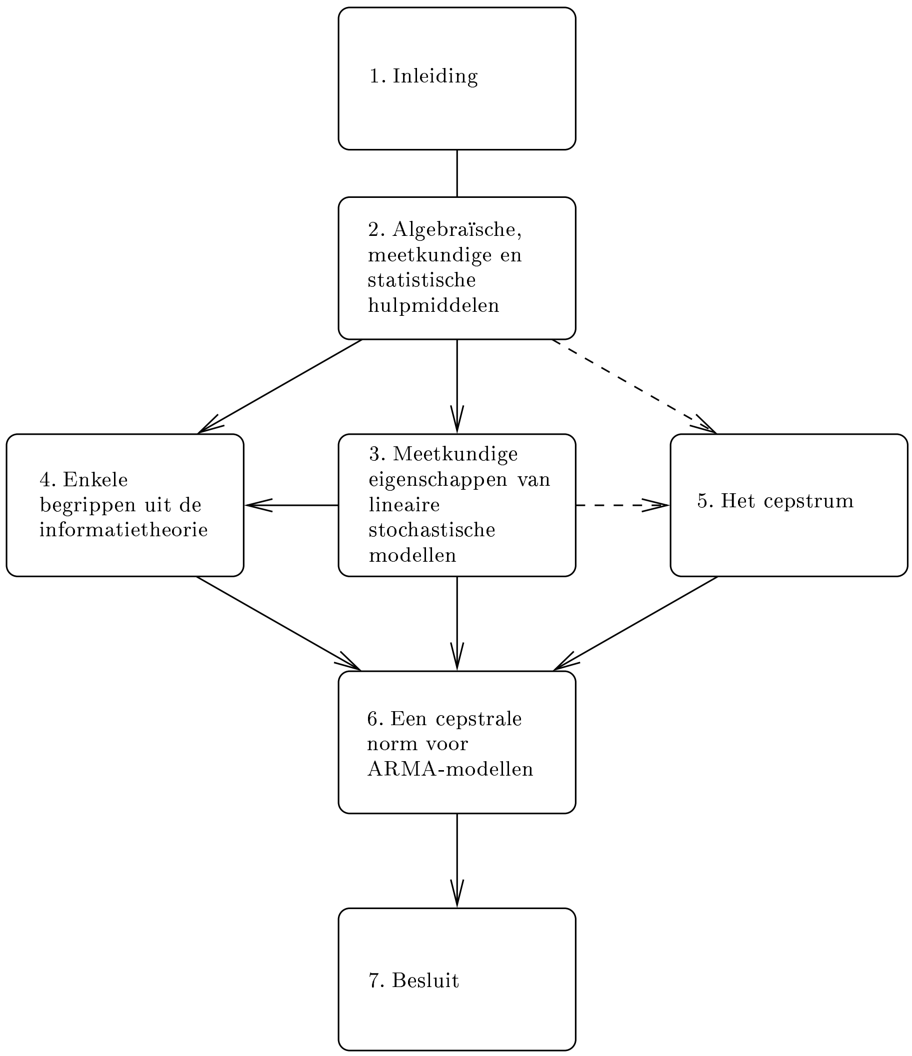
De wederzijdse informatie van twee Gaussische stochastische processen kan beschreven worden in termen van hun canonieke correlaties. Omdat

we uitdrukkingen hebben voor de canonieke correlaties kennen we dus ook de wederzijdse informatie van de processen in functie van de modelparameters.
Vooral de wederzijdse informatie van de verleden en de toekomstige uitgangsprocessen is een interessante grootheid, omdat die kan geschat worden uit geobserveerde uitgangsdata.

(2) **Een link tussen systeemtheorie en signaalverwerking: de canonieke hoeken tussen rijruimten van controleerbaarheidsmatrices.**
Het cepstrum is het geschikte hulpmiddel voor de definitie van afstandsmaten tussen signalen die afhangen van de verhouding van de spectra. We tonen hoe een recent gedefinieerde cepstrale afstand tussen twee lineaire modellen voor scalaire stochastische processen geïnterpreteerd kan worden in systeemtheoretische termen.
We tonen aan dat de overeenkomstige cepstrale norm van een lineair stochastisch model een functie is van de canonieke hoeken tussen de rijruimten van de controleerbaarheidsmatrix van het model en de controleerbaarheidsmatrix van het inverse model. Bovendien bewijzen we dat de cosinussen van deze hoeken gelijk zijn aan de n kleinste canonieke correlaties van de toekomstige ingangs- en uitgangsprocessen van het $n$de orde model en de andere canonieke correlaties zijn gelijk aan 1. We kunnen bijgevolg de cepstrale norm uitdrukken als een functie van deze canonieke correlaties.
Gebaseerd op deze resultaten en op de eigenschappen van lineaire stochastische modellen, leiden we verscheidene uitdrukkingen af voor de cepstrale norm in functie van observeerbaarheids- en controleerbaarheidsgramianen, die we vinden door Lyapunov-vergelijkingen op te lossen.
Door twee verschillende uitdrukkingen voor de canonieke correlaties van toekomstige ingang en uitgang aan elkaar gelijk te stellen, bekomen we als een nevenresultaat een nieuwe eigenschap voor Lyapunov-vergelijkingen.

(3) **Een link tussen informatietheorie en signaalverwerking: de wederzijdse informatie van verleden en toekomst**
Via de meetkundige inzichten van lineaire stochastische modellen, kunnen we een stap verder gaan. Met behulp van de relaties tussen de verschillende canonieke correlaties, die besproken werden in puntje (1), bewijzen we dat de cepstrale norm nauw verwant is aan de wederzijdse informatie, een begrip uit de informatietheorie.

De structuur van de thesis wordt in figuur 2 weergegeven. We zien dat hoofdstuk 3, waarin de canonieke hoeken tussen de verschillende paren van ingangs- en uitgangsruimten berekend worden, een centrale rol vervult en dat in hoofdstuk 6 de resultaten en inzichten van de hoofdstukken 3, 4 en 5 samenkomen om zo de cepstrale afstandsmaat met systeemtheorie enerzijds en informatietheorie anderzijds in verband te brengen.



Figuur 2: Structuur van het proefschrift.

Hoofdstuk 2 Hulpmiddelen uit de lineaire algebra, meetkunde en statistiek

In dit hoofdstuk geven we een overzicht van de algebraïsche, meetkundige en statistische technieken die we in de volgende hoofdstukken toepassen. Zoals we gezien hebben in figuur 1, is de belangrijkste van deze technieken het berekenen

van de canonieke hoeken tussen lineaire deelruimten. Dit concept werd reeds in de 19de eeuw door Camille Jordan [77] beschreven. Hotelling introduceerde in [68] de canoniekecorrelatieanalyse, die een statistische interpretatie geeft van de canonieke hoeken. Sindsdien is het een belangrijke statistische techniek met heel wat toepassingen (zie bijvoorbeeld [53]). De canonieke hoeken zelf worden onder andere toegepast in stochastische realisatie [2, 156, 159].
Hoewel de definitie van de canonieke hoeken tussen twee deelruimten en hun corresponderende canonieke vectoren in de deelruimten erg eenvoudig is en enkel gebaseerd op Euclidische meetkunde, steunen we voor hun berekening op technieken uit de numerieke lineaire algebra, zoals de singulierewaardenontbinding en orthogonale projectoren. Daarom beginnen we dit hoofdstuk met een bespreking van deze algebraïsche begrippen, waarbij we de nadruk leggen op hun geometrische interpretatie. Daarna behandelen we de canonieke hoeken en we eindigen met een korte bespreking van canoniekecorrelatieanalyse voor random variabelen en stochastische processen.

Algebraïsche technieken

De **singulierewaardenontbinding (SWO)** van een matrix wordt dikwijls toegepast in dit proefschrift. Niet alleen bij berekeningen, maar ook in theoretische afleidingen. We gebruiken zowel de volledige als de gereduceerde SWO, die als volgt gedefinieerd worden.
Elke reële $m \times n$ matrix A kan ontbonden worden als $A = USV^T$, waarbij $U \in \mathbb{R}^{m\times m}$ en $V \in \mathbb{R}^{n\times n}$ orthogonale matrices zijn en $S = \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix}$ met $S_1 = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, r is de rang van A en $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.
Als A niet van volle rang is ($r < \min(m, n)$), dan kunnen de kolommen van U en V in twee groepen verdeeld worden: $U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}$ en $V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}$, met $U_1 \in \mathbb{R}^{m\times r}$, $U_2 \in \mathbb{R}^{m\times(m-r)}$, $V_1 \in \mathbb{R}^{n\times r}$ en $V_2 \in \mathbb{R}^{n\times(n-r)}$. De matrix A kan dan ook ontbonden worden als $A = U_1 S_1 V_1^T$. Deze ontbinding wordt de gereduceerde singulierewaardenontbinding genoemd.
De kolommen van U_1 vormen een orthonormale basis voor de kolomruimte van A, en de kolommen van V_1 zijn een orthonormale basis voor de rijruimte van A.
Via de SWO van de matrix A kunnen we de Moore–Penrose veralgemeende inverse van A berekenen als $A^\dagger = V_1 S_1^{-1} U_1^T$.
De orthogonale projectie van een vector op een deelruimte kan beschreven worden door een matrixvectorproduct, waarbij de matrix een **orthogonale projector** is op de deelruimte. Een orthogonale projector Π is een idempotente matrix ($\Pi^2 = \Pi$) die symmetrisch is ($\Pi^T = \Pi$). Als de gereduceerde SWO van een matrix $A \in \mathbb{R}^{m\times n}$ gelijk is aan $A = U_1 S_1 V_1^T$, dan is de orthogonale projector op de rijruimte van A gelijk aan $\Pi_A = V_1 V_1^T$. De orthogonale projectie van de rijruimte van $B \in \mathbb{R}^{p\times n}$ op de rijruimte van $A \in \mathbb{R}^{m\times n}$ is dan gelijk aan de rijruimte van $B/A = B\Pi_A$.

Meetkundige technieken

De **canonieke hoeken tussen twee deelruimten** S_1 en S_2 in $\mathbb{R}^n$ zijn een veralgemening van de hoek tussen twee vectoren. Veronderstel dat S_1 dimensie d_1 heeft en S_2 dimensie d_2 en $d_1 \leq d_2$, dan kunnen d_1 canonieke hoeken en corresponderende canonieke vectorparen gedefinieerd worden. De eerste canonieke hoek, θ_1, is de kleinste hoek tussen een vector van S_1 en een vector van S_2 en de vectoren vormen na normalisatie tot lengte 1 het eerste paar canonieke vectoren: $u_1 \in S_1$ en $v_1 \in S_2$. Voor de tweede canonieke hoek zoeken we de vector u_2 met lengte 1 in S_1 die loodrecht staat op u_1 en v_2 in S_2 loodrecht op v_1 waarvoor geldt dat de hoek tussen u_2 en v_2 minimaal is. De vectoren u_2 en v_2 zijn het tweede paar canonieke vectoren en de hoek ertussen de tweede canonieke hoek θ_2. Zo gaat men door tot er d_1 hoeken en vectorparen gevonden zijn.
De canonieke hoeken tussen de rijruimten van twee matrices $A \in \mathbb{R}^{p\times n}$ en $B \in \mathbb{R}^{q\times n}$ met respectievelijk rang r_a en r_b kunnen als volgt berekend worden [57].

- Bereken de gereduceerde SWO van A en B
$$A = U_{a_1} S_{a_1} V_{a_1}^T \quad \text{en} \quad B = U_{b_1} S_{b_1} V_{b_1}^T \ .$$
- Bereken de volledige SWO van $V_{a_1}^T V_{b_1}$
$$V_{a_1}^T V_{b_1} = PSQ^T \ .$$
- De diagonaalelementen van S zijn de cosinussen van de canonieke hoeken tussen de rijruimten van A en B. De canonieke vectoren in de rijruimte van A zijn de eerste $\min(r_a, r_b)$ kolommen van $V_{a_1}P$ en de canonieke vectoren in de rijruimte van B zijn de eerste $\min(r_a, r_b)$ kolommen van $V_{b_1}Q$.

Door de canonieke hoeken tussen en canonieke vectoren in de rijruimten van A en B te karakteriseren als de singuliere waarden en singuliere vectoren van $\Pi_A\Pi_B$, met Π_A en Π_B de orthogonale projectoren op de rijruimten van respectievelijk A en B, kan men een aantal interessante eigenschappen afleiden voor de canonieke hoeken tussen **orthogonale projecties van deelruimten**. Veronderstel dat $r_{(A,B)}$ canonieke hoeken tussen de rijruimten van A en B gelijk zijn aan $\frac{\pi}{2}$, dan geldt

$$\begin{aligned}[A \sphericalangle B] &= [A \sphericalangle A/B]\,, \underbrace{\frac{\pi}{2}, \ldots, \frac{\pi}{2}}_{r_{(A,B)}}\,, \\ &= [B/A \sphericalangle B]\,, \underbrace{\frac{\pi}{2}, \ldots, \frac{\pi}{2}}_{r_{(A,B)}}\,,\end{aligned}$$

en ook

$$[A \sphericalangle B] = [B/A \sphericalangle A/B], \underbrace{\frac{\pi}{2}, \ldots, \frac{\pi}{2}}_{r_{(A,B)}} .$$

Als de matrices $A \in \mathbb{R}^{p\times n}$ en $B \in \mathbb{R}^{q\times n}$ van volle rijrang zijn, dan kunnen de canonieke hoeken tussen hun rijruimten ook bekomen worden als volgt. Veronderstel dat $p \le q$, dan zijn de vierkantswortels van de eigenwaarden van $(AA^T)^{-1}AB^T(BB^T)^{-1}BA^T$ gelijk aan de cosinussen van de canonieke hoeken. Deze eigenschap wordt dikwijls gebruikt in de afleidingen.

Statistische technieken

De **correlatiecoëfficiënt** van twee scalaire random variabelen X en Y met verwachte waarde 0 wordt gedefinieerd als $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$, waarbij σ_{xy} de covariantie van X en Y is ($\sigma_{xy} = E\{XY\}$) en σ_x^2 en σ_y^2 de variantie van respectievelijk X en Y ($\sigma_x^2 = E\{X^2\}$ en analoog voor σ_y^2). Het is een maat voor de lineaire afhankelijkheid van de twee variabelen. Door de vectoren $x^{(j)} = \frac{1}{\sqrt{j}}\begin{pmatrix} x_0 & x_1 & \ldots & x_{j-1}\end{pmatrix}$ en $y^{(j)} = \frac{1}{\sqrt{j}}\begin{pmatrix} y_0 & y_1 & \ldots & y_{j-1}\end{pmatrix}$ met j monsters van X en Y te definiëren, zien we dat de correlatiecoëffiënt ook gezien kan worden als de cosinus van de hoek tussen $x^{(j)}$ en $y^{(j)}$, voor $j \to \infty$:

$$\rho_{xy} = \lim_{j\to\infty} \cos\left[x^{(j)} \sphericalangle y^{(j)}\right] .$$

Zoals de hoek tussen twee vectoren veralgemeend kan worden tot een reeks van canonieke hoeken tussen twee deelruimten, zo wordt de correlatiecoëffiënt van twee scalaire random variabelen veralgemeend tot de **canonieke correlaties van twee random variabelen** met meerdere componenten. De canonieke correlaties van de random variabelen $X \in \mathbb{R}^p$ en $Y \in \mathbb{R}^q$ zijn dus gelijk aan de cosinussen van de canonieke hoeken tussen de rijruimten van de matrices $X^{(j)} = \frac{1}{\sqrt{j}}\begin{pmatrix} x_0 & \cdots & x_{j-1}\end{pmatrix} \in \mathbb{R}^{p\times j}$ en $Y^{(j)} = \frac{1}{\sqrt{j}}\begin{pmatrix} y_0 & \cdots & y_{j-1}\end{pmatrix} \in \mathbb{R}^{q\times j}$, voor $j \to \infty$.

Een stochastisch proces $\{x(k)\}$ kan beschouwd worden als een opeenvolging van random variabelen $X(0), X(1), \ldots$ De **canonieke correlaties van de stochastisch processen** $\{x(k)\}$ en $\{y(k)\}$ kunnen dan bekomen worden door de canonieke correlaties te berekenen van $\begin{pmatrix} X(0)\\ X(1)\\ \vdots\end{pmatrix}$ en $\begin{pmatrix} Y(0)\\ Y(1)\\ \vdots\end{pmatrix}$. Voor stationaire, ergodische processen betekent dit dat we de canonieke correlaties bekomen als de cosinussen van de canonieke hoeken tussen de rijruimten van de blok-Hankel-matrices $X^{(j)}_{0|i-1}$ en $Y^{(j)}_{0|i-1}$ (voor $i, j \to \infty$), met

$$X^{(j)}_{0|i-1} = \frac{1}{\sqrt{j}} \begin{pmatrix} x(0) & x(1) & x(2) & \cdots & x(j-1) \\ x(1) & x(2) & x(3) & \cdots & x(j) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x(i-1) & x(i) & x(i+1) & \cdots & x(i+j-2) \end{pmatrix} ,$$

en analoog voor $Y^{(j)}_{0|i-1}$. De elementen $x(0), \ldots, x(i+j-2)$ van $X^{(j)}_{0|i-1}$ zijn opeenvolgende monsters van het stochastische proces $\{x(k)\}$. Als een eindig aantal observaties gegeven is, kunnen we de canonieke correlaties tussen de processen dus benaderend berekenen als de cosinussen tussen de eindige matrices $X^{(j)}_{0|i-1}$ en $Y^{(j)}_{0|i-1}$.

Hoofdstuk 3 Meetkundige eigenschappen van lineaire stochastische modellen

Vele technische, economische, fysische, enz. processen kunnen gemodelleerd worden als de uitgang van een lineair tijdinvariant stochastisch model waarvan de ingang stationaire witte ruis is. De modellen worden gebruikt voor simulatie, predictie, analyse, datacompressie, ... Lineaire modellen maken het bovendien mogelijk om elegante meetkundige interpretaties te geven en ze lenen zich uitstekend tot efficiënte numerieke algoritmen.

De modelklasse

De representatie voor een lineair stochastisch model die het vaakst wordt gebruikt in dit proefschrift is de voorwaartse innovatievorm. De toestandsruimtevergelijkingen zijn

$$\begin{cases} x(k+1) &= Ax(k) + Ku(k) , \\ y(k) &= Cx(k) + u(k) , \end{cases}$$

met $\{x(k)\}$ het toestandsproces en $\{u(k)\} \in \mathbb{R}^m$ en $\{y(k)\} \in \mathbb{R}^m$ het respectievelijk ingangs- en uitgangsproces. Als de orde van het model gelijk is aan n, dan hebben de toestandsruimtematrices de volgende dimensies: $A \in \mathbb{R}^{n\times n}$, $K \in \mathbb{R}^{n\times m}$ en $C \in \mathbb{R}^{m\times n}$. De ingang is stationaire witte ruis met autocovariantiefunctie $R_u(\tau) = S_u\delta(\tau)$ en onafhankelijk van de begintoestand. Het toestandsproces is daarom onafhankelijk van de huidige en toekomstige ingang en de uitgang is onafhankelijk van alle toekomstige ingangen. Het model wordt volledig gekenmerkt door de systeemmatrices (A, K, C) en de ingangscovariantiematrix S_u.

De canonieke hoeken tussen de verleden en toekomstige ingangs- en uitgangsruimten

Voor deze representatie berekenen we de canonieke hoeken tussen de verleden en toekomstige ingangs- en uitgangsruimten in functie van de modelparameters, (A, K, C) en S_u. Daarmee bedoelen we de canonieke hoeken tussen de verschillende combinaties van rijruimten van de blok-Hankel-matrices $U_p = U^{(j)}_{0|i-1}$, $U_f = U^{(j)}_{i|2i-1}$, $Y_p = Y^{(j)}_{0|i-1}$ en $Y_f = Y^{(j)}_{i|2i-1}$. We veronderstellen daarbij dat $j \to \infty$. Omdat de ingang witte ruis is weten we dat de hoeken tussen de rijruimten van U_p en U_f gelijk zijn aan $\frac{\pi}{2}$. Ook de hoeken tussen de rijruimten van Y_p en U_f zijn recht, want de toekomstige ingang is onafhankelijk van de verleden uitgang. Verder zijn door de stationariteit de hoeken tussen de rijruimten van U_p en Y_p gelijk aan de hoeken tussen de rijruimten van U_f en Y_f. Voor de overblijvende paren bekomen we:

$$\begin{aligned}
\lim_{j\to\infty} \cos^2 [U_f \sphericalangle Y_f] &= \lambda\left((I_n + \mathcal{G}_{z_i}\Sigma)^{-1}\right), \underbrace{1,\ldots,1}_{mi-n} ,\\
\lim_{j\to\infty} \cos^2 [U_p \sphericalangle Y_f] &= \lambda\left(\mathcal{D}_i(\mathcal{G}_{z_i}^{-1} + \Sigma)^{-1}\right), \underbrace{0,\ldots,0}_{mi-n} ,
\end{aligned}$$

en

$$\begin{aligned}
\lim_{j\to\infty} \cos^2 [Y_p \sphericalangle Y_f] = \Big(&-A^i\Sigma R_i^T + \mathcal{D}_i - R_i T_i R_i^T + A^i\Sigma\mathcal{G}_{z_i}(\Sigma A^{i^T} + T_i R_i^T) \\
&+ \left(-R_i - A^i\Sigma\mathcal{G}_{z_i}T_i\mathcal{G}_{z_i} + R_i T_i \mathcal{G}_{z_i}\right)\Sigma A^{i^T}\Big)\left(\mathcal{G}_{z_i}^{-1} + \Sigma\right)^{-1}, \underbrace{0,\ldots,0}_{mi-n} .
\end{aligned}$$

De matrices in de uitdrukkingen kunnen als volgt berekend worden. De toestandscovariantiematrix Σ is de oplossing van de Lyapunov-vergelijking

$$\Sigma = A\Sigma A^T + K S_u K^T ,$$

en de andere matrices zijn gelijk aan

$$\begin{aligned}
\mathcal{G}_{z_i} &= \sum_{k=0}^{i-1} (A - KC)^{k^T} C^T S_u^{-1} C (A - KC)^k ,\\
\mathcal{D}_i &= \sum_{k=0}^{i-1} A^k K S_u K^T A^{k^T} ,\\
R_i &= -\sum_{k=0}^{i-1} A^{i-1-k} K C (A - KC)^k ,\\
T_i &= (\Sigma^{-1} + \mathcal{G}_{z_i})^{-1} .
\end{aligned}$$

Door i naar oneindig te laten gaan, bekomen we de canonieke correlaties tussen de verleden en toekomstige ingangs- en uitgangsprocessen in functie van de modelparameters. Hieruit kunnen de verbanden tussen de verschillende canonieke hoeken (en canonieke correlaties) gemakkelijk afgeleid worden:

$$[U_p \sphericalangle Y_f] = [Y_p \sphericalangle Y_f] = \frac{\pi}{2} - [Y_p \sphericalangle Y_f] \ . \tag{0.1}$$

De canonieke hoeken tussen de rijruimten van U_p en Y_f zijn gelijk aan de canonieke hoeken tussen de rijruimten van Y_p en Y_f en ze zijn complementair aan de hoeken tussen de rijruimten van U_f en Y_f.
Voor modellen met slechts één ingang en één uitgang (SISO-modellen), kunnen de uitdrukkingen voor de canonieke hoeken en canonieke correlaties vereenvoudigd worden. Ze zijn bijvoorbeeld onafhankelijk van de ingangsvariantie. Een andere eigenschap van de canonieke correlaties voor SISO-modellen is de volgende. Ze speelt een belangrijke rol in de afleidingen van de hoofdstukken die volgen.

Eigenschap 1. De canonieke correlaties van de toekomstige ingang en toekomstige uitgang van een SISO-model
De kleinste n canonieke correlaties van de toekomstige ingang en de toekomstige uitgang van een SISO-model zijn gelijk aan de cosinussen van de canonieke hoeken tussen de rijruimten van de oneindige controleerbaarheidsmatrix van het model en de oneindige controleerbaarheidsmatrix van het inverse model.

We bewijzen verder nog dat de canonieke hoeken tussen de verleden en toekomstige uitgangsruimten van een autoregressief (AR) model onafhankelijk zijn van i, van zodra $i \geq n$. De cosinussen zijn dan gelijk aan de vierkantswortels van de eigenwaarden van $P^{-1}A^nPA^{n^T}$ en dit zijn dus ook de canonieke correlaties tussen het verleden en de toekomst van $\{y(k)\}$. De matrix P is de controleerbaarheidsgramiaan van het model.

Hoofdstuk 4 Informatietheoretische begrippen

We bespreken twee belangrijke informatietheoretische begrippen: de entropie en de wederzijdse informatie.

De entropie van een random variabele

De entropie van een random variabele is een maat voor de onzekerheid die een waarnemer heeft over de uitkomst van een steekproef. Omdat na de steekproef de onzekerheid volledig is verdwenen, kan men de informatie die men bekomt door het uitvoeren van het experiment gelijkstellen aan de onzekerheid die men ervoor had. De entropie is dus ook een maat voor de informatie die de

waarnemer wint door een experiment uit te voeren.
Voor een discrete random variabele die slechts een eindig aantal waarden kan aannemen, wordt de entropie gedefinieerd als

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \ ,$$

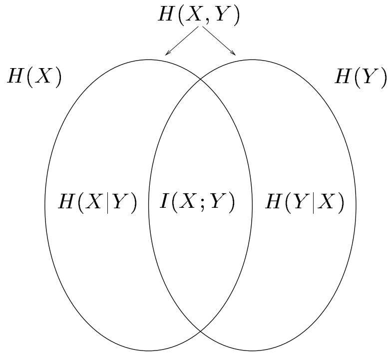
met $p(x)$ de kans op de uitkomst x en $\mathcal{X}$ de verzameling van mogelijke waarden. Ze wordt uitgedrukt in bits als voor de logaritmische functie basis 2 wordt gebruikt.
De entropie van een continue variabele kan gedefinieerd worden door de waarschijnlijkheidsverdelingsfunctie te discretiseren en van de bekomen discrete random variabele de entropie te berekenen. Als de discretisatie steeds fijner wordt gemaakt, kan de limiet de entropie van de continue variabele genoemd worden. Maar, deze limiet is oneindig groot, wat ingezien kan worden door te bedenken dat de onzekerheid over de waarde van een steekproef oneindig groot is omdat de variabele oneindig veel waarden kan aannemen. In de praktijk is men echter enkel geïnteresseerd in een benadering van de uitkomst. De onzekerheid hangt dan af van de nauwkeurigheid waarmee de uitkomst gemeten wordt, of dus van het discretisatieniveau van de waarschijnlijkheidsverdeling. Als men toch een entropiegrootheid wil definiëren voor continue variabelen die onafhankelijk is van het discretisatieniveau, kan men zijn toevlucht nemen tot de integraal $h(X) = -\int f(x) \log f(x)\, dx$, waarin $f(x)$ de verdelingsfunctie is van de variabele X. Dit wordt de differentiële entropie van de continue random variabele genoemd om een onderscheid te maken met de echte entropie, die gelijk is aan ∞.
De differentiële entropie van een Gaussisch verdeelde random variabele $X \in \mathbb{R}^m$ is enkel afhankelijk van haar covariantiematrix Q, ze is namelijk gelijk aan

$$h(X) = \frac{m}{2} \log(2\pi e) + \frac{1}{2} \log \det Q \ .$$

De wederzijdse informatie van twee random variabelen

Als we twee onafhankelijke variabelen X en Y observeren, is het gemakkelijk in te zien dat de hoeveelheid informatie die we bekomen door de uitkomst van X en Y te kennen, gelijk is aan de som van de informatiehoeveelheden die X en Y elk leveren. Dus, voor onafhankelijke variabelen is de gemeenschappelijke entropie gelijk aan de som van de afzonderlijke entropieën: $H(X,Y) = H(X) + H(Y)$. Wanneer X en Y echter afhankelijk zijn, geldt deze gelijkheid niet, want X observeren zegt ons dan ook iets over Y. De extra informatie die we toch nog bekomen bij het observeren van Y als X al gekend is, wordt gemeten met $H(Y|X)$, de conditionele entropie van Y, gegeven X. De gemeenschappelijke entropie van X en Y is in het algemeen dus gelijk aan $H(X,Y) = H(X) + H(Y|X)$. We zien ook dat de informatie die X geeft over Y, gelijk is aan $H(Y) - H(Y|X)$. Een analoge redenering voor het geval we

eerst Y te weten komen en dan pas X, geeft $H(X,Y) = H(Y) + H(X|Y)$. Dus volgt dat $H(Y) - H(Y|X) = H(X) - H(X|Y)$, de informatie die Y bevat over X is gelijk aan de informatie die X bevat over Y. Deze grootheid is de wederzijdse informatie van X en Y, voorgesteld met $I(X;Y)$.
In figuur 3 wordt het verband tussen de verschillende entropieën en de weder-



Figuur 3: Het verband tussen de wederzijdse informatie en de entropie van twee random variabelen X en Y.

zijdse informatie in een Venn-diagram weergegeven. De gemeenschappelijke entropie kan gezien worden als de unie van de verzamelingen en de wederzijdse informatie die X en Y over elkaar bevatten is de doorsnede. De onzekerheid over X als Y gekend is, is dan dat gedeelte van $H(X)$ dat niet in de doorsnede zit.
Ook voor continue random variabelen X en Y kan de wederzijdse informatie gedefinieerd worden en bekomen we het verband $h(X,Y) = h(X) + h(Y) - I(X;Y)$, waar $h(X)$ en $h(Y)$ de differentiële entropie van respectievelijk X en Y is en $h(X,Y)$ de gemeenschappelijke differentiële entropie van X en Y.
Indien de variabele (X,Y) normaal verdeeld is met covariantiematrix Q, dan verkrijgen we via de uitdrukking voor de differentiële entropie van een Gaussische variabele

$$I(X;Y) = -\frac{1}{2}\log\frac{\det Q}{\det Q_x \det Q_y} \ ,$$

met Q_x en Q_y de covariantiematrices van X en Y. Het is dan gemakkelijk aan te tonen dat de wederzijdse informatie van $X \in \mathbb{R}^p$ en $Y \in \mathbb{R}^q$ ook gelijk is aan

$$I(X;Y) = -\frac{1}{2}\log\prod_{k=1}^{\min(p,q)}\sin^2\phi_k \ ,$$

met $\phi_1, \ldots, \phi_{\min(p,q)}$ de canonieke hoeken tussen de rijruimten van $X^{(j)}$ en $Y^{(j)}$ voor $j \to \infty$.

De wederzijdse informatie van twee stochastische processen

Omdat we een stochastisch proces kunnen beschouwen als een opeenvolging van random variabelen, kunnen we de resultaten voor de random variabelen onmiddellijk toepassen op stochastische processen. De wederzijdse informatie van de stochastisch processen $\{x(k)\}$ en $\{y(k)\}$ kan gedefinieerd worden als de wederzijdse informatie van $X_{0|i-1} = \begin{pmatrix} X(0) \\ \vdots \\ X(i-1) \end{pmatrix}$ en $Y_{0|i-1} = \begin{pmatrix} Y(0) \\ \vdots \\ Y(i-1) \end{pmatrix}$, waarbij $i \to \infty$ (als de limiet bestaat). Voor Gaussische processen hebben we dus ook

$$I(x;y) = \lim_{i\to\infty} -\frac{1}{2}\log\frac{\det Q_i}{\det Q_{x_i}\det Q_{y_i}} ,$$

met $Q_i = \begin{pmatrix} Q_{x_i} & Q_{xy_i} \\ Q_{yx_i} & Q_{y_i} \end{pmatrix} = E\left\{\begin{pmatrix} X_{0|i-1} \\ Y_{0|i-1} \end{pmatrix}\begin{pmatrix} X_{0|i-1}^T & Y_{0|i-1}^T \end{pmatrix}\right\}$. En in functie van de canonieke hoeken ϕ_k $(k = 1,\ldots,i)$ tussen de rijruimten van de blok-Hankel-matrices $X_{0|i-1}^{(j)}$ en $Y_{0|i-1}^{(j)}$, geeft dit

$$I(x;y) = \lim_{i,j\to\infty} -\frac{1}{2}\log\prod_{k=1}^{i}\sin^2\phi_k \; . \tag{0.2}$$

Met de resultaten uit hoofdstuk 3, kunnen we dan de wederzijdse informatie van de verleden en toekomstige ingangs- en uitgangsprocessen van een lineair stochastisch model uitdrukken in functie van de modelparameters. Stel dat we een $n$de orde model hebben met toestandsruimtematrices (A, K, C) en ingangscovariantiematrix S_u. Dan geldt,

$$\begin{aligned} I(u_p;u_f) &= I(y_p;u_f) = 0 \; , \\ I(u_p;y_p) &= I(u_f;y_f) = \infty \; , \\ I(u_p;y_f) &= I(y_p;y_f) = -\frac{1}{2}\log\det(I_n - \Sigma(\mathcal{G}_z^{-1}+\Sigma)^{-1}) \; , \\ &= -\frac{1}{2}\log\det(I_n - \Sigma\bar{\Sigma}) \; . \end{aligned}$$

Hierin is Σ de toestandscovariantiematrix, $\mathcal{G}_z = \lim_{i\to\infty}\mathcal{G}_{z_i}$ en $\bar{\Sigma}$ de toestandscovariantiematrix van de achterwaartse innovatierepresentatie van het model. De matrix $\mathcal{G}_z$ is de oplossing van de Lyapunov-vergelijking

$$\mathcal{G}_z = (A-KC)^T\mathcal{G}_z(A-KC) + C^TS_u^{-1}C \; ,$$

en $\bar{\Sigma}$ is de oplossing van de algebraïsche Riccati-vergelijking

$$\bar{\Sigma} = A^T\bar{\Sigma}A + (C^T - A^T\bar{\Sigma}G)(C\Sigma C^T + S_u - G^T\bar{\Sigma}G)^{-1}(C - G\bar{\Sigma}A) \; ,$$

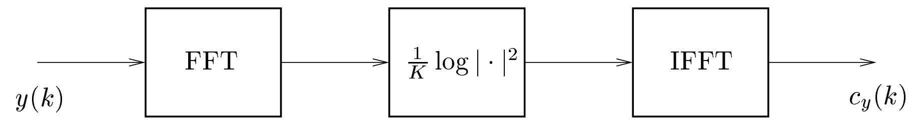
met $G = C\Sigma A^T + S_uK^T$.
Voor SISO-modellen kunnen de resultaten vereenvoudigd worden.

Hoofdstuk 5 Het cepstrum

Het woord 'cepstrum', anagram van spectrum, werd voor het eerst gebruikt in het artikel uit 1963 'The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking' van Bogert, Healy and Tukey [17]. Zij voerden het cepstrum in als techniek om in seismologische signalen echo's te detecteren. Het wordt ook veelvuldig toegepast in signaalverwerkingsalgoritmen voor spraaktoepassingen, zie [41] en referenties erin, en als basis voor afstandsmaten tussen signalen en modellen (zie [6] voor een overzicht) en hoofdstuk 6 van dit proefschrift. Daarnaast vinden we het ook terug in sommige technieken voor de detectie van fouten in roterende machines [161]. Andere toepassingen, o.a. radar en sonar, en het elektroëncefalogram, kunnen teruggevonden worden in [27]. Een nieuwe ontwikkeling zien we in recente publicaties van Byrnes, Enqvist en Lindquist [22, 23, 46, 47], waar de cepstrale parameters aangewend worden om lineaire modellen te vinden die een gegeven sequentie van covarianties genereren.
In dit proefschrift passen we zowel het vermogencepstrum als het complexe cepstrum toe. Ze worden in de volgende paragraaf gedefinieerd. Voor het cepstrum van een scalair ARMA-proces[1] bestaan eenvoudige uitdrukkingen die we in het volgende hoofdstuk nodig hebben. We bespreken ze hier kort.

Het vermogencepstrum en het complexe cepstrum

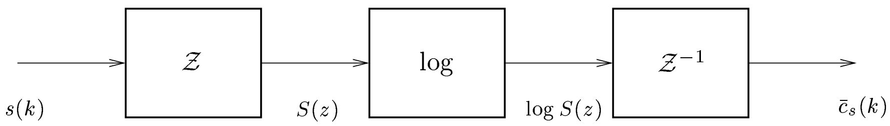
Het vermogencepstrum van een scalair stochastisch proces $\{y(k)\}$ wordt gedefinieerd als de inverse Fourier-transformatie van het logaritme van zijn spectrum $\Phi_y(e^{i\theta})$. Voor de schatting van het vermogencepstrum kunnen we gebruik maken van het 'snelle Fourier-algoritme' (FFT), zoals weergegeven in figuur 4.



Figuur 4: Schatting van het vermogencepstrum. FFT staat voor 'fast Fourier transform' en IFFT voor 'inverse fast Fourier transform'. De parameter K is het aantal datapunten.

Voor absoluut optelbare sequenties, zoals bijvoorbeeld de impulsrespons van een stabiel model, kan men het complexe cepstrum definiëren als de inverse Z-transformatie van het logaritme van hun Z-transformatie. Dit wordt schematisch voorgesteld in figuur 5.

[1] ARMA staat voor 'autoregressive moving average'



Figuur 5: De berekening van het complexe cepstrum van een absoluut optelbaar signaal $s(k)$ bestaat uit drie stappen. Transformeer $s(k)$ naar het Z-domein, neem het (complexe) logaritme van de Z-transformatie, $S(z)$, en transformeer $\log S(z)$ terug naar het tijdsdomein.

Een interessant verband tussen de absoluut optelbare sequentie sequentie $s(k)$ en zijn complexe cepstrum $\bar{c}_s(k)$ is de volgende recursiebetrekking

$$s(k) = \frac{1}{k} \sum_{n=-\infty}^{\infty} n\bar{c}_s(n) s(k-n), \;\; k \neq 0 \; . \tag{0.3}$$

Als $s(k)$ de impulsrespons is van een stabiel ARMA-model, bekomen we zo het verband tussen de Markov-parameters van het model en zijn complex cepstrum (zie verder).

Het cepstrum van een ARMA-proces

We bekijken het uitgangsproces $\{y(k)\}$ van een stabiel en minimum-fase ARMA-model met polen $\alpha_1, \ldots, \alpha_p$ en nullen $\beta_1, \ldots, \beta_q$, waarbij de ingang van het model witte ruis is met variantie σ^2. Het is gemakkelijk aan te tonen dat het vermogenspectrum van het ARMA-proces $\{y(k)\}$ gelijk is aan de volgende even functie

$$c_y(k) = \begin{cases} \log \sigma^2 & k = 0 \; , \\ \displaystyle\sum_{i=1}^{p} \frac{\alpha_i^{|k|}}{|k|} - \sum_{i=1}^{q} \frac{\beta_i^{|k|}}{|k|} & k \neq 0 \; . \end{cases}$$

Het complexe cepstrum van de impulsrespons $h(k)$ van het model is niet even. De waarden voor $k \leq 0$ zijn gelijk aan 0, terwijl het even gedeelte van de functie ($k > 0$) gelijk is aan het even gedeelte van het vermogencepstrum van $\{y(k)\}$:

$$\bar{c}_h(k) = \begin{cases} 0 & k \leq 0 \; , \\ \displaystyle\sum_{i=1}^{p} \frac{\alpha_i^{k}}{k} - \sum_{i=1}^{q} \frac{\beta_i^{k}}{k} & k > 0 \; . \end{cases}$$

Hieruit volgt dat we het complexe cepstrum van een ARMA-model ook kunnen schatten aan de hand van observaties van het uitgangsproces met de methode in figuur 4. Bovendien bekomen we de volgende betrekking tussen de

Markov-parameters van het ARMA-model en het vermogencepstrum van het uitgangsproces door (0.3) toe te passen op de impulsrespons $h(k)$

$$h(k) = \frac{1}{k} \sum_{n=1}^{k} n c_y(n) h(k-n) \text{ voor } k > 0 \; .$$

Dit verband kan bijvoorbeeld gebruikt worden om een ARMA-model te identificeren uitgaande van (schattingen van) het cepstrum.

Hoofdstuk 6 Een cepstrale norm voor ARMA-modellen

Zoals uit figuur 2 blijkt, komen heel wat concepten, inzichten en resultaten uit de hoofdstukken 3, 4 en 5 samen in dit hoofdstuk. We bestuderen hier de cepstrale afstandsmaat voor ARMA-modellen die in [105] gedefinieerd werd door Martin.

Gegeven twee stabiele en minimum-fase ARMA-modellen met transferfuncties $H^{(1)}(z)$ en $H^{(2)}(z)$ en complexe cepstra[2] $\bar{c}^{(1)}(k)$ en $\bar{c}^{(2)}(k)$, dan is de gekwadrateerde afstand tussen de modellen gedefinieerd als

$$d(\log H^{(1)}, \log H^{(2)})^2 = \sum_{k=1}^{\infty} k \left(\bar{c}^{(1)}(k) - \bar{c}^{(2)}(k) \right)^2 \; . \tag{0.4}$$

Vooraleer we een systeemtheoretische en informatietheoretische interpretatie afleiden voor de afstand (zie figuur 1), richten we eerst onze aandacht op de corresponderende norm.

Een cepstrale norm voor ARMA-modellen

Uit de definitie van de cepstrale afstand halen we onmiddellijk de definitie van de bijhorende norm:

$$\| \log H \|^2 = \sum_{k=1}^{\infty} k \bar{c}(k)^2 \; ,$$

voor een stabiel en minimum-fase ARMA-model met transferfunctie $H(z)$ en complex cepstrum $\bar{c}(k)$. Gebruiken we de uitdrukking van het complexe cepstrum in functie van de polen en nullen van het model en de reeksontwikkeling

[2] We noemen het complexe cepstrum van de impulsrespons van een ARMA model ook kortweg het complexe cepstrum van het model

$\sum_{k=1}^{\infty} \frac{x^k}{k} = -\log(1-x)$ voor $|x| < 1$, dan vinden we dat de norm gelijk is aan

$$\|\log H\|^2 = \log \frac{\displaystyle\prod_{i=1}^{p}\prod_{j=1}^{q} \left|1-\alpha_i\bar{\beta}_j\right|^2}{\displaystyle\prod_{i,j=1}^{p}(1-\alpha_i\bar{\alpha}_j)\prod_{i,j=1}^{q}(1-\beta_i\bar{\beta}_j)} ,$$

waar $\alpha_1, \ldots, \alpha_p$ de polen zijn van het model en $\beta_1, \ldots, \beta_q$ de nullen.
Voor een model van orde 1 met pool α en nul β geeft dit

$$\|\log H\|^2 = \log \frac{(1-\alpha\beta)^2}{(1-\alpha^2)(1-\beta^2)} ,$$

wat gelijk is aan $-\log\cos^2\theta$, met θ de hoek tussen de vectoren $\begin{pmatrix}1 & \alpha & \alpha^2 & \cdots\end{pmatrix}$ en $\begin{pmatrix}1 & \beta & \beta^2 & \cdots\end{pmatrix}$. Dit inzicht is te veralgemenen voor een model met grotere orde door de canonieke hoeken tussen twee bepaalde deelruimten te bekijken.

Eigenschap 2. De cepstrale norm van een ARMA-model in termen van canonieke hoeken
De cepstrale norm van het stabiele en minimum-fase ARMA-model van orde n met transferfunctie $H(z) = \frac{b(z)}{a(z)}$ kan als volgt uitgedrukt worden in functie van de canonieke hoeken $\theta_1, \ldots, \theta_n$ tussen de rijruimten van de oneindige controleerbaarheidsmatrices $\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)$ en $\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)$

$$\|\log H\|^2 = -\log\prod_{k=1}^{n}\cos^2\theta_k .$$

Omdat deze matrices gelijk zijn aan de controleerbaarheidsmatrices van respectievelijk het model en het inverse model, zijn door eigenschap 1 de hoeken $\theta_1, \ldots, \theta_n$ gelijk aan de grootste n hoeken tussen de toekomstige ingangs- en uitgangsruimte van het model. Bovendien hebben we gezien dat die hoeken complementair zijn aan de hoeken $\phi_1, \ldots, \phi_n$ tussen de verleden en toekomstige uitgangsruimten (zie vergelijking (0.1)). Het gevolg is dat de cepstrale norm ook gelijk is aan

$$\|\log H\|^2 = -\log\prod_{k=1}^{n}\sin^2\phi_k ,$$

en dit is exact tweemaal de wederzijdse informatie tussen het verleden en toekomstige uitgangsproces (zie (0.2)), op voorwaarde dat de uitgang een Gaussisch proces is (of dus de ingang Gaussische witte ruis).
De cepstrale norm kan ook in functie van Gramianen uitgedrukt worden:

$$\begin{aligned}\|\log H\|^2 &= = \log\det(I_n + Q_z P) ,\\ &= -\log\frac{(\det R)^2}{\det P \det Q_z} ,\end{aligned}$$

waarin P de controleerbaarheidsgramiaan is van het model, Q_z de observeerbaarheidsgramiaan van het inverse model en R het product van $\mathcal{C}_\infty$, de oneindige controleerbaarheidsmatrix van het model, en Γ_{z_∞}, de oneindige observeerbaarheidsmatrix van het inverse model. Als het ARMA-model beschreven wordt door het trio (A, K, C), dan kunnen de matrices P, Q_z en R berekend worden als de oplossing van volgende Lyapunov-vergelijking

$$\begin{pmatrix} P & R \\ R^T & Q_z \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & (A-KC)^T \end{pmatrix} \begin{pmatrix} P & R \\ R^T & Q_z \end{pmatrix} \begin{pmatrix} A^T & 0 \\ 0 & A-KC \end{pmatrix} + \begin{pmatrix} K \\ -C^T \end{pmatrix} \begin{pmatrix} K^T & -C \end{pmatrix} .$$

We geven ook een grafische interpretatie aan de cepstrale norm. De norm is namelijk gelijk is aan de Hilbert–Schmidt-norm van de Hankel-matrix met de cepstrale coëffiënten $\bar{c}(1), \bar{c}(2), \ldots$ Hanzon geeft in [65] een grafische interpretatie van de Hilbert–Schmidt-norm van de Hankel-matrix met de Markov-parameters. Zijn resultaat vertalen naar de cepstrumcoëfficiënten geeft volgende interpretatie. De cepstrale norm in het kwadraat van een stabiel minimum-fase ARMA-model met transferfunctie $H(z)$ is gelijk aan $\frac{1}{\pi}A(\delta)$, waarbij $A(\delta)$ de oppervlakte is die ingesloten wordt door de curve $\delta(\theta) = \log H(e^{i\theta})$, met $\theta \in [0, 2\pi)$ lopend van 0 tot 2π.

De cepstrale afstand tussen twee ARMA-modellen

Er kunnen ook uitdrukkingen afgeleid worden voor de cepstrale afstand tussen twee ARMA-modellen. We werken met de stabiele en minimum-fase ARMA-modellen met transferfunctie $H^{(1)}(z) = \frac{b^{(1)}(z)}{a^{(1)}(z)}$ en $H^{(2)}(z) = \frac{b^{(2)}(z)}{a^{(2)}(z)}$ respectievelijk. De cepstrale afstand tussen deze modellen, zoals gedefinieerd in (0.4), is gelijk aan de cepstrale norm van het stabiel en minimum-fase ARMA-model met transferfunctie

$$H(z) = \frac{H^{(1)}(z)}{H^{(2)}(z)} = \frac{b^{(1)}(z)a^{(2)}(z)}{a^{(1)}(z)b^{(2)}(z)} .$$

Door nu de uitdrukkingen voor de cepstrale norm toe te passen op dit ARMA-model, bekomen we onmiddellijk de informatietheoretische en systeemtheoretische interpretaties van de afstandsmaat.

Hoofdstuk 7 Besluiten en suggesties voor verder onderzoek

Besluiten

In dit proefschrift hebben we verbanden aangetoond tussen systeemtheorie, informatietheorie en signaalverwerking, waarbij we vooral steunen op geometrische inzichten die we bekomen door de canonieke hoeken tussen verschillende paren deelruimten te berekenen.

In hoofdstuk 2 hebben we de definitie van de canonieke hoeken tussen en corresponderende canonieke vectoren in twee deelruimten gegeven, we hebben getoond hoe ze berekend kunnen worden (via de singulierewaardenontbinding, de eigenwaardenontbinding, de LQ-ontbinding) en hoe ze gerelateerd zijn aan orthogonale projecties.
Verder hebben we beschreven hoe de canonieke correlaties van twee random variabelen en van twee stochastische processen gezien kunnen worden als de cosinussen van de canonieke hoeken tussen twee deelruimten.

We hebben vooral aandacht besteed aan stationaire, ergodische stochastische processen die kunnen gedacht worden als de uitgang van een lineair tijdinvariant stochastisch model. Daarom hebben we in hoofdstuk 3 een grondige beschrijving gegeven van deze modellen, met hun verschillende representaties. Voor de modellen met slechts één ingang en één uitgang (SISO-modellen), hebben we enkele nuttige eigenschappen afgeleid.
Daarnaast hebben we in hoofdstuk 3 uitdrukkingen bekomen voor de canonieke hoeken en canonieke correlaties tussen de verschillende verleden en toekomstige ingangs- uitgangsparen. Zoals we in de voorbeelden getoond hebben, kunnen de hoeken ook geschat worden uitgaande van observaties van de stochastische processen, zonder dat een model voor de processen gekend is. Deze hoeken zijn de canonieke hoeken tussen de rijruimten van $mi \times j$ blok-Hankel-matrices met de observaties. Dat is de niet-parametrische benadering. Door echter de hoeken uit te drukken in functie van de modelparameters, bekomt men de exacte waarden voor eindige i en $j \to \infty$, waaruit de formules voor de canonieke correlaties van de verschillende processen zijn afgeleid $(i, j \to \infty)$.
Voor de SISO-modellen hebben we vervolgens de uitdrukkingen vereenvoudigd.

- We hebben aangetoond dat de canonieke hoeken en canonieke correlaties van de verschillende combinaties van verleden en toekomstige ingangs- en uitgangsprocessen onafhankelijk zijn van de ingangsvariantie.

- We hebben bewezen dat de kleinste n canonieke correlaties van de toekomstige ingang en de toekomstige uitgang van een $n$de orde SISO-model gelijk zijn aan de cosinussen van de canonieke hoeken tussen de rijruimte van de controleerbaarheidsmatrix van het model en de rijruimte van de controleerbaarheidsmatrix van het inverse model. De andere canonieke correlaties zijn gelijk aan 1.

- Gebaseerd op de laatste eigenschap hebben we een conjectuur geformuleerd i.v.m. de eigenwaarden van matrices die bekomen worden door een Lyapunov-vergelijking op te lossen.

- We hebben aangetoond dat de canonieke hoeken tussen de rijruimten van de verleden en toekomstige $i \times j$ Hankel-matrices van een autoregressief proces onafhankelijk zijn van i als $j \to \infty$ en $i \geq n$.

In hoofdstuk 4 hebben we vooreerst de nodige achtergrondinformatie gegeven over de informatietheoretische begrippen entropie en wederzijdse informatie. De wederzijdse informatie van twee normaal verdeelde random variabelen kan uitgedrukt worden in functie van hun canonieke correlaties. Op dezelfde manier hebben we de wederzijdse informatie van twee Gaussische stochastische processen uitgedrukt in functie van hun canonieke correlaties.
We hebben ons vooral toegelegd op stationaire stochastische processen die gemodelleerd kunnen worden als de uitgang van een lineair, tijdinvariant stochastisch model. De resultaten uit hoofdstuk 3, namelijk de uitdrukkingen voor de canonieke correlaties van de verschillende paren in- en uitgangsprocessen, hebben we dan aangewend om de wederzijdse informatie van de processen te schrijven in functie van de modelparameters.

Via het voorbeeld van echodetectie en echoverwijdering hebben we in hoofdstuk 5 het cepstrum gemotiveerd. Zowel het vermogencepstrum als het complexe cepstrum werden aangebracht, met de belangrijkste eigenschappen die in hoofdstuk 6 van pas komen, o.a. het vermogencepstrum van een stabiel en minimum-fase ARMA-proces in functie van de polen en nullen en het verband tussen het vermogencepstrum van een ARMA-proces en het complexe cepstrum van de impulsrespons van het model. Daarnaast hebben we kort aangegeven hoe het cepstrum gebruikt kan worden voor de identificatie van een lineair model. Vooral het feit dat de nullen van het model gemakkelijker kunnen bekomen worden uit het cepstrum dan uitgaande van de autocovariantiefunctie, lijkt veelbelovend.

Het cepstrum, canonieke hoeken en correlaties, de wederzijdse informatie van twee stochastische processen, controleerbaarheids- en observeerbaarheidsmatrices, Gramianen, Lyapunov-vergelijkingen en de Hilbert–Schmidt–Hankel-norm, al deze begrippen hebben we in hoofdstuk 6 toegepast om verschillende interpretaties van een cepstrale norm voor stabiele en minimum-fase ARMA-modellen af te leiden.

- Via de uitdrukking van het cepstrum in functie van de polen en de nullen van het model, hebben we de norm ook in functie van deze modelparameters geschreven.
- We hebben aangetoond dat de cepstrale norm kan uitgedrukt worden als een functie van de canonieke hoeken tussen de rijruimte van de oneindige controleerbaarheidsmatrix van het model en de rijruimte van de oneindige controleerbaarheidsmatrix van het inverse model.

- In hoofdstuk 3 hebben we bewezen dat de cosinussen van deze hoeken gelijk zijn aan de kleinste n (n is de orde van het model) canonieke correlaties van het toekomstige ingangs- en uitgangsproces van het model. We kunnen de cepstrale norm dus ook zien als een functie van deze canonieke correlaties. De uitdrukkingen voor de canonieke correlaties in functie van de modelparameters, zoals we die in hoofdstuk 3 hebben afgeleid, hebben we ook op de cepstrale norm toegepast.

- Via de meetkundige verbanden van hoofdstuk 3 (de canonieke hoeken tussen toekomstige in- en uitgangsruimten zijn de complementen van de canonieke hoeken tussen de verleden en toekomstige uitgangsruimten) hebben we aangetoond dat het kwadraat van de cepstrale norm gelijk is aan het dubbel van de wederzijdse informatie van het verleden en het toekomstige uitgangsproces.

- Door in de cepstrale norm de Hilbert–Schmidt-norm van de Hankelmatrix met de cepstrale coëfficiënten te herkennen, hebben we de cepstrale norm ook een grafische interpretatie kunnen geven. We hebben bewezen dat de oppervlakte ingesloten door het polaire diagram van het logaritme van de transferfunctie, gelijk is aan π keer de cepstrale norm in het kwadraat.

Aan de cepstrale afstand tussen ARMA-modellen, zoals gedefinieerd in [105], hebben we via de uitdrukkingen voor de norm ook een systeemtheoretische en informatietheoretische interpretatie gegeven. Verder hebben we een overzicht gegeven van de manieren waarop de norm en de afstand kunnen berkend worden, waarbij zowel parametrische als niet-parametrische technieken behandeld zijn. Ook hebben we uitgelegd hoe de afstand tussen een model en een signaal berekend kan worden.

Verder onderzoek

We suggereren tot slot mogelijke onderwerpen voor verder onderzoek en vermelden enkele open problemen.
De integratie van informatietheorie in systeemtheorie, meer bepaald systeemidentificatie en regeltheorie, staat nog maar in haar kinderschoenen. We denken dat in die richting nog heel wat mogelijkheden voor verder onderzoek kunnen gevonden worden. Stoorvogel en van Schuppen hebben in [142] slechts de aanzet gegeven. We denken dat het bijvoorbeeld mogelijk moet zijn om te bepalen welke kostfunctie door deelruimte-identificatiemethoden wordt geminimaliseerd. Informatietheorie zou hierbij kunnen helpen, zoals de pogingen in [143, 144] tonen.
Een grondige statistische analyse van de canonieke hoeken die we gebruiken en de cepstrale afstandsmaat is nodig om deze concepten toe te passen in bijvoorbeeld foutdetectie en technieken voor de schatting van de modelorde.

In hoofdstuk 6 hebben we verschillende representaties en berekeningswijzen voorgesteld voor de cepstrale norm en cepstrale afstandsmaat. Een vergelijkende studie zou kunnen uitwijzen welke methode het betrouwbaarste resultaat geeft en welke methode snelst is.
De cepstrale afstandsmaat die in hoofdstuk 6 wordt bestudeerd, is enkel gedefinieerd voor stochastische modellen met slechts één uitgang. Kan een gelijkaardige afstandsmaat, met dezelfde interessante interpretaties, gedefinieerd worden voor deterministische modellen en modellen met meerdere ingangen en uitgangen?
Zoals we in hoofdstuk 5 kort hebben aangetoond, kunnen de polen en de nullen van een fractioneel model geïdentificeerd worden vertrekkend van de cepstrumcoëfficiënten. Deze ideeën moeten nog uitgewerkt worden tot een goed werkend algoritme dat toepasbaar is op realistische problemen.
We hebben niet veel aandacht besteed aan de concepten in Appendix B. De verbanden met de ideeën in dit proefschrift moeten verder uitgediept worden.
Uiteraard zijn we nog op zoek naar een bewijs voor Conjectuur 3.13. Een interpretatie van dit resultaat en mogelijke gevolgen moeten ook nog bestudeerd worden.

Chapter 1

Introduction

System theory, information theory and signal processing: these are the three pillars of this thesis and our main working tool is the geometric notion of principal angles between linear subspaces. By means of the principal angles we establish and reveal interesting connections between system theory, information theory and signal processing. For that purpose, we rely upon research results in many different areas. The most important disciplines are shown in Figure 1.1. More in particular, we have been highly inspired by the work of Van Overschee and De Moor on subspace identification methods [156, 157, 158, 159] and the papers by Basseville et al. on subspace based fault detection [7, 8] and distance measures for signal processing [6], while our developments were accel-

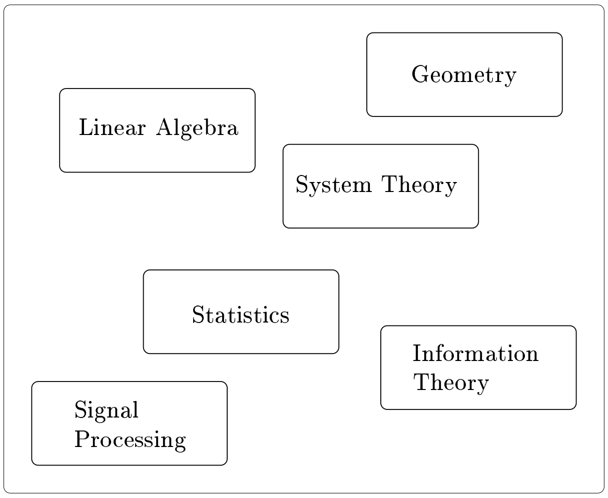


Figure 1.1: Our work draws upon several disciplines.

erated by the results of Martin [105]. We acquired background information on information theory from [29, 52, 123] and Shannon's seminal paper [130] and on cepstral signal processing from [17, 110, 113] and the recent work of Enqvist et al. [23, 47]. Of course, all of this is based on the work of many more persons, a few of whom will be named in the historical notes of Section 1.4.
The multidisciplinary character of the thesis is also illustrated in Figure 1.2, where a number of concepts that we will deal with, are enumerated.

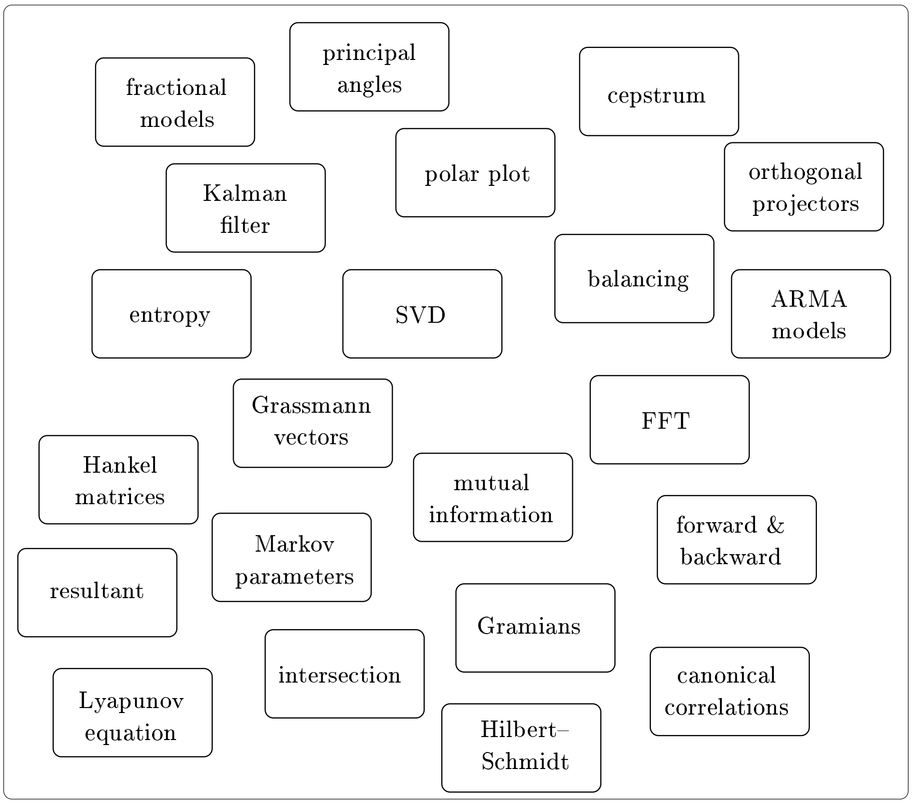


Figure 1.2: Notions that will be encountered throughout this thesis.

Overview of this introductory chapter
In Section 1.1 we explain our approach: by computing the principal angles between subspaces we highlight existing and establish new aspects of some relationships between system theory, information theory and signal processing. Many of the methods that we discuss throughout the thesis, can directly be applied to measurement data. Furthermore, we also derive exact expressions for the asymptotic values, which would be attained as the number of data goes to infinity. These expressions rely on the parameters of a presumed underlying

model. In Section 1.2 we classify many elements that are treated in the thesis, into, on the one hand, a data-driven or non-parametric group and on the other hand a model-driven or parametric group, while we also show how they are related to each other.
In Section 1.3 we enumerate our main contributions and in Section 1.4 a few historical notes are made. Finally, in Section 1.5 a chapter by chapter overview of the thesis is given.

1.1 Links between system theory, information theory and signal processing in this thesis

A schematic representation of our approach and results is given in Figure 1.3. Centrally located are the principal angles and their statistical counterparts,

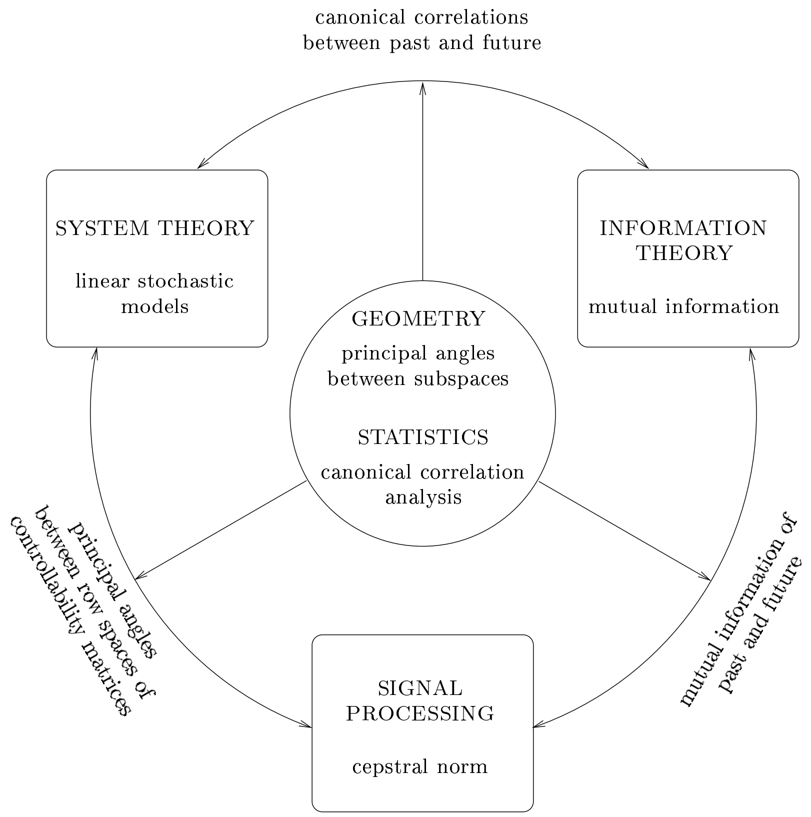


Figure 1.3: A schematic representation of the relations between system theory, information theory and signal processing that are treated in the thesis.

the canonical correlations, around which the three disciplines, system theory, information theory and signal processing are placed.
From **system theory** we use the linear stochastic models and their properties. The one concept of **information theory** that we are mostly interested in, is the mutual information of two stochastic processes. The cepstrum is a concept developed in **signal processing** that is relatively unknown to system theorists. We are in particular interested in a cepstral distance measure for linear models and its related cepstral norm.
By means of the geometric tools of principal angles between subspaces, we establish the following links between the three areas.

(1) **A link between system theory and information theory: the canonical correlations between the past and the future.**
We compute the principal angles between the different combinations of past and future input and output block Hankel spaces of a linear stochastic model. By using the properties of linear stochastic models, we express these angles in terms of the model parameters. The asymptotic values, as the number of observations goes to infinity, are equal to the canonical correlations of the respective processes. Hence, we have expressions for the canonical correlations as a function of the model parameters, from which we derive relations between the different canonical correlations.
Since the canonical correlations of Gaussian processes are related to the mutual information, we also obtain formulas for the mutual information in terms of the model parameters. The mutual information of the past and the future of the output process is particularly interesting, since it can be estimated from given output observations, without having access to the input process or knowledge of the model parameters.

(2) **A link between system theory and signal processing: the principal angles between row spaces of controllability matrices.**
The cepstrum is an appropriate tool in signal processing for the definition of distance measures between signals that only depend on the ratio of the signals' spectra. We show how one of the cepstral distances between two single input single output (SISO) linear stochastic models can be interpreted in system theoretic terms.
More specifically, we show that the related cepstral norm of an nth order model is a function of the n principal angles between the row spaces of the controllability matrices of the model and the inverse model (or the column spaces of their observability matrices).
We furthermore prove that the cosines of these angles are the n smallest canonical correlations of the future input and future output processes of the model. The other canonical correlations are equal to 1. Consequently, we can express the cepstral norm in terms of these canonical correlations. Based on these results and on the properties of linear stochastic models we derive several expressions of the cepstral norm as a function of observability and controllability Gramians, which are obtained by solving

Lyapunov equations.
As a side result, we obtain, by equating two different descriptions for the canonical correlations of the future input and output, a property for Lyapunov equations.
Note that from the properties of the cepstral norm, the corresponding properties of the cepstral distance can directly be derived.

(3) **A link between information theory and signal processing: the mutual information of the past and the future.**
Via the geometric insights that we have of linear stochastic models, we can go one step further. Using the relations between the different canonical correlations, discussed in item (1), we prove that the cepstral norm is tightly related to the information theoretic notion of mutual information, if Gaussian processes are considered. In fact, the squared cepstral norm is equal to twice the mutual information of the past and the future. Since the cepstral norm is a measure for the distance to white noise, we can also give this interpretation to the mutual information of the past and the future.

We also want to mention the following remarkable, but rather straightforward similarity between the geometry of linear subspaces and the concept of mutual information in information theory. Both the dimension theorem of Grassmann, which tells us how many principal angles between two subspaces are zero, and the definition of the mutual information of two probability measures, are based on this simple formula for the number of elements in the union of two finite sets A and B:

$$\#(A \cup B) = \#(A) + \#(B) - \#(A \cap B) \ ,$$

where $\#(A)$ denotes the cardinal number, i.e. the number of elements, of the set A. This relation is illustrated in Figure 1.4. In the dimension theorem of Grassmann, the sets A and B represent linear subspaces S_1 and S_2, so that we obtain

$$\dim(S_1 \cup S_2) = \dim(S_1) + \dim(S_2) - \dim(S_1 \cap S_2) \ ,$$

or with S_1 and S_2 the row spaces of the matrices A and B respectively[1],

$$\operatorname{rank}\begin{pmatrix} A \\ B \end{pmatrix} = \operatorname{rank}(A) + \operatorname{rank}(B) - \dim\left(\operatorname{row}(A) \cap \operatorname{row}(B)\right) \ .$$

In information theoretic terms, the area of the sets is viewed as the amount of information (entropy) that two random variables X and Y offer.[2] A non-empty intersection of the two sets indicates that the random variables are statistically dependent. They have information in common, which is the so-called mutual information of X and Y, denoted by $I(X;Y)$. The total information that is

[1] The row space of a matrix is defined as the span of its rows.
[2] We assume discrete-valued random variables with a finite alphabet, which ensures a finite entropy, see Chapter 4.

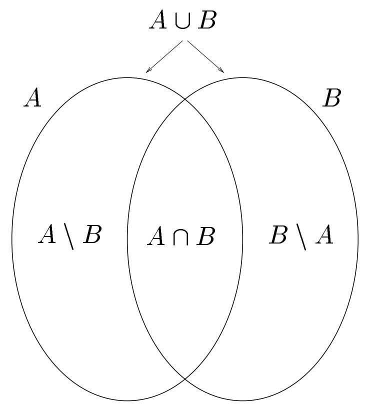


Figure 1.4: Venn-diagram to illustrate the similarity of the geometrical theorem of Grassmann, and the definition of the information theoretic notion of mutual information. The intersection of the sets represents the intersection of two linear subspaces in Grassmann's dimension theorem, while in information theory it stands for the information that two random variables have in common: their mutual information.

obtained by observing both X and Y, i.e. the joint entropy $H(X,Y)$, can be written as

$$H(X,Y) = H(X) + H(Y) - I(X;Y) \ ,$$

where $H(X)$ is the entropy of X and similarly for $H(Y)$.

1.2 Parametric versus non-parametric methods

The results derived and discussed in this thesis can also be categorized in another way. Most theoretical results rely on the underlying linear time invariant or state space model. This is the parametric or model-driven approach. However, at several occasions, we show how to estimate the quantities from observations only, i.e. without knowing a parametric description of the underlying system. This is the non-parametric or data-driven approach. Let us clarify the distinction by discussing a few examples, illustrated by Figure 1.5.
In Chapter 3 we discuss the principal angles between the past and future input and output spaces of a stochastic model. When measurements of the processes are given, the angles can be computed with methods described in Chapter 2 (non-parametric). Hence we have put 'principal angles' in the non-parametric set of Figure 1.5. On the other hand, if we want to know the canonical correlations of the processes, an infinite amount of measurements would be necessary. But if the model parameters are known, closed form expressions for the canonical correlations in terms of these parameters can be obtained (parametric).

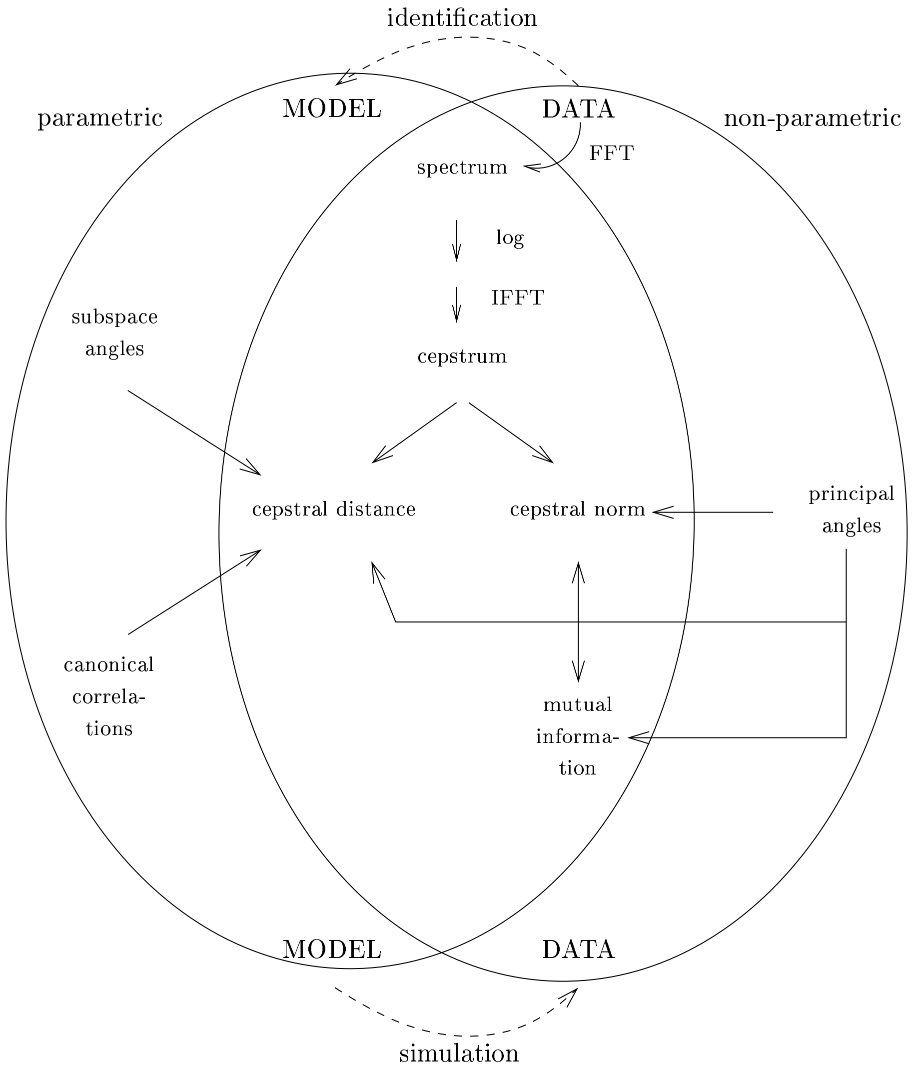


Figure 1.5: Parametric and non-parametric approaches for the computation or estimation of the quantities mentioned. The elements of the intersection can be obtained by both parametric as non-parametric methods.

In [36, 37, 39] we have defined the subspace angles between two models. They are determined by the model parameters and can therefore be found in the parametric set.
A well-known example in signal processing of a quantity for which parametric as well as non-parametric estimation methods exist, is the power spectral density (Chapter 3) and by extension the power cepstrum (Chapter 5) of a stochastic

process. Consequently, the spectrum and cepstrum lie in the intersection of the parametric and non-parametric set of Figure 1.5. A non-parametric estimate of the spectrum can be obtained via smoothing of the periodogram, which is the squared absolute value of the Fourier transform of the observed sequence. In the parametric approach, one first identifies a model from the observations and the spectrum of the model serves as estimate for the spectrum of the process. In Chapter 6 finally, we give several expressions for the cepstral norm of an ARMA process and the cepstral distance between two ARMA processes. Some of them are in terms of the model parameters (parametric), others are based on observations of the process (non-parametric), see Figure 1.5. For Gaussian stochastic processes, the mutual information of the past and the future is proportional to the cepstral norm and can thus be estimated in the same way as the cepstral norm. For this reason, also the mutual information is an element of the intersection.

1.3 Contributions

Our main contributions are illustrated in the diagrams of Figure 1.6, Figure 1.7 and Figure 1.8, which we now briefly discuss.

1.3.1 Parametric expressions for principal angles and canonical correlations (Figure 1.6)

Starting from the state space equations of a linear stochastic model in forward innovation form, we derive expressions for the principal angles between the row spaces of the $mi \times j$ input and output block Hankel matrices, where i is finite and $j \to \infty$. The parameter m denotes the number of outputs of the model, i is the number of block rows in the block Hankel matrices and j the number of columns. The order of the model will be denoted by n.
Although mi angles are defined between the respective subspaces, only n turn out to be non-trivial, i.e. they are not equal to 0 or to $\frac{\pi}{2}$. The cosines of these n angles can be obtained as eigenvalues of matrices derived from the solution of two Lyapunov equations. The expressions for the cosines depend on the parameter i. By taking the limit for $i \to \infty$ in these expressions, the cosines become equal to the canonical correlations of the corresponding processes.
The matrix expressions for the canonical correlations immediately reveal the complementarity of on the one hand the canonical correlations of the future input and future output and on the other hand the canonical correlations of the past and future output.
Furthermore, since the mutual information of the past and future of a Gaussian process only depends on the canonical correlations of the past and future, we obtain parametric formulas for this mutual information.

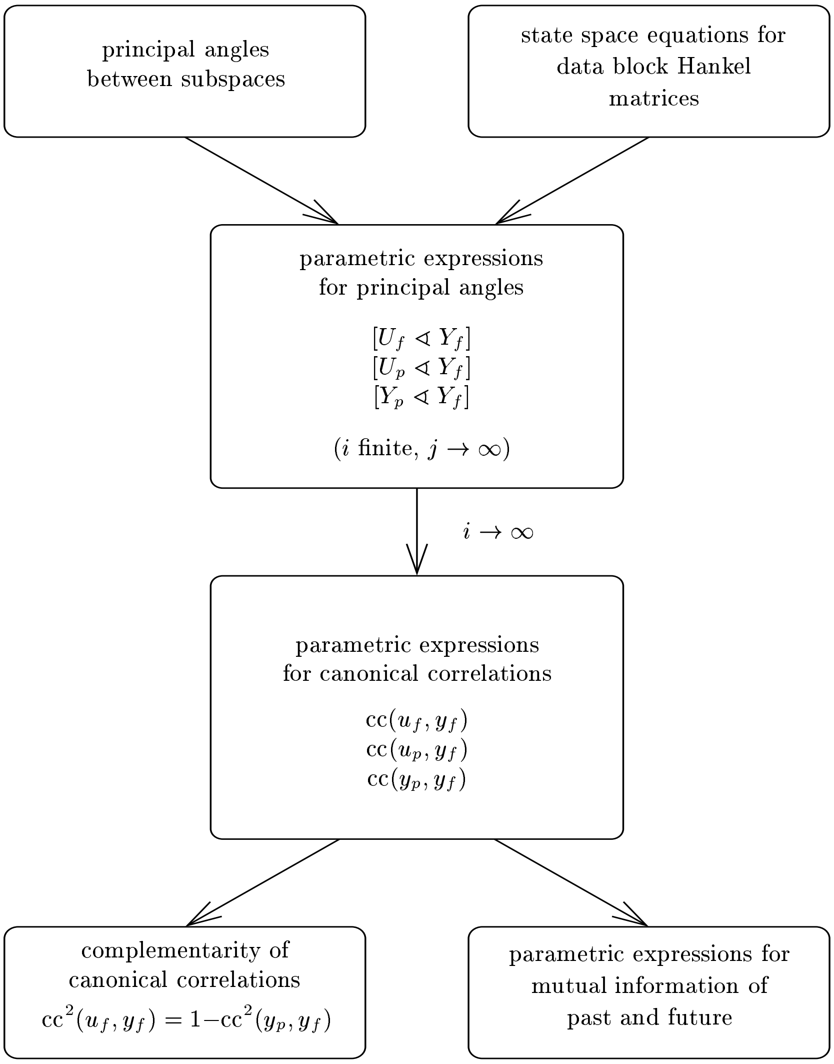


Figure 1.6: Based on the state space equations of a stochastic linear model in forward innovation form, we derive parametric expressions for the principal angles between the row spaces of input and output block Hankel matrices and for the canonical correlations of the corresponding processes. From these expressions follows a complementarity property and expressions for the mutual information of the past and the future of a Gaussian process.

1.3.2 New results for single input single output (SISO) models (Figure 1.7)

It is easy to prove that for single input single output (SISO) models, the results described above are independent of the variance of the input noise. A more canonical correlations of two AR processes

$$\mathrm{cc}(y^{(1)}, y^{(2)}) = \cos\left[\mathcal{C}_\infty(H^{(1)}) \sphericalangle \mathcal{C}_\infty(H^{(2)})\right], 1, 1, \ldots$$

canonical correlations of two ARMA processes

$$\mathrm{cc}(y^{(1)}, y^{(2)}) = \cos\left[\mathcal{C}_\infty(H^{(1)}H^{(2)^{-1}}) \sphericalangle \mathcal{C}_\infty(H^{(2)}H^{(1)^{-1}})\right], 1, 1, \ldots$$

canonical correlations of future input and output of ARMA model

$$\mathrm{cc}(u_f, y_f) = \cos\left[\mathcal{C}_\infty(H) \sphericalangle \mathcal{C}_\infty(H^{-1})\right], 1, 1, \ldots$$

Figure 1.7: By first considering the principal angles between the future output spaces of two AR models that are driven by the same white input process, we derive the canonical correlations of the future output processes of two ARMA models that are driven by the same input process. This leads to a new characterization of the canonical correlations of the future input and future output of a SISO ARMA model.

interesting and less obvious contribution is obtained as depicted in Figure 1.7. We give a new characterization for the canonical correlations of the future input and output process of a SISO autoregressive moving average (ARMA) model. This is obtained in the following way.

1. We show that the smallest n canonical correlations of the future output processes of two nth order autoregressive (AR) models that are driven by the same white noise process, are equal to the cosines of the principal angles between the row spaces of their infinite controllability matrices. The other canonical correlations are equal to 1. In Figure 1.7, the transfer functions of the AR models are denoted by $H^{(1)}(z)$ and $H^{(2)}(z)$ respectively.

2. When considering the future output processes of ARMA models that have the same input, the canonical correlations are related to the model parameters as follows. Let $H^{(1)}(z)$ denote the transfer function of the first model of order $n^{(1)}$ and $H^{(2)}(z)$ the transfer function of the second model of order $n^{(2)}$. Consider $H(z) = H^{(1)}(z)H^{(2)^{-1}}(z)$. We then prove that the smallest $n^{(1)} + n^{(2)}$ canonical correlations of the future output processes are equal to the cosines of the principal angles between the row spaces of the controllability matrix of $H(z)$ and the controllability matrix of the inverse transfer function $H^{-1}(z)$. The other canonical correlations are equal to 1.

3. Based on the previous properties, we prove that the n smallest canonical correlations of the future input and future output process of an nth order SISO autoregressive moving average (ARMA) model with transfer function $H(z)$ are equal to the cosines of the principal angles between the row spaces of the controllability matrix of the model and the controllability matrix of the inverse model $H^{(-1)}(z)$. The other canonical correlations are equal to 1.

The principal angles between the controllability matrices can be viewed as measures for the distance of the models under consideration. Therefore, we have called them in [36, 37, 38, 39] the subspace angles between the models.
We mention two consequences of these results.
First, by equating the parametric descriptions for the canonical correlations of the future input and output processes of a SISO ARMA model obtained in this section and Section 1.3.1, we have proven a new property for the eigenvalues of matrices that satisfy certain Lyapunov equations.
Secondly, we see that the canonical correlations of the processes do not depend on the difference of the transfer functions (or the difference of the power spectral densities), but on their ratio. Hence, when the transfer functions of both models are multiplied by the same third transfer function, then the canonical correlations remain the same. Since this is also a property of cepstral distance measures, this observation makes a nice link to the following group of contributions.

1.3.3 New expressions for a cepstral norm (Figure 1.8)

The cepstrum of a stochastic process is defined as the inverse Fourier transform of the logarithm of its spectrum. Cepstral distance measures for signals and models depend on the difference of the cepstra, and consequently on the ratio of the spectra and transfer functions. Since this is also the case for the canonical correlations of the future output processes of models that are driven by the same input process (see the last paragraph in Section 1.3.2), cepstral distances seem to be related to the canonical correlations. Indeed, we consider one particular cepstral distance and prove that it can be expressed in terms of cepstral norm

$$\|\log H\|^2 = \sum k\bar{c}^2(k)$$

poles and zeros
$H(z) = \prod \frac{(z-\beta_k)}{(z-\alpha_k)}$

$$\|\log H\|^2 = \log \prod \frac{\left|1-\alpha_i\bar{\beta}_j\right|^2}{(1-\alpha_i\bar{\alpha}_j)(1-\beta_i\bar{\beta}_j)}$$

$(\theta_1, \ldots, \theta_n) = [\mathcal{C}_\infty \sphericalangle \mathcal{C}_{z_\infty}]$

$$\|\log H\|^2 = -\log \prod \cos^2 \theta_k$$

$\mathrm{cc}(u_f, y_f) = \cos [\mathcal{C}_\infty \sphericalangle \mathcal{C}_{z_\infty}], 1, \ldots, 1$

$$\|\log H\|^2 = -\log \prod \tau_k^2$$

$\mathrm{cc}(y_p, y_f) = \sin [\mathcal{C}_\infty \sphericalangle \mathcal{C}_{z_\infty}], 1, \ldots, 1$

$$\|\log H\|^2 = -\log \prod (1-\rho_k^2)$$

$\{y(k)\}$ Gaussian
$I(y_p; y_f) = -\frac{1}{2} \log \prod_{k=1}^{n} (1-\rho_k^2)$

$$\|\log H\|^2 = 2I(y_p; y_f)$$

Figure 1.8: The different steps that lead to the equality, except for a factor two, of the squared cepstral norm of a SISO ARMA model and the mutual information of the past and the future of its Gaussian output process.

the canonical correlations.
For that purpose, we first study the related cepstral norm of an ARMA model. As indicated in Figure 1.8, we start from the expression of the cepstral norm in terms of the poles (α_k, $k = 1, \dots, n$) and zeros (β_k, $k = 1, \dots, n$) of the model. We subsequently prove that it is a function of the principal angles $\theta_1, \dots, \theta_n$ between the row spaces of the infinite controllability matrix of the model and the infinite controllability matrix of the inverse model. Since the cosines of these angles are the smallest n canonical correlations τ_k of the future input process and future output process of the model, we obtain the expression of the norm in terms of the canonical correlations. The complementarity of the canonical correlations of future input and output to the canonical correlations of the past and future output (ρ_k) leads to the expression of the squared cepstral norm in terms of the canonical correlations of past and future. This expression is exactly two times the mutual information of the past and the future of the output process, provided it is Gaussian.
We obtain formulas for the cepstral norm in terms of Gramians by using the parametric expressions for the canonical correlations that are discussed in Section 1.3.1 and Section 1.3.2. Furthermore, the cepstral norm can be seen as the Hilbert–Schmidt norm of an infinite Hankel matrix, which leads to a graphical interpretation of the norm. The squared cepstral norm of an ARMA model with transfer function $H(z)$ is proportional to the area enclosed by the polar plot of $\log H(z)$.
All these results for the cepstral norm of an ARMA model can be translated to the cepstral distance between two models. We also show how the cepstral norm and cepstral distance can be estimated from measurements. Estimating the distance between a given model and a measured signal can be done by mere Kalman filtering, i.e. without explicitly modeling the observed data. We obtain elegant formulas for this distance between a model and a signal, which can easily be updated. This is important for the applications that we envisage (see Remark 1.1).

Note that besides these new contributions, we try to give a didactical survey of the background material. Especially the Chapters 2, 4 and 5 have been written with this purpose in mind.

Remark 1.1. Applications
Although we do not explicitly treat applications in the thesis, we anticipate that the concepts and insights of this thesis offer a broad gamut of possible applications in different areas. In Figure 1.9 we enumerate the possibilities that we have in mind. We briefly discuss a few of them. Furthermore, Chapter 6 is concluded with a number of illustrative examples that indicate the applicability.

Monitoring and fault detection If at different time instants a model is identified for the same object, there will always be small or larger differences between the models. It is important to understand the causes

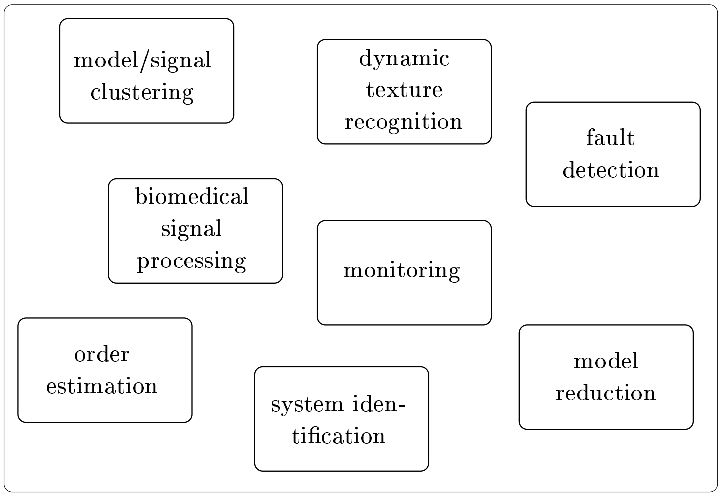


Figure 1.9: Possible applications of the concepts in this thesis.

of the differences. They can be uncertainties innate to the estimation method, a dysfunction of a sensor or real structural changes of the system.
Principal angles between subspaces are already being used for the detection of changes in finite element models [48, 49]. The subspace-based fault detection algorithm of Basseville et al. [7, 8], on the other hand, is based on linear dynamical models, i.e. the type of models that we deal with. Changes in the eigenmodes of the observed system are determined by hypothesis tests on residuals that describe the difference between the column spaces of the observability matrix of the nominal linear dynamical model and the observability matrix of the model that can be identified from the measurements. The difference between the column spaces could also be quantified by the principal angles between the subspaces. Since for a single input single output model the column space of the observability matrix is equal to the row space of its controllability matrix, it is clear that these are exactly the angles that we consider for AR models (see e.g. Figure 1.7). When dealing with ARMA models, we also take into account the zeros of the models. In Chapter 6 we consider a distance measure that is a function of the principal angles. We also show how it can be estimated from data. Monitoring the distance of measurements on the system to the nominal model, can be done by updating determinants of covariance matrices, for which easy numerical methods exist. An example of the application of our techniques to a monitoring problem is described in Example 6.6.

Clustering of models and signals Experimental evidence shows that the

modal parameters of constructions such as bridges are influenced by the environmental temperature [3, 116, 117, 125]. It is difficult to decide when e.g. a frequency change is due to little cracks in the bridge or to the temperature difference. To make a distinction between changing environmental factors as the temperature, and structural differences, one can determine clusters of models. Each cluster corresponds to e.g. a certain temperature range. Clustering is based on a certain distance measure. The distance measure that we study and which is based on principal angles between subspaces, might be a good candidate. When measurements are made, we will first compute their distance to the different clusters. If the data set ends up in the cluster corresponding to the temperature during measurement, then everything seems to be all right. When it falls outside this region, the system might be damaged.
Also in bio-informatics, problems show up where signals have to be classified. The clustering of genes, for example, is an important step in the analysis process of the interaction of genes. It is based on the expression behavior of genes, which is measured using micro-arrays.
Furthermore, based on our results in [36, 37], clustering has been performed to computer vision problems [15, 127].

Biomedical signal processing Some human impairments can be detected by two signals having a similar dynamical behavior. This is e.g. the case when the cerebrovascular autoregulation of premature babies is not functioning well, see [108, 109]. The subspace angles and distances that we study seem to be useful for determining the dynamic similarity of two signals. ○

1.4 Historical notes

Table 1.1 gives a survey of the persons whose name appears in theorems and concepts used in this thesis, or who are the founding fathers of one of the key notions that we apply. Most information was found in the MacTutor History of Mathematics Archive [103].

name	∘ – †	**related concept**	**page**
Euclid of Alexandria	-325 – -265	Euclidean norm	38
Jacopo Francesco Riccati	1676 – 1754	algebraic Riccati equation	71
Marc-Antoine Parseval des Chênes	1755 – 1836	Parseval's theorem	229

continued on next page

name	∘ †	related concept	page
Jean Baptiste Joseph Fourier	1768 – 1830	Fourier transform	64
Johann Carl Friedrich Gauss	1777 – 1855	Gaussian random variables	53
Augustin Louis Cauchy	1789 – 1857	Cauchy–Schwarz inequality	38
Hermann Günter Grassmann	1809 – 1877	dimension theorem of Grassmann, Grassmann vectors and coordinates	27 263
James Joseph Sylvester	1814 – 1897	Sylvester matrix, singular value decomposition	232 27
George Gabriel Stokes	1819 – 1903	theorem of Stokes	230
Marie Ennemond Camille Jordan	1838 – 1922	principal angles between subspaces, singular value decomposition	37 27
Hermann Hankel	1839 – 1873	Hankel matrix	59
Hermann Amandus Schwarz	1843 – 1921	Cauchy–Schwarz inequality	38
Jorgen Pedersen Gram	1850 – 1916	Gram–Schmidt orthogonalization, Gramian	48 74
Andrei Andreyevich Markov	1856 – 1922	Markov parameters, Markov modulated Poisson process	69 91
Aleksander Mikhailovich Lyapunov	1857 – 1918	Lyapunov equation	68
Eliakim Hastings Moore	1862 – 1932	Moore–Penrose inverse	32
David Hilbert	1862 – 1943	Hilbert–Schmidt norm	228
Erhard Schmidt	1876 – 1959	Gram–Schmidt orthogonalization Hilbert–Schmidt norm	48 228
Otto Toeplitz	1881 – 1940	Toeplitz matrix	59
Harry Nyquist	1889 – 1976	Nyquist diagram	229
Harold Hotelling	1895 – 1973	canonical correlation analysis	51
Hendrik W. Bode	1905 – 1982	Bode plot	196
Wallace Givens	1911 – 1993	Givens rotations	244

continued on next page

name	∘ †	related concept	page
John Wilder Tukey	1915 – 2000	Fast Fourier Transform, cepstrum	82 187
Claude Elwood Shannon	1916 – 2001	information theory	146
Rudolf Emil Kalman	1930 –	Kalman filter	69
Roger Penrose	1931 –	Moore–Penrose inverse	32

Table 1.1: Persons whose names are still alive in theorems, concepts and tools that are used in this thesis, or to whose work our results are much indebted (such as Jordan, Hotelling and Shannon). In the second column the year of birth and death are given, in the third column the concept with their name, and in the last column the page where it appears first.

We give a brief historical sketch of the main achievements on which our results are based and mention a few recent developments in the different research areas.

Linear estimation and numerical linear algebra[3] There is a long history of linear estimation problems, going back to Gauss at the beginning of the nineteenth century. Gauss developed least squares techniques for solving an overdetermined set of linear equations, see [51, 138]. Before the widespread use of matrices, he described his elimination algorithm in terms of determinants and bilinear and quadratic forms. This algorithm can be viewed as a precursor of the singular value decomposition (SVD), a technique in numerical linear algebra that we will apply throughout the thesis. Beltrami [12] and Jordan [75, 76] are the true progenitors of the SVD. Other contributions to the early development of the SVD come from Sylvester, Schmidt and Weyl, see [137], while more recent efforts, especially by Golub et al. [55, 56], have yielded efficient, stable algorithms to compute the SVD.
In 1875, a year after his article on the SVD, Jordan introduced the principal angles between linear subspaces [77]. This notion was translated in 1936 by Hotelling [68] into the statistical quantities of canonical correlations.
The least squares ideas were also cast into a stochastic framework. As for the deterministic case, a geometric formulation can be derived. The idea of regarding random variables as elements of a linear vector space apparently goes back to Fréchet in 1937. Wold [162] picked up this suggestion and noted that it would be convenient to represent a correlated set of random variables by an uncorrelated set, and that the Gram–Schmidt method provided a way of doing this. This gave birth to the innovation process. Kolmogorov developed the ideas further [88, 89, 90].

[3]For the notes on linear estimation we make use of the complements of Kailath et al. in [80], while the history of the singular value decomposition comes from Stewart [137]

System theory The breakthrough of vector-valued stationary process estimation came with the introduction of state space models and the work of Kalman in the early 1960s [82, 84, 85]. With these models, much of the analysis is indifferent to whether the processes are scalar- or vector-valued. Besides bringing the state space representation into the spotlights of system theory, and introducing the fundamental system theoretic notions of controllability, observability and minimality, Kalman's name especially remains attached to system theory and signal processing by the so-called Kalman filter. He introduced the filter in his famous paper 'A new approach to linear filtering and prediction problems' [83] as a recursive solution to the discrete-data linear filtering problem. The Kalman filter is one of the most well-known tools for stochastic estimation from noisy sensor measurements. Applications include the GPS (global positioning system), remote sensing, controlling the steam temperature in power plants, failure detection, ..., see [81].

Information theory In 1948, Shannon published his seminal paper, 'A Mathematical Theory of Communication' [130], giving birth to the new scientific discipline called information theory. The motivation was practical: how to efficiently transmit messages over noisy channels. To analyze this problem properly, he introduced the notion of entropy of a source, adapting a term from statistical mechanics, to quantify the source's information content. Furthermore, he defined the capacity of a channel, i.e. the highest rate of transmission of information for a channel with noise. His fundamental encoding theorems for which he gave outlines of proofs, were later rigorously proven by other mathematicians.
Kullback [91] emphasized the interrelation between information theory and statistics and introduced the divergence (the relative entropy or Kullback–Leibler distance) as a distance for probability measures. Gel'fand and Yaglom [52] generalized the notion of mutual information of random variables to stochastic processes. Shannon's ideas spread beyond the fields of communications engineering, taking root in computer science, cryptography, probability theory and even investment theory and biology [86, 87, 96, 129, 131].
Also in the domain of system identification, attempts have been made to use information theoretic criteria:

Maximum entropy models If the first $n + 1$ points of the autocovariance function of a Gaussian process are given, then, the model with these autocovariances that generates the maximum amount of entropy is an autoregressive (AR) model of order n. This was shown by Burg in [21].

Order estimation for linear models There exist several algorithms for the order estimation of linear, dynamical, state space systems. Akaike [1] has introduced the information criteria for maximum likelihood identification methods. These criteria make a trade-off between the model fit and the order of the model. The criterion proposed by Akaike (AIC) is based on a

minimization of the divergence between two probability measures, which reveals its relation to information theory.
For a second family of identification algorithms, the subspace identification methods (see e.g. [159]), there exist only few references dealing with the estimation of the model order. Three different approaches are presented and analyzed in [10], of which two are based on the canonical correlations of the past and the future output. These canonical correlations are related to the information theoretic concept of mutual information.

Model reduction Closely related to the order estimation problem is model reduction. Pal and Desai [42, 114] propose a model reduction technique that is also based on the canonical correlations of the past and the future, and is hence related to the mutual information of the past and the future. The subspace methods of Larimore [94] and more recently Dahlén et al. [30, 32, 104], incorporate this reduction step.
Stoorvogel and van Schuppen [143] prove that performing this reduction step to Gaussian random variables is equivalent to minimizing the divergence between the high-dimensional variable and the Gaussian variables of the lower dimension, which should be fixed beforehand. They assumed that the subspace identification methods based on this reduction method, might also minimize the divergence, but it turned out to be false hope [144].

System identification Stoorvogel and van Schuppen describe a different approach to joining system identification and information theory in [142]. The problem they tackle, is the approximation of a stationary Gaussian process by the output of a Gaussian system. To this end, they use approximation criteria that originate from information theory: the mutual information rate and the divergence rate. These criteria are new approximation measures in system identification, where one mainly uses the likelihood function, leading to a maximum likelihood method, or a quadratic cost function, which results in a least squares method.

Cepstral analysis Besides his contributions to pure mathematics and statistics (ANOVA, the jackknife) and many advisory tasks (see [19]), John Wilder Tukey introduced spectrum analysis to the engineering world. Together with Cooley he introduced in 1965 the Fast Fourier Transform algorithm (FFT) [28]. But before this, in 1963, he published together with Bogert and Healy a paper with the unusual title 'The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking', in which they introduced the cepstrum and showed that it is a useful tool for the detection of echoes in seismological data [17]. According to [19] the cepstrum was introduced to address the problem of distinguishing earthquakes from nuclear explosions. As can be seen from the title of their paper [17], Bogert, Healy and Tukey proposed a whole new terminology by anagrammatizing existing words.

Meanwhile, Oppenheim had been developing his theory of homomorphic systems [110], for which soon applications in signal processing were found, such as image enhancement [139], speech analysis [112] and blind deconvolution. A classic example of blind deconvolution is the restoration of the old phonograph recordings of Caruso [140], where the horn distortions are removed from the audio signal[4].
The cepstrum analysis as described by Bogert et al. turned out to be a special case of Oppenheim's homomorphic signal processing. Oppenheim takes the complex logarithm of the Z-transform of a signal, whereas Bogert et al. consider the logarithm of the power spectrum, which is a real function.
An interesting document to get to know the circumstances and spirit of the times in which these concepts were developed, is the interview with Oppenheim, conducted for the IEEE History Center [54].
Besides the above mentioned applications, the cepstrum is also used for machine diagnostics because of its ability to detect periodicities in the spectrum [161] and for measuring the distance between two signals, see [6] and references therein and [105].
Recently, Byrnes et al. [22, 23] and Enqvist [46, 47] showed that the cepstrum coefficients can be applied to find solutions for the rational covariance extension problem.

1.5 Chapter by chapter overview

A description of the contents by chapter follows. The relations between the different chapters are represented in Figure 1.10.

Chapter 2: Linear Algebraic, Geometric and Statistical Tools
In Chapter 2 we introduce the tools that are used throughout the thesis, the most important of which are the principal angles between and corresponding principal directions in two linear subspaces. However, for a good understanding of this geometric concept, we begin Chapter 2 with some linear algebraic notions, such as the singular value decomposition and orthogonal projectors, with emphasis on their geometric meaning and properties.
Next, we show how statistical properties like independence and correlation can be interpreted in a geometric framework. From this follows the connection of the canonical correlations between two random variables and the principal angles between two linear subspaces and the relation of the corresponding canonical variates and principal directions.

[4]The result of this restoration can be heard via the video of the Research Laboratory of Electronics of the MIT, available on `http://rleweb.mit.edu/video/video.htm`. The audio fragment is contained in video clip 10 'Digital Signal Processing – continued'.

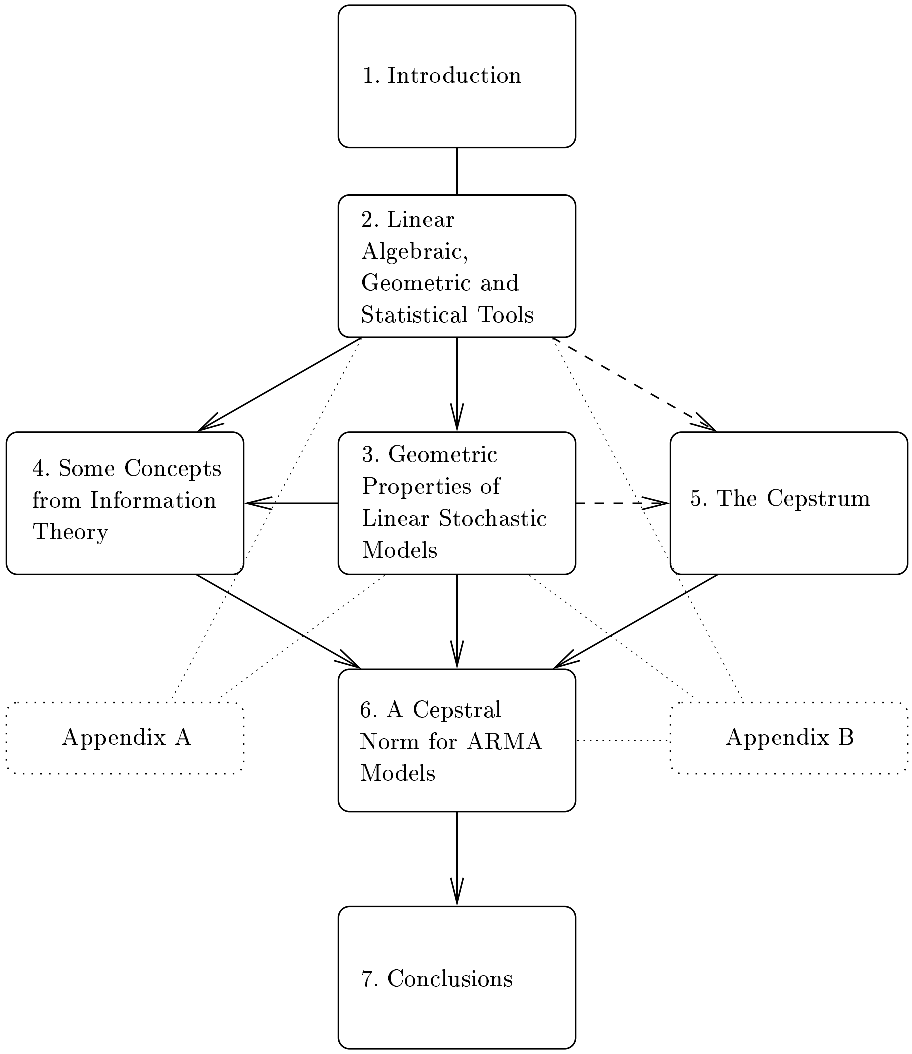


Figure 1.10: An overview of the different chapters and their interconnection. A dashed arrow between two chapters indicates a minor dependence.

Chapter 3: Geometric Properties of Linear Stochastic Models
We describe the model class which we work with throughout the thesis. Several state space representations of the model are given together with the statistical assumptions on the input, state and output processes. By introducing the data Hankel matrices and expressing the state space equations in terms of these matrices, we can interpret the statistical assumptions on the processes involved,

geometrically. Also other matrices that are important for the rest of the thesis are introduced here: the controllability and observability matrices and their corresponding Gramians.
We then specialize to the single-input single-output (SISO) linear stochastic models: the moving average (MA) model, the autoregressive (AR) model and the autoregressive moving average (ARMA) model. We give a state space representation for them and discuss two properties that do not hold for multiple-input multiple-output (MIMO) models and which will prove very useful further on in the thesis.
The real work starts in Section 3.4 where the principal angles between the different combinations of past and future input and output spaces are computed. We express the angles in terms of the model parameters, and looking at the limiting values, we obtain the canonical correlations of the corresponding processes. Next, we relate the angles of the different space pairs. For the assumptions made, the principal angles between the past input (U_p) and the future output (Y_f) are equal to the principal angles between the past output (Y_p) and the future output. Furthermore, we prove that the angles between the past and the future output are complementary to the angles between the future input and output. For SISO models, the expressions for the principal angles and the canonical correlations can be simplified.

- We show that they are all independent of the variance of the input process.
- We give a new characterization for the canonical correlations between the future input and the future output of a SISO ARMA model: they are equal to the cosines of the principal angles between the row spaces of the infinite controllability matrix of the model and the infinite controllability matrix of the inverse model.
- We give a simple expression for the canonical correlations of the past and the future of a SISO AR model.

All these parametric results are verified by examples, in which we compute the angles between the row spaces of the block Hankel matrices with the observed data samples.

Chapter 4: Some Concepts from Information Theory
In Chapter 4 we discuss two important information theoretic concepts: the entropy and the mutual information. We treat the entropy of a random variable as well as the entropy rate of a stochastic process. Similarly, we give definitions for the mutual information of two random variables and of two stochastic processes.
As can be seen in Figure 1.10, the insights of Chapter 3 are used in Chapter 4. Indeed, for most examples in Chapter 4 that are concerned with stochastic processes, we rely on the output processes of linear stochastic models, of which the

assumptions, parameterizations and properties are given in Chapter 3. But the relation of Chapter 3 and Chapter 4 goes deeper. The canonical correlations of two Gaussian random variables are closely connected to the mutual information of the random variables, and similarly for the canonical correlations and mutual information of two Gaussian stochastic processes. We can hence directly apply the results of Chapter 3 to obtain expressions for the mutual information of the past and future input and output processes of a linear stochastic model.

Chapter 5: The Cepstrum

We begin Chapter 5 with a motivational example, namely the detection of an echo in a signal. This was the original application for which the power cepstrum was introduced in [17]. Besides the power cepstrum of a stochastic process we also use the complex cepstrum of an absolutely summable sequence.

If a stochastic process is the output of a stable and minimum phase linear model driven by white noise, then the power cepstrum of the process can be expressed in terms of the poles and zeros of the model. The same can be done for the complex cepstrum of the impulse response of the model. Not surprisingly, the two cepstra are strongly related. In fact, their even parts ($k > 0$) are equal.

The cepstrum is not only used in echo detection applications. An interesting recent development is the use of the cepstrum for the covariance extension problem [22, 23, 46, 47]. We place this in the more general context of stochastic identification based on the cepstrum.

Chapter 6: A Cepstral Norm for ARMA Models

As can be seen from Figure 1.10, all the concepts, insights and results of Chapter 3, Chapter 4 and Chapter 5 come together in Chapter 6, where we discuss the cepstral distance measure for ARMA models that was defined by Martin [105].

Before interpreting the distance measure in a system theoretic and information theoretic context (see also Figure 1.3), we first study the properties of the corresponding cepstral norm for ARMA models. We show that this particular cepstral norm of an ARMA model can be expressed in terms of the principal angles between the row spaces of the controllability matrix of the model and the controllability matrix of the inverse model. From this and the results of Chapter 3 and Chapter 4 it follows that the squared cepstral norm is also proportional to the mutual information of the past and the future of the model's output process. Next, we give a graphical interpretation of the cepstral norm. By recognizing that the cepstral norm is the Hilbert-Schmidt norm of the Hankel matrix with the cepstral coefficients, we can show that the norm is proportional to the area enclosed by the polar plot of the logarithm of the transfer function.

These results are then applied to the cepstral distance between two ARMA models.

Chapter 7: Conclusions
In Chapter 7 we give an overview of our contributions and list the open problems and questions that we still have.

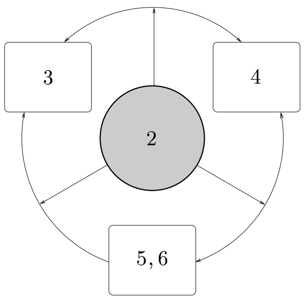
Appendices A and B
The appendices only play a minor role in the thesis. However, we have included them since they shed a new light on some aspects. In Appendix A we give the higher dimensional generalization of some geometric properties in the plane. Theorem A.1 can be used to prove the relations between the canonical correlations in Chapter 3 in a geometric way.
Appendix B deals with the interrelated concepts of compound matrices, Grassmann vectors and the volume of a matrix. As is indicated in Figure 1.10, they are connected to elements in Chapter 2, Chapter 3 and Chapter 6.

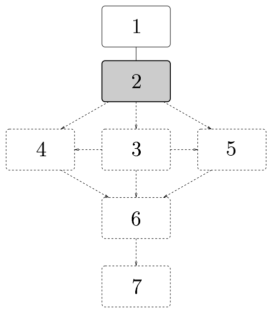
Remark 1.2. The schemes in Figure 1.10 and Figure 1.3, will be used throughout the thesis to show where each chapter should be situated in the whole. ○

Remark 1.3. Many of the methods and algorithms proposed in the thesis are undoubtedly not optimal in a numerical perspective. We give a mixture of theoretical observations that lend themselves easily for numerical computations. However, the thesis should not be read from a numerical analysis point of view. ○

Chapter 2

Linear Algebraic, Geometric and Statistical Tools





We begin this thesis with and introduction of the algebraic, geometrical and statistical properties that form the mathematical core of this thesis. The center of Figure 1.3 is treated in this chapter (see also the pictogram above left). As indicated in this figure, many derivations in the rest of the thesis rely on the foundations laid in this chapter.

The geometric notion of principal angles between and corresponding principal directions in two subspaces and its statistical counterpart, the canonical correlation analysis for random variables and stochastic processes, constitute the largest part of the chapter. However, in order to fully grasp their meaning and properties, we first give background information on the singular value decomposition and orthogonal projectors, of which the geometric interpretation is emphasized.

But first a few rank properties of matrices that will be applied throughout the thesis, are mentioned.

Chapter overview In Section 2.1 we give some well-known equalities and inequalities on the rank of matrices. Section 2.2 deals with the singular value decomposition (SVD) of matrices. The geometrical interpretation of the SVD is particularly important for the comprehension of the notions that follow. Based on the singular value decomposition, we give in Section 2.3 the definition and some properties of the Moore–Penrose inverse. This pseudo-inverse is used in Section 2.4, where we discuss projectors, both orthogonal and nonorthogonal. In Section 2.5 we arrive at the most important geometrical tool, the principal angles between and corresponding principal directions in two subspaces. The statistical version of it is called canonical correlation analysis, which is discussed in Section 2.6. We argument that the canonical correlations of two random variables, and by extension of two stochastic processes, can be viewed as cosines of principal angles between subspaces.

2.1 Rank properties

Throughout this thesis, we will sometimes analyze the rank of the product of matrices. For that purpose, we make use of the following basic lemma, which can e.g. be found in [67, p. 13].

Lemma 2.1. Sylvester's rank inequality
If $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, then

$$\operatorname{rank}(A) + \operatorname{rank}(B) - n \leq \operatorname{rank}(AB) \leq \min(\operatorname{rank}(A), \operatorname{rank}(B)) \ . \qquad (2.1)$$

We explain the inequality in geometrical terms by means of the scheme in Figure 2.1, where p, n and m are respectively equal to 4, 6 and 7. The crosses

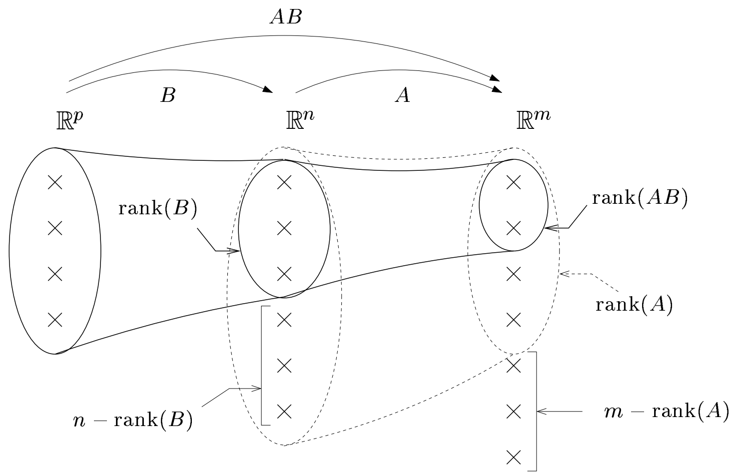


Figure 2.1: Graphical explanation of Sylvester's rank inequality.

represent the dimensions, or basis vectors. We start with 4 basis vectors in $\mathbb{R}^p$, which are transformed by B to 4 vectors in $\mathbb{R}^n$. However, the dimension of the subspace that these transformed vectors span, can be smaller (but of course never larger) than 4. This dimension is equal to the rank of B, which is 3 in Figure 2.1. We take 3 basis vectors in range(B) and transform them by A to 3 vectors in $\mathbb{R}^m$. The dimension of the resulting subspace, which is the range of AB is smaller than or equal to the rank of B. We have chosen $\text{rank}(AB) = 2$. A complete basis in $\mathbb{R}^n$ is transformed by A to the range of A, of which the dimension is at most equal to n. In Figure 2.1, this is depicted by the dashed lines, and $\text{rank}(A) = 4$. Furthermore, since only $n - \text{rank}(B) = 3$ extra vectors are transformed, compared to the transformation of the range of B to $\mathbb{R}^m$, the rank of A cannot be larger than $\text{rank}(AB)+n-\text{rank}(B) = \text{rank}(AB)+3$, which explains the left hand side inequality. And since $\text{rank}(AB) \leq \text{rank}(A)$ and $\text{rank}(AB) \leq \text{rank}(B)$, the right hand side inequality of (2.1) is also explained. If the matrix A is a nonsingular $n \times n$ matrix, then we see by a similar reasoning that the rank of AB is equal to the rank of B. Analogously, if B is a nonsingular $n \times n$ matrix, then the rank of AB is equal to the rank of A.
Two other rank properties:

- $\text{rank}(A^T A) = \text{rank}(AA^T) = \text{rank}(A) = \text{rank}(A^T)$.
- If $A \in \mathbb{R}^{m\times n}$ has rank k, then there exist matrices $X \in \mathbb{R}^{m\times k}$ and $Y \in \mathbb{R}^{k\times n}$ and a nonsingular matrix $B \in \mathbb{R}^{k\times k}$ so that
$$A = XBY \ .$$

When dealing with the principal angles between two subspaces, it will be very useful to compute the number of zero angles. This is equal to the dimension of the intersection of the subspaces, which is given by Grassmann's dimension theorem (also see Section 1.1, Figure 1.4).

Lemma 2.2. Grassmann's dimension theorem
Let $A \in \mathbb{R}^{m\times p}$ and $B \in \mathbb{R}^{m\times q}$. Then

$$\dim(\text{range}(A) \cap \text{range}(B)) = \text{rank}(A) + \text{rank}(B) - \text{rank}(\begin{pmatrix} A & B \end{pmatrix}) \ .$$

The range of a matrix A, denoted by range(A) and also called the column space of A, is defined in (2.4).

2.2 The singular value decomposition

In this section, we introduce the singular value decomposition, which will be abbreviated throughout this thesis as SVD. It is a matrix factorization whose computation is a step in many 'modern' algorithms. For an interesting account

on the early history of the SVD, we refer to [137].
In Section 2.2.1 we explain the full and reduced singular value decomposition of complex and real matrices. The nomenclature related to the SVD is given in Section 2.2.2. Its relation to the eigenvalue decomposition is briefly dealt with in Section 2.2.3. In Section 2.2.4 we demonstrate how the SVD is a convenient numerical tool to establish certain geometrical properties of matrices.

2.2.1 Main theorem

The SVD is applicable to both real and complex matrices. Theorem 2.3 gives the properties of the SVD of a complex matrix. From this full decomposition we can deduce a reduced SVD, which is explained in Corollary 2.4. For real matrices, the factor matrices in the decomposition can be kept real. This is given in Theorem 2.5.

Theorem 2.3. The (full) singular value decomposition for complex matrices
Every complex $m \times n$ matrix A can be factorized as[1]

$$A = USV^H \ ,$$

where the matrices $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$ are unitary, i.e. $U^H U = I_m = UU^H$ and $V^H V = I_n = VV^H$. Furthermore, the $m \times n$ matrix S is real and has the form

$$S = \begin{array}{cc} \begin{array}{cc} r & n-r \end{array} & \\ \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} & \begin{array}{l} r \\ m-r \end{array} \end{array} \ ,$$

where $r = \operatorname{rank}(A)$, $S_1 = \operatorname{diag}(\sigma_i)$, $i = 1, 2, \ldots, r$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.

In Figure 2.2 the SVD of the matrix A is schematically depicted. The white parts in the matrix S are filled with zeros. The matrices U and V are divided into two parts $U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}$ and $V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}$ with dimensions as indicated. We use the convention that U_2 is the empty matrix when $m = r$ and V_2 is empty if $n = r$.
From the SVD of A, we also obtain its dyadic decomposition

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^H \ ,$$

where u_i and v_i are the ith columns of U and V, respectively, which correspond to the ith singular value σ_i. The dyadic decomposition is the rationale for data reduction in many applications. It provides a canonical description of a matrix as a sum of r rank-one matrices of decreasing importance, as measured by the

[1] Throughout the thesis, A^H denotes the complex conjugate transpose of the matrix A.

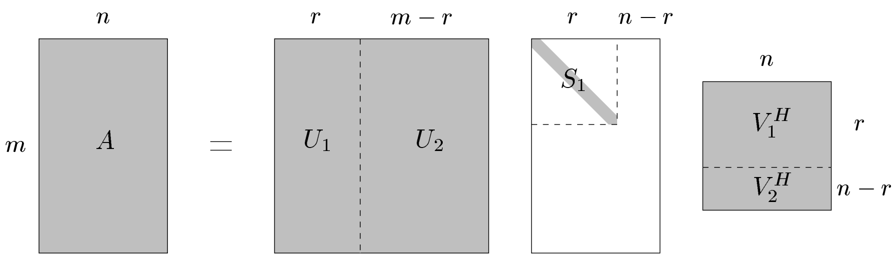


Figure 2.2: Schematic representation of the singular value decomposition of the matrix A. The white parts in the matrix S are filled with zeros

singular values.
From Theorem 2.3 and the scheme in Figure 2.2 it immediately follows that we can reduce the matrices involved in the SVD. The last $m-r$ columns of U and the last $n-r$ rows of V^H, i.e. the submatrices U_2 and V_2, are multiplied by the zeros in S. By leaving out these columns and rows, we obtain the reduced SVD[2], which is shown in Figure 2.3.

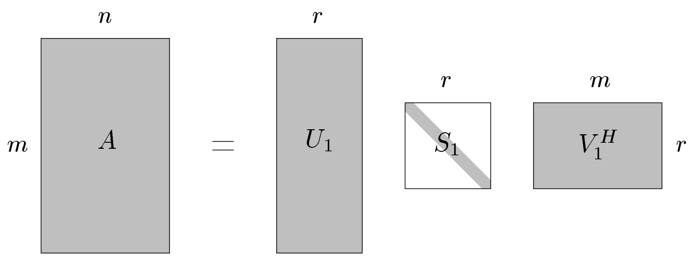


Figure 2.3: The reduced SVD of the matrix A. White parts in the matrix S are filled with zeros.

Corollary 2.4. The reduced singular value decomposition
Let A be a complex $m \times n$ matrix of rank r. Then, there exist matrices $U_1 \in \mathbb{C}^{m\times r}$, $S_1 \in \mathbb{R}^{r\times r}$ and $V_1 \in \mathbb{C}^{n\times r}$ so that

$$A = U_1 S_1 V_1^H \ , \tag{2.2}$$

where the columns of U_1 and V_1 are orthonormal, $U_1^H U_1 = V_1^H V_1 = I_r$, and S_1 is defined as in Theorem 2.3.

If A is real, all matrices in the SVD can be kept real as given in the following Theorem.

[2]Note, however, that U_2 and V_2 do play a role, as they form an orthonormal basis for the null spaces of A^T and A, respectively (see Property 2.7).

Theorem 2.5. The singular value decomposition for real matrices *For every real $m \times n$ matrix A, there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ ($U^TU = I_m = UU^T$ and $V^TV = I_n = VV^T$) such that*

$$A = USV^T \ ,$$

where $S \in \mathbb{R}^{m\times n}$ is given by $S = \begin{array}{cc} \begin{array}{cc} r & n-r \end{array} & \\ \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} & \begin{array}{l} r \\ m-r \end{array} \end{array}$ *, where $r = \operatorname{rank}(A)$, $S_1 = \operatorname{diag}(\sigma_i)$, $i = 1, 2, \ldots, r$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.*

From now on, unless otherwise specified, we will always work with real matrices.

2.2.2 Some nomenclature

For $A \in \mathbb{R}^{m\times n}$ of rank r and with SVD $A = USV^T$, we use the following notions.

Singular values: The $\min(m, n)$ diagonal elements of S are called the singular values of the matrix A. There are r non-zero and $\min(m, n) - r$ zero singular values. The non-increasing ordering is a widely accepted convention; σ_i is called the ith singular value.

Singular spectrum: The singular spectrum $\sigma(A)$ of a matrix A is the ordered set of its singular values (in non-increasing order):

$$\sigma(A) = (\sigma_1, \sigma_2, \cdots, \sigma_r, \underbrace{0, \cdots, 0}_{\min(m,n)-r}) \ ,$$

where r is the rank of A. A singular value σ_i is said to be of multiplicity μ_i if it is repeated μ_i times in the singular spectrum. The zero singular values are called 'ghost singular values'.

Singular vectors: The columns of U are the left singular vectors while those of V are the right singular vectors. The ith column u_i is the ith left singular vector while the ith column v_i is the ith right singular vector.

Singular vector matrices: The left singular vector matrix is the matrix U with the left singular vectors as its columns. The right singular vector matrix is the matrix V with the right singular vectors as its columns:

$$U = \begin{pmatrix} u_1 & u_2 & \cdots & u_m \end{pmatrix}, \ V = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix} \ .$$

Singular triplet: The set (u_i, σ_i, v_i) is called the ith singular triplet. It consists of the ith left singular vector, the ith singular value and the ith right singular vector.

Singular subspaces: The ranges of the left and right singular matrices, $\operatorname{range}(U)$ and $\operatorname{range}(V)$, are the left and right singular subspaces.

2.2.3 Relation to the eigenvalue decomposition

The SVD is in several ways related to the eigenvalue decomposition. We briefly list the relations that are used in the rest of the thesis.

- The left singular vectors of a matrix A are the eigenvectors of AA^T. The right singular vectors are the eigenvectors of A^TA. The non-zero singular values are the square roots of the non-zero eigenvalues of A^TA and AA^T.
- The non-zero eigenvalues of the matrix $\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$ are $\pm$ the non-zero singular values of A. The singular triplets (u_i, σ_i, v_i) of A satisfy the following equations, which Lanczos in [93, p. 117] called the shifted eigenvalue problem:
$$\begin{aligned} Av_i &= u_i\sigma_i \\ A^Tu_i &= v_i\sigma_i \ . \end{aligned} \tag{2.3}$$

Remark 2.1. Uniqueness of the SVD
It can be proven that the singular values of a matrix are uniquely determined by its elements. This follows easily from the fact that the squares of the singular values are the eigenvalues of A^TA and hence the roots of $\det(A^TA - \sigma^2 I_n) = 0$. The converse is however not true: it is very easy to find two completely different matrices that have the same singular spectrum. However, matrices with the same singular spectrum are related as follows:

Property 2.6. Relation of matrices with the same singular spectrum
The matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, where $p \leq m$ and $q \leq n$ have the same singular spectrum except for $\min(m,n) - \min(p,q)$ zero singular values if and only if there exist matrices $U \in \mathbb{R}^{m\times p}$ and $V \in \mathbb{R}^{n\times q}$ with orthonormal columns, so that $A = UBV^T$. ○

2.2.4 Geometrical properties

In this section, we demonstrate how the SVD is a convenient numerical tool to establish certain geometrical properties of matrices. First we need some definitions.
Consider a real $m \times n$ matrix A of rank r. Its domain, $\mathrm{dom}(A)$, is $\mathbb{R}^n$, while its codomain, The range of A, $\mathrm{range}(A)$, is defined as

$$\mathrm{range}(A) = \{y \in \mathbb{R}^m \mid \exists x \in \mathbb{R}^n : y = Ax\} \ . \tag{2.4}$$

The row space of A, denoted by $\mathrm{row}(A)$, is the range of A^T.
The null space $\mathrm{null}(A)$ of A is defined as

$$\mathrm{null}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\} \ .$$

The domain and codomain of a matrix can be decomposed orthogonally as:

$$\begin{aligned} \mathrm{dom}(A) &= \mathrm{row}(A) + \mathrm{null}(A) \ , \\ \mathrm{codom}(A) &= \mathrm{range}(A) + \mathrm{null}(A^T) \ . \end{aligned} \tag{2.5}$$

Since $\mathrm{range}(A)$ and $\mathrm{null}(A^T)$ are orthogonal complements in $\mathbb{R}^m$, we also denote $\mathrm{null}(A^T)$ by $\mathrm{range}(A)^\perp$. Analogously, $\mathrm{null}(A)$ is also denoted by $\mathrm{row}(A)^\perp$.

Property 2.7. Geometrical interpretation of the SVD
Let the SVD of a rank r matrix $A \in \mathbb{R}^{m\times n}$ be $A = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}$*, where $U_1 \in \mathbb{R}^{m\times r}$, $U_2 \in \mathbb{R}^{m\times(m-r)}$, $S_1 \in \mathbb{R}^{r\times r}$, $V_1 \in \mathbb{R}^{n\times r}$ and $V_2 \in \mathbb{R}^{n\times(n-r)}$. Then,*

$$\begin{aligned} \mathrm{range}(U_1) &= \mathrm{range}(A) \ , \\ \mathrm{range}(U_2) &= \mathrm{null}(A^T) = \mathrm{range}(A)^\perp \ , \\ \mathrm{range}(V_1) &= \mathrm{range}(A^T) = \mathrm{row}(A) \ , \\ \mathrm{range}(V_2) &= \mathrm{null}(A) = \mathrm{range}(A^T)^\perp = \mathrm{row}(A)^\perp \ . \end{aligned}$$

Hence, the columns of U_1, which are left singular vectors corresponding to the non-zero singular values, form an orthonormal basis for the range of A, the columns of V_1 are an orthonormal basis for the row space of A, the columns of U_2 are an orthonormal basis for the null space of A^T and the columns of V_2 are an orthonormal basis for the null space of A.

2.3 The Moore–Penrose inverse

The standard definition for the inverse of a matrix fails if the matrix is not square or if it is singular. Using the singular value decomposition, however, it is possible to define a pseudo-inverse for non-square matrices or for singular square ones.
There exist various types of generalized inverses of a matrix [14], all of which satisfy one or more of the so-called Penrose conditions (see below).

Definition 2.1. Generalized inverses
For $A \in \mathbb{R}^{m\times n}$, a matrix $G \in \mathbb{R}^{n\times m}$ is called an (i,j,k)-inverse for A if G satisfies the ith, jth and kth $(i,j,k = 1,\ldots,4)$ Penrose conditions:

1. $AGA = A$,
2. $GAG = G$,
3. $(AG)^T = AG$,

4. $(GA)^T = GA$.

For example, G is a $(1,3)$-inverse for A if $AGA = A$ and AG is symmetric. The order of the equations in Definition 2.1 is of course important for the naming convention.
The unique matrix G that satisfies all four Penrose conditions is the so-called Moore–Penrose pseudo-inverse of A [119]. It is denoted by $A^\dagger$ and can be computed by means of the SVD as follows.
Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank r with reduced SVD $A = U_1 S_1 V_1^T$, then the Moore–Penrose inverse of A is equal to

$$A^\dagger = V_1 S_1^{-1} U_1^T \ .$$

Note that

$$\begin{aligned} A^\dagger &= (A^T A)^\dagger A^T \ , \\ A^\dagger &= A^T (A A^T)^\dagger \ . \end{aligned}$$

This reduces to

$$\begin{aligned} A^\dagger &= (A^T A)^{-1} A^T && \text{if } A \text{ is of full column rank,} && (2.6) \\ A^\dagger &= A^T (A A^T)^{-1} && \text{if } A \text{ is of full row rank,} && (2.7) \\ A^\dagger &= A^{-1} && \text{if } A \text{ is square and nonsingular.} && (2.8) \end{aligned}$$

2.4 Orthogonal projectors

One of the geometrical tools that will be used throughout this thesis, is the orthogonal projection into subspaces. In Section 2.4.1 we first define projectors. Then, we turn to the special class of projectors that will be of primary interest to us, namely orthogonal projectors. They are discussed in Section 2.4.2.

2.4.1 Projectors

Although we will mainly use orthogonal projectors, we first define the projectors, which form a more general class of matrices.

Definition 2.2. Projector
A projector is a square matrix Π *that satisfies*

$$\Pi^2 = \Pi \ . \qquad (2.9)$$

Matrices that satisfy (2.9) are also said to be idempotent.
Definition 2.2 includes both orthogonal projectors, to be discussed in Section 2.4.2, and nonorthogonal ones. To avoid confusion one may use the term oblique projector in the nonorthogonal case.
If Π is a projector, $I - \Pi$ is also a projector, for it is also idempotent:

$$(I - \Pi)^2 = I - 2\Pi + \Pi^2 = I - 2\Pi + \Pi = I - \Pi \ .$$

If $\Pi \in \mathbb{R}^{m \times m}$ is a projector, then one can prove that $\text{range}(\Pi)$ and $\text{range}(I_m - \Pi)$ are complementary spaces in $\mathbb{R}^m$, i.e.,

$$\begin{aligned} \text{range}(\Pi) \cap \text{range}(I_m - \Pi) &= \{0\} \ , \\ \text{range}(\Pi) + \text{range}(I_m - \Pi) &= \mathbb{R}^m \ . \end{aligned}$$

The matrix $I_m - \Pi$ is called the complementary projector to Π. It projects onto the null space of Π. Therefore, for any projector Π,

$$\text{range}(I_m - \Pi) = \text{null}(\Pi) \ .$$

Due to the complementarity of $\text{range}(\Pi)$ and $\text{range}(I_m - \Pi)$, an arbitrary vector $x \in \mathbb{R}^m$ can always be decomposed uniquely as (see Figure 2.4)

$$x = \Pi x + (I_m - \Pi)x \ . \tag{2.10}$$

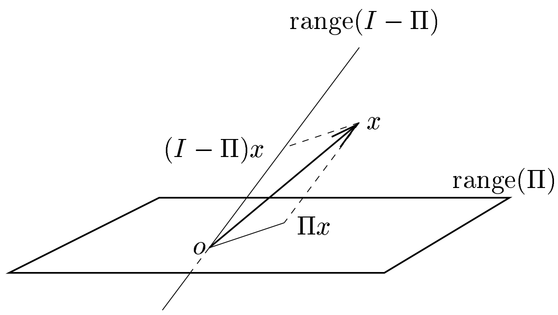


Figure 2.4: The projection of the vector $x \in \mathbb{R}^3$ by the projector $\Pi \in \mathbb{R}^{3\times 3}$. The projection of x onto the range of Π, which is a two-dimensional subspace of $\mathbb{R}^3$, is Πx. Its projection onto the complementary subspace $\text{range}(I - \Pi)$ of dimension one, is $(I - \Pi)x$.

2.4.2 Orthogonal projectors

A special class of projectors are the orthogonal projectors. They are defined as follows.

Definition 2.3. Orthogonal projector
A projection matrix $\Pi \in \mathbb{R}^{m\times m}$ *is called orthogonal if* Πx *is orthogonal to* $(I_m - \Pi)y$ *for all vectors* x *and* $y \in \mathbb{R}^m$ *:*

$$\forall x, y \in \mathbb{R}^m : (\Pi x)^T (I_m - \Pi) y = 0 \ . \tag{2.11}$$

From (2.11) it follows that

$$x^T(\Pi^T - \Pi^T\Pi)y = 0 \qquad \forall x, y$$

and thus

$$\Pi^T = \Pi^T\Pi \ .$$

Hence, an orthogonal projection matrix is symmetric. In fact, the matrix Π is an orthogonal projection matrix if and only if $\Pi^2 = \Pi$ and $\Pi^T = \Pi$.

The orthogonal projection onto the column space of a matrix

The orthogonal projection matrix $\Pi_{\text{range}(U)}$ that projects onto the column space range(U) of a matrix $U \in \mathbb{R}^{m\times r}$ of which the columns form an orthonormal basis for range(U) is a symmetric matrix defined as

$$\Pi_{\text{range}(U)} = UU^T \qquad (U^TU = I_r) \ . \tag{2.12}$$

It satisfies indeed the following three conditions for an orthogonal projection matrix on the range of U:

$$\begin{aligned}
\text{range}(\Pi_{\text{range}(U)}) &= \text{range}(U) \ , \\
\Pi^2_{\text{range}(U)} &= \Pi_{\text{range}(U)} \ , \\
\Pi^T_{\text{range}(U)} &= \Pi_{\text{range}(U)} \ .
\end{aligned}$$

The orthogonal projection matrix $\Pi_{\text{range}(A)}$ onto the column space of a given matrix $A \in \mathbb{R}^{m\times n}$ with rank r can for instance be obtained from its reduced SVD, $A = U_1S_1V_1^T$:

$$\Pi_{\text{range}(A)} = U_1U_1^T \ .$$

From the SVD, it can also be shown that

$$\Pi_{\text{range}(A)} = AA^\dagger \ ,$$

where $A^\dagger$ is the Moore–Penrose inverse of A. Hence, if A is of full column rank, then by (2.6), $\Pi_{\text{range}(A)} = A(A^TA)^{-1}A^T$. If $x \in \mathbb{R}^m$, then the vector $\Pi_{\text{range}(A)}x$ is the orthogonal projection of x onto range(A).
If $\Pi_{\text{range}(A)} \in \mathbb{R}^{m\times m}$ is the orthogonal projection matrix that projects a vector onto the column space of a matrix $A \in \mathbb{R}^{m\times n}$, then $I_m - \Pi_{\text{range}(A)}$ is the

orthogonal projection matrix onto the orthogonal complement of range(A). This can be seen from

$$I_m - \Pi_{\text{range}(A)} = I_m - U_1 U_1^T = U_2 U_2^T = \Pi_{\text{range}(U_2)} ,$$

and we know from Property 2.7 that range(U_2) = range$(A)^\perp$. The unique decomposition (see (2.10)) of a vector $x \in \mathbb{R}^m$ into

$$x = \Pi_{\text{range}(A)} x + (I_m - \Pi_{\text{range}(A)}) x ,$$

therefore yields the orthogonal decomposition described in (2.5). This is also shown in Figure 2.5.

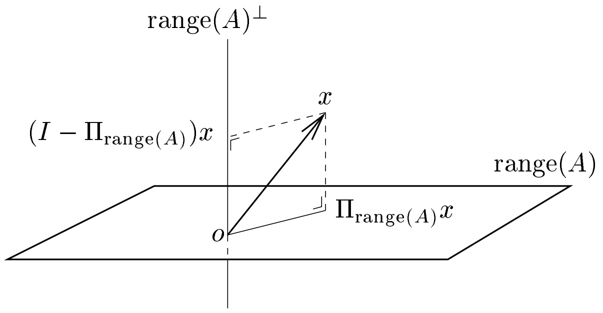


Figure 2.5: The decomposition of the vector $x \in \mathbb{R}^3$ into its orthogonal projection onto the range of A and its orthogonal projection onto the orthogonal complement of range(A), which is the range of $I - \Pi_{\text{range}(A)}$.

Remark 2.2. An orthogonal projector is not necessarily orthogonal
The fact that $\Pi_{\text{range}(A)}$ is an orthogonal projection matrix does not imply that $\Pi_{\text{range}(A)}$ is an orthogonal matrix. This is only true if A is a square non-singular matrix. ○

The orthogonal projection onto the row space of a matrix

We now consider the orthogonal projection of the row space of a matrix $B \in \mathbb{R}^{q\times n}$ onto the row space of a matrix $A \in \mathbb{R}^{p\times n}$. Adapting the formulas of the previous paragraph is obvious. Nevertheless, we prefer to write them down explicitly since we will often need them.
Let the reduced SVD of the matrix $A \in \mathbb{R}^{p\times n}$ be $A = U_1 S_1 V_1^T$. Then, the orthogonal projection matrix onto the row space of A can be obtained as

$$\Pi_{\text{row}(A)} = V_1 V_1^T \in \mathbb{R}^{n\times n} , \tag{2.13}$$

and also

$$\Pi_{\mathrm{row}(A)} = A^{\dagger} A \ , \tag{2.14}$$

which is for a full row rank matrix

$$\Pi_{\mathrm{row}(A)} = A^T (AA^T)^{-1} A \ .$$

The orthogonal projection of a vector $x \in \mathbb{R}^n$ onto the row space of a matrix $A \in \mathbb{R}^{p \times n}$ is $\Pi_{\mathrm{row}(A)} x$. The orthogonal projection of the row space of $B \in \mathbb{R}^{q \times n}$ onto the row space of A is the row space of $B\Pi_{\mathrm{row}(A)}$. This projection is denoted by $\mathrm{row}(B/A)$. For a full row rank matrix A, we have

$$B/A = BA^T (AA^T)^{-1} A \ . \tag{2.15}$$

SVD-related orthogonal projectors

There are several important orthogonal projections associated with the singular value decomposition. Suppose the SVD of the matrix $A \in \mathbb{R}^{m \times n}$ of rank r is

$$A = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} \ ,$$

where $U_1 \in \mathbb{R}^{m \times r}$, $U_2 \in \mathbb{R}^{m \times (m-r)}$, $S_1 \in \mathbb{R}^{r \times r}$, $V_1 \in \mathbb{R}^{n \times r}$ and $V_2 \in \mathbb{R}^{n \times (n-r)}$. Then, from Property 2.7 and Equation (2.12) it follows that

$$\begin{aligned}
U_1 U_1^T &= \Pi_{\mathrm{range}(A)} \ , \\
U_2 U_2^T &= \Pi_{\mathrm{range}(A)^{\perp}} = I_m - \Pi_{\mathrm{range}(A)} \ , \\
V_1 V_1^T &= \Pi_{\mathrm{row}(A)} \ , \\
V_2 V_2^T &= \Pi_{\mathrm{row}(A)^{\perp}} = I_n - \Pi_{\mathrm{row}(A)} \ .
\end{aligned}$$

2.5 The principal angles between subspaces

The concept of principal angles between two subspaces goes back to Jordan in the nineteenth century [77]. This notion was translated into the statistical quantities of canonical correlations by Hotelling [68]. Applications include data analysis [53], random processes [52, 79] and stochastic realization [2, 25, 156, 159] (and references herein). Numerically stable methods to compute the principal angles and vectors via a singular value decomposition have been proposed in [16, 58] and can also be found in [57, pp. 603–604].
In the sequel of the thesis, we will quite often compute the angles between the

row spaces of two matrices. Therefore, we deal in this section mainly with the principal angles between row spaces of matrices. The formulas for the principal angles between the column spaces of two matrices can easily be obtained by considering the principal angles between the row spaces of the transposed matrices.
The principal angles between two subspaces are a generalization of an angle between two vectors. This is briefly recalled in Section 2.5.1. In Section 2.5.2 we then define the principal angles between and the principal directions in two subspaces. We also show how they can be computed by using the SVD. Section 2.5.3 deals with the different kinds of principal angles (zero, acute and right angles). In Section 2.5.4 the relation between principal angles and orthogonal projectors is explained. Throughout the thesis we will also need an expression for the cosines of the principal angles as the eigenvalues of a certain matrix. This is described in Section 2.5.5. Section 2.5.6 is devoted to an efficient implementation that is based on the LQ decomposition.

2.5.1 The angle between two vectors

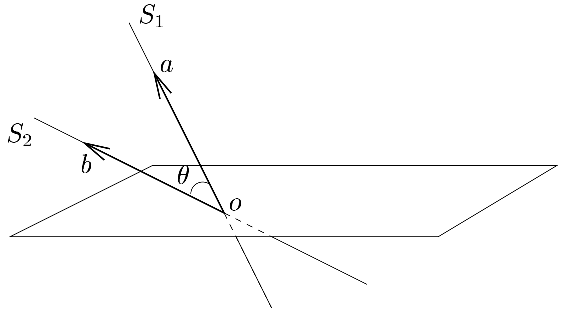


Figure 2.6: The angle between the one-dimensional subspaces S_1 and S_2 in $\mathbb{R}^3$ is θ. Since S_1 is spanned by $a \in \mathbb{R}^3$ and S_2 by $b \in \mathbb{R}^3$, $\theta = [a \sphericalangle b]$ and its cosine can be computed as in (2.16).

Consider two real non-zero vectors $a, b \in \mathbb{R}^n$. Based on the inequality of Cauchy–Schwarz, the cosine of the angle between a and b can be defined as

$$\cos [a \sphericalangle b] = \frac{|a^T b|}{||a|| ||b||} \, , \tag{2.16}$$

where $||a||$ is the Euclidean norm of a: $||a||^2 = \sum_{i=1}^n a_i^2$. Observe that we use the notation $[a \sphericalangle b]$ for the angle between the vectors a and b. On purpose we choose the absolute value of $a^T b$ in order to make $[a \sphericalangle b]$ belong to $[0, \frac{\pi}{2}]$. The vectors a and b each span a one-dimensional linear subspace. The angle between two one-dimensional subspaces is thus defined as the angle in $[0, \frac{\pi}{2}]$ between two vectors that span the respective subspaces (see Figure 2.6).

2.5.2 Principal angles and directions

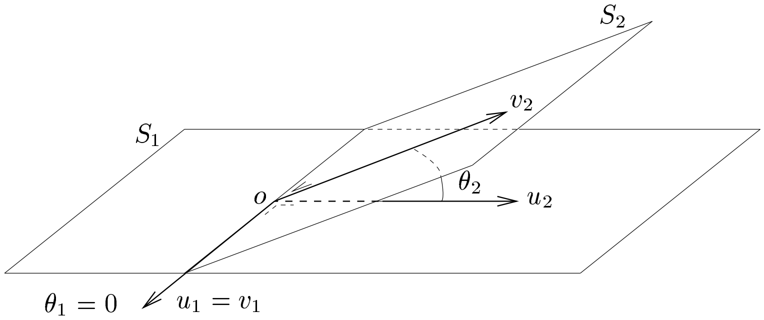


Figure 2.7: The principal angles between the two-dimensional subspaces S_1 and S_2 in $\mathbb{R}^3$. The first principal angle, θ_1, is zero and the first principal vectors, $u_1 \in S_1$ and $v_1 \in S_2$, coincide, revealing a one-dimensional intersection of S_1 and S_2. The second principal directions are orthogonal to the first principal directions: $u_2 \perp u_1$ and $v_2 \perp v_1$. The angle between $u_2 \in S_1$ and $v_2 \in S_2$ is θ_2.

The notion of an angle can be generalized to angles between higher dimensional subspaces. Suppose we are given two linear subspaces S_1 and S_2 in $\mathbb{R}^n$ of dimension d_1 and d_2, respectively. A natural extension of the one-dimensional case is to choose a unit vector u_1 from S_1 and a unit vector v_1 from S_2 such that the angle between u_1 and v_1 is minimized. The vectors u_1 and v_1 are the first principal directions and the angle between them is the first principal angle θ_1. Next, choose a unit vector $u_2 \in S_1$ orthogonal to u_1 and $v_2 \in S_2$ orthogonal to v_1 and minimize the angle θ_2 between them. This is the second principal angle and u_2 and v_2 the corresponding principal directions. Continue in this way until $\min(d_1, d_2)$ angles and corresponding principal vectors have been found. The result of this procedure is shown in Figure 2.7, where the principal angles between two two-dimensional subspaces in $\mathbb{R}^3$ are obtained. This informal description is now formalized.

Definition 2.4. Principal angles and directions
The principal angles $0 \leq \theta_1 \leq \theta_2 \leq \ldots \theta_{\min(d_1,d_2)} \leq \pi/2$ *between the subspaces* S_1 *and* S_2 *of dimension* d_1 *and* d_2*, respectively, and the corresponding principal directions* $u_i \in S_1$ *and* $v_i \in S_2$ *are defined recursively as*

$$\begin{aligned}
\cos\theta_1 &= \max_{\substack{u\in S_1\\ v\in S_2}} u^T v = u_1^T v_1 \ , \\
\cos\theta_k &= \max_{\substack{u\in S_1\\ v\in S_2}} u^T v = u_k^T v_k \ , \ \textit{for } k = 2, \ldots, \min(d_1, d_2) \ ,
\end{aligned}$$

subject to

$$\|u\| = \|v\| = 1 \ ,$$

and for $k > 1$

$$u^T u_i = 0 \quad i = 1, \ldots, k-1 \ ,$$
$$v^T v_i = 0 \quad i = 1, \ldots, k-1 \ .$$

Let $A \in \mathbb{R}^{p \times n}$ be of rank r_a and $B \in \mathbb{R}^{q \times n}$ of rank r_b. Then, the ordered set of $\min(r_a, r_b)$ principal angles between the row spaces of A and B is denoted by

$$\left(\theta_1, \theta_2, \ldots, \theta_{\min(r_a, r_b)}\right) = [A \sphericalangle B] \ .$$

Several numerical algorithms have been proposed for the computation of the principal angles and the corresponding principal directions. One of them uses the SVD as in the following property.

Property 2.8. The principal angles and directions via the SVD
Let $V_{a_1} \in \mathbb{R}^{n \times r_a}$ and $V_{b_1} \in \mathbb{R}^{n \times r_b}$ be the right singular vector matrices in the reduced SVD of $A \in \mathbb{R}^{p \times n}$ of rank r_a and $B \in \mathbb{R}^{q \times n}$ of rank r_b, respectively. Then, the principal angles between and principal directions in the row space of A and the row space of B can be obtained from the SVD of $V_{a_1}^T V_{b_1}$:

$$V_{a_1}^T V_{b_1} = PSQ^T \ , \tag{2.17}$$

as follows. The cosines of the principal angles are the diagonal elements of S. The principal directions in $\mathrm{row}(A)$ *are the first* $\min(r_a, r_b)$ *columns of* $V_{a_1} P$ *and the principal directions in* $\mathrm{row}(B)$ *are the first* $\min(r_a, r_b)$ *columns of* $V_{b_1} Q$.

2.5.3 Three kinds of principal angles

Let the rank of $V_{a_1}^T V_{b_1} \in \mathbb{R}^{r_a \times r_b}$ be denoted by r_{ab}. Then, there are r_{ab} non-zero singular values and $\min(r_a, r_b) - r_{ab}$ zero ones. In the set of non-zero singular values we can distinguish between two kinds: singular values equal to one and singular values smaller than one. The number of singular values equal to one is denoted by t. The matrix S in the SVD of $V_{a_1}^T V_{b_1}$ (see Equation (2.17)), for the case $r_a \leq r_b$ schematically looks as in Figure 2.8. Since $\sigma_i(V_{a_1}^T V_{b_1}) = \cos\theta_i$, where θ_i is the ith principal angle between $\mathrm{row}(A)$ and $\mathrm{row}(B)$, we can distinguish between three kinds of principal angles: t zero principal angles, $r_{ab} - t$ acute principal angles and $\min(r_a, r_b) - r_{ab}$ principal angles equal to $\frac{\pi}{2}$. These three groups will now be discussed in some more detail.

1. t zero principal angles: $\theta_1 = \cdots = \theta_t = 0$
   The principal directions corresponding to the t zero principal angles between $\mathrm{row}(A)$ and $\mathrm{row}(B)$ are common to the row spaces of A and B (see e.g. u_1 and v_1 in Figure 2.7). This means that the dimension of the intersection of $\mathrm{row}(A)$ and $\mathrm{row}(B)$ is equal to t:

$$\dim(\mathrm{row}(A) \cap \mathrm{row}(B)) = t \ .$$

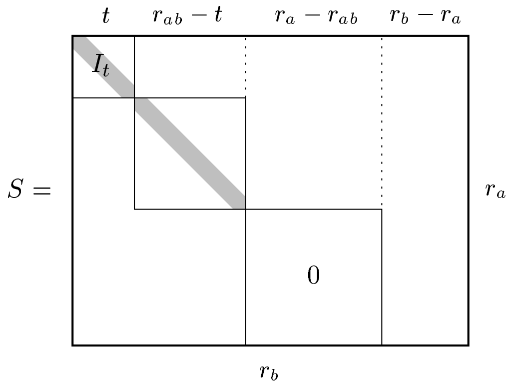


Figure 2.8: Three kinds of singular values of $V_{a_1}^T V_{b_1}$ for the case $r_a \leq r_b$.

The number of zero angles between row(A) and row(B) follows from Grassmann's dimension theorem (Theorem 2.2):

$$t = \operatorname{rank}(A) + \operatorname{rank}(B) - \operatorname{rank}\begin{pmatrix} A \\ B \end{pmatrix} .$$

2. $r_{ab} - t$ acute principal angles: $0 < \theta_{t+1} \leq \cdots \leq \theta_{r_{ab}} < \frac{\pi}{2}$
   The number r_{ab}, which was defined as $\operatorname{rank}(V_{a_1}^T V_{b_1})$, can also be obtained as $r_{ab} = \operatorname{rank}(AB^T)$.

3. $\min(r_a, r_b) - r_{ab}$ right principal angles: $\theta_{r_{ab}+1} = \cdots = \theta_{\min(r_a,r_b)} = \frac{\pi}{2}$
   The principal directions in row(A) corresponding to these principal angles are orthogonal to row(B) and conversely, the corresponding principal directions in row(B) are orthogonal to row(A).

2.5.4 Principal angles and orthogonal projectors

The concept of principal angles between two subspaces is closely related to orthogonal projectors. In this section we first show in Property 2.9 how the principal angles between and the principal directions in row(A) and row(B) can be obtained from the SVD of $\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}$. Then, we consider the plane spanned by two corresponding principal vectors, which reveals some interesting geometric properties about the principal angles between the different combinations of row(A), row(B), row$(A)^{\perp}$ and row$(B)^{\perp}$, which are proven in Theorem A.1 (see Appendix A).

Computing the principal angles and principal directions from the orthogonal projectors

Property 2.9. The principal angles and directions via orthogonal projectors (see e.g. [160])
Let $\Pi_{\mathrm{row}(A)}$ *and* $\Pi_{\mathrm{row}(B)}$ *be the orthogonal projection matrices onto the row spaces of two real matrices* $A \in \mathbb{R}^{p\times n}$ *of rank* r_a *and* $B \in \mathbb{R}^{q\times n}$ *of rank* r_b*. Then, the cosines of the principal angles are the first* $\min(r_a, r_b)$ *singular values and the principal directions are the first* $\min(r_a, r_b)$ *left and right singular vectors of the* $n \times n$ *matrix* $\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}$*.*

Proof.
From (2.13) it follows that

$$\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)} = V_{a_1}V_{a_1}^T V_{b_1}V_{b_1}^T \ ,$$

where $V_{a_1} \in \mathbb{R}^{n\times r_a}$ and $V_{b_1} \in \mathbb{R}^{n\times r_b}$ are the right singular matrices in the reduced SVD of A and B, respectively.
Let the full SVD of $V_{a_1}^T V_{b_1}$ be PSQ^T. Then,

$$\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)} = V_{a_1} PSQ^T V_{b_1}^T \ .$$

Since the columns of $V_{a_1}P \in \mathbb{R}^{n\times r_a}$ and the columns of $V_{b_1}Q \in \mathbb{R}^{n\times r_b}$ are orthonormal, we know from Property 2.6 that the singular spectra of the matrices $\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}$ and S are equal, except for $n - \min(r_a, r_b)$ zero singular values. Hence,

$$\begin{aligned}\sigma(\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}) &= \sigma(V_{a_1}^T V_{b_1}), \ \underbrace{0,\ldots,0}_{n-\min(r_a,r_b)} \ , \\ &= \cos\left[A \sphericalangle B\right], \ \underbrace{0,\ldots,0}_{n-\min(r_a,r_b)} \qquad \text{by Property 2.8} \ .\end{aligned}$$

Furthermore, from Property 2.8 we know that the first $\min(r_a, r_b)$ columns of $V_{a_1}P$ are the principal directions in row(A) and the first $\min(r_a, r_b)$ columns of $V_{b_1}Q$ are the principal directions in row(B). □

In the scheme in Figure 2.9 we show part of the SVD of $\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}$ for the case $r_a \le r_b$. Only the first r_a columns of the singular vector matrices are considered (the other singular vectors correspond to zero singular values that are not cosines of principal angles). The principal directions in row(A) and row(B) corresponding to the zero principal angles are the columns of U_0 and V_0, respectively. Those corresponding to the acute angles are the columns of U_{ac} and V_{ac}. Finally, the principal directions corresponding to the angles equal to $\frac{\pi}{2}$ are the columns of $U_{\frac{\pi}{2}}$ and $V_{\frac{\pi}{2}}$.

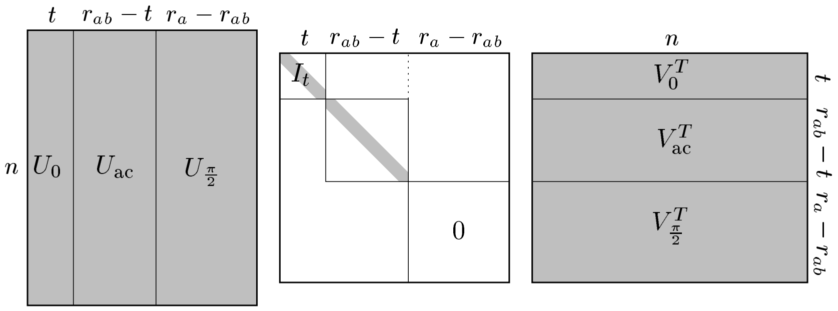


Figure 2.9: Part of the SVD of $\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}$. The principal directions in row(A) and row(B) corresponding to the zero principal angles are the columns of U_0 and V_0, respectively. Those corresponding to the acute angles are the columns of $U_{\rm ac}$ and $V_{\rm ac}$. The principal directions corresponding to the angles equal to $\frac{\pi}{2}$ are the columns of $U_{\frac{\pi}{2}}$ and $V_{\frac{\pi}{2}}$.

Corollary 2.10. Projections of principal directions [160]
Consider the principal angles between and principal directions in row(A) *and* row(B)*. If u_i and v_i are the ith principal directions in* row(A) *and* row(B)*, respectively, corresponding to the ith principal angle θ_i, then*

$$\begin{array}{rcl} \Pi_{\mathrm{row}(A)}v_i & = & u_i\cos\theta_i \ , \\ \Pi_{\mathrm{row}(B)}u_i & = & v_i\cos\theta_i \ . \end{array} \tag{2.18}$$

Proof.
From Property 2.9 it follows that

$$\begin{aligned} \Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}v_i &= u_i\sigma_i \ , \\ (\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)})^T u_i &= v_i\sigma_i \ , \end{aligned}$$

where (u_i,σ_i,v_i) is the ith singular triplet of $\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}$ (see also Equation (2.3)).
Note that $(\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)})^T = \Pi_{\mathrm{row}(B)}\Pi_{\mathrm{row}(A)}$.
Since $v_i \in \mathrm{row}(B)$ and $u_i \in \mathrm{row}(A)$, the orthogonal projection of v_i onto row(B) is equal to v_i and analogously for u_i:

$$\begin{aligned} \Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}v_i &= \Pi_{\mathrm{row}(A)}v_i \ , \\ \Pi_{\mathrm{row}(B)}\Pi_{\mathrm{row}(A)}u_i &= \Pi_{\mathrm{row}(B)}u_i \ . \end{aligned}$$

Together with $\sigma_i = \cos\theta_i$, this leads to

$$\begin{aligned} \Pi_{\mathrm{row}(A)}v_i &= u_i\cos\theta_i \ , \\ \Pi_{\mathrm{row}(B)}u_i &= v_i\cos\theta_i \ . \end{aligned}$$

□

This means that we can obtain the ith principal direction v_i in row(B) by projecting the corresponding principal direction u_i in row(A) orthogonally onto row(B) and vice versa, if $\theta_i \neq \frac{\pi}{2}$. In Figure 2.7 e.g. the orthogonal projection of v_2 onto S_1 is equal to $u_2 \cos\theta_2$.
It follows from Corollary 2.10 that the ith principal angle between row(A) and row(B) is equal to the angle between u_i and the orthogonal projection of u_i onto row(B) and also equal to the angle between v_i and its orthogonal projection onto row(A), provided the angle is not $\frac{\pi}{2}$:

$$\theta_i = \begin{cases} \frac{\pi}{2} & \text{if } \Pi_{\text{row}(A)} v_i = 0 \ , \\ \left[u_i \lessdot \Pi_{\text{row}(B)} u_i\right] = \left[v_i \lessdot \Pi_{\text{row}(A)} v_i\right] & \text{otherwise.} \end{cases}$$

Thus, if $r_{(A,B)}$ is the number of principal angles between row(A) and row(B) that are equal to $\frac{\pi}{2}$, then,

$$\begin{aligned} [A \lessdot B] &= [A \lessdot A/B]\,, \underbrace{\frac{\pi}{2}, \ldots, \frac{\pi}{2}}_{r_{(A,B)}} \\ &= [B/A \lessdot B]\,, \underbrace{\frac{\pi}{2}, \ldots, \frac{\pi}{2}}_{r_{(A,B)}} \ , \end{aligned}$$

and also

$$[A \lessdot B] = [B/A \lessdot A/B]\,, \underbrace{\frac{\pi}{2}, \ldots, \frac{\pi}{2}}_{r_{(A,B)}} \ .$$

Recall that $r_{(A,B)}$ can be obtained as

$$r_{(A,B)} = \min(r_a, r_b) - \text{rank}(AB^T) \ , \tag{2.19}$$

where r_a and r_b are the rank of A and B, respectively.

The subspace spanned by u_i and v_i

Let us now study the subspace spanned by u_i and v_i when θ_i is acute. This two-dimensional subspace is denoted by S_i:

$$S_i = \text{range}\begin{pmatrix} u_i & v_i \end{pmatrix} \ .$$

It has two important properties:

1. It is invariant under $\Pi_{\text{row}(A)}$ and $\Pi_{\text{row}(B)}$:

   $$x \in S_i \Rightarrow \Pi_{\text{row}(A)} x \in S_i \text{ and } \Pi_{\text{row}(B)} x \in S_i \ .$$

   This is a direct consequence of (2.18).

2. It is orthogonal to all other subspaces of this kind:

$$S_i \perp S_j \text{ if } j \neq i \ .$$

This can be seen as follows. Let u_j and v_j span the subspace S_j. The vector $u_i \in S_i$ is orthogonal to $u_j \in S_j$ (by definition) and $u_i \in S_i$ is also orthogonal to $v_j \in S_j$:

$$\begin{aligned} u_i^T v_j &= (\Pi_{\mathrm{row}(A)} u_i)^T v_j \ , \\ &= u_i^T \Pi_{\mathrm{row}(A)} v_j \ , \\ &= u_i^T u_j \cos\theta_j = 0 \ . \end{aligned}$$

Since S_i is invariant under $\Pi_{\mathrm{row}(A)}$ and $\Pi_{\mathrm{row}(B)}$, it is also invariant under $I - \Pi_{\mathrm{row}(A)} = \Pi_{\mathrm{row}(A)^\perp}$ and $I - \Pi_{\mathrm{row}(B)} = \Pi_{\mathrm{row}(B)^\perp}$. Hence, $\Pi_{\mathrm{row}(A)^\perp} v_i \in S_i$ and $\Pi_{\mathrm{row}(B)^\perp} u_i \in S_i$. The angles between u_i, v_i and their orthogonal projections onto the orthogonal complements of $\mathrm{row}(B)$ and $\mathrm{row}(A)$, respectively, can be deduced from geometric properties in the plane S_i. This plane is shown in Figure 2.10, in which θ_i is the ith principal angle between $\mathrm{row}(A)$ and $\mathrm{row}(B)$. By normalizing the orthogonal projection of the principal vector v_i onto the row space of $A^\perp$, we obtain x_i. The normalized orthogonal projection of u_i onto $\mathrm{row}(B)^\perp$ is y_i. From Figure 2.10 it immediately follows that

$$\begin{aligned} \Pi_{\mathrm{row}(A)} y_i &= u_i \sin\theta_i \\ \Pi_{\mathrm{row}(B)} x_i &= v_i \sin\theta_i \end{aligned} \quad \text{and} \quad \begin{aligned} \Pi_{\mathrm{row}(A)^\perp} v_i &= x_i \sin\theta_i \\ \Pi_{\mathrm{row}(B)^\perp} u_i &= y_i \sin\theta_i \ . \end{aligned}$$

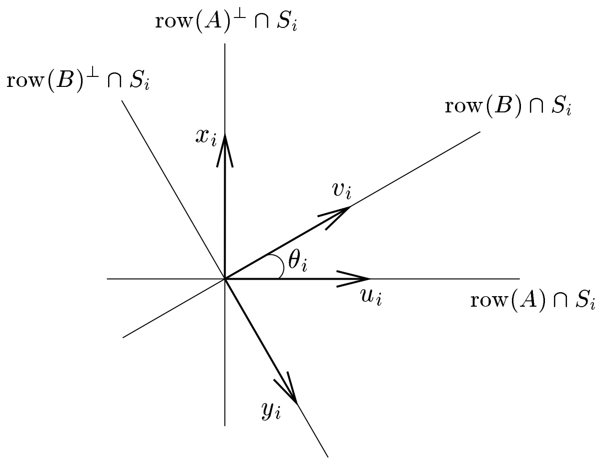


Figure 2.10: The subspace S_i, spanned by u_i and v_i. From this figure we can deduce some important properties of the principal angles between $\mathrm{row}(A)$ and $\mathrm{row}(B)$, between $\mathrm{row}(A)$ and $\mathrm{row}(B)^\perp$, between $\mathrm{row}(B)^\perp$ and $\mathrm{row}(A)$ and between $\mathrm{row}(A)^\perp$ and $\mathrm{row}(B)^\perp$.

Furthermore,

$$\cos\left[x_i \sphericalangle y_i\right] = |x_i^T y_i| = \cos\theta_i \ .$$

We also see that

$$\begin{aligned}
\Pi_{\mathrm{row}(A)^\perp}\Pi_{\mathrm{row}(B)}v_i &= x_i \sin\theta_i \ , \\
\Pi_{\mathrm{row}(B)^\perp}\Pi_{\mathrm{row}(A)}u_i &= y_i \sin\theta_i \ , \\
\Pi_{\mathrm{row}(A)^\perp}\Pi_{\mathrm{row}(B)^\perp}y_i &= -x_i \cos\theta_i \ ,
\end{aligned}$$

which leads to (see also Figure 2.11)

- $(x_i, \sin\theta_i, v_i)$ is a singular triplet of $\Pi_{\mathrm{row}(A)^\perp}\Pi_{\mathrm{row}(B)}$ (see (2.3)) and thus is $\frac{\pi}{2} - \theta_i$ one of the principal angles between $\mathrm{row}(A)^\perp$ and $\mathrm{row}(B)$, with corresponding principal directions $x_i \in \mathrm{row}(A)^\perp$ and $v_i \in \mathrm{row}(B)$ (see Property 2.9),
- $(y_i, \sin\theta_i, u_i)$ is a singular triplet of $\Pi_{\mathrm{row}(B)^\perp}\Pi_{\mathrm{row}(A)}$ and thus is $\frac{\pi}{2} - \theta_i$ one of the principal angles between $\mathrm{row}(B)^\perp$ and $\mathrm{row}(A)$, with corresponding principal directions y_i and u_i,
- $(x_i, \cos\theta_i, -y_i)$ is a singular triplet of $\Pi_{\mathrm{row}(A)^\perp}\Pi_{\mathrm{row}(B)^\perp}$ and thus is θ_i one of the principal angles between $\mathrm{row}(A)^\perp$ and $\mathrm{row}(B)^\perp$, with corresponding principal directions x_i and $-y_i$.

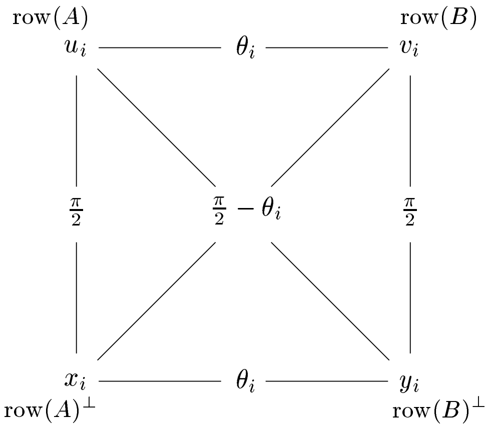


Figure 2.11: Schematic representation of the relations between the principal angles and directions of the following combinations: $[A \sphericalangle B]$, $\left[A \sphericalangle B^\perp\right]$, $\left[A^\perp \sphericalangle B\right]$ and $\left[A^\perp \sphericalangle B^\perp\right]$ in S_i. The vectors u_i and v_i are the ith principal directions in $\mathrm{row}(A)$ and $\mathrm{row}(B)$. The angle between them is θ_i, the ith principal angle between $\mathrm{row}(A)$ and $\mathrm{row}(B)$. The vectors x_i and y_i are the principal directions in $\mathrm{row}(A)^\perp$ and $\mathrm{row}(B)^\perp$, respectively.

We can generalize the above results to all the principal angles between and principal directions in row(A) and row(B). This is done in Appendix A.

2.5.5 The cosines of the principal angles and the principal directions as eigenvalues and eigenvectors

Let $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{q\times n}$ have rank r_a and r_b, respectively. It can be shown (see e.g. [58]) that the principal angles between and the principal directions in row(A) and row(B) follow from the symmetric generalized eigenvalue problem:

$$\begin{pmatrix} 0 & AB^T \\ BA^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} AA^T & 0 \\ 0 & BB^T \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \lambda \ , \qquad (2.20)$$
$$\text{subject to } x^T AA^T x = 1 \text{ and } y^T BB^T y = 1.$$

Assume that the $p+q$ (real) eigenvalues λ_i are sorted in non-ascending order as

$$\lambda_1 \geq \ldots \geq \lambda_{p+q} \ ,$$

then one can show that

$$\lambda_1 = \cos\theta_1, \ldots, \lambda_{\min(r_a,r_b)} = \cos\theta_{\min(r_a,r_b)} \geq 0 \ , \qquad (2.21a)$$
$$\lambda_{p+q} = -\cos\theta_1, \ldots, \lambda_{p+q-\min(r_a,r_b)+1} = -\cos\theta_{\min(r_a,r_b)} \ , \qquad (2.21b)$$
$$\lambda_{\min(r_a,r_b)+1} = \lambda_{\min(r_a,r_b)+2} = \cdots = \lambda_{p+q-\min(r_a,r_b)} = 0 \ . \qquad (2.21c)$$

The vectors $A^T x_i$ and $B^T y_i$, for $i = 1, \ldots, \min(r_a, r_b)$ where x_i and y_i satisfy (2.20) with $\lambda = \lambda_i$, are the principal directions corresponding to the principal angle θ_i.
For quite some derivations, the following formula will prove to be useful.
If $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{q\times n}$ are of full row rank and assume $p \leq q$, then the squared cosines of the principal angles between row(A) and row(B) are equal to the eigenvalues of $(AA^T)^{-1}AB^T(BB^T)^{-1}BA^T$:

$$\boxed{\cos^2[A \sphericalangle B] = \lambda\left((AA^T)^{-1}AB^T(BB^T)^{-1}BA^T\right) \ .} \qquad (2.22)$$

This follows from the characterization of the cosines as eigenvalues (see (2.20) and (2.21a)), but it can also be deduced from Property 2.9, where the cosines are expressed as singular values of the product of orthogonal projectors. Indeed, the eigenvalues of $(AA^T)^{-1}AB^T(BB^T)^{-1}BA^T$ are the largest p eigenvalues of $A^T(AA^T)^{-1}AB^T(BB^T)^{-1}B$. Since A and B are full row rank matrices, the latter matrix is equal to $\Pi_{\text{row}(A)}\Pi_{\text{row}(B)}$. By applying the defining properties of orthogonal projectors, it is easy to show that the largest p eigenvalues of $\Pi_{\text{row}(A)}\Pi_{\text{row}(B)}$ are equal to its largest p squared singular values, which are the squared cosines of the principal angles between the row spaces of A and B by Property 2.9.

2.5.6 An efficient implementation based on the LQ factorization

Since we will have to compute the principal angles between subspaces in $\mathbb{R}^n$, where n is a large number, e.g. 10 000, it is useful to have an efficient algorithm. We present here an algorithm that is based on the LQ decomposition, which is first defined.

Definition 2.5. The LQ factorization of a matrix
The LQ factorization of a real $m \times n$ matrix A is given by

$$A = LQ^T \ ,$$

where $Q \in \mathbb{R}^{n \times n}$ is orthogonal and $L \in \mathbb{R}^{m \times n}$ is lower triangular.

A schematic representation is given in Figure 2.12.
Note, that if the matrix A is of full row rank ($\mathrm{rank}(A) = m$), then the first m columns of Q are an orthonormal basis for $\mathrm{row}(A)$. Let us denote this submatrix by Q_1. Then, the orthogonal projector onto the row space of A can be written as (see (2.13))

$$\Pi_{\mathrm{row}(A)} = Q_1 Q_1^T \ . \tag{2.23}$$

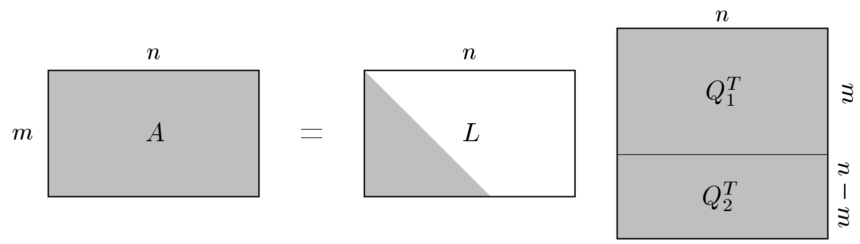


Figure 2.12: The LQ factorization of a matrix A. If A is of full row rank, then the columns of Q_1 are an orthonormal basis for the row space of A.

Remark 2.3. The LQ decomposition as a QR decomposition
The LQ decomposition of a matrix A boils down to the QR decomposition of A^T, which is the numerical version of the Gram-Schmidt orthogonalization (see e.g. [146]). If $A^T = QR$, where Q is orthogonal and R upper triangular, then $A = R^T Q^T$, which gives a factorization of A into a lower triangular matrix R^T and the transpose of an orthogonal matrix, as in the LQ factorization. ○

In the following theorem, which is hidden in the function `sto_stat.m` in [159], it will become clear that for the computation of the principal angles between two subspaces in $\mathbb{R}^n$, we only need the triangular part of the L matrix of two LQ decompositions and we never have to calculate the Q matrices in the decompositions. Since for the computation of the principal angles, the LQ decomposition is made of matrices with usually many more columns than rows, this leads to an important reduction of the data storage and to computational efficiency. Moreover, it is fairly simple to update the L factors whenever a new row or column is appended to the A matrix [57, pp. 610–611].

Theorem 2.11. An efficient computational scheme for the principal angles between two subspaces
Let $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{q\times n}$, where $p \leq q$ and $p+q \leq n$, be matrices of full row rank. The principal angles between $\mathrm{row}(A)$ *and* $\mathrm{row}(B)$ *can be computed as follows.*

1. *Compute the triangular part of the LQ factorization of the matrix* $\begin{pmatrix} A \\ B \end{pmatrix}$. *The triangular part is denoted by*
$$\begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \in \mathbb{R}^{(p+q)\times(p+q)} ,$$
where $L_{11} \in \mathbb{R}^{p\times p}$, $L_{21} \in \mathbb{R}^{q\times p}$ *and* $L_{22} \in \mathbb{R}^{q\times q}$.
2. *Compute the triangular part of the LQ factorization of* $\begin{pmatrix} L_{21} & L_{22} \end{pmatrix}$:
$$\begin{pmatrix} L_{21} & L_{22} \end{pmatrix} = \begin{pmatrix} S & 0 \end{pmatrix} T .$$
The resulting lower triangular matrix $S \in \mathbb{R}^{q\times q}$ *is non-singular.*
3. *The cosines of the principal angles between* $\mathrm{row}(A)$ *and* $\mathrm{row}(B)$ *are the singular values of* $S^{-1}L_{21}$.

Proof.
Let the LQ factorization of $\begin{pmatrix} A \\ B \end{pmatrix}$ be
$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \end{pmatrix} \begin{pmatrix} Q_1^T \\ Q_2^T \\ Q_3^T \end{pmatrix} ,$$
where $Q_1 \in \mathbb{R}^{n\times p}, Q_2 \in \mathbb{R}^{n\times q}, L_{11} \in \mathbb{R}^{p\times p}, L_{21} \in \mathbb{R}^{q\times p}$ and $L_{22} \in \mathbb{R}^{q\times q}$ and the matrix $Q = \begin{pmatrix} Q_1 & Q_2 & Q_3 \end{pmatrix}$ is orthogonal. The matrices A and B can thus be written as
$$A = L_{11}Q_1^T , \tag{2.24}$$
$$B = L_{21}Q_1^T + L_{22}Q_2^T . \tag{2.25}$$
Since A and B are full row rank matrices and the columns of Q are orthogonal, we know that L_{11} and $\begin{pmatrix} L_{21} & L_{22} \end{pmatrix}$ also have full row rank (p and q, respectively).

In order to compute the principal angles between the row spaces of A and B, we consider the singular values of $\Pi_{\mathrm{row}(B)}\Pi_{\mathrm{row}(A)}$ (see Property 2.9), which can be written as

$$\begin{aligned}\Pi_{\mathrm{row}(B)}\Pi_{\mathrm{row}(A)} &= B^{\dagger}B\Pi_{\mathrm{row}(A)} && \text{by (2.14)} ,\\ &= B^{\dagger}(B/A) ,\end{aligned} \tag{2.26}$$

where $B^{\dagger}$ is the Moore–Penrose pseudo-inverse of B and $\mathrm{row}(B/A)$ is the orthogonal projection of the row space of B onto the row space of A.
We first compute this orthogonal projection. Next, we derive an expression for $B^{\dagger}$ and subsequently multiply the two factors.
From (2.23) it follows that the orthogonal projector onto $\mathrm{row}(A)$ is

$$\Pi_{\mathrm{row}(A)} = Q_1Q_1^T .$$

Using (2.25) leads to

$$B\Pi_{\mathrm{row}(A)} = (L_{21}Q_1^T + L_{22}Q_2^T)Q_1Q_1^T ,$$

which is, due to the orthonormality of the columns of Q, equal to

$$B\Pi_{\mathrm{row}(A)} = L_{21}Q_1^T . \tag{2.27}$$

The first factor in (2.26) is $B^{\dagger}$, which can be computed as $B^T(BB^T)^{-1}$ since B is of full row rank (see (2.7)). Hence,

$$B^{\dagger} = (Q_1L_{21}^T + Q_2L_{22}^T)(L_{21}L_{21}^T + L_{22}L_{22}^T)^{-1} . \tag{2.28}$$

This can be simplified by applying the LQ decomposition to $\begin{pmatrix}L_{21} & L_{22}\end{pmatrix}$:

$$\begin{pmatrix}L_{21} & L_{22}\end{pmatrix} = \begin{pmatrix}S & 0\end{pmatrix}\begin{pmatrix}T_{11}^T & T_{21}^T\\ T_{12}^T & T_{22}^T\end{pmatrix} = S\begin{pmatrix}T_{11}^T & T_{21}^T\end{pmatrix} , \tag{2.29}$$

where $T_{11} \in \mathbb{R}^{p\times q}$, $T_{21} \in \mathbb{R}^{q\times q}$, $S \in \mathbb{R}^{q\times q}$ is a full rank lower-triangular matrix[3] and the columns of $\begin{pmatrix}T_{11}\\ T_{21}\end{pmatrix}$ are orthonormal. Substituting L_{21} and L_{22} in (2.28) by the expressions in (2.29) gives

$$\begin{aligned}B^{\dagger} &= (Q_1T_{11} + Q_2T_{21})S^T\left(S\begin{pmatrix}T_{11}^T & T_{21}^T\end{pmatrix}\begin{pmatrix}T_{11}\\ T_{21}\end{pmatrix}S^T\right)^{-1} ,\\ &= (Q_1T_{11} + Q_2T_{21})S^{-1} .\end{aligned} \tag{2.30}$$

Multiplying (2.27) and (2.30) gives the following expression for the product of the orthogonal projectors on $\mathrm{row}(A)$ and $\mathrm{row}(B)$:

$$\Pi_{\mathrm{row}(B)}\Pi_{\mathrm{row}(A)} = (Q_1T_{11} + Q_2T_{21})S^{-1}L_{21}Q_1^T .$$

[3]The lower-triangular matrix S is nonsingular because $\begin{pmatrix}L_{21} & L_{22}\end{pmatrix}$ has full row rank.

Since the columns of $Q_1T_{11}+Q_2T_{21}$ and Q_1 are orthonormal, the singular values of $\Pi_{\text{row}(A)}\Pi_{\text{row}(B)} \in \mathbb{R}^{n\times n}$ are equal to the singular values of $S^{-1}L_{21} \in \mathbb{R}^{q\times p}$ except for $n-p$ zero singular values (see Property 2.6). The cosines of the p principal angles between row(A) and row(B) are thus the p singular values of $S^{-1}L_{21}$. $\square$

The above described computational scheme leads to a very simple Matlab program, which is given in Table 2.1.

```
function cosines = cosines_lq(A,B)

p = size(A,1);
q = size(B,1);
L = triu(qr([A;B]'));
L = L(1:p+q,p+1:p+q);
S = triu(qr(L));
S = S(1:q,:);
L = L(1:p,:);
cosines = svd(L/S);
```

Table 2.1: The Matlab program `cosines_lq.m` for the computation of the principal angles between the row spaces of the matrices A and B via the method described in Section 2.5.6.

2.6 Canonical correlation analysis

Canonical correlation analysis (CCA) is a tool in statistics that is used for measuring the linear relationship between two sets of random variables, which was developed by Hotelling [68]. It is the statistical interpretation of the principal angles and principal directions of two subspaces. A wide variety of applications exists in econometrics, biology [53], meteorology, etc. Quite recently, this technique has also been introduced in the fields of learning and signal processing [18, 154]. Larimore has developed a subspace identification method that is based on CCA [94].

In the book by Anderson [4] a clear description of the method is given as well as the statistical properties for sample estimates.

As we have mentioned before, CCA is very closely related to the principal angles between and principal directions in two linear subspaces. Nevertheless, we want to describe it briefly, not only because we will encounter it further on, but also because we introduce in this section some notation and concepts that will be used in the following chapters. We also seize the opportunity to show how statistical properties like independence and correlation can be interpreted in a geometric framework. This is done in Section 2.6.1. In Section 2.6.2 the canonical correlation analysis for random variables is explained and in Section 2.6.3 this is extended to stationary stochastic processes.

2.6.1 Statistical properties in a geometric framework

Let X be a single-component random variable[4] with probability density function $f(x)$. Then, the expected value of X is defined as $E\{X\} = \int x f(x)\, dx$. By the law of large numbers (see e.g. [123, pp. 263–264]), the expected value can also be obtained as follows:

$$E\{X\} = \lim_{j\to\infty} \frac{1}{j} \sum_{k=0}^{j-1} x_k \ ,$$

where x_k ($k = 0, 1, \ldots$) are independent realizations of the random variable X. Henceforth we will assume that the mean has been subtracted, so we will work with zero-mean random variables.
The *variance* of the zero-mean random variable X is

$$E\left\{X^2\right\} = \lim_{j\to\infty} \frac{1}{j} \sum_{k=0}^{j-1} x_k^2 \ .$$

We denote it by σ_x^2. Consider the row vector

$$x^{(j)} = \frac{1}{\sqrt{j}} \begin{pmatrix} x_0 & \cdots & x_{j-1} \end{pmatrix} \ , \tag{2.31}$$

of j normalized (by $\frac{1}{\sqrt{j}}$) realizations of X. Then, the variance of X can be interpreted as the squared length of this vector, for $j \to \infty$:

$$\sigma_x^2 = \lim_{j\to\infty} \|x^{(j)}\|^2 \ .$$

The *covariance* of two zero-mean random variables X and Y is equal to the expected value of XY and can thus be obtained as

$$E\{XY\} = \sigma_{xy} = \lim_{j\to\infty} \frac{1}{j} \sum_{k=0}^{j-1} x_k y_k \ .$$

This can be seen as the inner product of the two row vectors $x^{(j)}$ and $y^{(j)}$ of j normalized realizations of X and Y, respectively, where $j \to \infty$:

$$\sigma_{xy} = \lim_{j\to\infty} x^{(j)} {y^{(j)}}^T \ .$$

[4] Random variables will be denoted by capitals, even if they consist of only one component.

The *correlation coefficient* ρ_{xy} between the zero-mean variables X and Y is defined as

$$\rho_{xy} = \frac{E\{XY\}}{\sqrt{E\{X^2\}}\sqrt{E\{Y^2\}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \ , \tag{2.32}$$

which is in geometrical terms the cosine of the angle between the vectors $x^{(j)}$ and $y^{(j)}$, for $j \to \infty$:

$$\rho_{xy} = \lim_{j\to\infty} \cos\left[x^{(j)} \sphericalangle\, y^{(j)}\right] \ .$$

The independence of X and Y implies[5] that $\rho_{xy} = 0$ and hence that the vectors of realizations $x^{(j)}$ and $y^{(j)}$ are orthogonal for $j \to \infty$:

$$X \text{ and } Y \text{ are independent } \Rightarrow \left[x^{(j)} \sphericalangle\, y^{(j)}\right] = \frac{\pi}{2} \ .$$

The converse is not true in general, as is shown in the following example.

Example 2.1. Correlation is not equivalent to dependence
Let the random variables X and Y be equal to $X = \cos\theta$ and $Y = \sin\theta$, where θ has a uniform distribution on $[0, 2\pi)$. The covariance of X and Y is equal to

$$E\{XY\} = \int_0^{2\pi} \frac{1}{2\pi} \cos\theta \sin\theta \, d\theta = 0 \ .$$

Therefore, the variables X and Y are uncorrelated (see (2.32)). However, since $X^2 = 1 - Y^2$, it is clear that they are not independent. This can also be seen in the scatter diagram of X and Y, which is given in the left picture of Figure 2.13. We take 100 sample pairs (x, y) from the described random variables and plot their values, which all lie on a circle, while the scatter diagram of 100 samples of two independent variables with uniform distribution on $[-1, 1]$ looks like the one in the right picture of Figure 2.13.
We illustrate the geometric interpretation of the correlation of two random variables by taking samples θ_k from a uniform distribution on $[0, 2\pi)$ and computing the angle between the vectors $x^{(j)} = \frac{1}{\sqrt{j}}\begin{pmatrix}\cos\theta_0 & \cos\theta_1 & \cdots & \cos\theta_{j-1}\end{pmatrix}$ and $y^{(j)} = \frac{1}{\sqrt{j}}\begin{pmatrix}\sin\theta_0 & \sin\theta_1 & \cdots & \sin\theta_{j-1}\end{pmatrix}$ for increasing j ($j = 1, 2, \ldots, 1000$). In Figure 2.14 the result is shown for 100 trials. Note that the variance on the estimate decreases for increasing j and that for $j \to \infty$, the angle between $x^{(j)}$ and $y^{(j)}$ is equal to $\frac{\pi}{2}$. $\Diamond$

The correlation coefficient is a measure for the *linear* interdependence between the two variables. Any non-linear relationship, as e.g. in Example 2.1, is not taken into account. For Gaussian random variables, however, one can show that independence is equivalent to $\rho_{xy} = 0$ and hence to the orthogonality of

[5] If the zero-mean random variables X and Y are independent, then $E\{XY\} = E\{X\}E\{Y\}$, which is equal to 0.

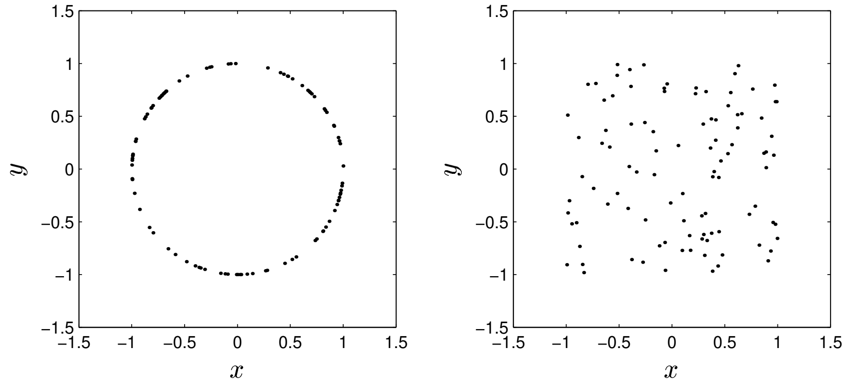


Figure 2.13: In the left picture the scatter diagram is given of 100 realizations of X and Y, where $X = \cos\theta$ and $Y = \sin\theta$ and θ has a uniform distribution over $[0, 2\pi)$. The right picture shows the scatter diagram of two independent random variables that are uniformly distributed over $[-1, 1]$.

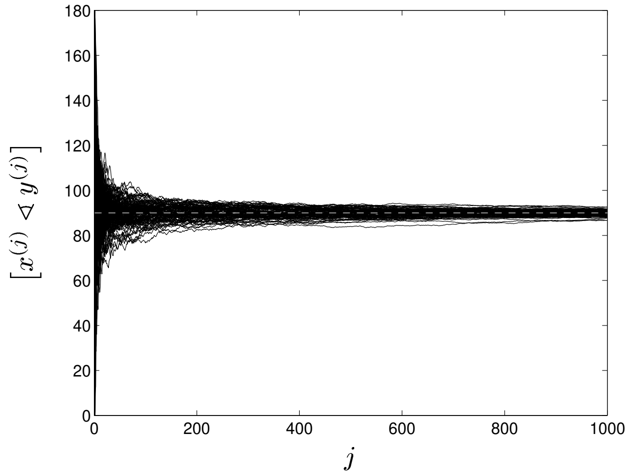


Figure 2.14: The angle between $x^{(j)}$ and $y^{(j)}$ for increasing j and 100 different trials. The angle is expressed in degrees.

$x^{(j)}$ and $y^{(j)}$ for $j \to \infty$ (see also Section 4.3.4, where we discuss the mutual information of Gaussian variables).

For the zero-mean multivariate random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, the covariance matrices $Q_x = E\left\{XX^T\right\} \in \mathbb{R}^{p\times p}$, $Q_{xy} = E\left\{XY^T\right\} \in \mathbb{R}^{p\times q}$ and $Q_y = E\left\{YY^T\right\} \in \mathbb{R}^{q\times q}$ can be viewed as matrices of inner products (also called Gramians). Let the matrix $X^{(j)} \in \mathbb{R}^{p\times j}$ consist of j realizations of the random variable X, normalized by $\frac{1}{\sqrt{j}}$ and let $Y^{(j)} \in \mathbb{R}^{q\times j}$ contain j realizations of Y, normalized by $\frac{1}{\sqrt{j}}$:

$$X^{(j)} = \frac{1}{\sqrt{j}} \begin{pmatrix} x_0 & \cdots & x_{j-1} \end{pmatrix} \text{ and } Y^{(j)} = \frac{1}{\sqrt{j}} \begin{pmatrix} y_0 & \cdots & y_{j-1} \end{pmatrix} . \tag{2.33}$$

Then, we obtain

$$\begin{aligned} Q_x &= \lim_{j\to\infty} X^{(j)} X^{(j)^T} , \\ Q_y &= \lim_{j\to\infty} Y^{(j)} Y^{(j)^T} , \\ Q_{xy} &= \lim_{j\to\infty} X^{(j)} Y^{(j)^T} , \\ Q_{yx} &= \lim_{j\to\infty} Y^{(j)} X^{(j)^T} . \end{aligned} \tag{2.34}$$

2.6.2 Canonical correlation analysis for random variables

In canonical correlation analysis the interrelation of two sets of random variables is studied. The aim is to find two bases, one in each set, that are optimal with respect to the correlations between the two sets. The resulting basis vectors are called the canonical variates and the correlation coefficients between the canonical variates are the canonical correlations. Canonical correlation analysis is particularly useful if one studies the correlation structure of large sets of variables. By taking only the most highly correlated linear combinations from each set, a considerable dimensionality reduction can be made to both sets and their correlation structure.

Canonical correlation analysis is the statistical interpretation of the geometric tool of principal angles between and principal directions in linear subspaces. Moreover, if we have j observations of two random variables X and Y, then their canonical correlations can be estimated by computing the principal angles between the row spaces of the observation matrices $X^{(j)}$ and $Y^{(j)}$ (see (2.33)). The cosines of the angles between the principal directions (the principal angles) correspond to the correlation coefficients of the canonical variates (the canonical correlations). The different canonical variates of X are uncorrelated, which also holds for the canonical variates of Y, while the correlation of a canonical variate of X and corresponding canonical variate of Y is maximal.

Let X be a zero-mean p-component random variable and Y a zero-mean q-component random variable with joint covariance matrix $Q \in \mathbb{R}^{(p+q)\times(p+q)}$:

$$Q = E\left\{ \begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} X^T & Y^T \end{pmatrix} \right\} = \begin{pmatrix} Q_x & Q_{xy} \\ Q_{yx} & Q_y \end{pmatrix} ,$$

The matrices Q_x and Q_y are the within-sets covariance matrices of X and Y, respectively, and the matrix $Q_{xy} = Q_{yx}^T$ is the between-sets covariance matrix. Let r_x be the rank of Q_x and r_y the rank of Q_y. The canonical correlations of the random variables X and Y are the largest $\min(r_x, r_y)$ eigenvalues in the following generalized eigenvalue problem

$$\begin{pmatrix} 0 & Q_{xy} \\ Q_{yx} & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} Q_x & 0 \\ 0 & Q_y \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \lambda \,, \qquad \text{subject to } x^T Q_x x = 1 \text{ and } y^T Q_y y = 1. \tag{2.35}$$

Assume that the eigenvalues are ordered in non-ascending order. The canonical variates are then the random variables $x_k^T X$ and $y_k^T Y$, where x_k and y_k satisfy (2.35), with λ the kth canonical correlation. Of course, if Q_x and Q_y are full rank matrices, and assume $p \leq q$, then the squared canonical correlations of X and Y, which we denote by $\mathrm{cc}(X, Y)$, can also be obtained as the eigenvalues of $Q_x^{-1} Q_{xy} Q_y^{-1} Q_{yx}$:

$$\boxed{\mathrm{cc}^2(X, Y) = \lambda(Q_x^{-1} Q_{xy} Q_y^{-1} Q_{yx}) \ .}$$

Remark 2.4. The estimation of the canonical correlations and canonical variates
If only a finite number j of measurements of both random variables are given, the canonical correlations and canonical variates can be estimated as follows.

1. Estimate the mean of the variables X and Y:

$$E\{X\} \approx \bar{X} = \frac{1}{j} \sum_{k=0}^{j-1} x_k \quad \text{and} \quad E\{Y\} \approx \bar{Y} = \frac{1}{j} \sum_{k=0}^{j-1} y_k \ ,$$

   and subtract it from the measurements: $x_k^* = x_k - \bar{X}$, $y_k^* = y_k - \bar{Y}$ for $k = 0, \ldots, j-1$.

2. Estimate the covariance matrices Q_x, Q_y and Q_{xy}:
   $Q_x \approx \frac{1}{j} \sum_{k=0}^{j-1} x_k^* x_k^{*^T}$, $Q_y \approx \frac{1}{j} \sum_{k=0}^{j-1} y_k^* y_k^{*^T}$, $Q_{xy} \approx \frac{1}{j} \sum_{k=0}^{j-1} x_k^* y_k^{*^T}$.

3. Solve the generalized eigenvalue problem (2.35).

It is readily seen that these estimates can also be obtained by computing the cosines of the principal angles between and principal directions in the row spaces of $X^{(j)^*} = \frac{1}{\sqrt{j}} \begin{pmatrix} x_0^* & \cdots & x_{j-1}^* \end{pmatrix}$ *and* $Y^{(j)^*} = \frac{1}{\sqrt{j}} \begin{pmatrix} y_0^* & \cdots & y_{j-1}^* \end{pmatrix}$ *via the generalized eigenvalue problem in (2.20).* ○

Example 2.2. The canonical correlations of two Gaussian random variables
We estimate the canonical correlations of a 3-component Gaussian random variable X and a 4-component Gaussian random variable Y by computing the principal angles between the row spaces of the matrices $X^{(j)}$ and $Y^{(j)}$. The

covariance matrices of the random variables are equal to

$$Q_x = \begin{pmatrix} 0.6244 & -0.0484 & -0.1320 \\ -0.0484 & 0.5605 & 0.0952 \\ -0.1320 & 0.0952 & 0.5217 \end{pmatrix}$$

$$Q_y = \begin{pmatrix} 0.3274 & -0.0581 & -0.0658 & 0.0303 \\ -0.0581 & 0.2727 & -0.0930 & 0.0456 \\ -0.0658 & -0.0930 & 0.4621 & -0.0753 \\ 0.0303 & 0.0456 & -0.0753 & 0.1904 \end{pmatrix}$$

and the cross-covariance matrix of X and Y is

$$Q_{xy} = \begin{pmatrix} 0.2314 & -0.2131 & -0.0346 & -0.0922 \\ 0.1078 & -0.1217 & -0.2077 & 0.0469 \\ 0.0158 & -0.1479 & 0.1821 & -0.0932 \end{pmatrix}.$$

We take samples x_k and y_k from the distributions and compute the principal angles between the row spaces of the matrices $X^{(j)}$ and $Y^{(j)}$, defined in (2.33), for increasing j ($j = 10, 11, \ldots, 1000$). In Figure 2.15 the result is shown for 100 trials. Via the given covariance matrices we can also compute the exact canonical correlations, which are equal to the cosines of the principal angles for $j \to \infty$. The corresponding angles are equal to $19.1377\,^\circ$, $49.4039\,^\circ$ and $79.0149\,^\circ$, which are depicted in Figure 2.15 by the dashed lines. The variance of the estimates clearly decreases as j grows. $\Diamond$

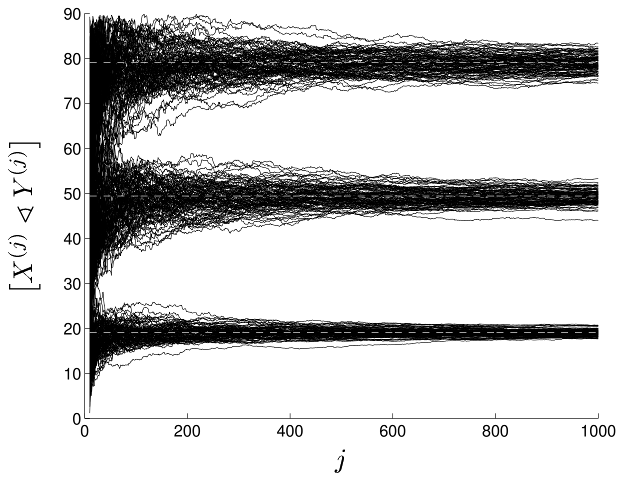


Figure 2.15: The principal angles between the row spaces of $X^{(j)}$ and $Y^{(j)}$ for increasing j (from 10 till 1000). The values for $j = \infty$ are depicted by the dashed lines. The cosines of these angles are the canonical correlations of the random variables X and Y. The angles are expressed in degrees.

2.6.3 Canonical correlation analysis for stationary and ergodic stochastic processes

A stochastic process $\{x(k)\}$ can be viewed as an infinite sequence of random variables $X(0), X(1), X(2), \ldots$ The canonical correlations and canonical variates of the stochastic processes $\{x(k)\}$ and $\{y(k)\}$ can therefore be obtained by applying canonical correlation analysis to the random variables $X_{0|\infty} = \begin{pmatrix} X(0) \\ X(1) \\ \vdots \end{pmatrix}$ and $Y_{0|\infty} = \begin{pmatrix} Y(0) \\ Y(1) \\ \vdots \end{pmatrix}$. The geometrical interpretation of the canonical correlations and the canonical variates of two random variables as given in the previous section, can now be extended to stochastic processes. Assume that $\{x(k)\}$ and $\{y(k)\}$ are zero-mean stochastic processes. Let $X_{0|\infty}^{(j)}$ and $Y_{0|\infty}^{(j)}$ be matrices of j normalized realizations of the random variables $X_{0|\infty}$ and $Y_{0|\infty}$:

$$X_{0|\infty}^{(j)} = \frac{1}{\sqrt{j}} \begin{pmatrix} x(0)_1 & x(0)_2 & \cdots & x(0)_j \\ x(1)_1 & x(1)_2 & \cdots & x(1)_j \\ x(2)_1 & x(2)_2 & \cdots & x(2)_j \\ \vdots & \vdots & \ldots & \vdots \end{pmatrix} ,$$

and analogously for $Y_{0|\infty}^{(j)}$. Then, the canonical correlations of $\{x(k)\}$ and $\{y(k)\}$ are the cosines of the principal angles between the row spaces of $X_{0|\infty}^{(j)}$ and $Y_{0|\infty}^{(j)}$, for $j \to \infty$. Of course, we usually do not have several realizations of the stochastic processes, only the samples of one realization. However, if the processes are stationary and ergodic, then one realization suffices to compute the canonical correlations and canonical variates.

Stationarity If $\{x(k)\}$ is a stationary stochastic process, then its *autocovariance function*, which is defined as

$$R_x(k, l) = E\left\{X(k)X(l)^T\right\} ,$$

only depends on the time difference $k - l$:

$$R_x(\tau) = E\left\{X(k+\tau)X(k)^T\right\} ,$$

and thus $R_x(-\tau) = R_x(\tau)^T$. Similarly, the *cross-covariance function* of the stationary processes $\{x(k)\}$ and $\{y(k)\}$, defined as

$$C_{xy}(k, l) = E\left\{X(k)Y(l)^T\right\} ,$$

is only dependent on the time lag $k - l$:

$$C_{xy}(\tau) = E\left\{X(k+\tau)Y(k)^T\right\} ,$$

and $C_{xy}(-\tau) = C_{yx}(\tau)^T$. The covariance matrix of the random variable $X_{0|\infty}$ is then equal to the following symmetric Toeplitz matrix (or block Toeplitz matrix if $\{x(k)\}$ is a multivariate process):

$$Q_{x_\infty} = E\left\{X_{0|\infty}X_{0|\infty}^T\right\} = \begin{pmatrix} R_x(0) & R_x(1)^T & R_x(2)^T & \cdots \\ R_x(1) & R_x(0) & R_x(1)^T & \cdots \\ R_x(2) & R_x(1) & R_x(0) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} ,$$

and analogously for $Q_{y_\infty} = E\left\{Y_{0|\infty}Y_{0|\infty}^T\right\}$. The cross-covariance matrix of $X_{0|\infty}$ and $Y_{0|\infty}$ is the (block) Toeplitz matrix

$$Q_{xy_\infty} = E\left\{X_{0|\infty}Y_{0|\infty}^T\right\} = \begin{pmatrix} C_{xy}(0) & C_{xy}(-1) & C_{xy}(-2) & \cdots \\ C_{xy}(1) & C_{xy}(0) & C_{xy}(-1) & \cdots \\ C_{xy}(2) & C_{xy}(1) & C_{xy}(0) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} .$$

Ergodicity We will always assume that $\{x(k)\}$ and $\{y(k)\}$ are ergodic. Then, the autocovariance functions may be computed as time averages:

$$R_x(\tau) = \lim_{j\to\infty} \frac{1}{j} \sum_{k=0}^{j-1} x(k+\tau)x(k)^T ,$$

and analogously for R_y. Similarly, the cross-covariance function of the processes $\{x(k)\}$ and $\{y(k)\}$ is

$$C_{xy}(\tau) = \lim_{j\to\infty} \frac{1}{j} \sum_{k=0}^{j-1} x(k+\tau)y(k)^T .$$

Consider now the (block) Hankel matrices

$$X_{0|\infty}^{(j)} = \frac{1}{\sqrt{j}} \begin{pmatrix} x(0) & x(1) & x(2) & \cdots & x(j-1) \\ x(1) & x(2) & x(3) & \cdots & x(j) \\ x(2) & x(3) & x(4) & \cdots & x(j+1) \\ \vdots & \vdots & \vdots & \cdots & \vdots \end{pmatrix} , \tag{2.36}$$

$$Y_{0|\infty}^{(j)} = \frac{1}{\sqrt{j}} \begin{pmatrix} y(0) & y(1) & y(2) & \cdots & y(j-1) \\ y(1) & y(2) & y(3) & \cdots & y(j) \\ y(2) & y(3) & y(4) & \cdots & y(j+1) \\ \vdots & \vdots & \vdots & \cdots & \vdots \end{pmatrix} . \tag{2.37}$$

Due to the stationarity and ergodicity, the covariance matrices Q_{x_∞}, Q_{y_∞} and Q_{xy_∞} can be obtained as

$$
\begin{aligned}
Q_{x_\infty} &= \lim_{j\to\infty} X_{0|\infty}^{(j)} X_{0|\infty}^{(j)^T} , \\
Q_{y_\infty} &= \lim_{j\to\infty} Y_{0|\infty}^{(j)} Y_{0|\infty}^{(j)^T} , \\
Q_{xy_\infty} &= \lim_{j\to\infty} X_{0|\infty}^{(j)} Y_{0|\infty}^{(j)^T} .
\end{aligned}
$$

As a result, the canonical correlations and canonical variates of $\{x(k)\}$ and $\{y(k)\}$, denoted by $\mathrm{cc}(x,y)$, can be computed as the limiting values for the cosines of the principal angles between and the principal directions in the row spaces of $X_{0|\infty}^{(j)}$ and $Y_{0|\infty}^{(j)}$ for $j \to \infty$:

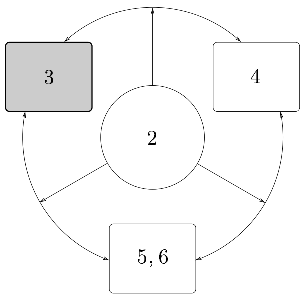
$$
\mathrm{cc}(x,y) = \lim_{j\to\infty} \cos \left[X_{0|\infty}^{(j)} \sphericalangle Y_{0|\infty}^{(j)} \right] .
$$

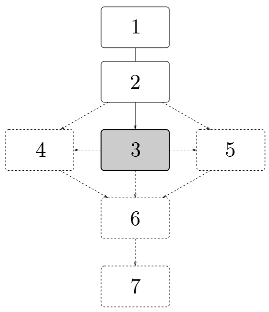
2.7 Conclusions

To facilitate the understanding of the concepts and their properties of this chapter, we have first given a short survey of the singular value decomposition and orthogonal projectors. Next, we have introduced the main geometric and statistical tools that are used in the thesis: the principal angles and canonical correlations. We have given the definition and most important properties of the principal angles between and corresponding principal vectors in two subspaces. We have shown how they can be computed (via the singular value decomposition, the eigenvalue decomposition and the LQ-factorization) and how they are related to orthogonal projections. Furthermore, we have described how the canonical correlations of two random variables and of two stochastic processes can be seen as the cosines of the principal angles between two subspaces. As a side result, we have seen that the data Hankel matrices that we will deal with in the next chapter, naturally arise when considering stationary ergodic stochastic processes.

Chapter 3

Geometric Properties of Linear Stochastic Models





After the introduction of the main geometric and statistical tools, which form the core in the pictogram left, we now arrive at the first block in the picture: the study of linear stochastic models. We give in this chapter insight into the geometrical structure of linear time invariant models by computing the principal angles between the different combinations of past and future input and output spaces. The results will often be referred to in the following chapters. Indeed, as can be seen in Figure 1.10 (or the above right pictogram), this chapter too plays a central role in the thesis.

Contributions The main contributions in this chapter have been described in Section 1.3.1 and Section 1.3.2, illustrated by the diagrams in Figure 1.6 and Figure 1.7.

Section 3.2.5 When discussing parametric descriptions for single input single output (SISO) models, we prove two properties that will be useful later on.

- The row space of the controllability matrix of a SISO model is equal to the column space of the observability matrix and only depends on the poles of the model.
- The row space of the output Hankel matrix of a SISO autoregressive (AR) model has a special structure.

Section 3.4.2 We derive parametric expressions for the principal angles between input and output spaces and the canonical correlations of the corresponding processes.

- We obtain parametric expressions for the principal angles between the $mi \times j$ input and output spaces for finite i and $j \to \infty$. The parameters involved are the state space matrices of the forward innovation model, (A, K, C) and the input covariance matrix S_u.
- By taking $i \to \infty$ in the expressions of the cosines of the principal angles, we obtain the canonical correlations of the corresponding stochastic processes, which are again a function of (A, K, C) and S_u.
- The canonical correlations of the past and future output processes can also be expressed in terms of the forward and backward state covariance matrix. By comparing the two expressions for these canonical correlations, we obtain a simple characterization of the backward state covariance matrix.

Section 3.4.3 Based on the matrix expressions, we prove that the canonical correlations of the future input and output are complementary to the canonical correlations of the past and future output. We show that this property is easily explained when the model is in balanced form. We discuss two balanced realizations: a stochastically balanced form and a minimum phase balanced form.

Section 3.5.1 For SISO models, the parametric expressions for the principal angles and canonical correlations turn out to be independent of the input variance.

Section 3.5.2 We derive formulas for the canonical correlations of the output processes of two autoregressive (AR) models with the same input, and for two autoregressive moving average (ARMA) models with the same input process.

Section 3.5.3 The results are then used to give a new characterization for the canonical correlations of the input and output process of a SISO ARMA model, which leads to a conjecture for the eigenvalues of matrices obtained by solving a Lyapunov equation.

Section 3.5.4 Finally, we derive the canonical correlations of the past and future of an AR model.

Chapter overview We informally introduce linear models in Section 3.1. Next, the model class and assumptions made are presented in Section 3.2. We briefly describe four state space representations: forward, forward innovation, backward and backward innovation. We also introduce in Section 3.2 the past and future input and output block Hankel matrices, as well as the state sequence matrix and some model related matrices, such as the controllability and observability matrices and Gramians, that will be needed. The state space matrices for the inverse model are derived and Section 3.2 is concluded with an overview of parametric descriptions for single input single output (SISO) models. We discuss the moving average (MA) model, the autoregressive (AR) model and the autoregressive moving average (ARMA) model, for which we give a state space description and derive properties. Section 3.3 is a brief description of system identification, which we will refer to in Chapter 5. In Section 3.4 we compute the principal angles between the past and future input and output spaces in terms of the model parameters. From these expressions we derive the canonical correlations of the corresponding processes and their relations. In Section 3.5 we show that for single input single output models, i.e. models that describe scalar stochastic processes, the results of Section 3.4 can be simplified. Moreover, via new expressions for the canonical correlations of two ARMA processes that are driven by the same input process, we prove a new characterization for the canonical correlations of the input and output process of an ARMA model. Finally, a formula for the principal angles between the past and future output spaces of an AR process is derived.

3.1 Introduction

Many technical, economical, physical, biological processes can be modeled as the output of a linear time invariant (LTI) model where the input is stationary white noise, schematically represented in Figure 3.1. Throughout this chapter, the number of components of the stochastic process $\{y(k)\}$ is denoted by m. White noise is a stochastic process of which the subsequent random variables are independent. The autocovariance function of a scalar ($m = 1$) stationary

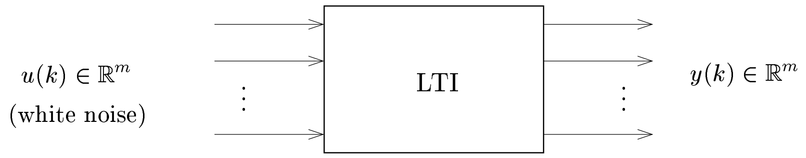


Figure 3.1: Schematic representation of an m-component stochastic process $\{y(k)\}$ that is the output of a linear time invariant (LTI) model driven by a white noise input process $\{u(k)\}$.

white noise process $\{u(k)\}$, e.g., is equal to

$$R_u(\tau) = E\{U(k+\tau)U(k)\} = \sigma_u^2 \delta(\tau) ,$$

where σ_u^2 is the variance of the process and $\delta(\tau)$ is the Kronecker delta, defined as

$$\delta(\tau) = \begin{cases} 1 & \text{for } \tau = 0 , \\ 0 & \text{for } \tau \neq 0 . \end{cases}$$

The power spectral density of $\{u(k)\}$, which is the Fourier transform of its autocovariance function, is constant over the whole frequency axis ($\Phi_u(e^{i\theta}) = \sigma_u^2$). This explains the term 'white', since the contribution to the spectral distribution of the process power is equal for all frequencies, just as for white light.

Remark 3.1. The notions that are used in this introduction, such as the power spectral density of a stochastic process, are explained in detail in the rest of this chapter. ○

By sending the white noise through a linear time invariant model, as in Figure 3.1, the output process $\{y(k)\}$ becomes colored, i.e. its spectral density is not flat any more. An example of such a spectrum can be seen in Figure 3.4 (bottom picture). Also, the random variables $Y(k)$ and $Y(l)$ may be correlated, even if $k \neq l$.
Linear time invariant models are typically used for simulation, prediction, analysis, data compression, etc. In addition, LTI models allow elegant and often intricate geometrical interpretations, and not in the least lead to manageable numerical algorithms.

Example 3.1. The vibration modes in a steel transmitter mast
In this example we consider measurements obtained from an accelerometer on a steel mast that is part of a cellular phone network. Wind is the main cause

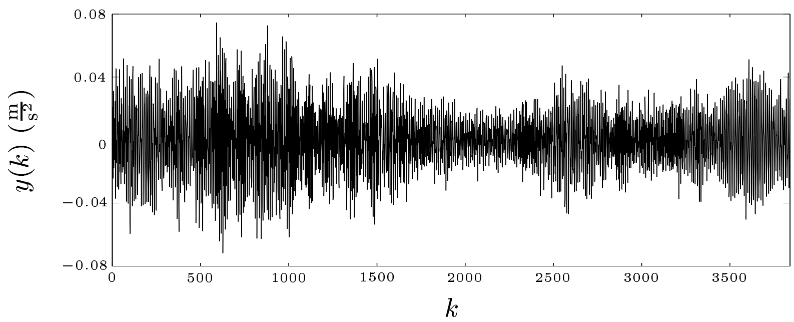


Figure 3.2: Measurements made by an accelerometer on a steel mast for the cellular phone network in Belgium. The acceleration measurements $y(k)$ are expressed in $\frac{\mathrm{m}}{\mathrm{s}^2}$.

of the mast's vibrations. To prevent malfunctioning of the antennae, the rotation at the top should be small. A good model of the vibration modes is an important help in the design process.
Since it is very difficult, if not impossible, to measure the true wind load acting on the structure, only response measurements, namely the accelerations at nine points of the mast, are recorded. It is then assumed that the whole system can be modeled reasonably well as a black box driven by white noise.
The measurements were used for the identification of the vibrational modes of the mast, as described in [118]. For more information, photographs and pictures about this project, see [115].
In Figure 3.2, the 3 839 measurements of one accelerometer are shown as a function of the time index k. Let us check whether the data can be approximated by the output of a finite dimensional LTI model driven by white noise. We estimate the first 60 points of the autocovariance function of $\{y(k)\}$ as follows:

$$R(\tau) = \frac{1}{N-\tau} \sum_{k=0}^{N-\tau-1} y(k)y(k+\tau) \quad \text{for } \tau = 0, 1, \ldots, 59 ,$$

where $N = 3839$, the number of data. The values $R(1), \ldots, R(59)$ are put in the Hankel matrix $\mathcal{R}_{30}$, the rank of which gives an indication of the model order (see Section 3.3.2):

$$\mathcal{R}_{30} = \begin{pmatrix} R(1) & R(2) & \cdots & R(30) \\ R(2) & R(3) & \cdots & R(31) \\ \vdots & \vdots & \ldots & \vdots \\ R(30) & R(31) & \cdots & R(59) \end{pmatrix} .$$

In order to determine the numerical rank of $\mathcal{R}_{30}$, we look at its singular values σ_j $(j = 1, \ldots, 30)$ in Figure 3.3. Although the matrix is of full rank, the

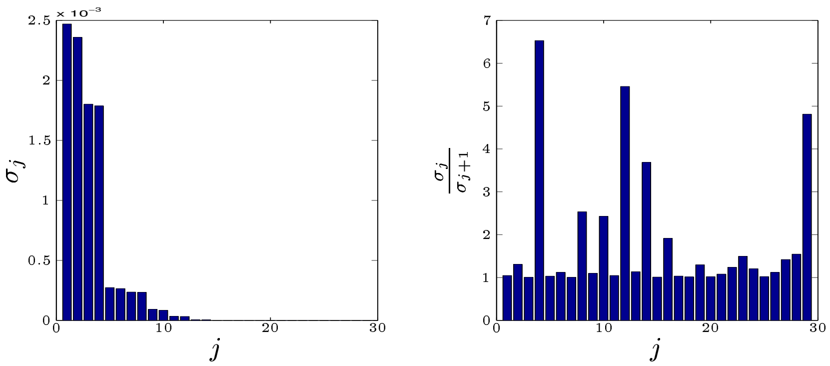


Figure 3.3: The left picture shows the singular values of the 30×30 Hankel matrix $\mathcal{R}_{30}$ with the autocovariance estimates $R(1), \ldots, R(59)$ of the measurements $y(k)$. Right we give the ratio's $\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_3}, \ldots, \frac{\sigma_{29}}{\sigma_{30}}$.

stochastic LTI model will be fairly good if we take the order equal to 12. Hence, we identify an autoregressive moving average (ARMA) model of order 12. By sending white noise through the model, we can simulate the output ARMA process. Of course, this simulated output sequence is not equal to the measurements in Figure 3.2. However, the second order statistics of the model's output process approximate the estimations of the second order statistics of the real system well. The autocovariance and the spectrum of model and measurements are given in Figure 3.4. We see that the output process of the ARMA model driven by white noise, is correlated over time and also fact that the model colors the spectrum is clearly visible. ◊

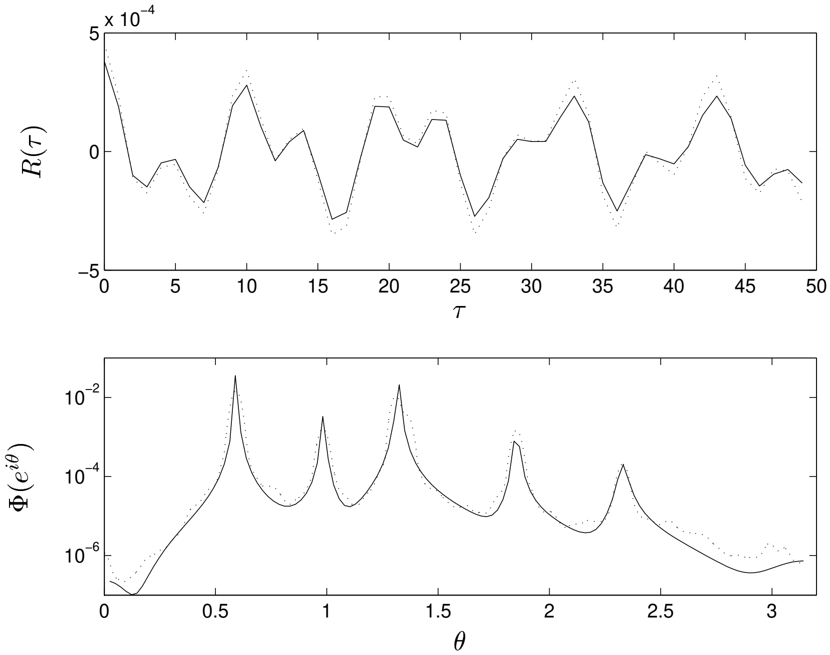


Figure 3.4: The top picture shows the autocovariance function estimated from the data (full line) and the autocovariance function of the identified ARMA model (dotted line). The spectrum of data (full line) and model (dotted line) are given in the bottom picture. Note the logarithmic scale.

Remark 3.2. The input process
Normally, when working with stochastic models, the input process is not observed, but for some of the theoretical derivations and simulations in this chapter, we will assume that observations of both $\{u(k)\}$ and $\{y(k)\}$ are available. However, the innovation sequence of a stochastic process for which a linear model is known, can always be computed by means of the Kalman filter, which leads to the so-called forward innovation representation. The innovation process can then be considered as the input of the model, see Section 3.2.1. ○

3.2 Model class

We describe in this section the model class that we will work with throughout the thesis. In Section 3.2.1 four state space representations are given, of which we will discuss the forward innovation and backward innovation models in more detail. We also give the assumptions on the different stochastic processes involved. For the forward innovation model we define the controllability and observability matrices and Gramians in Section 3.2.2 and the past and future input and output block Hankel matrices in Section 3.2.3. The state space equations and assumptions made in Section 3.2.1 are then reformulated in terms of the block Hankel matrices. In Section 3.2.4 we give the state space matrices for the inverse model. Since we will need them in the following chapters, we briefly describe in Section 3.2.5 the main properties of three classes of single-input single-output (SISO) models: the moving average (MA) model, the autoregressive (AR) model and the autoregressive moving average (ARMA) model.

3.2.1 State space representations and assumptions

There are many state space representations for a stochastic process $\{y(k)\}$, all of which are equivalent, in the sense that the second order statistics of the outputs generated by the different models are the same, i.e., the autocovariance sequences $E\left\{Y(k+\tau)Y(k)^T\right\}$ of the outputs of the different models are identical. We briefly discuss four forms: the forward form, the forward innovation form, the backward form and the backward innovation form. We will mainly use the forward innovation form. In some derivations, the backward innovation form will be helpful. The two other forms help to introduce the innovation representations.

The forward form

The forward representation of a given m-component stationary stochastic process $\{y(k)\}$ is the following

$$\left\{ \begin{array}{rcl} x_f(k+1) & = & Ax_f(k) + w(k) \ , \\ y(k) & = & Cx_f(k) + v(k) \ , \end{array} \right. \tag{3.1}$$

where $\{w(k)\} \in \mathbb{R}^n$ and $\{v(k)\} \in \mathbb{R}^m$ are zero-mean white noise processes with joint covariance matrix

$$E\left\{ \begin{pmatrix} W(k) \\ V(k) \end{pmatrix} \begin{pmatrix} W(l)^T & V(l)^T \end{pmatrix} \right\} = \begin{pmatrix} S_w & S_{wv} \\ S_{wv}^T & S_v \end{pmatrix} \delta(k-l) \ .$$

The process $\{x_f(k)\} \in \mathbb{R}^n$ is the state process associated to this model, where n is the model order, and $A \in \mathbb{R}^{n\times n}, C \in \mathbb{R}^{m\times n}$ are the system and output matrix, respectively. The first equation in (3.1) is referred to as the state equation,

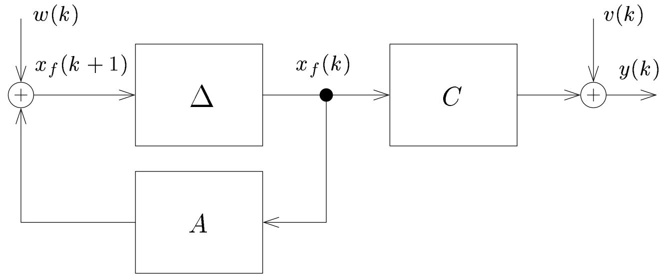


Figure 3.5: A linear time invariant stochastic system with outputs $y(k)$ and states $x_f(k)$. The white noise processes $\{w(k)\}$ and $\{v(k)\}$ are called the process noise and output noise, respectively. The system is described by the system matrix A, the output matrix C and the noise covariance matrices S_w, S_v, S_{wv}. The symbol Δ represents a delay and the subscripts f refer to 'forward'. This is the so-called forward representation.

while the second equation is the output equation. A schematic representation of this model is given in Figure 3.5.
It is assumed that the system matrix $A \in \mathbb{R}^{n\times n}$ is *asymptotically stable*, i.e. all its eigenvalues, which are the poles of the system, lie strictly inside the unit circle.
The state process is zero-mean and stationary, so that the state covariance matrix, denoted by Σ_f, is independent of the time:

$$E\left\{X_f(k)X_f(k)^T\right\} = \Sigma_f \ .$$

The noise processes $\{w(k)\}$ and $\{v(k)\}$ are assumed to be independent of the initial state. Hence,

$$\begin{aligned} E\left\{X_f(k)V(l)^T\right\} &= 0 \quad l \geq k \ , \\ E\left\{X_f(k)W(l)^T\right\} &= 0 \quad l \geq k \ . \end{aligned}$$

From the state equation in (3.1) we then find the Lyapunov equation for the state covariance matrix Σ_f:

$$\Sigma_f = A\Sigma_f A^T + S_w \ .$$

Furthermore, the autocovariance function $R_y(\tau) = E\left\{Y(k+\tau)Y(k)^T\right\}$ of the output process is

$$R_y(\tau) = \begin{cases} C\Sigma_f C^T + S_v & \tau = 0 \ , \\ CA^{\tau-1}G_f & \tau > 0 \ , \end{cases} \tag{3.2}$$

where

$$G_f = E\left\{X_f(k+1)Y(k)^T\right\} = A\Sigma_f C^T + S_{wv} \ .$$

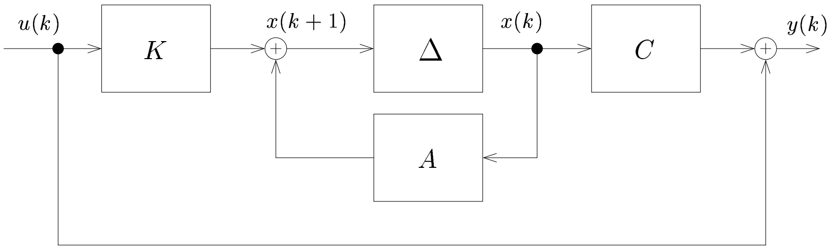


Figure 3.6: The forward innovation representation of the forward model in Figure 3.5. The model is determined by A, K, C and S_u, the input covariance matrix.

The forward innovation representation

The forward model in (3.1) can be converted into a so-called forward innovation model (see e.g. [114]). The forward innovation model is obtained by applying the steady state Kalman filter to the stochastic system (3.1), which results in

$$\left\{ \begin{array}{rcl} x(k+1) & = & Ax(k) + Ku(k) \ , \\ y(k) & = & Cx(k) + u(k) \ , \end{array} \right. \tag{3.3}$$

where $K \in \mathbb{R}^{n \times m}$ is the Kalman gain

$$K = (G_f - A\Sigma C^T)(R_y(0) - C\Sigma C^T)^{-1} \ , \tag{3.4}$$

and Σ is the state covariance matrix (see further). The forward innovation model is shown in Figure 3.6. This is the model that we will use mostly. Note that we are still describing the same process $\{y(k)\}$ and the system matrices A and C also remained the same. However, the state process is different from the state process in the forward form. Furthermore, the noise processes in the state equation and the output equation are now equal. Therefore, the model satisfies the input–output description of Figure 3.1.
We will denote the model (3.3) by the threesome (A, K, C). Its transfer function

$$H(z) = C(zI_n - A)^{-1}K + I_m \ ,$$

can be developed in a series as $H(z) = \sum_{k=0}^{\infty} H(k)z^{-k} = I_m + CKz^{-1} + CAKz^{-2} + \cdots$, which yields the impulse response $H(0), H(1), \ldots$ of the model. The matrices $H(k)$ $(k \geq 0)$, which are equal to

$$\left\{ \begin{array}{rcl} H(0) & = & I_m \ , \\ H(k) & = & CA^{k-1}K \quad \text{for } k > 0 \ , \end{array} \right. \tag{3.5}$$

are also known as the Markov parameters of the model.
The following assumptions[1] are made on the input, state and output processes of (3.3).

[1] If the forward innovation form is obtained from a forward form via Kalman filtering, these assumptions are in fact properties, once the assumptions of the forward form are satisfied.

Assumption 1. The input process (forward innovation process)

(a) The input process $\{u(k)\}$ is an *m-component zero-mean, stationary, ergodic and white stochastic process* with full rank covariance matrix $S_u \in \mathbb{R}^{m\times m}$, i.e. the input is a sequence of zero-mean, independent random variables $U(k)$, with autocovariance function

$$\begin{aligned} R_u(\tau) &= E\left\{U(k+\tau)U(k)^T\right\} \\ &= S_u\delta(\tau) \end{aligned} \tag{3.6}$$

(b) The input process is assumed to be independent of the initial state, which implies

$$E\{X(0)U(k)^T\} = 0 \quad \forall k \geq 0 \ .$$

Assumption 2. The state process
The state process $\{x(k)\}$ is a *zero-mean, stationary and ergodic stochastic process* with covariance matrix Σ:

$$\begin{aligned} E\{X(k)\} &= 0 \ , \\ E\{X(k)X(k)^T\} &= \Sigma \in \mathbb{R}^{n\times n} \ . \end{aligned} \tag{3.7}$$

From the state space equations (3.3) and Assumptions 1 and 2 it follows that Σ satisfies the following Lyapunov equation:

$$\Sigma = A\Sigma A^T + KS_uK^T \ . \tag{3.8}$$

From Assumption 1 and the model description (3.3) it follows that $X(k)$ is independent of the present and all future inputs. Consequently,

$$E\{X(k)U(k+\tau)^T\} = 0 \quad \text{for } \tau \geq 0 \ , \tag{3.9}$$

Assumption 1 also implies that $Y(k)$ is independent of all future inputs. Thus,

$$E\{Y(k)U(k+\tau)^T\} = 0 \quad \text{for } \tau > 0 \ . \tag{3.10}$$

Assumption 3. The output process
The output process is a *zero-mean, stationary and ergodic process*

By using the state space description (3.3) and the assumptions above, the autocovariance function $R_y(\tau)$ of $\{y(k)\}$ can be expressed in terms of the system matrices as follows:

$$\begin{aligned} R_y(\tau) &= E\left\{Y(k+\tau)Y(k)^T\right\} \ , \\ &= \begin{cases} C\Sigma C^T + S_u & \text{for } \tau = 0 \ , \\ CA^{\tau-1}G & \text{for } \tau > 0 \ , \end{cases} \end{aligned} \tag{3.11}$$

and $R_y(-\tau) = R_y(\tau)^T$ as usual. The matrix G in (3.11) is the covariance of $Y(k)$ and $X(k+1)$:

$$G = E\left\{X(k+1)Y(k)^T\right\} = A\Sigma C^T + KS_u \ ,$$

and

$$G_f = G \ . \tag{3.12}$$

Under fairly general conditions [80], the forward innovation model is *strictly minimum phase*. This means that all the zeros of the model, which are the eigenvalues of $A - KC$, are less than one in modulus. Furthermore, the system is assumed to be minimal, so that the state covariance matrix Σ is of full rank.

Remark 3.3. The state covariance matrix is the solution of the algebraic Riccati equation
We know that Σ satisfies the Lyapunov equation

$$\Sigma = A\Sigma A^T + KS_uK^T \ .$$

Since the Kalman gain K is equal to $(G - A\Sigma C^T)(R_y(0) - C\Sigma C^T)^{-1}$ and since from (3.11) it follows that

$$S_u = R_y(0) - C\Sigma C^T \ , \tag{3.13}$$

we have

$$\Sigma = A\Sigma A^T + (G - A\Sigma C^T)(R_y(0) - C\Sigma C^T)^{-1}(G^T - C\Sigma A^T) \ , \tag{3.14}$$

which is the algebraic Riccati equation. It can be solved for Σ by solving a generalized eigenvalue problem, see [159, p. 62].
This expression can be used to obtain the state covariance matrix when the matrices A, C, G and $R_y(0)$ are known, while K and S_u are unknown, as is e.g. the case in the stochastic subspace procedure in [159], see also Section 3.3.2, where we give a brief description of the stochastic identification problem. Once Σ is determined, one obtains K from (3.4) and S_u from (3.13). Hence, a complete parameterization of the model is known, namely (A, K, C) and S_u. ○

The backward form

Associated with the forward model in (3.1), there is a backward[2] model. Its state space representation is

$$\left\{ \begin{array}{rcl} \bar{x}_b(k-1) & = & \bar{A}\bar{x}_b(k) + \bar{w}(k) \ , \\ y(k) & = & \bar{C}\bar{x}_b(k) + \bar{v}(k) \ , \end{array} \right. ,$$

[2]Backward means that the iterative state space formulas of the model are running backward in time.

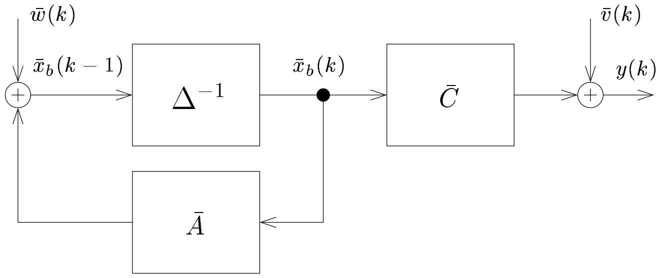


Figure 3.7: The backward representation for the model in Figure 3.5. It is described by the model parameters $\bar{A} = A^T, \bar{C} = G^T$ and the noise covariance matrices $S_{\bar{w}}, S_{\bar{v}}, S_{\bar{w}\bar{v}}$. Note that this representation is equal to the one in Figure 3.5, except for the delay Δ, which is replaced by the forward shift Δ^{-1}.

where $\bar{A} = A^T$ and $\bar{C} = G^T$. The covariance matrix of the noise processes is denoted by

$$E\left\{\begin{pmatrix}\bar{W}(k)\\ \bar{V}(k)\end{pmatrix}\begin{pmatrix}\bar{W}(l)^T & \bar{V}(l)^T\end{pmatrix}\right\} = \begin{pmatrix}S_{\bar{w}} & S_{\bar{w}\bar{v}}\\ S_{\bar{w}\bar{v}}^T & S_{\bar{v}}\end{pmatrix}\delta(k-l) \ .$$

The derivation of the backward model, which is schematically shown in Figure 3.7, can be found in e.g. in [159].
The backward state $\bar{x}_b(k)$ is defined as

$$\bar{x}_b(k) = \Sigma_f^{-1} x_f(k+1) \ , \tag{3.15}$$

and thus is the backward state covariance matrix equal to the inverse of the forward state covariance matrix:

$$\bar{\Sigma}_b = E\left\{\bar{X}_b(k)\bar{X}_b(k)^T\right\} = \Sigma_f^{-1} \ .$$

Since in the backward model $\bar{W}(k)$ and $\bar{V}(k)$ and their pasts are independent of $\bar{X}(k)$, the backward state covariance matrix satisfies the Lyapunov equation

$$\bar{\Sigma}_b = A^T\bar{\Sigma}_b A + S_{\bar{w}} \quad \text{or} \quad \Sigma_f^{-1} = A^T\Sigma_f^{-1}A + S_{\bar{w}} \ ,$$

and the output autocovariance is

$$R_y(\tau) = \begin{cases}\bar{C}\bar{\Sigma}_b\bar{C}^T + S_{\bar{v}} = G^T\Sigma_f^{-1}G + S_{\bar{v}} & \tau = 0 \ ,\\ \bar{C}\bar{A}^{\tau-1}\bar{G} = G^T(A^{\tau-1})^T\bar{G} & \tau < 0 \ ,\end{cases} \tag{3.16}$$

where

$$\bar{G} = E\left\{\bar{X}_b(k-1)Y(k)^T\right\} = \bar{A}\bar{\Sigma}_b\bar{C}^T + S_{\bar{w}\bar{v}} \ .$$

Due to (3.15), the matrix $\bar{G}$ can be written as $\Sigma_f^{-1}E\left\{X_f(k)Y(k)^T\right\}$, and consequently,

$$\bar{G} = C^T \ .$$

The backward innovation representation

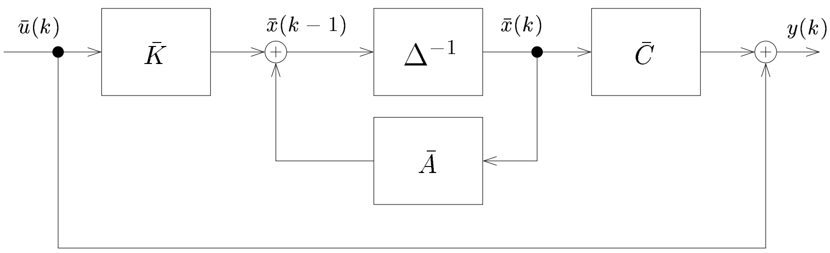


Figure 3.8: The backward innovation representation. The model is described by $\bar{A} = A^T$, $\bar{C} = G^T$, $\bar{K}$ and the input covariance matrix $S_{\bar{u}}$.

Also for the backward model we can construct an innovation representation. It is of the form

$$\left\{ \begin{array}{rcl} \bar{x}(k-1) & = & \bar{A}\bar{x}(k) + \bar{K}\bar{u}(k) \ , \\ y(k) & = & \bar{C}\bar{x}(k) + \bar{u}(k) \ , \end{array} \right. \tag{3.17}$$

where $\bar{A} = A^T$, and $\bar{C} = G^T$ and $\bar{K}$ is the backward Kalman gain.
The covariance matrix of the backward innovation process $\{\bar{u}(k)\}$ is denoted by $S_{\bar{u}}$:

$$E\left\{\bar{U}(k+\tau)\bar{U}(k)^T\right\} = S_{\bar{u}}\delta(\tau) \ .$$

Since it is a backward model, $\bar{X}(k)$ is independent of the past and present input and $Y(k)$ is independent of the past input:

$$\begin{aligned} E\left\{\bar{U}(k-\tau)\bar{X}(k)^T\right\} &= 0 \quad \text{for } \tau \geq 0 \ , \\ E\left\{\bar{U}(k-\tau)Y(k)^T\right\} &= 0 \quad \text{for } \tau > 0 \ , \end{aligned} \tag{3.18}$$

whereas in the forward model, the output and state processes were independent of the future input.
The state covariance matrix of the backward innovation model is denoted by $\bar{\Sigma}$:

$$E\left\{\bar{X}(k)\bar{X}(k)^T\right\} = \bar{\Sigma} \ ,$$

which satisfies the Lyapunov equation

$$\bar{\Sigma} = A^T\bar{\Sigma}A + \bar{K}S_{\bar{u}}\bar{K}^T \ . \tag{3.19}$$

By using the state space description (3.17) and (3.18), we obtain

$$R_y(\tau) = \begin{cases} \bar{C}\bar{\Sigma}\bar{C}^T + S_{\bar{u}} = G^T\bar{\Sigma}G + S_{\bar{u}} & \tau = 0 \ , \\ \bar{C}\bar{A}^{\tau-1}\bar{G} = G^T(A^{\tau-1})^TC^T & \tau < 0 \ . \end{cases}$$

Filling out the Kalman gain $\bar{K} = (C^T - A^T\bar{\Sigma}G)(R_y(0) - G^T\bar{\Sigma}G)^{-1}$ and the input variance $S_{\bar{u}} = R_y(0) - G^T\bar{\Sigma}G$ in the Lyapunov equation (3.19), leads to the algebraic Riccati equation for the backward state covariance matrix:

$$\bar{\Sigma} = A^T\bar{\Sigma}A + (C^T - A^T\bar{\Sigma}G)(R_y(0) - G^T\bar{\Sigma}G)^{-1}(C - G^T\bar{\Sigma}A) \ . \tag{3.20}$$

In Table 3.1 we give a schematic overview of the different model representations and their relations.
Note that the backward innovation form is not the backward version of the forward innovation form, as is the backward form for the forward form. Also note that therefore, $\bar{\Sigma}$ is not the inverse of Σ. This will be important in Section 3.4.2, more specifically in Property 3.5. We show in Property 3.5 that the largest n canonical correlations of the past and the future output process of an nth order model are equal to the eigenvalues of $\Sigma\bar{\Sigma}$. If $\bar{\Sigma}$ were equal to the inverse of Σ, then these canonical correlations would be equal to 1, which is impossible for the models that we consider (see also Property 3.6).

forward form (A, C, S_w, S_{wv}, S_v) Σ_f	Kalman filter $\longrightarrow$	forward innovation form (A, K, C, S_u) Σ
$\downarrow$		
backward form $(A^T, G, S_{\bar{w}}, S_{\bar{w}\bar{v}}, S_{\bar{v}})$ $\bar{\Sigma}_b = \Sigma_f^{-1}$	$\longrightarrow$ backward Kalman filter	backward innovation form $(A^T, \bar{K}, G, S_{\bar{u}})$ $\bar{\Sigma} \neq \Sigma^{-1}$

Table 3.1: Schematic overview of the different representation forms. Note that $\bar{\Sigma}_b = \Sigma_f^{-1}$, but $\bar{\Sigma} \neq \Sigma^{-1}$. In Section 3.4.2 (Property 3.5 and Property 3.6) we will come back to this asymmetry.

3.2.2 The controllability and observability matrices and Gramians

The controllability matrix $\mathcal{C}_i$ of the forward innovation model (3.3) is defined as

$$\mathcal{C}_i = \begin{pmatrix} K & AK & A^2K & \cdots & A^{i-1}K \end{pmatrix} ,$$

and its observability matrix Γ_i is

$$\Gamma_i = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{i-1} \end{pmatrix} . \tag{3.21}$$

The controllability Gramian P of the forward innovation model (3.3) is defined as the solution of the controllability Lyapunov equation

$$P = APA^T + KK^T , \tag{3.22}$$

while the observability Gramian Q follows from the observability Lyapunov equation

$$Q = A^T Q A + C^T C \ . \tag{3.23}$$

Since the model is stable and minimal, the matrices P and Q are the unique and positive definite solutions of the respective equations. The explicit solution for P is of the form

$$P = \sum_{k=0}^{\infty} A^k K K^T (A^k)^T = \mathcal{C}_\infty \mathcal{C}_\infty^T \ ,$$

where $\mathcal{C}_\infty$ is the infinite controllability matrix. Similarly, the observability Gramian Q can be obtained as

$$Q = \sum_{k=0}^{\infty} (A^k)^T C^T C A^k = \Gamma_\infty^T \Gamma_\infty \ ,$$

where Γ_∞ is the infinite observability matrix of the model.

Remark 3.4. Comparing P and Σ
We have seen in (3.8) that the state covariance matrix of the forward innovation model (A, K, C) satisfies the Lyapunov equation $\Sigma = A\Sigma A^T + K S_u K^T$. Similarly to the controllability Gramian P, we can write Σ explicitly as

$$\Sigma = \sum_{k=0}^{\infty} A^k K S_u K^T A^{k^T} = \mathcal{C}_\infty Q_u \mathcal{C}_\infty^T \ ,$$

where $Q_u = \mathrm{diag}(S_u, S_u, \ldots)$.
Compared to P, the only difference is the input covariance matrix S_u. If this matrix is equal to I_m, then the state covariance matrix is equal to the controllability Gramian.
For single input single output (SISO) models the relation of Σ and P is simplified. If the input variance of the model is equal to σ^2, then

$$\Sigma = \sigma^2 P \ . \tag{3.24}$$

Equation (3.24) also holds for a multiple input multiple output (MIMO) model where the different inputs are uncorrelated and have the same variance σ^2, i.e. the matrix S_u reduces to $\mathrm{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2)$. ○

3.2.3 The block Hankel matrices

We define the input and output block Hankel matrices U and Y. These matrices play an important role in the computation of the canonical correlations since the canonical correlations between two stationary and ergodic stochastic processes

are equal to the cosines of the principal angles between the row spaces of the associated double infinite block Hankel matrices (see Section 2.6.3). In practice, only a finite amount of data is available. By computing the principal angles between the row spaces of the finite dimensional block Hankel matrices that are constructed from the given data, we obtain approximate values for the canonical correlations.
The output block Hankel matrix Y is defined as[3]

$$Y = \frac{1}{\sqrt{j}} \begin{pmatrix} y(0) & y(1) & y(2) & \cdots & y(j-1) \\ y(1) & y(2) & y(3) & \cdots & y(j) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ y(i-1) & y(i) & y(i+1) & \cdots & y(i+j-2) \\ \hline y(i) & y(i+1) & y(i+2) & \cdots & y(i+j-1) \\ y(i+1) & y(i+2) & y(i+3) & \cdots & y(i+j) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ y(2i-1) & y(2i) & y(2i+1) & \cdots & y(2i+j-2) \end{pmatrix} \tag{3.25}$$

$$= Y_{0|2i-1} = \begin{pmatrix} Y_{0|i-1} \\ \hline Y_{i|2i-1} \end{pmatrix} = \begin{pmatrix} Y_p \\ \hline Y_f \end{pmatrix} \in \mathbb{R}^{2mi \times j} \ , \tag{3.26}$$

where

- The number of columns (j) is typically equal to $K-2i+1$, where K is the total number of data samples, which implies that all given data samples are used. Throughout this chapter, for statistical reasons, we will often assume that $j, K \to \infty$. This means that the covariance matrices in the canonical correlation analysis are exact.
- The subscripts of $Y_{0|2i-1}, Y_{0|i-1}, Y_{i+1|2i-1}$ denote the subscript of the first and last element of the first column in the block Hankel matrix. The subscript 'p' stands for 'past' and the subscript 'f' for 'future'. The reason for this notation will become clear in Section 3.4.

The input block Hankel matrices $U_{0|2i-1}, U_p, U_f$ are defined in a similar way. We now describe how the state space equations of the forward innovation form and the backward innovation form can be expressed in terms of the block Hankel matrices.

Equations for the forward innovation model

The state sequence matrix is defined as

$$X_i = \frac{1}{\sqrt{j}} \begin{pmatrix} x(i) & x(i+1) & \cdots & x(i+j-1) \end{pmatrix} \ ,$$

[3] We will not indicate the number of columns, as we did in Section 2.6.3 since the number of columns will always be assumed to be j.

where the subscript i denotes the subscript of the first element of the state sequence. Analogously to the past inputs and outputs, we denote the past state sequence by X_p and the future state sequence by X_f:

$$X_p = X_0 \in \mathbb{R}^{n\times j} \text{ and } X_f = X_i \in \mathbb{R}^{n\times j} \ .$$

The state space equations (3.3) of the forward innovation model can now be written as

$$X_f = A^i X_p + \Delta_i U_p \ , \tag{3.27}$$

$$Y_p = \Gamma_i X_p + H_i U_p \ , \tag{3.28}$$

$$Y_f = \Gamma_i X_f + H_i U_f \ , \tag{3.29}$$

where $\Delta_i \in \mathbb{R}^{n\times mi}$ is the reversed controllability matrix:

$$\Delta_i = \begin{pmatrix} A^{i-1}K & A^{i-2}K & \cdots & AK & K \end{pmatrix} \ ,$$

the matrix $\Gamma_i \in \mathbb{R}^{mi\times n}$ is the observability matrix of the model (see (3.21)) and the matrix $H_i \in \mathbb{R}^{mi\times mi}$ is a block lower triangular and block Toeplitz matrix with the Markov parameters of the model (the impulse response sequence) as its elements:

$$H_i = \begin{pmatrix} I_m & 0 & 0 & \cdots & 0 \\ CK & I_m & 0 & \cdots & 0 \\ CAK & CK & I_m & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{i-2}K & CA^{i-3}K & CA^{i-4}K & \cdots & I_m \end{pmatrix} \ . \tag{3.30}$$

We can now reformulate (3.6), (3.7), (3.9) and (3.10) in terms of the above defined matrices.
The white noise assumption on the input process (Assumption 1 (a)) leads to the block diagonal structure of the input covariance matrix, which will be denoted by Q_{u_i}:

$$Q_{u_i} = \lim_{j\to\infty} U_p U_p^T = \lim_{j\to\infty} U_f U_f^T = \begin{pmatrix} S_u & 0 & \cdots & 0 \\ 0 & S_u & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_u \end{pmatrix} \in \mathbb{R}^{mi\times mi} \ . \tag{3.31}$$

Since S_u is assumed to be of full rank, we see here that U is a full rank matrix too (for $j \to \infty$):

$$\text{rank}(U) = 2mi \text{ for } j \to \infty \ . \tag{3.32}$$

By using the state sequence matrices, we can write the state covariance matrix Σ as

$$\Sigma = \lim_{j\to\infty} X_p X_p^T = \lim_{j\to\infty} X_f X_f^T \ . \tag{3.33}$$

Due to the minimality of the model, Σ and hence X_p and X_f are full rank matrices:

$$\text{rank}(X_p) = \text{rank}(X_f) = n \text{ for } j \to \infty \ . \tag{3.34}$$

The fact that the states are uncorrelated with the present and future inputs (see (3.9)) and that the output is uncorrelated with the future inputs, translates into

$$\begin{array}{rcl} \lim_{j\to\infty} X_p U^T & = & 0 \\ \lim_{j\to\infty} X_f U_f^T & = & 0 \end{array} \tag{3.35}$$

and

$$\lim_{j\to\infty} Y_p U_f^T = 0 \tag{3.36}$$

respectively. Hence, the input block Hankel matrix is orthogonal to the state sequence matrix and the future input block Hankel matrix is orthogonal to the past output clock Hankel matrix (for $j \to \infty$).
We can now show that the output block Hankel matrix has full rank (for $j \to \infty$).

Proposition 3.1. The output block Hankel matrix Y is of full rank
The block Hankel matrix of stochastic process $\{y(k)\}$ generated by a linear model that satisfies the above assumptions, has full row rank if $j \to \infty$:

$$\text{rank}(Y) = 2mi \textit{ for } j \to \infty \ . \tag{3.37}$$

Proof.
From (3.28) and (3.29) it follows that Y can be written as

$$Y = \begin{pmatrix} \Gamma_{2i} & H_{2i} \end{pmatrix} \begin{pmatrix} X_p \\ U \end{pmatrix} \ .$$

Sylvester's rank inequality (see Lemma 2.1) says that

$$\begin{aligned} \text{rank} \begin{pmatrix} \Gamma_{2i} & H_{2i} \end{pmatrix} + \text{rank} \begin{pmatrix} X_p \\ U \end{pmatrix} - (n + 2mi) &\leq \text{rank} Y \\ &\leq \min \left(\text{rank} \begin{pmatrix} \Gamma_{2i} & H_{2i} \end{pmatrix}, \text{rank} \begin{pmatrix} X_p \\ U \end{pmatrix} \right) \ . \end{aligned} \tag{3.38}$$

The matrix H_{2i} is of full rank, which is $2mi$. Consequently,

$$\text{rank} \begin{pmatrix} \Gamma_{2i} & H_{2i} \end{pmatrix} = 2mi \ . \tag{3.39}$$

On the other hand, the rank of $\begin{pmatrix} X_p \\ U \end{pmatrix}$ is equal to the sum of the ranks of X_p and U (if $j \to \infty$), due to the independence of $\{x(k)\}$ and $\{u(k)\}$. And since both matrices are of full rank (see (3.34) and (3.32)), we have

$$\text{rank} \begin{pmatrix} X_p \\ U \end{pmatrix} = 2mi + n \text{ for } j \to \infty \ . \tag{3.40}$$

Filling (3.39) and (3.40) out in (3.38) gives

$$\operatorname{rank}(Y) = 2mi \text{ for } j \to \infty \ .$$

$\square$

Equations for the backward innovation form

The backward innovation state space equations (3.17) lead to the following relations between the past and future data block Hankel matrices:

$$\begin{aligned} \bar{X}_p &= (A^i)^T \bar{X}_f + \bar{\Delta}_i \bar{U}_f \ , \\ Y_p &= \bar{\Gamma}_i \bar{X}_p + \bar{H}_i \bar{U}_p \ , \\ Y_f &= \bar{\Gamma}_i \bar{X}_f + \bar{H}_i \bar{U}_f \ , \end{aligned}$$

where the past and future state matrices are defined as

$$\begin{array}{rclcl} \bar{X}_p & = & \bar{X}_{i-1} & = & \frac{1}{\sqrt{j}} \begin{pmatrix} \bar{x}(i-1) & \bar{x}(i) & \cdots & \bar{x}(i+j-2) \end{pmatrix} \ , \\ \bar{X}_f & = & \bar{X}_{2i-1} & = & \frac{1}{\sqrt{j}} \begin{pmatrix} \bar{x}(2i-1) & \bar{x}(2i) & \cdots & \bar{x}(2i+j-2) \end{pmatrix} \ . \end{array}$$

The matrix $\bar{\Delta}_i$ is the backward controllability matrix:

$$\bar{\Delta}_i = \begin{pmatrix} \bar{K} & A^T \bar{K} & \cdots & (A^{i-1})^T \bar{K} \end{pmatrix} \ ,$$

the matrix $\bar{\Gamma}_i$ is the reversed backward observability matrix:

$$\bar{\Gamma}_i = \begin{pmatrix} G^T (A^{i-1})^T \\ G^T (A^{i-2})^T \\ \vdots \\ G^T A^T \\ G^T \end{pmatrix} \ , \tag{3.41}$$

and the matrix $\bar{H}_i$ is the following upper block triangular Toeplitz matrix:

$$\bar{H}_i = \begin{pmatrix} I_m & G^T \bar{K} & G^T A^T \bar{K} & \cdots & G^T (A^{i-2})^T \bar{K} \\ 0 & I_m & G^T \bar{K} & \cdots & G^T (A^{i-3})^T \bar{K} \\ 0 & 0 & I_m & \cdots & G^T (A^{i-4})^T \bar{K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & I_m \end{pmatrix} \ .$$

Note the symmetry of the forward innovation and the backward innovation forms. The matrices $\bar{\Delta}_i$, $\bar{\Gamma}_i$ and $\bar{H}_i$ are obtained by replacing the matrices A, K and C in Δ_i, Γ_i and H_i by $\bar{A} = A^T$, $\bar{K}$ and $\bar{C} = G^T$, respectively and reversing the order (due to the reversed direction of the time).

3.2.4 The inverse model

For the stable and minimum phase model (A, K, C), the inverse model is also stable and minimum phase. Its state space description is readily derived from (3.3):

$$\left\{\begin{array}{rcl} x(k+1) &=& (A-KC)x(k)+Ky(k)\ , \\ u(k) &=& -Cx(k)+y(k)\ . \end{array}\right.$$

The state space matrices of the inverse model are denoted by (A_z, K_z, C_z):

$$(A_z, K_z, C_z) = (A - KC, K, -C)\ . \tag{3.42}$$

The transfer function of the inverse model is then equal to

$$H_z(z) = -C\left(zI_n - (A-KC)\right)^{-1} K + I_m\ ,$$

and the poles of $H_z(z)$, which are the eigenvalues of the system matrix $A-KC$, are the zeros of $H(z)$. In Figure 3.2.4 this is shown schematically.
The Markov parameters of the inverse model are denoted by $H_z(k)$ and they are equal to

$$\left\{\begin{array}{rcl} H_z(0) &=& I_m\ , \\ H_z(k) &=& -C(A-KC)^{k-1}K \quad \text{for } k>0\ . \end{array}\right. \tag{3.43}$$

We will also need the observability matrix of the inverse model, denoted by Γ_{z_i}:

$$\Gamma_{z_i} = \begin{pmatrix} -C \\ -C(A-KC) \\ -C(A-KC)^2 \\ \vdots \\ -C(A-KC)^{i-1} \end{pmatrix}\ ,$$

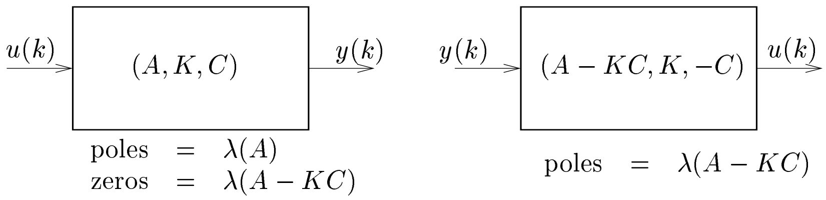


Figure 3.9: Inverting the stochastic model (A, K, C) can be seen as creating a whitening filter for $\{y(k)\}$ since $(A-KC, K, -C)$ generates the white innovation process $\{u(k)\}$ if $\{y(k)\}$ is the input process. The eigenvalues of A are the poles of the model, while the eigenvalues of $A-KC$ are the zeros of the model. From Kalman filtering theory it follows that the zeros lie inside the unit circle (the model is minimum phase).

where the subscript i in Γ_{z_i} denotes the number of block rows. From (3.28) or (3.29) it immediately follows that

$$\Gamma_{z_i} = -H_i^{-1}\Gamma_i \ . \tag{3.44}$$

The observability Gramian of the inverse model is denoted by Q_z and it is equal to

$$Q_z = \Gamma_{z_\infty}^T \Gamma_{z_\infty} \ . \tag{3.45}$$

It is the solution of the observability Lyapunov equation for the inverse model

$$Q_z = (A - KC)^T Q_z (A - KC) + C^T C \ . \tag{3.46}$$

3.2.5 Parametric descriptions for single input single output (SISO) models

Models that satisfy the description and assumptions of Section 3.2.1 and for which $m = 1$, are models for scalar stochastic processes. We describe three classes of parametric single-input single-output (SISO) models: moving average (MA) models, autoregressive (AR) models and the combination of these two: autoregressive moving average (ARMA) models. Next, we give a few properties that are valid for SISO models but not for MIMO (multiple input multiple output). We start, however, with a general description of the power spectral density[4] of scalar stochastic processes.

The spectral density of a scalar, zero-mean, stationary stochastic process

Let $\{y(k)\}$ be a stationary scalar stochastic process with autocovariance function $R(\tau) = E\{y(k+\tau)y(k)\}$. We will always assume that the autocovariance is absolutely summable, i.e. $\sum_{\tau=-\infty}^{\infty} |R(\tau)| < \infty$. Then the spectral density of $\{y(k)\}$ can be defined as the Z-transform of the autocovariance function:

$$\Phi(z) = \sum_{\tau=-\infty}^{\infty} R(\tau) z^{-\tau} \ .$$

In the special case that $z = e^{i\theta}$, $\theta \in [0, 2\pi)$, the spectral density is given by the Fourier transform of the autocovariance sequence

$$\Phi(e^{i\theta}) = \sum_{\tau=-\infty}^{\infty} R(\tau) e^{-i\tau\theta} \ ,$$

[4] In the course of the thesis we will sometimes refer to the power spectral density as the spectral density, the power spectrum or shortly the spectrum.

and the autocovariance is thus the inverse Fourier transform of the spectral density

$$R(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(e^{i\theta}) e^{i\tau\theta} \, d\theta \ .$$

In short,

$$\Phi = \mathcal{F}(R) \ .$$

Example 3.2. The spectral density of the output of a linear model
If $\{y(k)\}$ is the output of a linear model with stable transfer function $H(z)$, driven by white noise with variance σ^2, then the spectral density is related to the transfer function as

$$\Phi(e^{i\theta}) = \sigma^2 H(e^{i\theta}) H(e^{-i\theta}) = \sigma^2 |H(e^{i\theta})|^2 \ ,$$

from which it is clear that the spectral density has to be nonnegative on the unit circle. Since $H(e^{i\theta})$ is the Fourier transform of the impulse response $\{h(k)\}$ of the model, we also have

$$\Phi = \sigma^2 |\mathcal{F}(h)|^2 \ .$$

$\diamondsuit$

Remark 3.5. The estimation of the power spectral density of a stationary stochastic process
Two classes of methods exist for the estimation of the spectral density of a stochastic process, given a finite number of observations: parametric and non-parametric methods [122].
A parametric method first identifies a model from the data and uses the spectrum of the model as an estimate of the process' spectrum. Popular models are autoregressive (AR), moving average (MA) and autoregressive moving average models (ARMA), which will be discussed below.
Non-parametric methods are based on the periodogram, which is the normalized squared absolute value of the discrete Fourier transform (DFT) of the sequence of observations, which can be computed via the Fast Fourier Transform (FFT) algorithm [28] (see also Section 5.2 where we discuss the estimation of the cepstrum, which is based on an estimate of the spectrum). Several types of windows can be applied to the periodogram to obtain smoothing.
Some authors have a strong preference for one method. Broersen [20], for example, states: 'It is a fact that even the best windowed periodogram is less accurate than the spectrum of the automatically selected single time series model.' $\bigcirc$

Moving average model (MA)

A moving average process $\{y(k)\}$ is modeled as

$$y(k) = \sum_{i=0}^{n} b_i u(k-i) \ , \tag{3.47}$$

where $\{u(k)\}$ is a white noise process, with zero mean and variance σ^2. By convention $b_0 = 1$. If $b_n \neq 0$, then the order of the model is n and (3.47) is denoted by MA(n).
The transfer function of the MA(n) model is

$$H(z) = \frac{b(z)}{z^n} ,$$

where $b(z) = z^n + b_1 z^{n-1} + \cdots + b_n$. The zeros of the model, which are the roots of the polynomial $b(z)$, are denoted by $\beta_1, \beta_2, \ldots, \beta_n$. Since we only consider minimum phase models, they should all lie inside the unit circle.
The impulse response coefficients (or Markov parameters) of the model are $b_0, b_1, \ldots, b_n, 0, 0, \ldots$ Therefore, an MA model is also called a finite impulse response (FIR) model.
The spectral density of an MA(n) process is

$$\Phi(e^{i\theta}) = \sigma^2 |b(e^{i\theta})|^2 .$$

Autoregressive model (AR)

An autoregressive (AR) process $\{y(k)\}$ is modeled as

$$y(k) = -\sum_{i=1}^{n} a_i y(k-i) + u(k) , \tag{3.48}$$

where $\{u(k)\}$ is a white noise process with zero mean and variance σ^2. Further, for $a_n \neq 0$ the model (3.48) will be referred to as an AR(n) model, where n is the order of the model.
The transfer function of the model is

$$H(z) = \frac{z^n}{a(z)} ,$$

where $a(z) = z^n + a_1 z^{n-1} + \cdots + a_n$. We require the polynomial $a(z)$ to be stable, i.e., all its roots should lie inside the unit circle. These roots are the poles of the model and they are denoted by $\alpha_1, \alpha_2, \ldots, \alpha_n$.
The spectral density of the AR(n) process (3.48) is

$$\Phi(e^{i\theta}) = \frac{\sigma^2}{|a(e^{i\theta})|^2} .$$

Autoregressive moving average (ARMA) model

In general, the spectral density of a stationary stochastic process may have high peaks as well as deep valleys. The spectral densities of AR models typically have high peaks. These peaks are produced by roots of $a(z)$ (the poles of

the model) that are located near the unit circle. On the other hand, the spectral densities of MA models are characterized by deep valleys, which arise from roots of $b(z)$ (the zeros) that are close to the unit circle. Qualitatively, AR models have difficulties in reproducing deep valleys and MA models have difficulties in reproducing high peaks, when each of these model classes are used to approximate an arbitrary process. To achieve a good agreement with the spectrum of an arbitrary process, it might be necessary to use approximating AR or MA models of very high order. This motivates the introduction of a third model description, the ARMA model, which combines the properties of the AR and MA classes. The AR and MA models can be considered as special cases of the more general ARMA models.
An ARMA process $\{y(k)\}$ is modeled as

$$y(k) = -\sum_{i=1}^{p} a_i y(k-i) + \sum_{i=0}^{q} b_i u(k-i) \ , \tag{3.49}$$

where $\{u(k)\}$ is a white noise process, with zero mean and variance σ^2. By convention $b_0 = 1$. If a_p and b_q are not equal to zero, then the ARMA process in (3.49) is denoted by ARMA(p,q).
The transfer function of the ARMA(p,q) model has the following form

$$H(z) = z^{p-q}\frac{b(z)}{a(z)} \ , \tag{3.50}$$

where $b(z) = z^q + b_1 z^{q-1} + \cdots + b_q$ and $a(z) = z^p + a_1 z^{p-1} + \cdots + a_p$. Again, we require the model to be stable and minimum phase. Hence, the zeros $\beta_1, \ldots, \beta_q$, which are the roots of $b(z)$, and the poles $\alpha_1, \ldots, \alpha_p$ (the roots of $a(z)$) should lie inside the unit circle. Note that, if $p \neq q$, then in addition to these poles and zeros, there are also $|p-q|$ poles or zeros in the origin, depending on whether $q > p$ or $p > q$, respectively. The transfer function $H(z)$ can also be written as

$$H(z) = z^{p-q}\frac{\prod_{i=1}^{q}(z-\beta_i)}{\prod_{i=1}^{p}(z-\alpha_i)} = \frac{\prod_{i=1}^{q}(1-\beta_i z^{-1})}{\prod_{i=1}^{p}(1-\alpha_i z^{-1})} \ .$$

The spectral density of the ARMA(p,q) process is

$$\Phi(e^{i\theta}) = \sigma^2\frac{|b(e^{i\theta})|^2}{|a(e^{i\theta})|^2} \ .$$

It is clear from the above description that an MA(n) model can be seen as an ARMA$(0,n)$ model, of which all poles lie at zero. An AR(n) model, on the other hand, is an ARMA$(n,0)$ model, with all zeros in zero. Conversely, a stable and minimum phase ARMA(p,q) model can be written as an MA model of infinite order as well as an AR model of infinite order (see e.g. [50]).

Remark 3.6. The parametric SISO models do not have a gain
Some authors prefer to work with transfer functions of the form $A\frac{b(z)}{a(z)}$, where

$a(z)$ and $b(z)$ are monic polynomials, i.e. $a_0 = b_0 = 1$. We will always assume that $A = 1$, and $a_0 = b_0 = 1$. This does not restrict the model class since the input variance σ^2 acts as a gain factor. For the output process, this is equivalent to having models with transfer function $H(z) = A\frac{b(z)}{a(z)}$, where A is the gain, and restricting the input process to white noise with variance 1. Note, however, that the zeroth Markov parameter of the SISO models that we consider, is always equal to 1. ○

State space representations for the parametric SISO models

The models presented above can be transformed into an equivalent state space model

$$\begin{cases} x(k+1) &= Ax(k) + Ku(k)\,, \\ y(k) &= Cx(k) + u(k)\,. \end{cases} \tag{3.51}$$

For an ARMA(p,q) model this is done in the following way. Let $n = \max(p,q)$ and define $a_{p+1} = a_{p+2} = \cdots = a_n = 0$ if $p < q$ and $b_{q+1} = b_{q+2} = \cdots = b_n = 0$ if $q < p$, so that the factor z^{p-q} of the transfer function in (3.50) is absorbed in the polynomials. Hence, $H(z) = \frac{b(z)}{a(z)}$, where $b(z) = z^n + b_1 z^{n-1} + \cdots + b_n$ and $a(z) = z^n + a_1 z^{n-1} + \cdots + a_n$. Then, the matrices of (3.51) can be chosen as

$$A = \begin{pmatrix} -a_1 & -a_2 & \cdots & -a_{n-1} & -a_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \quad K = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.52}$$

and

$$C = \begin{pmatrix} b_1 - a_1 & b_2 - a_2 & \cdots & b_n - a_n \end{pmatrix}\,. \tag{3.53}$$

This particular state space representation for the ARMA model is called the controller canonical form[5]. The controller canonical form for an MA(n) is obtained by putting the coefficients a_i $(i = 1, \ldots, n)$ in (3.52) and (3.53) equal to zero. Similarly, the controller canonical form for an AR(n) model is obtained by putting the b_i's in (3.53) to zero.

Remark 3.7. An ARMA model of order n
By an nth order ARMA model, we mean a model with transfer function $H(z) = \frac{b(z)}{a(z)}$, where $a(z)$ and $b(z)$ are polynomials of order n. Its state space matrices have dimensions $A \in \mathbb{R}^{n\times n}$, $K \in \mathbb{R}^n$ and $C \in \mathbb{R}^{1\times n}$. Note that an nth order ARMA model is not always an ARMA(n,n) model. If the model has s poles in the origin, it is an ARMA$(n-s,n)$ model. Similarly, if s zeros lie in the origin, the model is an ARMA$(n,n-s)$ model. ○

[5] There is disagreement in the literature on the precise definition for these canonical forms. We use the Matlab convention.

Properties of SISO models

We now give two properties that hold for SISO models but not for MIMO models, which will prove useful further on. First, in Proposition 3.2 we show that the row space of the controllability matrix of a SISO ARMA models is only dependent on the poles of the model. This will be used in the proofs of Proposition 3.9 and Proposition 3.10, in Section 3.5.3 and in Chapter 6. Proposition 3.3 describes the special structure of the row space of the output Hankel matrix of a SISO AR process. It will be useful in derivations in Section 3.5.1 and in Section 3.5.2.

Proposition 3.2. The row space of the controllability matrix of a SISO ARMA model
Consider a minimal SISO ARMA model with distinct poles $\alpha_1, \ldots, \alpha_n$. The row space of the $n \times j$ $(j \geq n)$ controllability matrix of the model, denoted by $\mathcal{C}_j$, is equal to the row space of $\mathcal{C}_j^*$, where $\mathcal{C}_j^*$ is defined as

$$\mathcal{C}_j^* = \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots & \alpha_1^{j-1} \\ 1 & \alpha_2 & \alpha_2^2 & \cdots & \alpha_2^{j-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \alpha_n & \alpha_n^2 & \cdots & \alpha_n^{j-1} \end{pmatrix}. \tag{3.54}$$

Proof.
Let (A, K, C) be the state space representation of the model and T the similarity transformation that makes the system matrix A diagonal. The transformed matrices are denoted by $\Lambda = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) = T^{-1}AT$, $C^* = CT$ and $K^* = T^{-1}K$. Then, the controllability matrix of the model can be written as

$$\begin{aligned} \mathcal{C}_j &= T \begin{pmatrix} K^* & \Lambda K^* & \Lambda^2 K^* & \cdots & \Lambda^{j-1} K^* \end{pmatrix}, \\ &= T \operatorname{diag}(k_1^*, \ldots, k_n^*) \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots & \alpha_1^{j-1} \\ 1 & \alpha_2 & \alpha_2^2 & \cdots & \alpha_2^{j-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \alpha_n & \alpha_n^2 & \cdots & \alpha_n^{j-1} \end{pmatrix}, \end{aligned}$$

where k_i^* $(i = 1, \ldots, n)$ are the components of K^*. Since T and $\operatorname{diag}(k_1^*, \ldots, k_n^*)$ are full rank matrices, the row space of $\mathcal{C}_j$ is equal to the row space of $\mathcal{C}_j^*$. □

Remark 3.8. Poles with multiplicity larger than 1
If one or more poles of the model have multiplicity larger than 1, the Jordan form of the matrix A should be used instead of the diagonal matrix Λ. One can then show in an analogous way as the proof of Proposition 3.2, that each pole α with multiplicity $m \geq 1$ contributes an $m \times j$ block of the following form to

the matrix $\mathcal{C}_j^*$

$$\begin{pmatrix} \binom{0}{0} & \binom{1}{0}\alpha & \binom{2}{0}\alpha^2 & \cdots & \binom{m-1}{0}\alpha^{m-1} & \binom{m}{0}\alpha^m & \cdots & \binom{j-1}{0}\alpha^{j-1} \\ 0 & \binom{1}{1} & \binom{2}{1}\alpha & \cdots & \binom{m-1}{1}\alpha^{m-2} & \binom{m}{1}\alpha^{m-1} & \cdots & \binom{j-1}{1}\alpha^{j-2} \\ 0 & 0 & \binom{2}{2} & \cdots & \binom{m-1}{2}\alpha^{m-3} & \binom{m}{2}\alpha^{m-2} & \cdots & \binom{j-1}{2}\alpha^{j-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \binom{m-1}{m-1} & \binom{m}{m-1}\alpha & \cdots & \binom{j-1}{m-1}\alpha^{j-m} \end{pmatrix},$$

where $\binom{k}{l}$ is the binomial coefficient: $\binom{k}{l} = \frac{k!}{l!(k-l)!}$ ○

Remark 3.9. The column space of the observability matrix of a SISO ARMA model

It is easily seen that the column space of the observability matrix of a SISO ARMA model is equal to the row space of the controllability matrix of the model:

$$\text{range}(\Gamma_j) = \text{row}(\mathcal{C}_j) \ .$$

○

Proposition 3.3. The row space of the output Hankel matrix of a SISO AR process

The row space of the output Hankel matrix of an nth order AR process is equal to

$$\text{row}(Y_{0|i-1}) = \text{row}\begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix} \quad \textit{for } j \to \infty \textit{ and } i \geq n \ .$$

Proof.

From the controller canonical form (3.52–3.53) it follows immediately that for an AR model ($b_1 = \cdots = b_n = 0$) we can take the n past outputs as state components:

$$x(k) = \begin{pmatrix} y(k-1) \\ y(k-2) \\ \vdots \\ y(k-n) \end{pmatrix} .$$

Consequently, the state sequence matrix

$$X_n = \begin{pmatrix} x(n) & x(n+1) & \cdots & x(n+j-1) \end{pmatrix}$$

is a Toeplitz matrix that is obtained by interchanging the rows of the output

Hankel matrix $Y_{0|n-1}$:

$$X_n = \begin{pmatrix} y(n-1) & y(n) & \cdots & y(n+j-2) \\ y(n-2) & y(n-1) & \cdots & y(n+j-3) \\ \vdots & \vdots & \cdots & \vdots \\ y(0) & y(1) & \vdots & y(j-1) \end{pmatrix} \in \mathbb{R}^{n \times j} \ .$$

Therefore, the row space of $Y_{0|n-1}$ is equal to the row space of X_n:

$$\text{row}(Y_{0|n-1}) = \text{row}(X_n) \ . \tag{3.55}$$

Furthermore, it is easy to see from the state space equation (3.28) that

$$\text{row}(Y_{n|i-1}) \subseteq \text{row} \begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix} \ . \tag{3.56}$$

Combining (3.55) and (3.56), we obtain

$$\text{row}(Y_{0|i-1}) \subseteq \text{row} \begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix} \ .$$

The matrices $Y_{0|i-1}$ and $\begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix}$ are both of rank i for $j \to \infty$ (see Proposition 3.1). Consequently,

$$\text{row}(Y_{0|i-1}) = \text{row} \begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix} \quad \text{for } j \to \infty.$$

□

3.3 System identification

In the previous section we have given representations for linear stochastic models. In this section we briefly describe how these models can be obtained from a given sequence of measurements. This is the black-box approach to modeling, in contrast to white-box techniques, where a model is found by writing down the physical equations that govern the system under consideration.

In Chapter 5 some new identification techniques are briefly discussed which are based on the cepstrum of the measurements. Here, we only deal with identification based on the Markov parameters (deterministic identification) or based on the covariances (stochastic identification).

3.3.1 Deterministic identification

In deterministic realization theory, a state space model is constructed from the Markov parameters. This can for example be done in the following way, which was described by Kung [92]. Let $H(0), \ldots, H(N)$ be a finite sequence of Markov parameters of a linear model with state space matrices (A, K, C) (see Section 3.2.1).

- Put the Markov parameters $H(1), \ldots, H(N) \in \mathbb{R}^m$ in a block Hankel matrix
$$\mathcal{H}_i = \begin{pmatrix} H(1) & H(2) & \ldots & H(i) \\ H(2) & H(3) & \ldots & H(i+1) \\ \vdots & \vdots & & \vdots \\ H(i) & H(i+1) & \ldots & H(2i-1) \end{pmatrix} ,$$
where $2i - 1 \leq N$. The parameter i should be larger than or equal to the model order. If the model order is not known, i should be large enough, so that $\mathcal{H}_i$ is rank deficient.

- Since the Markov parameters are equal to
$$H(0) = I_m \text{ and } H(k) = CA^{k-1}K \text{ for } k > 0 , \tag{3.57}$$
the block Hankel matrix $\mathcal{H}_i$ can be factorized as
$$\mathcal{H}_i = \Gamma_i \mathcal{C}_i ,$$
where Γ_i is the $mi \times n$ observability matrix and $\mathcal{C}_i$ the $n \times mi$ controllability matrix of the system. Given the matrix $\mathcal{H}_i$, the factorization can be done by the reduced SVD (Corollary 2.4). Let $\mathcal{H}_i = U_1 S_1 V_1^T$, then up to a similarity transformation, the matrices Γ_i and $\mathcal{C}_i$ can be chosen to be
$$\Gamma_i = U_1 S_1^{1/2} \text{ and } \mathcal{C}_i = S^{1/2} V_1^T .$$

- The state space matrices are obtained as follows. The matrix C is formed by the first m rows of Γ_i, while K is obtained as the first m columns of $\mathcal{C}_i$. Furthermore, the matrix A satisfies
$$\underline{\Gamma}_i A = \overline{\Gamma}_i ,$$
where $\underline{\Gamma}_i$ and $\overline{\Gamma}_i$ are the $m(i-1) \times n$ matrices constructed from Γ_i by omission of the last and first m rows, respectively. The matrix A can thus be estimated as $A = \underline{\Gamma}_i^{\dagger} \overline{\Gamma}_i$.

Other algorithms are given in e.g. [107, 164].
Identifying a model based on measurements could be done by first estimating the Markov parameters and then applying the above procedure.

3.3.2 Stochastic identification

In stochastic identification the input is assumed to be an unknown sequence of white noise and the only data available for determining a model is the corresponding sequence of outputs, which can be viewed as the outcome of a stochastic process. The aim of stochastic system identification is thus to approximate a stochastic process as the output of a linear model that is driven by a white noise process. Most stochastic identification algorithms are based on an estimate of the autocovariance sequence of the process. However, as we will briefly indicate in Chapter 5, for SISO systems one can also rely on the cepstral coefficients, or make a combination of the two parameter sets, autocovariance estimates and cepstral coefficients. The cepstrum is discussed in detail in Chapter 5.

There is a vast literature on (stochastic) system identification and we shall not go into describing the various methods. We refer the reader to the standard books [64, 101, 134, 159]. We only mention some important aspects. There are two main classes of identification methods for linear models: prediction error methods [101] and subspace identification methods [94, 156, 159].

In prediction error methods the parametric models (AR, MA, ARMA), are identified via the solution of an optimization problem. For AR models this can be done very efficiently. On the other hand, if an ARMA model is identified, the problem becomes highly nonlinear and is solved as a non-convex optimization problem, which may converge to a local optimum.

Other methods do not have these drawbacks, e.g. the methods proposed by Stoica et al. [141] in which a convex minimization problem is solved, which leads to a computationally fast algorithm.

Rather recently, so-called subspace identification methods have become very popular. They are based on similar principles as Kung's deterministic realization algorithm, which was described above. This can be seen as follows. The autocovariance function of a stochastic linear model is equal to $R(\tau) = CA^{\tau-1}G$ for $\tau > 0$, where A is the system matrix and C the output matrix of the model and $G = E\left\{X(k+1)Y(k)^T\right\}$. By comparing it to the Markov parameters of a linear model in (3.57), we see that the autocovariance matrices can be viewed as the Markov parameters of the linear model with state space matrices (A, G, C). The block Hankel covariance matrix $\mathcal{R}_i$ thus be factorized into two rank n matrices:

$$\mathcal{R}_i = \begin{pmatrix} CG & CAG & \cdots & CA^{i-1}G \\ CAG & CA^2G & \cdots & CA^iG \\ \vdots & \vdots & \cdots & \vdots \\ CA^{i-1}G & CA^iG & \cdots & CA^{2i-2}G \end{pmatrix} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{i-1} \end{pmatrix} \begin{pmatrix} G & AG & \cdots & A^{i-1}G \end{pmatrix} .$$

This explains why we looked in Example 3.1 at the singular values of the matrix $\mathcal{R}_{30}$ to determine the order of the model. Just as in the deterministic case, the factorization of the block Hankel covariance matrix can be carried out by the SVD and from the factors, the matrices A, G and C can be determined analogously as in the deterministic realization problem in Section 3.3.1. From this triplet (A, G, C), the complete forward innovation model can be obtained as follows. First, the state covariance matrix Σ is computed as the solution of the Riccati equation (3.14). Then, the Kalman gain K is obtained as $(G - A\Sigma C^T)(R_y(0) - C\Sigma C^T)^{-1}$ and the input covariance matrix S_u is equal to $R_y(0) - C\Sigma C^T$.

The big advantage of the subspace methods over traditional methods is that optimization is avoided and instead numerically stable linear algebraic operations are used[6]. Moreover, the multiple output case is solved as easily as the scalar case. However, there is a price to be paid. The models obtained are not guaranteed to be stable. Furthermore, it is possible that the identified matrices A, C and G do not produce a valid covariance function $CA^{\tau-1}G$, i.e. one that leads to a spectral density that is positive on the unit circle, see [24, 31, 99, 150]. This is often referred to as the problem of positive realness. The regularization technique in [152, 153] remedies the stability problem and one of the algorithms in [159] leads to a model with spectral density that is guaranteed to be positive on the unit circle, but at the expense of a bias. Furthermore, the statistical properties of the subspace algorithms are not so well known as those of the prediction error methods, although also on this aspect progress has been made [9, 11]. Recently, an alternative strategy that has the advantages of subspace identification but none of its shortcomings, has been proposed [30, 32, 104]. It consists of covariance extension, followed by stochastically balanced model reduction.

Stochastic identification problems are found in many applications, even where one does not directly expect them. For the modeling of the traffic in asynchronous transmission mode (ATM) networks, for example, Markov modulated Poisson process (MMPP) models have been proposed [66]. They do not belong to the LTI model class. However, the autocovariance function of an nth order MMPP is of the form [155]

$$R(\tau) = \theta D_\gamma P^\tau D_\gamma e \ ,$$

where $P \in \mathbb{R}^{n \times n}$ is the transition probability matrix of the model, $D_\gamma \in \mathbb{R}^{n \times n}$ is a diagonal matrix with the Poisson parameters $\gamma_1, \ldots, \gamma_n$ on the diagonal, $\theta \in \mathbb{R}^{1 \times n}$ is the row vector with the state probabilities and $e \in \mathbb{R}^{n \times 1}$ is equal to $\begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}^T$. By denoting θD_γ by θ_γ and $PD_\gamma e$ by $e_{P\gamma}$, we see that the autocovariance function of the MMPP is equal to

$$R(\tau) = \theta_\gamma P^{\tau-1} e_{P\gamma} \ ,$$

which can be viewed as the impulse response of the SISO LTI model with state space matrices $(P, e_{P\gamma}, \theta_\gamma)$. Nevertheless, none of the identification algorithms presented above, can be applied right away. This is due to constraints on the

[6]The derivation above was only a sketch of the basic principles. In the real subspace algorithms one uses projections, principal angles between subspaces, ... Also, one works with the measured data directly, without explicit estimation of the covariances.

model parameters: P should be a stochastic matrix, i.e. a matrix of which the elements on each row sum to 1, θ should be the left eigenvector of P corresponding to the eigenvalue 1 and the elements of γ should be positive. Therefore, other identification strategies have been worked out in [26, 33, 34, 35, 40, 97, 98, 151, 163].

3.4 The principal angles between the past and future input and output spaces

In this section we compute the principal angles between the past and future input and output spaces and the canonical correlations of the corresponding processes. In Section 3.4.1 we first describe the future and the past of a stochastic process. We show how the canonical correlations of the processes will be computed and indicate how they are related. In Section 3.4.2, we derive expressions for the principal angles between different combinations of past and future input and output spaces, i.e. the row spaces of the $mi \times j$ data block Hankel matrices, where we assume $j \to \infty$. The expressions are in terms of the system matrices (A, K, C) and the input covariance matrix S_u. As we will see, the principal angles converge for $i \to \infty$. The cosines of the limiting principal angles are the canonical correlations of the corresponding processes. In Section 3.4.3 the relations between the different canonical correlations are derived.

3.4.1 Introduction

Past and future input and output processes of a linear model

Let $\{u(k)\ (k \in \mathbb{Z})\}$ and $\{y(k)\ (k \in \mathbb{Z})\}$ denote the input and output process of a linear stochastic model in forward innovation form. We assume that the processes are zero-mean, stationary and ergodic. The past output process is defined as

$$y_p = \{y(k)\ (k < 0)\}\ , \tag{3.58}$$

and the future output process is

$$y_f = \{y(k)\ (k \geq 0)\}\ . \tag{3.59}$$

Analogous definitions hold for the past and the future input process, u_p and u_f, respectively.

The canonical correlations of the past and future input and output processes

In line with the definitions of Section 2.6.3, the canonical correlations of the past and future input and output processes are defined as the canonical correlations of the corresponding random variables $U_{-1|-\infty}$, $U_{0|\infty}$, $Y_{-1|-\infty}$ and $Y_{0|\infty}$, where

$$Y_{-1|-\infty} = \begin{pmatrix} Y(-1) \\ Y(-2) \\ \vdots \end{pmatrix} \quad \text{and} \quad Y_{0|\infty} = \begin{pmatrix} Y(0) \\ Y(1) \\ \vdots \end{pmatrix} ,$$

and analogously for $U_{-1|-\infty}$ and $U_{0|\infty}$. For example, the canonical correlations of the past and future of the output process are equal to

$$\mathrm{cc}(y_p, y_f) = \mathrm{cc}(Y_{-1|-\infty}, Y_{0|\infty}) \ . \tag{3.60}$$

Or, in terms of the data block Hankel matrices as defined in (2.36–2.37), this becomes

$$\mathrm{cc}(y_p, y_f) = \lim_{j\to\infty} \cos\left[Y^{(j)}_{-1|-\infty} \sphericalangle Y^{(j)}_{0|\infty}\right] \ .$$

We can already treat three trivial cases:

1. Due to the independence of the past and future input processes, the canonical correlations of u_p and u_f are all equal to 0:

$$\boxed{\mathrm{cc}(u_p, u_f) = 0, 0, \ldots}$$

2. The future input is also independent of the past output process. Consequently, their canonical correlations are all equal to 0:

$$\boxed{\mathrm{cc}(y_p, u_f) = 0, 0, \ldots}$$

3. The output at a certain time step k is a linear combination of the present input and all past inputs:

$$y(k) = \sum_{i=1}^{\infty} CA^{i-1}Ku(k-i) + u(k) = \sum_{i=0}^{\infty} H(i)u(k-i) \ ,$$

where $H(i)$ is the ith Markov parameter of the linear model (see (3.5)). Consequently, all random variables in $Y_{-1|-\infty}$ can be obtained as linear combinations of the random variables in $U_{-1|-\infty}$. Otherwise formulated, the row space of $Y^{(j)}_{-1|-\infty}$ is contained in the row space of $U^{(j)}_{-1|-\infty}$:

$$\mathrm{row}(Y^{(j)}_{-1|-\infty}) \subseteq \mathrm{row}(U^{(j)}_{-1|-\infty}) \ . \tag{3.61}$$

The canonical correlations of the past input and past output processes are consequently all equal to 1:

$$\boxed{\mathrm{cc}(u_p, y_p) = 1, 1, \ldots} \tag{3.62}$$

Moreover, by applying the same reasoning to the inverse model, we obtain $u(k)$ as a linear combination of the present and past outputs:

$$u(k) = -\sum_{i=1}^{\infty} C(A - KC)^{i-1} K y(k-i) + y(k) = \sum_{i=0}^{\infty} H_z(i) y(k-i) ,$$

where $H_z(i)$ are the Markov parameters of the inverse model (see (3.43)). This leads to

$$\mathrm{row}(U^{(j)}_{-\infty|-1}) \subseteq \mathrm{row}(Y^{(j)}_{-\infty|-1}) . \tag{3.63}$$

It follows from (3.61) and (3.63) that the past input and the past output process span the same space:

$$\mathrm{row}(Y^{(j)}_{-\infty|-1}) = \mathrm{row}(U^{(j)}_{-\infty|-1}) . \tag{3.64}$$

The canonical correlations of the other combinations of past and future input and output processes can be obtained as the following limits (see also Remark 3.12 on page 117):

$$\mathrm{cc}(u_f, y_f) = \lim_{i\to\infty} \mathrm{cc}(U_{0|i-1}, Y_{0|i-1}) , \tag{3.65a}$$

$$\mathrm{cc}(u_p, y_f) = \lim_{i\to\infty} \mathrm{cc}(U_{-i|-1}, Y_{0|i-1}) , \tag{3.65b}$$

$$\mathrm{cc}(y_p, y_f) = \lim_{i\to\infty} \mathrm{cc}(Y_{-i|-1}, Y_{0|i-1}) , \tag{3.65c}$$

where

$$Y_{-i|-1} = \begin{pmatrix} Y(-i) \\ Y(-i+1) \\ \vdots \\ Y(-1) \end{pmatrix} \text{ and } Y_{0|i-1} = \begin{pmatrix} Y(0) \\ Y(1) \\ \vdots \\ Y(i-1) \end{pmatrix} ,$$

and analogously for the input random variables. The parameter i describes how far we go back into the past ($k = -i$) and forward into the future ($k = i-1$), where the present is at $k = 0$.

Since the processes are stationary, we can as well take the present at time instant $k = i$, the past from $k = 0$ to $k = i - 1$ and the future from $k = i$ to $k = 2i - 1$. This is only a convention that allows us to estimate the canonical correlations from measured data sequences. The canonical correlations of the past and future output processes, e.g., can then be computed as

$$\mathrm{cc}(y_p, y_f) = \lim_{i\to\infty} \mathrm{cc}(Y_{0|i-1}, Y_{i|2i-1}) .$$

As explained in Section 2.6.3, the canonical correlations of $Y_{0|i-1}$ and $Y_{i|2i-1}$ can be obtained as the cosines of the principal angles between the row spaces of the $mi \times j$ data block Hankel matrices $Y^{(j)}_{0|i-1}$ and $Y^{(j)}_{i|2i-1}$, provided $j \to \infty$. These block Hankel matrices are equal to the past and future output block Hankel matrices Y_p and Y_f, which are defined in (3.26). Consequently, we can compute the canonical correlations of the combinations of past and future processes in (3.65a–3.65c) as follows:

$$\mathrm{cc}(u_f, y_f) = \lim_{i\to\infty} \lim_{j\to\infty} \cos\left[U_f \sphericalangle Y_f\right] \;, \tag{3.66a}$$

$$\mathrm{cc}(u_p, y_f) = \lim_{i\to\infty} \lim_{j\to\infty} \cos\left[U_p \sphericalangle Y_f\right] \;, \tag{3.66b}$$

$$\mathrm{cc}(y_p, y_f) = \lim_{i\to\infty} \lim_{j\to\infty} \cos\left[Y_p \sphericalangle Y_f\right] \;. \tag{3.66c}$$

This explains why we denote the first i block rows of the output block Hankel matrix by Y_p and the following i block rows by Y_f (see (3.26)), and similarly for U_p and U_f.

Overview of the relations between the canonical correlations

From Equation (3.64) we can already deduce that the canonical correlations of the past input and future output are equal to the canonical correlations of the past and future output:

$$\mathrm{cc}(u_p, y_f) = \mathrm{cc}(y_p, y_f) \;.$$

The parametric expressions that we derive in Section 3.4.2, will also reveal that the canonical correlations of the future input and future output are related to the canonical correlations of the past and future output in the following way:

$$\mathrm{cc}^2(u_f, y_f) = 1 - \mathrm{cc}^2(y_p, y_f) \;.$$

	u_p	u_f	y_p	y_f
u_p	1	0	1	ρ_k
u_f	0	1	0	$\sqrt{1-\rho_k^2}$
y_p	1	0	1	ρ_k
y_f	ρ_k	$\sqrt{1-\rho_k^2}$	ρ_k	1

Table 3.2: Overview of the relations between the different sets of canonical correlations.

The corresponding principal angles are complementary:

$$[U_f \sphericalangle Y_f] = \frac{\pi}{2} - [Y_p \sphericalangle Y_f] \quad \text{for } i, j \to \infty \ .$$

An overview of the relations of the canonical correlations of the different combinations of processes is given in Table 3.2. The canonical correlations of the past and the future output are denoted by ρ_k in this table.

3.4.2 Computation of the principal angles between the past and future input and output spaces

In this section, the structure of which is given in Table 3.3, we compute the principal angles between the row spaces of the following combinations of input and output matrices in terms of the system parameters: $[U_f \sphericalangle Y_f]$, $[U_p \sphericalangle Y_f]$ and $[Y_p \sphericalangle Y_f]$. Note that we will always implicitly assume that $j \to \infty$ in the theoretical derivations of the principal angles. The canonical correlations of

1. future input and future output
   - 1.a principal angles: $[U_f \sphericalangle Y_f]$
     - intersection and orthogonal directions
     - factors of $(U_f U_f^T)^{-1} U_f Y_f^T (Y_f Y_f^T)^{-1} Y_f U_f^T$
     - computing the eigenvalues
   - 1.b canonical correlations: $\mathrm{cc}(u_f, y_f)$
2. past input and future output
   - 2.a principal angles: $[U_p \sphericalangle Y_f]$
     - intersection and orthogonal directions
     - factors of $(U_p U_p^T)^{-1} U_p Y_f^T (Y_f Y_f^T)^{-1} Y_f U_p^T$
     - computing the eigenvalues
   - 2.b canonical correlations: $\mathrm{cc}(u_p, y_f)$
3. past output and future output
   - 3.a principal angles: $[Y_p \sphericalangle Y_f]$
     - intersection and orthogonal directions
     - factors of $(Y_p Y_p^T)^{-1} Y_p Y_f^T (Y_f Y_f^T)^{-1} Y_f Y_p^T$
     - computing the eigenvalues
   - 3.b canonical correlations: $\mathrm{cc}(y_p, y_f)$
4. summary

Table 3.3: The structure of Section 3.4.2.

the corresponding processes are derived by taking the limit for $i \to \infty$ of the cosines of the principal angles.
The procedure is the same for the three cases. First, we determine the number of zero principal angles and the number of angles that are equal to $\frac{\pi}{2}$. Then, we compute the factors of the matrix of which the eigenvalues are the squared cosines of the principal angles between the row spaces (see (2.22)). Finally, the canonical correlations of the processes are obtained by taking $i \to \infty$.
Throughout this section, we will need the following expression for the inverse of a sum of matrices. It is referred to as the matrix inversion lemma.

Lemma 3.4. The matrix inversion lemma
If $A \in \mathbb{R}^{q\times q}$ and $R \in \mathbb{R}^{p\times p}$ are non-singular matrices, then the following equality holds

$$(A + BRC)^{-1} = A^{-1} - A^{-1}B(R^{-1} + CA^{-1}B)^{-1}CA^{-1} \ .$$

1.a The principal angles between the row spaces of U_f and Y_f

Since for $j \to \infty$, $U_f \in \mathbb{R}^{mi\times j}$ and $Y_f \in \mathbb{R}^{mi\times j}$ are of full row rank (see (3.32) and (3.37)), the total number of principal angles between $\text{row}(U_f)$ and $\text{row}(Y_f)$ is mi.

Intersection and orthogonal directions
We compute the dimension $t_{(U_f,Y_f)}$ of the intersection of the row spaces of U_f and Y_f, which is equal to the number of zero principal angles between the two subspaces, by applying Grassmann's dimension theorem (Theorem 2.2). The number of orthogonal principal directions is given by $r_{(U_f,Y_f)}$, which can be computed as in (2.19).

1. The number of zero principal angles between $\text{row}(U_f)$ and $\text{row}(Y_f)$ is equal to
$$t_{(U_f,Y_f)} = \text{rank}(U_f) + \text{rank}(Y_f) - \text{rank}\begin{pmatrix} U_f \\ Y_f \end{pmatrix} \ .$$
The rank of U_f and Y_f is mi, and the third matrix can be written as
$$\begin{pmatrix} U_f \\ Y_f \end{pmatrix} = \begin{pmatrix} I_i & 0 \\ H_i & \Gamma_i \end{pmatrix} \begin{pmatrix} U_f \\ X_f \end{pmatrix} \ .$$
The matrix $\begin{pmatrix} I_i & 0 \\ H_i & \Gamma_i \end{pmatrix} \in \mathbb{R}^{2mi\times(mi+n)}$ is of rank $mi + n$ and since the rows of U_f are linearly independent of the rows of X_f, the rank of $\begin{pmatrix} U_f \\ X_f \end{pmatrix}$ is equal to $mi + n$. Hence,
$$\text{rank}\begin{pmatrix} U_f \\ Y_f \end{pmatrix} = mi + n \ ,$$

which implies that

$$t_{(U_f,Y_f)} = mi - n \text{ for } j \to \infty \ . \tag{3.67}$$

2. The number of principal angles equal to $\frac{\pi}{2}$ is computed as follows:

$$r_{(U_f,Y_f)} = \min(\text{rank}(U_f), \text{rank}(Y_f)) - \text{rank}(U_f Y_f^T) \ .$$

From the output equation (3.29) and Assumptions 1 and 2, we know that

$$\lim_{j\to\infty} U_f Y_f^T = Q_{u_i} H_i^T \in \mathbb{R}^{mi\times mi} \ ,$$

of which the rank is mi. Consequently,

$$r_{(U_f,Y_f)} = 0 \text{ for } j \to \infty \ . \tag{3.68}$$

It follows from (3.67) and (3.68) that for $j \to \infty$, n principal angles θ_k between row(U_f) and row(Y_f) satisfy

$$0 < \theta_k < \frac{\pi}{2} \qquad (k = mi - n + 1, \ldots, mi) \ ,$$

and the other $mi - n$ principal angles are all zero.

The factors of $(U_f U_f^T)^{-1} U_f Y_f^T (Y_f Y_f^T)^{-1} Y_f U_f^T$

In order to compute the principal angles between row(U_f) and row(Y_f) we will compute the eigenvalues of the matrix

$$M_{(U_f,Y_f)} = (U_f U_f^T)^{-1} U_f Y_f^T (Y_f Y_f^T)^{-1} Y_f U_f^T$$

(see (2.22)), where we assume that the number of columns in U_f and Y_f is very large ($j \to \infty$). The four factors are first expressed in terms of the system matrices, the state covariance matrix Σ and the input covariance matrix S_u. In the following paragraph we will then further simplify the expression of the matrix $M_{(U_f,Y_f)}$.
From the state space equations (3.27–3.29), the assumptions in Section 3.2, the definition of Q_{u_i} in (3.31) and expression (3.33), it follows that

$$\begin{aligned}
\lim_{j\to\infty} U_f U_f^T &= Q_{u_i} \ , \\
\lim_{j\to\infty} U_f Y_f^T &= Q_{u_i} H_i^T \ , \\
\lim_{j\to\infty} Y_f Y_f^T &= \Gamma_i \Sigma \Gamma_i^T + H_i Q_{u_i} H_i^T \ , \\
\lim_{j\to\infty} Y_f U_f^T &= H_i Q_{u_i} \ .
\end{aligned}$$

Consequently, the squared cosines of the principal angles between the row spaces of U_f and Y_f are the mi eigenvalues of

$$M_{(U_f,Y_f)} = H_i^T (\Gamma_i \Sigma \Gamma_i^T + H_i Q_{u_i} H_i^T)^{-1} H_i Q_{u_i} \ .$$

Computation of the eigenvalues
By using the observability matrix of the inverse system, Γ_{z_i}, which is equal to $-H_i^{-1}\Gamma_i$, the matrix $M_{(U_f,Y_f)}$ can be expressed as

$$M_{(U_f,Y_f)} = (\Gamma_{z_i}\Sigma\Gamma_{z_i}^T + Q_{u_i})^{-1}Q_{u_i} \ .$$

Working out the inversion of the sum by means of the matrix inversion lemma (Lemma 3.4), leads to

$$M_{(U_f,Y_f)} = I_{mi} - Q_{u_i}^{-1}\Gamma_{z_i}(\Sigma^{-1} + \Gamma_{z_i}^T Q_{u_i}^{-1}\Gamma_{z_i})^{-1}\Gamma_{z_i}^T \ ,$$

where I_{mi} is the $mi \times mi$ identity matrix. Since the matrix $Q_{u_i}^{-1}\Gamma_{z_i}(\Sigma^{-1} + \Gamma_{z_i}^T Q_{u_i}^{-1}\Gamma_{z_i})^{-1}\Gamma_{z_i}^T \in \mathbb{R}^{mi\times mi}$ is of rank n, $mi-n$ eigenvalues of $M_{(U_f,Y_f)}$ are equal to 1. This is indeed what we predicted in (3.67). The other eigenvalues of $M_{(U_f,Y_f)}$ are equal to the eigenvalues of

$$I_n - \Gamma_{z_i}^T Q_{u_i}^{-1}\Gamma_{z_i}(\Sigma^{-1} + \Gamma_{z_i}^T Q_{u_i}^{-1}\Gamma_{z_i})^{-1} \ , \tag{3.69}$$

where I_n is the $n \times n$ identity matrix. Let $\Gamma_{z_i}^T Q_{u_i}^{-1}\Gamma_{z_i}$ be denoted by $\mathcal{G}_{z_i}$, which is equal to

$$\mathcal{G}_{z_i} = \sum_{k=0}^{i-1}(A-KC)^{k^T}C^T S_u^{-1}C(A-KC)^k \ , \tag{3.70}$$

then, (3.69) reduces to

$$I_n - \mathcal{G}_{z_i}(\Sigma^{-1} + \mathcal{G}_{z_i})^{-1} = (I_n + \mathcal{G}_{z_i}\Sigma)^{-1} \ .$$

The principal angles between row($\boldsymbol{U_f}$) and row($\boldsymbol{Y_f}$)
The squared cosines of the largest n principal angles between row(U_f) and row(Y_f) for $j \to \infty$ and finite i are the eigenvalues of $(I_n + \mathcal{G}_{z_i}\Sigma)^{-1}$. The other $mi-n$ principal angles are equal to zero:

$$\lim_{j\to\infty} \cos^2[U_f \sphericalangle Y_f] = \lambda\left((I_n + \mathcal{G}_{z_i}\Sigma)^{-1}\right), \underbrace{1,\ldots,1}_{mi-n} \ . \tag{3.71}$$

Remark 3.10. $\boldsymbol{\mathcal{G}_{z_i}}$ as the solution of a Lyapunov equation
If the state space matrices (A,K,C) and the input covariance matrix S_u are known, then the matrix $\mathcal{G}_{z_i}$ can be computed by making the sum in (3.70). However, $\mathcal{G}_{z_i}$ is also the solution of the following Lyapunov equation:

$$\mathcal{G}_{z_i} = (A-KC)^T\mathcal{G}_{z_i}(A-KC) + C^T S_u^{-1}C \\ - (A-KC)^{i^T}C^T S_u^{-1}C(A-KC)^i \ . \tag{3.72}$$

Since the models that we consider are minimum phase and the eigenvalues of $(A - KC)$ are the zeros of the model (A, K, C), we see that the contribution of the last term $(A - KC)^{i^T} C^T S_u^{-1} C (A - KC)^i$ diminishes when i becomes larger. For $i \to \infty$, which we will discuss in the next paragraph, we obtain

$$\mathcal{G}_z = (A - KC)^T \mathcal{G}_z (A - KC) + C^T S_u^{-1} C \ ,$$

where $\mathcal{G}_z = \lim_{i\to\infty} \mathcal{G}_{z_i}$. Equation (3.72) shows that the zeros of the model (A, K, C) determine the convergence rate of $\mathcal{G}_{z_i}$ to $\mathcal{G}_z$ and hence of the cosines of the principal angles between row(U_f) and row(Y_f) to the canonical correlations of the future input and output processes. This will be illustrated in Example 3.5. ○

Example 3.3. The principal angles between row(U_f) and row(Y_f)
We compute the principal angles between the future input and the future output space of the 2-input 2-output model of order 3 with the following state space matrices:

$$A = \begin{pmatrix} -0.1504 & -0.2074 & -0.2789 \\ 0.7811 & 0.0754 & -0.7154 \\ 0.4670 & 0.2294 & -1.2245 \end{pmatrix} \qquad K = \begin{pmatrix} 1.6469 & 0.0883 \\ 1.5047 & -0.5377 \\ 0.0505 & 1.0689 \end{pmatrix}$$

$$C = \begin{pmatrix} -0.3558 & 0.0236 & 0.2204 \\ 1.4693 & 0.0985 & -1.7360 \end{pmatrix} \tag{3.73}$$

The input covariance matrix S_u is equal to

$$S_u = \begin{pmatrix} 0.3813 & -0.1512 \\ -0.1512 & 1.3302 \end{pmatrix} \ . \tag{3.74}$$

This model will be used in all examples throughout this section.
The datasets are obtained as follows. The input sequence $u(k)$ consists of 15 000 white Gaussian noise samples. The output data $y(k)$ are obtained by simulating the model in (3.73) with input $u(k)$ and zero initial state. The input and output block Hankel matrices U_f and Y_f have $i = 20$ block rows (40 rows) and the number of columns is increased from 100 till 10 000 with steps of 100. The principal angles between the row spaces of U_f and Y_f are then computed with the method described in Section 2.5.6, which is based on the LQ decomposition. Besides the principal angles between the row spaces of the $40 \times j$ data block Hankel matrices, we also compute the asymptotic values for $j \to \infty$ by means of (3.71). This corresponds to the case that the covariance matrices of $\{u(k)\}$ and $\{y(k)\}$ and their cross-covariance matrices are estimated exactly. Note that these asymptotic values are obtained for finite i, namely $i = 20$. Their numerical values are $\cos\theta_1 = \cdots = \cos\theta_{37} = 1$, $\cos\theta_{38} = 0.7334$, $\cos\theta_{39} = 0.3205$ and $\cos\theta_{40} = 0.0539$.
In Figure 3.10 the results are shown. The cosines of the principal angles between the row spaces of the data block Hankel matrices are in full lines, the asymptotic

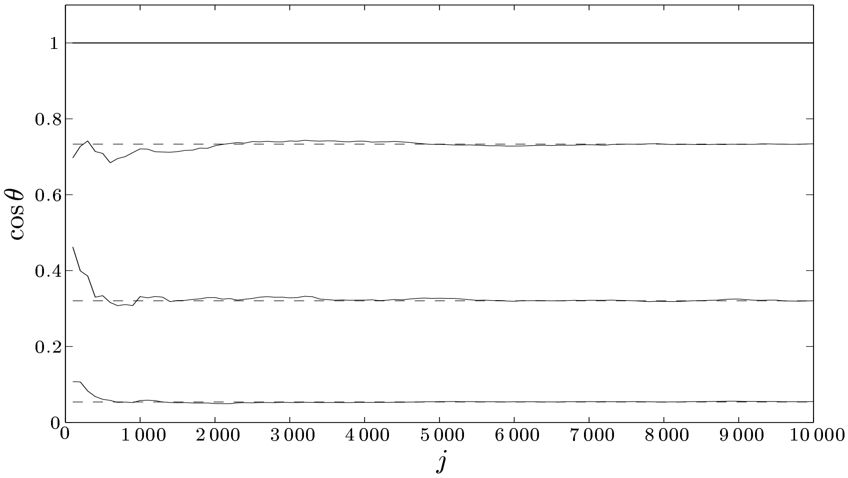


Figure 3.10: The cosines of the principal angles between the row spaces of U_f and Y_f are shown in full line. The block Hankel matrices have $i = 20$ block rows and an increasing number of columns: $j = 100, 200, \ldots, 10\,000$. The dashed lines represent the asymptotic values for $i = 20$ and $j = \infty$.

values in dashed lines. Although only three angles seem to be shown, there are in fact 40 lines. The other 37 lines are at 1, which means that 37 angles are equal to zero, even for finite j. This can be explained as follows. In this example, U_f and Y_f are of full row rank for each value of j. Furthermore, the rows of U_f and X_f are linearly independent for all j. This implies that $t_{(U_f,Y_f)}$, the number of zero principal angles between row(U_f) and row(Y_f), is equal to $mi - n = 37$, for all values of j. This will always be the case for generic data sets. $\Diamond$

1.b The canonical correlations of the future input and the future output processes

The canonical correlations of the future input and output processes of the linear model are the cosines of the principal angles between row(U_f) and row(Y_f) for $i, j \to \infty$.
Let $\mathcal{G}_z$ be the limit of $\mathcal{G}_{z_i}$ for $i \to \infty$

$$\mathcal{G}_z = \lim_{i\to\infty} \Gamma_{z_i}^T Q_{u_i}^{-1} \Gamma_{z_i} \ . \tag{3.75}$$

As we have seen in Remark 3.10, $\mathcal{G}_z$ satisfies the following discrete Lyapunov equation

$$\lim_{i\to\infty} \mathcal{G}_{z_i} = \mathcal{G}_z = (A - KC)^T \mathcal{G}_z (A - KC) + C^T S_u C \ . \tag{3.76}$$

From (3.71) we know that there are only n canonical correlations different from 1. They are the square roots of the eigenvalues of $(I_n + \mathcal{G}_z\Sigma)^{-1}$. If the state space matrices (A, K, C) and the input covariance matrix S_u are given, then the canonical correlations between the input and the output process are computed by solving two Lyapunov equations and an eigenvalue problem. This is summarized in Scheme 3.1.

The canonical correlations of the future input and the future output processes of a linear system

Given: (A, K, C) and S_u

1. Compute the state covariance matrix Σ by solving the Lyapunov equation
$$\Sigma = A\Sigma A^T + KS_uK^T \ .$$
2. Compute the matrix $\mathcal{G}_z$ by solving the Lyapunov equation
$$\mathcal{G}_z = (A - KC)^T\mathcal{G}_z(A - KC) + C^TS_u^{-1}C \ .$$
3. The smallest n canonical correlations of u_f and y_f are the square roots of the eigenvalues of $(I_n + \mathcal{G}_z\Sigma)^{-1}$ and the other canonical correlations are equal to 1:
$$\mathrm{cc}^2(u_f, y_f) = \lambda\left((I_n + \mathcal{G}_z\Sigma)^{-1}\right), 1, 1, \ldots \tag{3.77}$$

Scheme 3.1: Overview of the computation of the canonical correlations between the input and the output process of a linear model.

Example 3.4. The canonical correlations of u_f and y_f
Let us return to the model described in Example 3.3. We again compute the principal angles between the row spaces of the input and output block Hankel matrices, but now the number of columns is kept constant ($j = 10\,000$) and the number of block rows increases from 5 till 50 in steps of 5. The cosines of these angles are the full lines in Figure 3.11. The dashed lines represent the cosines of the principal angles for the given values of i and for $j = \infty$. They are obtained via (3.71). The asymptotic values for $i \to \infty$ and $j \to \infty$, i.e. the canonical correlations of u_f and y_f, which are computed as in (3.77), are given as dash-dot lines. The numerical values of the canonical correlations that are different from 1, are 0.7309, 0.2647 and 0.0439. $\Diamond$

Example 3.5. Influence of the zeros on the convergence
In order to illustrate the role of the zeros in the convergence of the principal angles, we create a new model, which has the same poles as (A, K, C), but zeros that lie closer to the origin. Via pole placement we change the matrix C so that the eigenvalues of $A - KC$ are shifted towards the origin. The new matrix, denoted by $C^{(2)}$, is equal to $\begin{pmatrix} -0.2832 & -0.0103 & 0.1460 \\ 1.2982 & 0.1504 & -1.6356 \end{pmatrix}$. The zeros of

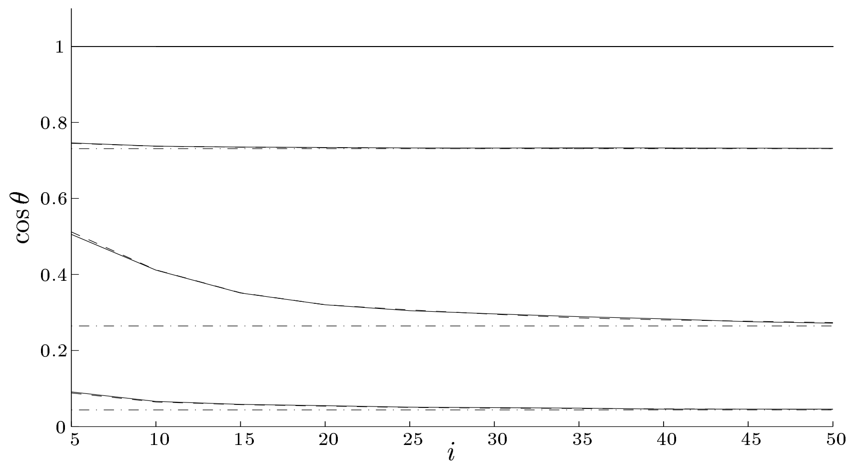


Figure 3.11: The cosines of the principal angles between the row spaces of U_f and Y_f. The full lines represent the cosines of the angles between the row spaces of the $2i \times 10\,000$ data block Hankel matrices, where i increases from 5 to 50 in steps of 5. The dashed lines show the cosines of the angles for the same values of i, but $j = \infty$. The cosines for $i = \infty$ and $j = \infty$ are given by the dash-dot lines. These values are the canonical correlations of the future input and the future output process of the model.

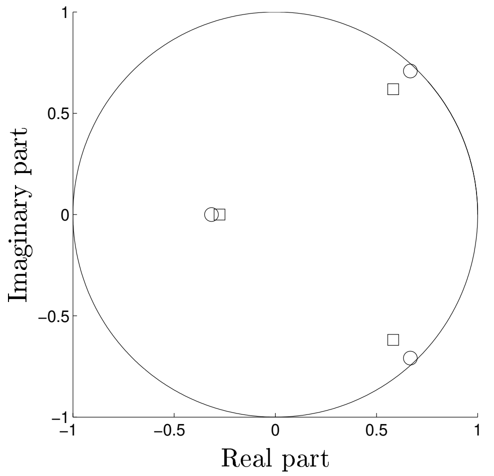


Figure 3.12: The zeros of the two third order models with two inputs and outputs. The ∘-marks are the zeros of the model (A, K, C) in (3.73), the squares are the zeros of the new model $(A, K, C^{(2)})$.

the original model (A, K, C) and of the new model are shown in Figure 3.12. By using the same input sequence as before, we simulate the new output $y^{(2)}$ and compute the principal angles between the row spaces of U_f and $Y_f^{(2)}$, where $j = 10\,000$ and $i = 5, 10, \ldots, 50$. The results are shown in Figure 3.13. Note that the convergence is faster than in Figure 3.11 because the zeros of $(A, K, C^{(2)})$ lie closer to the origin than the zeros of (A, K, C). Of course, the limiting values are different, namely 0.7240, 0.6296 and 0.1024, since the model has changed. ◊

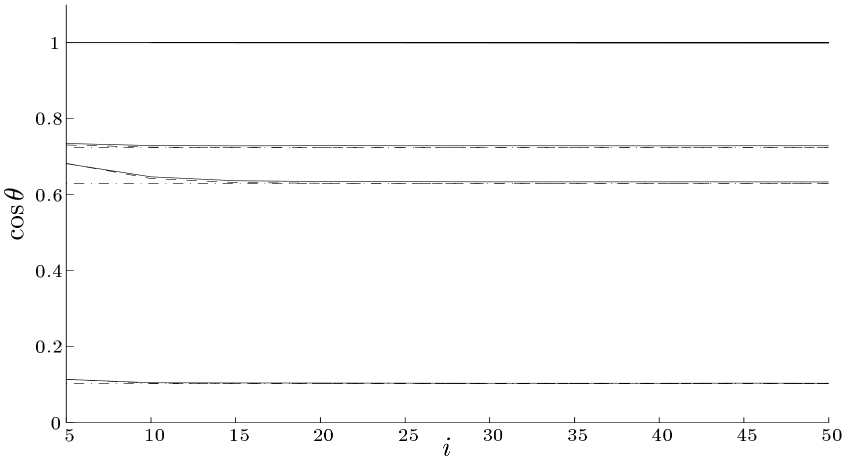


Figure 3.13: The cosines of the principal angles between the row spaces of the future input and output block Hankel matrices of the new model $(A, K, C^{(2)})$. The full lines represent the cosines of the angles between the row spaces of the $2i \times j$ data block Hankel matrices, where $j = 10\,000$ and i increases from 5 to 50 in steps of 5. The dashed lines show the cosines of the angles for the same values of i, but $j = \infty$. The cosines for $i = \infty$ and $j = \infty$ are given by the dash-dot lines. Note that the convergence is faster than in Figure 3.11 because the zeros of $(A, K, C^{(2)})$ lie closer to the origin than the zeros of (A, K, C).

2.a The principal angles between the row spaces of U_p and Y_f

Since for $j \to \infty$, $U_p \in \mathbb{R}^{mi \times j}$ and $Y_f \in \mathbb{R}^{mi \times j}$ are of full row rank, the total number of principal angles between $\operatorname{row}(U_p)$ and $\operatorname{row}(Y_f)$ is mi.

Intersection and orthogonal directions

1. The number of zero principal angles between $\operatorname{row}(U_p)$ and $\operatorname{row}(Y_f)$ is

$$t_{(U_p, Y_f)} = \operatorname{rank}(U_p) + \operatorname{rank}(Y_f) - \operatorname{rank}\begin{pmatrix} U_p \\ Y_f \end{pmatrix} .$$

The rank of U_p and Y_f is mi, and the third matrix can be written as

$$\begin{pmatrix} U_p \\ Y_f \end{pmatrix} = \begin{pmatrix} I_i & 0 & 0 \\ \Gamma_i \Delta_i & H_i & \Gamma_i A^i \end{pmatrix} \begin{pmatrix} U_p \\ U_f \\ X_p \end{pmatrix} ,$$

where

$$\begin{pmatrix} I_i & 0 & 0 \\ \Gamma_i \Delta_i & H_i & \Gamma_i A^i \end{pmatrix} \in \mathbb{R}^{2mi \times (2mi+n)} \text{ is of rank } 2mi ,$$

and

$$\begin{pmatrix} U_p \\ U_f \\ X_p \end{pmatrix} \in \mathbb{R}^{(2mi+n)\times j} \text{ has rank } 2mi+n .$$

Consequently, the rank of $\begin{pmatrix} U_p \\ Y_f \end{pmatrix}$ is $2mi$, which implies that

$$t_{(U_p,Y_f)} = 0 \text{ for } j \to \infty . \tag{3.78}$$

2. The number of principal angles equal to $\frac{\pi}{2}$ is computed as follows:

$$r_{(U_p,Y_f)} = \min(\text{rank}(U_p), \text{rank}(Y_f)) - \text{rank}(U_p Y_f^T) .$$

From the state space equations (3.27) and (3.29) and the assumptions made in Section 3.2, we know that

$$\lim_{j\to\infty} U_p Y_f^T = Q_{u_i} \Delta_i^T \Gamma_i^T ,$$

of which the rank is n. Consequently,

$$r_{(U_p,Y_f)} = mi - n \text{ for } j \to \infty . \tag{3.79}$$

From (3.78) and (3.79) it follows that for $j \to \infty$, n principal angles ω_k between $\text{row}(U_p)$ and $\text{row}(Y_f)$ satisfy

$$0 < \omega_k < \frac{\pi}{2} \qquad (k = 1, \dots, n) ,$$

and the other principal angles are all equal to $\frac{\pi}{2}$.

The factors of $(U_p U_p^T)^{-1} U_p Y_f^T (Y_f Y_f^T)^{-1} Y_f U_p^T$

From the state space equations (3.27–3.29), the assumptions in Section 3.2, the definition of Q_{u_i} in (3.31) and expression (3.33), it follows that

$$\begin{aligned} \lim_{j\to\infty} U_p U_p^T &= Q_{u_i} , \\ \lim_{j\to\infty} U_p Y_f^T &= Q_{u_i} \Delta_i^T \Gamma_i^T , \\ \lim_{j\to\infty} Y_f Y_f^T &= \Gamma_i \Sigma \Gamma_i^T + H_i Q_{u_i} H_i^T , \\ \lim_{j\to\infty} Y_f U_p^T &= \Gamma_i \Delta_i Q_{u_i} . \end{aligned}$$

Consequently, the squared cosines of the principal angles between the row spaces of U_p and Y_f are the mi eigenvalues of

$$M_{(U_p,Y_f)} = Q_{u_i}^{-1} Q_{u_i} \Delta_i^T \Gamma_i^T (\Gamma_i \Sigma \Gamma_i^T + H_i Q_{u_i} H_i^T)^{-1} \Gamma_i \Delta_i Q_{u_i} \ .$$

Computation of the eigenvalues

By substituting $-H_i^{-1}\Gamma_i$ by Γ_{z_i}, the observability matrix of the inverse model (see Equation 3.44), we obtain

$$M_{(U_p,Y_f)} = \Delta_i^T \Gamma_{z_i}^T (\Gamma_{z_i} \Sigma \Gamma_{z_i}^T + Q_{u_i})^{-1} \Gamma_{z_i} \Delta_i Q_{u_i} \ .$$

Since $M_{(U_p,Y_f)}$ is of rank n, $mi-n$ eigenvalues are equal to 0, which is in agreement with (3.68). The non-zero eigenvalues are the eigenvalues of

$$\Delta_i Q_{u_i} \Delta_i^T (\Gamma_{z_i}^T Q_{u_i}^{-1} \Gamma_{z_i} - \Gamma_{z_i}^T Q_{u_i}^{-1} \Gamma_{z_i} (\Sigma^{-1} + \Gamma_{z_i}^T Q_{u_i}^{-1} \Gamma_{z_i})^{-1} \Gamma_{z_i}^T Q_{u_i}^{-1} \Gamma_z) \ , \quad (3.80)$$

for which the matrix inversion lemma (Lemma 3.4) was used. Denote $\Delta_i Q_{u_i} \Delta_i^T$ by $\mathcal{D}_i$, which is equal to

$$\mathcal{D}_i = \sum_{k=0}^{i-1} A^k K S_u K^T A^{k^T} \quad (3.81)$$

and satisfies the Lyapunov equation

$$\mathcal{D}_i = A \mathcal{D}_i A^T + K S_u K^T - A^i K S_u K^T A^{i^T} \ . \quad (3.82)$$

Together with $\mathcal{G}_{z_i}$ as defined in (3.70), the matrix in (3.80) reduces to

$$\mathcal{D}_i \left(\mathcal{G}_{z_i} - \mathcal{G}_{z_i} (\Sigma^{-1} + \mathcal{G}_{z_i})^{-1} \mathcal{G}_{z_i} \right) = \mathcal{D}_i (\mathcal{G}_{z_i}^{-1} + \Sigma)^{-1} \ .$$

The principal angles between row(U_p) and row(Y_f)

The squared cosines of the smallest n principal angles between row(U_p) and row(Y_f) for $j \to \infty$ and finite i are the eigenvalues of $\mathcal{D}_i(\mathcal{G}_{z_i}^{-1} + \Sigma)^{-1}$. The other $mi-n$ principal angles are equal to $\frac{\pi}{2}$:

$$\lim_{j\to\infty} \cos^2 [U_p \sphericalangle Y_f] = \lambda \left(\mathcal{D}_i (\mathcal{G}_{z_i}^{-1} + \Sigma)^{-1} \right), \underbrace{0, \ldots, 0}_{mi-n} \ . \quad (3.83)$$

Example 3.6. The principal angles between row(U_p) and row(Y_f)

In Figure 3.14 the cosines of the principal angles between the row spaces of U_p and Y_f are shown in full line for the model described in Example 3.3. The block Hankel matrices have $i = 20$ block rows and an increasing number of columns: $j = 100, 200, \ldots, 10\,000$. The dashed lines are the values of the cosines for

$i = 20$ and $j = \infty$. Observe that the largest $n = 3$ cosines converge to the asymptotic values, which are the eigenvalues of $\mathcal{D}_i(\mathcal{G}_{z_i}^{-1} + \Sigma)^{-1}$. The numerical values are $\cos\omega_1 = 0.98$, $\cos\omega_2 = 0.9471$ and $\cos\omega_3 = 0.6689$. The other lines in Figure 3.14 are converging to 0. They represent the cosines of the 37 angles that are equal to $\frac{\pi}{2}$ for $j \to \infty$. $\Diamond$

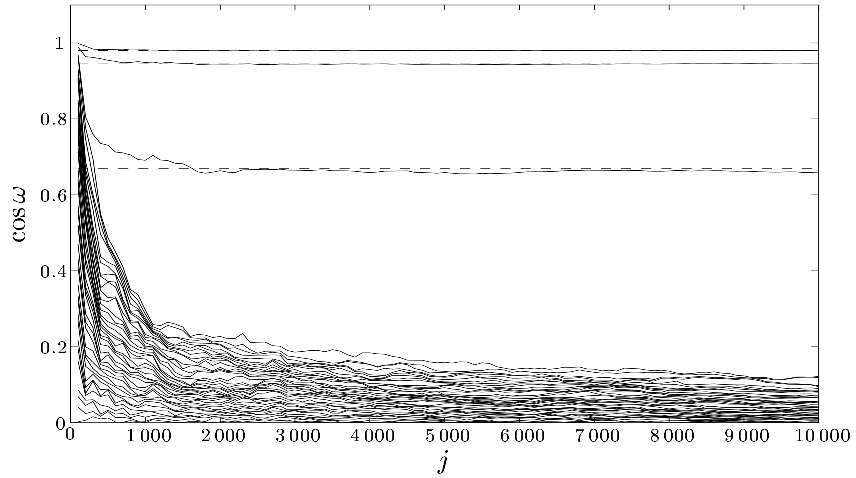


Figure 3.14: The cosines of the principal angles between the row spaces of U_p and Y_f are shown in full line. The block Hankel matrices have $i = 20$ block rows and an increasing number of columns: $j = 100, 200, \ldots, 10\,000$. The dashed lines represent the asymptotic values for $i = 20$ and $j = \infty$.

2.b The canonical correlations of the past input and the future output processes

The canonical correlations of the past input and the future output processes of the linear model are the cosines of the principal angles between $\text{row}(U_p)$ and $\text{row}(Y_f)$ for $i, j \to \infty$.
Again we take $\mathcal{G}_z = \lim_{i\to\infty} \mathcal{G}_{z_i}$. From (3.82) it follows that $\lim_{i\to\infty} \mathcal{D}_i$ satisfies the same Lyapunov equation as Σ (see Equation (3.8)). Hence,

$$\lim_{i\to\infty} \mathcal{D}_i = \Sigma \ .$$

From (3.83) it follows that there are only n non-zero canonical correlations. They are the square roots of the eigenvalues of $\Sigma(\mathcal{G}_z^{-1} + \Sigma)^{-1}$. If the state space matrices (A, K, C) and the input covariance matrix S_u are given, then the canonical correlations between the past input and the future output are computed by solving two Lyapunov equations and an eigenvalue problem. An overview of the computation is given in Scheme 3.2.

The canonical correlations of the past input and the future output processes of a linear system

Given: (A, K, C) and S_u

1. Compute the state covariance matrix by solving the Lyapunov equation:
$$\Sigma = A\Sigma A^T + KS_uK^T \ .$$
2. Compute the matrix $\mathcal{G}_z$ by solving the Lyapunov equation:
$$\mathcal{G}_z = (A - KC)^T\mathcal{G}_z(A - KC) + C^TS_u^{-1}C \ .$$
3. The largest n canonical correlations of the past input and the future output are the square roots of the eigenvalues of $\Sigma(\mathcal{G}_z^{-1} + \Sigma)^{-1}$ and the other canonical correlations are equal to 0:
$$\mathrm{cc}^2(u_p, y_f) = \lambda\left(\Sigma(\mathcal{G}_z^{-1} + \Sigma)^{-1}\right), 0, 0, \ldots \tag{3.84}$$

Scheme 3.2: Summary of the computation of the canonical correlations between the past input and the future output process of a linear model.

Example 3.7. The canonical correlations of u_p and y_f
We compute the principal angles between the row spaces of U_p and Y_f of the model in Example 3.3 for an increasing number of block rows $i = 5, 10, \ldots, 50$ and $j = 10\,000$ columns. The cosines of these angles are shown in full line in Figure 3.15. The cosines of the principal angles for $j = \infty$ are given in dashed lines, while the canonical correlations, i.e. the cosines for $i = \infty$ and $j = \infty$ are represented by the dash-dot lines. The numerical values of the canonical correlations of y_p and y_f are 0.9990, 0.9643 and 0.6825. For each increment of i there are 10 additional angles, which are asymptotically ($j \to \infty$) equal to $\frac{\pi}{2}$.
$\diamond$

3.a The principal angles between the row spaces of Y_p and Y_f

There are mi principal angles between the row spaces of $Y_p \in \mathbb{R}^{mi\times j}$ and $Y_f \in \mathbb{R}^{mi\times j}$ since the matrices are of full row rank (for $j \to \infty$).

Intersection and orthogonal directions

1. The number of zero principal angles between row(Y_p) and row(Y_f) is equal to
$$t_{(Y_p,Y_f)} = \mathrm{rank}(Y_p) + \mathrm{rank}(Y_f) - \mathrm{rank}\begin{pmatrix} Y_p \\ Y_f \end{pmatrix} \ .$$

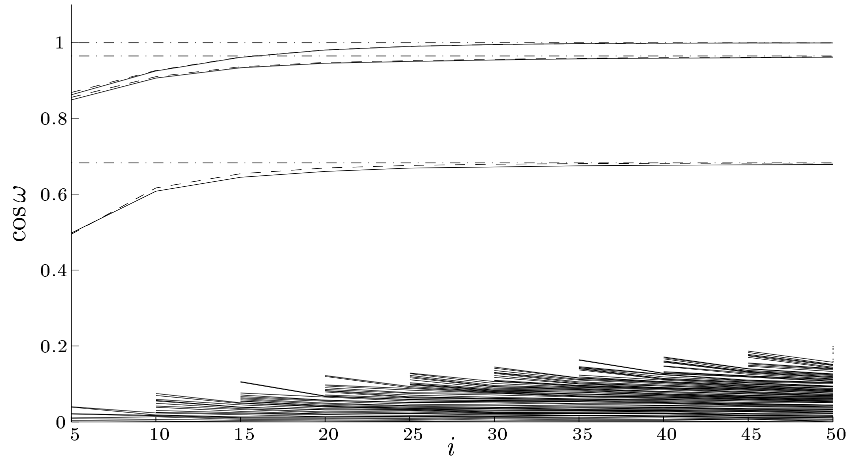


Figure 3.15: The cosines of the principal angles between the row spaces of U_p and Y_f. The full lines represent the cosines of the angles between the row spaces of the $2i \times 10\,000$ data block Hankel matrices, where i increases from 5 to 50 in steps of 5. The dashed lines show the cosines of the angles for the same values of i, but $j = \infty$. The cosines for $i = \infty$ and $j = \infty$ are given by the dash-dot lines. These values are the canonical correlations of the past input and the future output process of the model.

The rank of Y_p and Y_f is mi and the third matrix can be written as

$$\begin{pmatrix} Y_p \\ Y_f \end{pmatrix} = \begin{pmatrix} \Gamma_i & H_i & 0 \\ \Gamma_i A^i & \Gamma_i \Delta_i & H_i \end{pmatrix} \begin{pmatrix} X_p \\ U_p \\ U_f \end{pmatrix} ,$$

where

$$\begin{pmatrix} \Gamma_i & H_i & 0 \\ \Gamma_i A^i & \Gamma_i \Delta_i & H_i \end{pmatrix} \in \mathbb{R}^{2mi \times (2mi+n)} \text{ is of rank } 2mi ,$$

and

$$\begin{pmatrix} X_p \\ U_p \\ U_f \end{pmatrix} \in \mathbb{R}^{(2mi+n) \times j} \text{ has rank } 2mi + n .$$

Consequently, the rank of $\begin{pmatrix} Y_p \\ Y_f \end{pmatrix}$ is $2mi$, which implies that

$$t_{(Y_p, Y_f)} = 0 \text{ for } j \to \infty . \tag{3.85}$$

2. The number of principal angles equal to $\frac{\pi}{2}$ is computed as follows:

$$r_{(Y_p, Y_f)} = \min(\text{rank}(Y_p), \text{rank}(Y_f)) - \text{rank}(Y_p Y_f^T) .$$

From the state space equations (3.27–3.29) and the assumptions made in Section 3.2, we know that

$$\lim_{j\to\infty} Y_p Y_f^T = (\Gamma_i \Sigma A^{i^T} + H_i Q_{u_i} \Delta_i^T)\Gamma_i^T \ ,$$

of which the rank is n. Consequently,

$$r_{(Y_p,Y_f)} = mi - n \text{ for } j \to \infty \ . \tag{3.86}$$

From (3.85) and (3.86) it follows that for $j \to \infty$, there are n principal angles ϕ_k that satisfy

$$0 < \phi_k < \frac{\pi}{2} \ (k = 1, \ldots, n) \ , \tag{3.87}$$

and the other principal angles are all equal to $\frac{\pi}{2}$.

The factors of $(Y_p Y_p^T)^{-1} Y_p Y_f^T (Y_f Y_f^T)^{-1} Y_f Y_p^T$

In order to compute the principal angles between row(Y_p) and row(Y_f) we will compute the eigenvalues of the matrix

$$M_{(Y_p,Y_f)} = (Y_p Y_p^T)^{-1} Y_p Y_f^T (Y_f Y_f^T)^{-1} Y_f Y_p^T$$

(see (2.22)), where we assume that the number of columns is very large ($j \to \infty$).

The four factors are first expressed in terms of the system matrices, the state covariance matrix Σ and the input covariance matrix S_u. In the following paragraph we will then further simplify the expression of the matrix $M_{(Y_p,Y_f)}$.

From the state space equations (3.27–3.29), the assumptions in Section 3.2, the definition of Q_{u_i} in (3.31) and expression (3.33), it follows that

$$\begin{aligned}
Y_p Y_p^T &= \Gamma_i \Sigma \Gamma_i^T + H_i Q_{u_i} H_i^T \ , \\
Y_p Y_f^T &= \Gamma_i \Sigma A^{i^T} \Gamma_i^T + H_i Q_{u_i} \Delta_i^T \Gamma_i^T \ , \\
Y_f Y_f^T &= \Gamma_i \Sigma \Gamma_i^T + H_i Q_{u_i} H_i^T \ , \\
Y_f Y_p^T &= \Gamma_i A^i \Sigma \Gamma_i^T + \Gamma_i \Delta_i Q_{u_i} H_i^T \ .
\end{aligned}$$

Consequently, the squared cosines of the principal angles between the row spaces of Y_p and Y_f are the mi eigenvalues of

$$\begin{aligned}
M_{(Y_p,Y_f)} = {} & (\Gamma_i \Sigma \Gamma_i^T + H_i Q_{u_i} H_i^T)^{-1} (\Gamma_i \Sigma A^{i^T} \Gamma_i^T + H_i Q_{u_i} \Delta_i^T \Gamma_i^T) \\
& (\Gamma_i \Sigma \Gamma_i^T + H_i Q_{u_i} H_i^T)^{-1} (\Gamma_i A^i \Sigma \Gamma_i^T + \Gamma_i \Delta_i Q_{u_i} H_i^T) \ .
\end{aligned}$$

Computation of the eigenvalues
In an analogous way as the derivation of the principal angles between the row spaces of U_f and Y_f, i.e. by using $\Gamma_{z_i} = -H_i^{-1}\Gamma_i$ and $\mathcal{G}_{z_i} = \Gamma_{z_i}^T Q_{u_i}^{-1}\Gamma_{z_i}$ and applying the matrix inversion lemma (Lemma 3.4) we obtain

$$M_{(Y_p,Y_f)} = H_i^{-T}\left(Q_{u_i}^{-1} - Q_{u_i}^{-1}\Gamma_{z_i}(\Sigma^{-1}+\mathcal{G}_{z_i})^{-1}\Gamma_{z_i}^T Q_{u_i}^{-1}\right)\left(\Gamma_{z_i}\Sigma A^{i^T} - Q_{u_i}\Delta_i^T\right)$$
$$\left(\mathcal{G}_{z_i} - \mathcal{G}_{z_i}(\Sigma^{-1}+\mathcal{G}_{z_i})^{-1}\mathcal{G}_{z_i}\right)\left(A^i\Sigma\Gamma_{z_i}^T - \Delta_i Q_{u_i}\right)H_i^T \ .$$

Since M is of rank n, $mi-n$ eigenvalues are equal to 0. This is in agreement with (3.86). The non-zero eigenvalues are the eigenvalues of

$$\left(A^i\Sigma\Gamma_{z_i}^T - \Delta_i Q_{u_i}\right)\left(Q_{u_i}^{-1} - Q_{u_i}^{-1}\Gamma_{z_i}(\Sigma^{-1}+\mathcal{G}_{z_i})^{-1}\Gamma_{z_i}^T Q_{u_i}^{-1}\right)$$
$$(\Gamma_{z_i}\Sigma A^{i^T} - Q_{u_i}\Delta_i^T)(\mathcal{G}_{z_i}^{-1}+\Sigma)^{-1} \quad (3.88)$$

By using the matrices $\mathcal{G}_{z_i}$ and $\mathcal{D}_i$, defined in (3.70) and (3.82), respectively, and the following two matrices

$$T_i = (\Sigma^{-1}+\mathcal{G}_{z_i})^{-1} \ , \quad (3.89)$$
$$R_i = \Delta_i\Gamma_{z_i} = -\sum_{k=0}^{i-1} A^{i-1-k}KC(A-KC)^k \ , \quad (3.90)$$

the matrix in (3.88) can be rewritten as the following expression, in which only matrices of dimension $n\times n$ appear,

$$\Big(-A^i\Sigma R_i^T + \mathcal{D}_i - R_iT_iR_i^T + A^i\Sigma\mathcal{G}_{z_i}(\Sigma A^{i^T} + T_iR_i^T)$$
$$+\left(-R_i - A^i\Sigma\mathcal{G}_{z_i}T_i\mathcal{G}_{z_i} + R_iT_i\mathcal{G}_{z_i}\right)\Sigma A^{i^T}\Big)\left(\mathcal{G}_{z_i}^{-1}+\Sigma\right)^{-1} \ .$$

The principal angles between row($\boldsymbol{Y_p}$) and row($\boldsymbol{Y_f}$)
The squared cosines of the smallest n principal angles between row(Y_p) and row(Y_f) for $j\to\infty$ can be computed as the eigenvalues of

$$\Big(-A^i\Sigma R_i^T + \mathcal{D}_i - R_iT_iR_i^T + A^i\Sigma\mathcal{G}_{z_i}(\Sigma A^{i^T} + T_iR_i^T)$$
$$+\left(-R_i - A^i\Sigma\mathcal{G}_{z_i}T_i\mathcal{G}_{z_i} + R_iT_i\mathcal{G}_{z_i}\right)\Sigma A^{i^T}\Big)\left(\mathcal{G}_{z_i}^{-1}+\Sigma\right)^{-1} \ . \quad (3.91)$$

The other $mi-n$ angles are equal to $\frac{\pi}{2}$.

Example 3.8. The principal angles between row($\boldsymbol{Y_p}$) and row($\boldsymbol{Y_f}$)
In Figure 3.16 the cosines of the principal angles between the row spaces of Y_p and Y_f are shown in full line for the model described in Example 3.3. The block Hankel matrices have $i=20$ block rows and an increasing number of columns:

$j = 100, 200, \ldots, 10\,000$. The dashed lines are the values of the cosines for $i = 20$ and $j = \infty$. Observe that the largest $n = 3$ cosines converge to the asymptotic values: $\cos\phi_1 = 0.9937$, $\cos\phi_2 = 0.9308$ and $\cos\phi_3 = 0.6740$, which are the eigenvalues of the matrix in (3.91). The other lines in Figure 3.16 are the cosines of the 37 angles that are equal to $\frac{\pi}{2}$ for $j \to \infty$. $\Diamond$

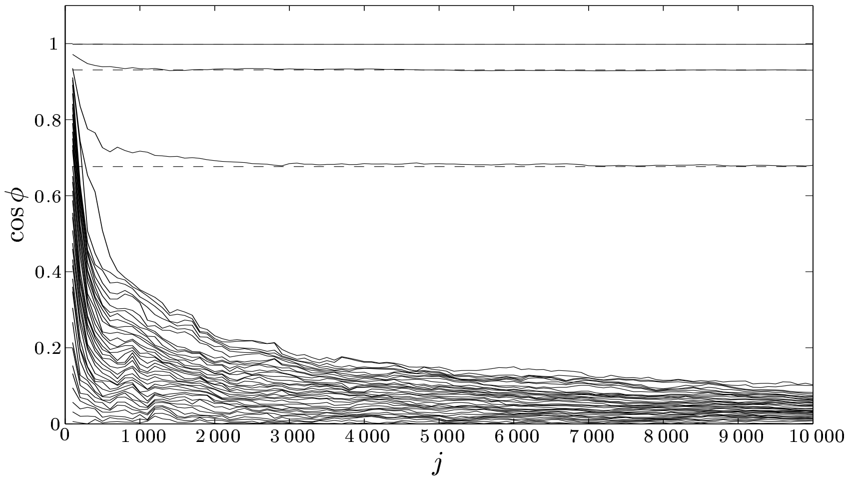


Figure 3.16: The cosines of the principal angles between the row spaces of Y_p and Y_f are shown in full line. The block Hankel matrices have $i = 20$ block rows and an increasing number of columns: $j = 100, 200, \ldots, 10\,000$. The dashed lines represent the asymptotic values for $i = 20$ and $j = \infty$.

3.b The canonical correlations of the past output and the future output processes

The canonical correlations of the past output and the future output processes of the linear model are the cosines of the principal angles between $\text{row}(Y_p)$ and $\text{row}(Y_f)$ for $i, j \to \infty$.
We know that $\mathcal{G}_z = \lim_{i\to\infty} \mathcal{G}_{z_i}$ and $\lim_{i\to\infty} \mathcal{D}_i = \Sigma$. Furthermore, we can make the following reductions in (3.91).

- Since the model is assumed to be stable

$$\lim_{i\to\infty} A^i = 0$$

- Since $R_i = -\sum_{k=0}^{i-1} A^{i-1-k} KC(A - KC)^k$ and the model is assumed to be stable ($\lim\limits_{i\to\infty} A^i = 0$) and minimum phase ($\lim\limits_{i\to\infty} (A - KC)^i = 0$), we obtain

$$\lim_{i\to\infty} R_i = 0 \ .$$

By filling this out in (3.91) we see that

The canonical correlations of the past and the future output processes of a linear stochastic system

The largest n squared canonical correlations of the past and the future output are the eigenvalues of $\Sigma(\mathcal{G}_z^{-1}+\Sigma)^{-1}$ and the other canonical correlations are equal to 0:

$$\mathrm{cc}^2(y_p, y_f) = \lambda\left(\Sigma(\mathcal{G}_z^{-1}+\Sigma)^{-1}\right), 0, 0, \ldots \tag{3.92}$$

The eigenvalues of this matrix are also the non-zero squared canonical correlations of the past input and the future output (see (3.84)). Indeed, u_p and y_p span the same space (see (3.64)), as shown in Section 3.4.1.

Example 3.9. The canonical correlations of y_p and y_f

We compute the principal angles between the row spaces of Y_p and Y_f of the model in Example 3.3 for an increasing number of block rows $i = 5, 10, \ldots, 50$ and $j = 10\,000$ columns. The cosines of these angles are shown in full line in Figure 3.17. The cosines of the principal angles for $j = \infty$ are given in dashed lines, while the canonical correlations, i.e. the cosines for $i = \infty$ and $j = \infty$ are represented by the dash-dot lines. Their values are 0.9990, 0.9643 and 0.6825, the same as for the canonical correlations of u_p and y_f. For each increment of i there are 10 additional angles, which are asymptotically ($j \to \infty$) equal to $\frac{\pi}{2}$. $\diamondsuit$

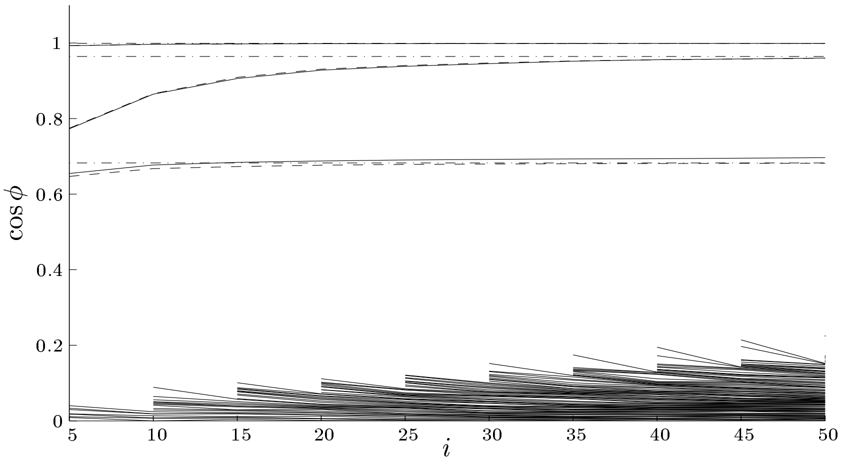


Figure 3.17: The cosines of the principal angles between the row spaces of Y_p and Y_f. The full lines represent the cosines of the angles between the row spaces of the $2i \times 10\,000$ data block Hankel matrices, where i increases from 5 to 50 in steps of 5. The dashed lines show the cosines of the angles for the same values of i, but $j = \infty$. The cosines for $i = \infty$ and $j = \infty$ are given by the dash-dot lines. These values are the canonical correlations of the past and the future output process of the model.

The canonical correlations of the past and the future output can also be characterized by means of the state covariance matrices of the forward innovation and backward innovation representations. This is shown in the following property.

Property 3.5. The canonical correlations of the past and future output process in terms of $\boldsymbol{\Sigma}$ and $\bar{\boldsymbol{\Sigma}}$ [114]
The largest n squared canonical correlations of the past and the future output of an nth order linear stochastic model are equal to the eigenvalues of $\Sigma\bar{\Sigma}$, where Σ and $\bar{\Sigma}$ are the state covariance matrices of the forward, respectively backward innovation representations. The other canonical correlations are equal to zero:

$$\mathrm{cc}^2(y_p, y_f) = \lambda\left(\Sigma\bar{\Sigma}\right), 0, 0, \ldots$$

Proof.
As shown in [159, Theorem 6 and Theorem 7] and [99], the state covariance matrix of the forward and backward innovation forms can be expressed as follows:

$$\Sigma = \lim_{i\to\infty} \bar{\Gamma}_i^T Q_{y_i}^{-1} \bar{\Gamma}_i \ , \tag{3.93}$$

$$\bar{\Sigma} = \lim_{i\to\infty} \Gamma_i^T Q_{y_i}^{-1} \Gamma_i \ , \tag{3.94}$$

where

$$Q_{y_i} = \lim_{j\to\infty} Y_p Y_p^T = \lim_{j\to\infty} Y_f Y_f^T \ ,$$

and the matrices Γ_i and $\bar{\Gamma}_i$ are the forward observability matrix and the reversed backward observability matrix, as defined in (3.21) and (3.41), respectively.
By noting that

$$\begin{aligned}
\Gamma_i \bar{\Gamma}_i^T &= \begin{pmatrix} CA^{i-1}G & CA^{i-2}G & \cdots & CG \\ CA^{i}G & CA^{i-1}G & \cdots & CAG \\ \vdots & \vdots & \ddots & \vdots \\ CA^{2i-2}G & CA^{2i-3}G & \cdots & CA^{i-1}G \end{pmatrix} \\
&= \begin{pmatrix} R_y(i) & R_y(i-1) & \cdots & R_y(1) \\ R_y(i+1) & R_y(i) & \cdots & R_y(2) \\ \vdots & \vdots & \ddots & \vdots \\ R_y(2i-1) & R_y(2i-2) & \cdots & R_y(i) \end{pmatrix} \qquad \text{by (3.11)},
\end{aligned}$$

we see that

$$\Gamma_i \bar{\Gamma}_i^T = \lim_{j\to\infty} Y_f Y_p^T \ .$$

Further, we know from (3.93–3.94) that

$$\lambda(\Sigma\bar{\Sigma}) = \lim_{i\to\infty} \lambda(\bar{\Gamma}_i^T Q_{y_i}^{-1} \bar{\Gamma}_i \Gamma_i^T Q_{y_i}^{-1} \Gamma_i) \ ,$$

and the eigenvalues of the right hand side matrix are equal to the eigenvalues of $Q_{y_i}^{-1}\bar{\Gamma}_i\Gamma_i^T Q_{y_i}^{-1}\Gamma_i\bar{\Gamma}_i^T = \lim_{j\to\infty}(Y_pY_p^T)^{-1}Y_pY_f^T(Y_fY_f^T)^{-1}Y_fY_p^T$, except for $i-n$ zero eigenvalues. Consequently, for $i\to\infty$, the largest n eigenvalues of

$$\lim_{j\to\infty}(Y_pY_p^T)^{-1}Y_pY_f^T(Y_fY_f^T)^{-1}Y_fY_p^T \ ,$$

which are the largest n squared canonical correlations of y_p and y_f, are equal to the eigenvalues of $\Sigma\bar{\Sigma}$ and the other canonical correlations are equal to zero. □

It follows from (3.92) and Property 3.5 that the eigenvalues of $\Sigma\bar{\Sigma}$ are equal to the eigenvalues of $\Sigma(\mathcal{G}_z^{-1}+\Sigma)^{-1}$, which reveals a close relation of $\bar{\Sigma}$ and $(\mathcal{G}_z^{-1}+\Sigma)^{-1}$. Indeed, these matrices turn out to be equal, which is formulated in the property below. For the continuous case, the proof of the equality has been given in [106] and for the discrete case in [128, Lemma 1], where it is used to show that minimum phase and stochastically balanced realizations are related to each other by a diagonal transformation.

Property 3.6. The backward state covariance matrix [128]
For a linear stochastic model with forward innovation representation (A,K,C) and input covariance matrix S_u, the backward innovation state covariance matrix is equal to

$$\bar{\Sigma} = (\mathcal{G}_z^{-1}+\Sigma)^{-1} \ , \tag{3.95}$$

where Σ is the forward innovation state covariance matrix and $\mathcal{G}_z$ satisfies the Lyapunov equation

$$\mathcal{G}_z = (A-KC)^T\mathcal{G}_z(A-KC) + C^TS_u^{-1}C \ .$$

Proof.
Recall from (3.75) that

$$\mathcal{G}_z = \lim_{i\to\infty}\Gamma_{z_i}^T Q_{u_i}^{-1}\Gamma_{z_i} \ ,$$

where

$$Q_{u_i} = \lim_{j\to\infty}U_pU_p^T = \lim_{j\to\infty}U_fU_f^T = \mathrm{diag}(\underbrace{S_u,\ldots,S_u}_{i\ \text{blocks}}) \ .$$

The observability matrix of the inverse model, Γ_{z_i}, is equal to $-H_i^{-1}\Gamma_i$ (see Equation (3.44)), and thus we have

$$\mathcal{G}_z = \lim_{i\to\infty}\Gamma_i^T(H_iQ_{u_i}H_i^T)^{-1}\Gamma_i \ .$$

Since

$$Q_{y_i} = \lim_{j\to\infty}Y_pY_p^T = \lim_{j\to\infty}Y_fY_f^T = \Gamma_i\Sigma\Gamma_i^T + H_iQ_{u_i}H_i^T \ ,$$

we can compute $(H_iQ_{u_i}H_i^T)^{-1}$ as $(Q_{y_i}-\Gamma_i\Sigma\Gamma_i^T)^{-1}$. By the matrix inversion lemma, $\mathcal{G}_z$ is then equal to

$$\mathcal{G}_z = \lim_{i\to\infty}\left(\Gamma_i^TQ_{y_i}^{-1}\Gamma_i - \Gamma_i^TQ_{y_i}^{-1}\Gamma_i(-\Sigma^{-1}+\Gamma_i^TQ_{y_i}^{-1}\Gamma_i)^{-1}\Gamma_i^TQ_{y_i}^{-1}\Gamma_i\right) \ . \tag{3.96}$$

Filling out (3.94) in (3.96), we obtain

$$\mathcal{G}_z = \bar{\Sigma} - \bar{\Sigma}(-\Sigma^{-1} + \bar{\Sigma})^{-1}\bar{\Sigma} \ .$$

Applying the inversion lemma once again leads to

$$\mathcal{G}_z = (\bar{\Sigma}^{-1} - \Sigma)^{-1} \ ,$$

from which (3.95) straightforwardly follows. □

Consequently, Property 3.5 could also be proven starting from (3.92) and filling out (3.95).
In Scheme 3.3 an overview is given of the computation of the canonical correlations of y_p and y_f via the forward and backward state covariance matrices.

The canonical correlations of the past and future output processes of a linear system

Given: (A, K, C) and S_u

1. Compute the forward state covariance matrix Σ by solving the Lyapunov equation:
$$\Sigma = A\Sigma A^T + KS_uK^T \ .$$
2. Compute the backward state covariance matrix $\bar{\Sigma}$ by solving the Riccati equation
$$\bar{\Sigma} = A^T\bar{\Sigma}A + (C^T - A^T\bar{\Sigma}G)(C\Sigma C^T + S_u - G^T\bar{\Sigma}G)^{-1}(C - G^T\bar{\Sigma}A) \ ,$$
where $G = A\Sigma C^T + KS_u$.
3. The largest n canonical correlations of y_p and y_f are the square roots of the eigenvalues of $\Sigma\bar{\Sigma}$ and the other canonical correlations are equal to 0:
$$\mathrm{cc}^2(y_p, y_f) = \lambda(\Sigma\bar{\Sigma}), 0, 0, \ldots$$

Scheme 3.3: Overview of the computation of the canonical correlations of the past and the future output process of an nth order linear model via the forward and backward state covariance matrices.

Remark 3.11. $\Sigma \neq \bar{\Sigma}^{-1}$
As we already noted in Section 3.2.1, the state covariance matrix of the forward innovation form is not equal to the inverse state covariance matrix of the backward innovation form. From Property 3.6 we see that $\mathcal{G}_z^{-1}$ can be viewed as a sort of correction term:

$$\bar{\Sigma}^{-1} = \Sigma + \mathcal{G}_z^{-1} \ .$$

forward form (A, C, S_w, S_{wv}, S_v) Σ_f	Kalman filter $\longrightarrow$	forward innovation form (A, K, C, S_u) Σ
$\downarrow$		
backward form $(A^T, G, S_{\bar{w}}, S_{\bar{w}\bar{v}}, S_{\bar{v}})$ $\bar{\Sigma}_b^{-1} = \Sigma_f$	$\longrightarrow$ backward Kalman filter	backward innovation form $(A^T, \bar{K}, G, S_{\bar{u}})$ $\bar{\Sigma}^{-1} = \Sigma + \mathcal{G}_z^{-1}$

Table 3.4: Schematic overview of the different representation forms, as in Table 3.1. We have now replaced the inequality $\bar{\Sigma}^{-1} \neq \Sigma$ by $\bar{\Sigma}^{-1} = \Sigma + \mathcal{G}_z^{-1}$.

Since for the models that we consider, i.e. stable and minimum phase models, the elements of $\mathcal{G}_z$ are always finite, the correction term $\mathcal{G}_z^{-1}$ cannot become equal to 0. In Table 3.4 an updated overview of the different representation forms is given. ○

Remark 3.12. A note on the computation of the canonical correlations
In the following discussion we take the canonical correlations of the future input and the future output as an example. The canonical correlations of the other pairs of processes can be treated similarly.
Recall that the canonical correlations of u_f and y_f are defined as the canonical correlations of the random variables $U_{0|\infty}$ and $Y_{0|\infty}$. We have computed the canonical correlations in the way explained in (3.66a): first, we compute the principal angles between the row spaces of the $mi \times \infty$ block Hankel matrices U_f and Y_f and then we let $i \to \infty$. In fact, this is a special way to obtain the canonical correlations of the infinite random variables $U_{0|\infty}$ and $Y_{0|\infty}$, since in the limiting process, the matrices U_f and Y_f have always the same number of rows. We could also look at what happens if this is not the case.
It turns out that it does not matter whether we take an equal number of rows or not for the three cases considered in Section 3.4.2 (see below). However, the computation of the canonical correlations of the past input and the past output process *is* sensitive to the way of taking limits. The following results are obtained in an analogous way as the computations above.

1. **Future input and future output**
   We consider the future input random variable $U_{0|k-1}$ and the future output random variable $Y_{0|i-1}$. The squared canonical correlations of these random variables are equal to

$$\mathrm{cc}^2(U_{0|k-1}, Y_{0|i-1}) = \begin{cases} \lambda\left(I_n - \mathcal{G}_{z_i}(\Sigma^{-1} + \mathcal{G}_{z_i})^{-1}\right), \underbrace{1, \ldots, 1}_{i-n} & k \geq i \geq n, \\ \lambda\left(I_n - \mathcal{G}_{z_k}(\Sigma^{-1} + \mathcal{G}_{z_i})^{-1}\right), \underbrace{1, \ldots, 1}_{k-n} & i \geq k \geq n, \end{cases}$$

where $\mathcal{G}_{z_i}$ is defined in (3.70), $\mathcal{G}_{z_k}$ is defined similarly and n is the model order. Of course, both $\mathcal{G}_{z_i}$ as $\mathcal{G}_{z_k}$ converge to $\mathcal{G}_z$ for $i \to \infty$, respectively $k \to \infty$ (see (3.76)). Consequently, we can first let k grow and then i or the other way around, the result is the same:

$$\begin{aligned}\lim_{i\to\infty}\lim_{k\to\infty} \mathrm{cc}^2(U_{0|k-1}, Y_{0|i-1}) &= \lim_{k\to\infty}\lim_{i\to\infty} \mathrm{cc}^2(U_{0|k-1}, Y_{0|i-1}) \ ,\\ &= \lambda\left((I_n + \mathcal{G}_z\Sigma)^{-1}\right), 1, 1, \dots \ ,\end{aligned}$$

and this is also equal to $\lim\limits_{i\to\infty}\lim\limits_{j\to\infty} \cos^2[U_f \sphericalangle Y_f]$ (see (3.77)).

2. **Past input and future output**
The largest n squared canonical correlations of $U_{-k|-1}$ and $Y_{0|i-1}$ are the eigenvalues of $\mathcal{D}_k(\mathcal{G}_{z_i}^{-1} + \Sigma)^{-1}$, irrespective of whether $i < k$, or $i \geq k$. The other canonical correlations are equal to 0. The matrix $\mathcal{D}_k$ is defined in (3.81) and $\lim_{k\to\infty} \mathcal{D}_k = \Sigma$. Consequently,

$$\begin{aligned}\lim_{i\to\infty}\lim_{k\to\infty} \mathrm{cc}^2(U_{-k|-1}, Y_{0|i-1}) &= \lim_{k\to\infty}\lim_{i\to\infty} \mathrm{cc}^2(U_{-k|-1}, Y_{0|i-1}) \ ,\\ &= \lambda\left(\Sigma(\mathcal{G}_z^{-1} + \Sigma)^{-1}\right), 0, 0, \dots \ ,\end{aligned}$$

which is equal to $\lim\limits_{i\to\infty}\lim\limits_{j\to\infty} \cos^2[U_p \sphericalangle Y_f]$ (see 3.84).

3. **Past and future output**
The largest n squared canonical correlations of $Y_{-k|-1}$ and $Y_{0|i-1}$ are the eigenvalues of

$$\begin{aligned}&\Big(-A^k\Sigma R_k^T + \mathcal{D}_k - R_k T_k R_k^T + A^k\Sigma\mathcal{G}_{z_k}(\Sigma A^{k^T} + T_k R_k^T)\\ &\quad + \left(-R_k - A^k\Sigma\mathcal{G}_{z_k}T_k\mathcal{G}_{z_k} + R_k T_k \mathcal{G}_{z_k}\right)\Sigma A^{k^T}\Big)\left(\mathcal{G}_{z_i}^{-1} + \Sigma\right)^{-1} \ ,\end{aligned} \tag{3.97}$$

and the other canonical correlations are equal to 0. The definitions of the matrices T_k and R_k can be found in (3.89) and (3.90), respectively. Since the expression in (3.97) holds for $i < k$ and $i \geq k$, we again see that

$$\begin{aligned}\lim_{i\to\infty}\lim_{k\to\infty} \mathrm{cc}(Y_{-k|-1}, Y_{0|i-1}) &= \lim_{k\to\infty}\lim_{i\to\infty} \mathrm{cc}(Y_{-k|-1}, Y_{0|i-1})\\ &= \lim_{i\to\infty}\lim_{j\to\infty} \cos[U_f \sphericalangle Y_f] \ .\end{aligned}$$

4. **Past input and past output**
The smallest n squared canonical correlations of the random variables $U_{-k|-1}$ and $Y_{-i|-1}$ are the eigenvalues of the following matrices, depending on whether $i > k$, $i = k$ or $i < k$, respectively:

$$\begin{cases} M_1 = I_n - (A-KC)^{(i-k)^T}\mathcal{G}_{z_k}(A-KC)^{(i-k)}T_i & i > k \geq n \ ,\\ M_2 = I_n - \mathcal{G}_{z_i}T_i & i = k \geq n \ ,\\ M_3 = I_n - \mathcal{G}_{z_i}\left((T_i\mathcal{G}_{z_i} - I_n)\mathcal{D}_{k-i} + T_i\right) & k > i \geq n \ .\end{cases}$$

The other canonical correlations are equal to 1. Note that the matrix M_1 converges to I_n for $k \to \infty$ and M_3 converges to I_n for $i \to \infty$. Consequently,

$$\lim_{i\to\infty}\lim_{k\to\infty} \mathrm{cc}(U_{-k|-1}, Y_{-i|-1}) = \lim_{k\to\infty}\lim_{i\to\infty} \mathrm{cc}(U_{-k|-1}, Y_{-i|-1}) = 1, 1, \dots ,$$

which is in agreement with (3.62). However, if $U_{-k|-1}$ and $Y_{-i|-1}$ grow in the same way (i.e. $k = i$), then there are n canonical correlations different from 1:

$$\lim_{i\to\infty} \mathrm{cc}^2(U_{-i|-1}, Y_{-i|-1}) = \lambda\left((I_n + \mathcal{G}_z\Sigma)^{-1}\right), 1, 1, \dots$$

These are the canonical correlations of the future input and future output. Indeed, due to stationarity, we always have

$$\mathrm{cc}(U_{-i|-1}, Y_{-i|-1}) = \mathrm{cc}(U_{0|i-1}, Y_{0|i-1}) .$$

The discrepancy can be understood as follows: the components of $Y_{-i|-1}$ are linear combinations of the components of $U_{-\infty|-1}$ and the other way round: $U_{-k|-1} \subset Y_{-\infty|-1}$. However, $Y_{-i|-1}$ contains information about the past that is not contained in $U_{-i|-1}$, namely the state $X(-i)$. This remains valid, no matter how large i is.

Summarizing, the canonical correlations of u_f and y_f, of u_p and y_f and of y_p and y_f can be obtained as the following limits

$$\begin{aligned}
\mathrm{cc}(u_f, y_f) &= \lim_{i\to\infty} \mathrm{cc}(U_{0|i-1}, Y_{0|i-1}) = \lim_{i\to\infty}\lim_{j\to\infty} \cos\left[U_f \sphericalangle Y_f\right] , \\
\mathrm{cc}(u_p, y_f) &= \lim_{i\to\infty} \mathrm{cc}(U_{-i|-1}, Y_{0|i-1}) = \lim_{i\to\infty}\lim_{j\to\infty} \cos\left[U_p \sphericalangle Y_f\right] , \\
\mathrm{cc}(y_p, y_f) &= \lim_{i\to\infty} \mathrm{cc}(Y_{-i|-1}, Y_{0|i-1}) = \lim_{i\to\infty}\lim_{j\to\infty} \cos\left[Y_p \sphericalangle Y_f\right] ,
\end{aligned}$$

but this computation method cannot be applied to the past input and the past output processes:

$$\mathrm{cc}(u_p, y_p) \neq \lim_{i\to\infty} \mathrm{cc}(U_{-i|0}, Y_{-i|0}) = \lim_{i\to\infty}\lim_{j\to\infty} \cos\left[U_p \sphericalangle Y_p\right] .$$

○

4. Summary of the expressions for the canonical correlations of the different process pairs

The canonical correlations of u_f and y_f

The smallest n canonical correlations of u_f and y_f are the square roots of the eigenvalues of

$$(I_n + \mathcal{G}_z\Sigma)^{-1} , \tag{3.98}$$

where Σ is the state covariance matrix, which can be found by solving the Lyapunov equation (3.8) and $\mathcal{G}_z$ satisfies the Lyapunov equation

$$\mathcal{G}_z = (A - KC)^T \mathcal{G}_z (A - KC) + C^T S_u^{-1} C \ . \tag{3.99}$$

The other canonical correlations are equal to 1.

The canonical correlations of y_p and y_f / of u_p and y_f

The largest n canonical correlations are the square roots of the eigenvalues of

$$\Sigma(\mathcal{G}_z^{-1} + \Sigma)^{-1} = \Sigma\bar{\Sigma} \ , \tag{3.100}$$

where Σ and $\mathcal{G}_z$ are the solutions of (3.8) and (3.99), respectively and $\bar{\Sigma}$ is the solution of the Riccati equation (3.20). The other canonical correlations are equal to 0.

Remark 3.13. Influence of a similarity transformation on the canonical correlations.

Since the canonical correlations of the input and output processes are characteristics of the processes, it is obvious that they should be independent of the state space basis of the model. We will briefly show that this can be deduced from their parametric description. As is shown in Table 3.5, the influence of a similarity transformation T of the model on the state covariance matrices Σ and $\bar{\Sigma}$ and on the matrix $\mathcal{G}_z$ is a contragredient transformation: Σ is transformed to $T\Sigma T^T$, the matrix $\bar{\Sigma}$ to $T^{-T}\bar{\Sigma}T^{-1}$ and $\mathcal{G}_z$ to $T^{-T}\mathcal{G}_z T^{-1}$. Consequently, the product $\Sigma\bar{\Sigma}$ is transformed to $T\Sigma\bar{\Sigma}T^{-1}$ and hence the eigenvalues of $\Sigma\bar{\Sigma}$ are not influenced. The same holds true for the eigenvalues of $\Sigma\left(\mathcal{G}_z^{-1} + \Sigma\right)^{-1}$. Consequently, the canonical correlations of y_p (or u_p) and y_f and the canonical correlations of u_f and y_f are independent of the state space basis. ○

(A, K, C)	$\longrightarrow$	(TAT^{-1}, TK, CT^{-1})
Σ	$\longrightarrow$	$T\Sigma T^T$
$\bar{\Sigma}$	$\longrightarrow$	$T^{-T}\bar{\Sigma}T^{-1}$
$\mathcal{G}_z$	$\longrightarrow$	$T^{-T}\mathcal{G}_z T^{-1}$
$\Sigma\bar{\Sigma}$	$\longrightarrow$	$T\Sigma\bar{\Sigma}T^{-1}$
$\Sigma\left(\mathcal{G}_z^{-1} + \Sigma\right)^{-1}$	$\longrightarrow$	$T\Sigma\left(\mathcal{G}_z^{-1} + \Sigma\right)^{-1} T^{-1}$

Table 3.5: The influence of a similarity transformation on Σ, $\bar{\Sigma}$ and $\mathcal{G}_z$: a contragredient transformation. The second part of the table shows that the canonical correlations of y_p (or u_p) and y_f, which are the eigenvalues of $\Sigma\bar{\Sigma}$, nor the canonical correlations of u_f and y_f, the eigenvalues of $\Sigma\left(\mathcal{G}_z^{-1} + \Sigma\right)^{-1}$, are changed by the similarity transformation T.

3.4.3 Relation of the canonical correlations between the different processes

The canonical correlations of the different pairs of processes (or the principal angles between the pairs of subspaces) are closely related, as we have already indicated in Table 3.2. Here, we show that the canonical correlations of the future input and output are complementary[7] to the canonical correlations of the past and future output (or past input and future output). The relation is straightforwardly proven by means of the matrices given in Section 3.4.

Property 3.7. Complementarity of cc(u_f, y_f) and cc(y_p, y_f)
The canonical correlations of u_f and y_f are complementary to the canonical correlations of y_p and y_f (u_p and y_f).

Proof.
The smallest n squared canonical correlations of u_f and y_f are the eigenvalues of $(I_n + \mathcal{G}_z\Sigma)^{-1} = I_n - (\mathcal{G}_z^{-1} + \Sigma)^{-1}\Sigma$ and the other canonical correlations are equal to 1. The eigenvalues of $I_n - (\mathcal{G}_z^{-1} + \Sigma)^{-1}\Sigma$ are equal to the eigenvalues of $I_n - \Sigma(\mathcal{G}_z^{-1} + \Sigma)^{-1}$. Since the eigenvalues of $\Sigma(\mathcal{G}_z^{-1} + \Sigma)^{-1}$ are the largest n squared canonical correlations of y_p and y_f (u_p and y_f) and the other canonical correlations are equal to 0, we have proven that the canonical correlations of u_f and y_f are complementary to the canonical correlations of y_p and y_f (u_p and y_f). □

Remark 3.14. Minimum phase and stochastic balancing
As we have seen in Remark 3.13, the canonical correlations are not influenced by a similarity transformation of the model. We can consequently choose a special state space basis so that the relations between the different canonical correlations are very clearly revealed. This is the case when considering balanced state space representations, which are very useful in certain model reduction algorithms. We will briefly discuss two balancing methods for stochastic models and their relation: minimum phase balancing and stochastic balancing. For more information on the methods, their properties and applications, we refer to [43, 106, 128].

- Minimum phase balancing: A state space model is in minimum phase balanced form if the forward innovation state covariance matrix Σ and the matrix $\mathcal{G}_z$ are equal and diagonal. We can always bring a state space model in this form by choosing the right similarity transformation. This scheme has been introduced in [106].
  Assume that the state space model (A, K, C) is in minimum phase balanced form and let Σ and $\mathcal{G}_z$ be equal to $\text{diag}(\sigma_1, \dots, \sigma_n)$, where $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ are called the minimum phase singular values of the system.

[7] Two canonical correlations ρ_1 and ρ_2 are complementary if $\rho_1^2 = 1 - \rho_2^2$. For the corresponding principal angles, θ_1 and θ_2 holds: $\theta_1 = \frac{\pi}{2} - \theta_2$.

From (3.98) we then obtain the canonical correlations of the future input and future output:

$$\mathrm{cc}^2(u_f, y_f) = \frac{1}{1+\sigma_1^2}, \frac{1}{1+\sigma_2^2}, \dots, \frac{1}{1+\sigma_n^2}, 1, 1, \dots\ , \tag{3.101}$$

while the squared canonical correlations of the past and future output follow from the left hand side in Equation (3.100):

$$\mathrm{cc}^2(y_p, y_f) = \frac{\sigma_1^2}{1+\sigma_1^2}, \frac{\sigma_2^2}{1+\sigma_2^2} \dots, \frac{\sigma_n^2}{1+\sigma_n^2}, 0, 0, \dots \tag{3.102}$$

From (3.101) and (3.102) it immediately follows that the canonical correlations of u_f and y_f are complementary to the canonical correlations of y_p and y_f, since

$$1 - \frac{1}{1+\sigma_k^2} = \frac{\sigma_k^2}{1+\sigma_k^2}\ .$$

We can now interpret the elements $\sigma_1, \dots, \sigma_n$ of $\mathcal{G}_z$ and Σ in geometrical terms. They are the tangents of the largest n principal angles between $\mathrm{row}(U_f)$ and $\mathrm{row}(Y_f)$, where $i, j \to \infty$. Also note that the backward state covariance matrix of the minimum phase balanced model is equal to

$$\bar{\Sigma} = \mathrm{diag}\left(\frac{\sigma_1}{1+\sigma_1^2}, \frac{\sigma_2}{1+\sigma_2^2}, \dots, \frac{\sigma_n}{1+\sigma_n^2}\right)\ .$$

- Stochastic balancing: A state space model is in stochastically balanced form if the forward innovation state covariance matrix Σ and the backward innovation state covariance matrix $\bar{\Sigma}$ are equal and diagonal. This scheme has been proposed in [43].
  Assume that the state space model (A, K, C) is in stochastically balanced form and let Σ and $\bar{\Sigma}$ be equal to $\mathrm{diag}(\rho_1, \dots, \rho_n)$, where $1 \geq \rho_1 \geq \dots \geq \rho_n \geq 0$. From the right hand side in (3.100) it follows that they are the largest n canonical correlations of the past and future output:

$$\mathrm{cc}^2(y_p, y_f) = \rho_1^2, \rho_2^2, \dots, \rho_n^2, 0, 0, \dots\ , \tag{3.103}$$

  while by using (3.95) and (3.98) we obtain

$$\mathrm{cc}^2(u_f, y_f) = 1-\rho_1^2, 1-\rho_2^2, \dots, 1-\rho_n^2, 1, 1, \dots\ , \tag{3.104}$$

  from which the complementarity even more clearly follows.
  The matrix $\mathcal{G}_z$ of a stochastically balanced model is equal to

$$\mathcal{G}_z = \mathrm{diag}\left(\frac{\rho_1}{1-\rho_1^2}, \frac{\rho_2}{1-\rho_2^2}, \dots, \frac{\rho_n}{1-\rho_n^2}\right)\ .$$

By equating the expressions for the canonical correlations of the past and future output in (3.102) and (3.103), or alternatively, the expressions for $\mathrm{cc}(u_f, y_f)$ in (3.101) and (3.104), which can be done since the canonical correlations are independent of the state space basis (see Remark 3.13), we find the relation of the minimum phase singular values σ_k and the canonical correlations ρ_k $(i = 1, \ldots, n)$:

$$\rho_k^2 = \frac{\sigma_k^2}{1 + \sigma_k^2} \ .$$

Furthermore, since the backward state covariance matrix $\bar{\Sigma}$ of the minimum phase balanced model is diagonal, the minimum phase and stochastically balanced realizations are related to each other by a diagonal similarity transformation. It is then easy to derive the n diagonal elements $\lambda_1, \ldots, \lambda_n$ of the transformation matrix that transforms the minimum phase balanced form to the stochastically balanced form:

$$\lambda_k^2 = \frac{\rho_k}{\sigma_k} = \sin \phi_k \ ,$$

where ϕ_k is the kth principal angle between $\mathrm{row}(Y_p)$ and $\mathrm{row}(Y_f)$ for $i, j \to \infty$. ○

3.5 The principal angles between the past and future input and output spaces of single input single output (SISO) models

For models that satisfy the description and assumptions in Section 3.2.1, but with only one input and one output $(m = 1)$, the results of Section 3.4 can be simplified. This is done in Section 3.5.1. In Section 3.5.2 we consider two SISO models that are driven by the same input process and we compute the canonical correlations of their future output processes. Based on this result, we prove in Section 3.5.3 a new characterization of the canonical correlations of the future input and future output processes of a SISO ARMA model. And we end this Chapter with the canonical correlations of the past and future of an AR process (Section 3.5.4).

3.5.1 Simplifications of the results obtained for MIMO models

The expressions for the canonical correlations that were derived in Section 3.4.2, all implicitly depend on the input covariance matrix S_u. For example, the largest n canonical correlations of the past and future output of an nth order model are equal to the square roots of the eigenvalues of $\Sigma\left(\mathcal{G}_z^{-1} + \Sigma\right)^{-1}$ and the

other canonical correlations are equal to 0. Since Σ and $\mathcal{G}_z$ are the solutions of $\Sigma = A\Sigma A^T + KS_uK^T$ and $\mathcal{G}_z = (A-KC)^T\mathcal{G}_z(A-KC) + C^TS_u^{-1}C$, respectively, the canonical correlations will in general depend on S_u.
For SISO models, however, the canonical correlations are independent of the input variance. Since for SISO models the input covariance matrix S_u reduces to the variance $\sigma^2 = E\{u(k)u(k)^T\}$, the state covariance matrix Σ is equal to $\sigma^2 P$ (see Remark 3.4), where the matrix P is the controllability Gramian of the model. Furthermore, for SISO models the matrix $\mathcal{G}_z = \Gamma_{z\infty}^T S_u^{-1}\Gamma_{z\infty}$ can be written as

$$\mathcal{G}_z = \frac{1}{\sigma^2}\Gamma_{z\infty}^T\Gamma_{z\infty} = \frac{1}{\sigma^2}Q_z \ ,$$

where Q_z is the observability Gramian of the inverse model (see (3.45)). By substituting $\Sigma = \sigma^2 P$ and $\mathcal{G}_z = \frac{1}{\sigma^2}Q_z$ in the matrices of Scheme 3.1 and 3.2 and Equation (3.92), we see that the input variance σ^2 vanishes from the expressions for the canonical correlations of the different processes. The results for SISO models can therefore be summarized as follows.

- The smallest n squared canonical correlations of the future input and future output are the eigenvalues of $(I_n + Q_zP)^{-1}$ and the other canonical correlations are equal to 1:

$$\mathrm{cc}^2(u_f, y_f) = \lambda\left((I_n + Q_zP)^{-1}\right), 1, 1, \ldots \qquad (3.105)$$

- The largest n squared canonical correlations of the past input and the future output and the largest n squared canonical correlations of the past and future output are the eigenvalues of $P(Q_z^{-1} + P)^{-1}$ and the other canonical correlations are equal to 0:

$$\mathrm{cc}^2(u_p, y_f) = \mathrm{cc}^2(y_p, y_f) = \lambda\left(P(Q_z^{-1} + P)^{-1}\right), 0, 0, \ldots \qquad (3.106)$$

Remark 3.15. MIMO models with independent input processes with the same variance
For models with $m > 1$ independent input processes that have the same variance, i.e., models for which the input covariance matrix S_u has the following form:

$$S_u = \mathrm{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2) \in \mathbb{R}^{m\times m} \ , \qquad (3.107)$$

the matrices Σ and $\mathcal{G}_z$ also reduce to $\sigma^2 P$ and $\frac{1}{\sigma^2}Q_z$, respectively. Consequently, the expressions for the canonical correlations in (3.105) and (3.106) also hold for these MIMO models. ○

Remark 3.16. The principal angles between the row spaces of the Hankel matrices
From the expressions in Section 3.4.2 it follows that for SISO models (and MIMO models with input covariance as in (3.107)), the principal angles between the row spaces of the corresponding Hankel matrices are independent of σ^2, even for finite i. ○

3.5.2 The canonical correlations of the future output processes of two ARMA models that are driven by the same input

The main objective of this section is the computation of the canonical correlations between the future output processes of single-input single-output autoregressive moving average (ARMA) models driven by the same input. Therefore, we compute the principal angles between the corresponding output Hankel matrices. Since there is an exact intersection between these spaces, a number of angles are zero (the corresponding canonical correlations are equal to one). We show that the other principal angles are equal to the principal angles between subspaces that can easily be derived from the model parameters.
Before we can treat ARMA processes, we first study the simpler case of two nth order AR processes that are driven by the same input. Two different input processes are treated: white noise and AR processes. Using the results obtained for the AR processes, we can compute the canonical correlations of the future of two ARMA processes of order $n^{(1)}$ and $n^{(2)}$, respectively, generated by the same input process.

The canonical correlations of AR processes

We compute the canonical correlations between the future outputs of two nth order AR models that are driven by the same input. We show that the smallest n canonical correlations are equal to the cosines of the principal angles between the row spaces of the infinite controllability matrices of the AR models and all other canonical correlations are equal to one. First, the input is assumed to be white noise and in the next paragraph we consider more general input sequences.

The input is white noise
We compute the principal angles between the row spaces of the $i \times j$ output

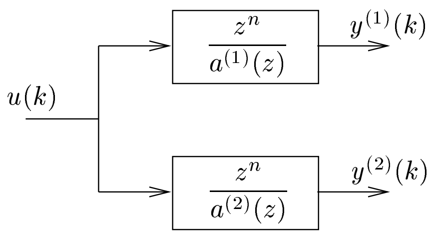


Figure 3.18: Setup for the computation of the canonical correlations between the future output of two AR models driven by the same white noise process $\{u(k)\}$.

Hankel matrices $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ of the nth order AR processes $\{y^{(1)}(k)\}$ and $\{y^{(2)}(k)\}$, respectively, when the input of the corresponding AR models with transfer function $H^{(1)}(z) = \frac{z^n}{a^{(1)}(z)}$ and $H^{(2)}(z) = \frac{z^n}{a^{(2)}(z)}$, is the same white noise process $\{u(k)\}$ with variance σ^2. This situation is shown in Figure 3.18. The infinite controllability matrices of the two models are denoted by $\mathcal{C}^{(1)}_\infty$ and $\mathcal{C}^{(2)}_\infty$, respectively.

Proposition 3.8. The principal angles between the row spaces of the future output Hankel matrices of two nth order AR processes driven by the same white noise input
Let $\{y^{(1)}(k)\}$ and $\{y^{(2)}(k)\}$ be the output processes of two nth order AR models that are driven by the same white noise $\{u(k)\}$ with variance σ^2 and let $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ be the $i \times j$ output Hankel matrices of these outputs, where $i \geq n$. Then, for $j \to \infty$, the largest n principal angles between the row spaces of $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ are equal to the principal angles between the row spaces of the infinite controllability matrices of the AR models. The remaining $i-n$ angles are equal to zero:

$$\lim_{j\to\infty}\left[Y^{(1)}_{0|i-1} \sphericalangle Y^{(2)}_{0|i-1}\right] = \left[\mathcal{C}^{(1)}_\infty \sphericalangle \mathcal{C}^{(2)}_\infty\right], \underbrace{0,\ldots,0}_{i-n} \quad \textit{for } i \geq n \ .$$

Proof.
We assume that $j \to \infty$ throughout the proof.
In Proposition 3.3 we have proven that the row spaces of the output Hankel matrices of the nth order AR processes are equal to

$$\text{row}(Y^{(1)}_{0|i-1}) = \text{row}\begin{pmatrix} X^{(1)}_n \\ U_{n|i-1} \end{pmatrix} \text{ and } \text{row}(Y^{(2)}_{0|i-1}) = \text{row}\begin{pmatrix} X^{(2)}_n \\ U_{n|i-1} \end{pmatrix} ,$$

which immediately reveals that the intersection between the row spaces of $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ is a subspace of dimension $\geq i-n$ (the rank of $U_{n|i-1}$). Consequently, there are at least $i-n$ angles between $\text{row}(Y^{(1)}_{0|i-1})$ and $\text{row}(Y^{(2)}_{0|i-1})$ equal to zero. Since the row space of $U_{n|i-1}$ is orthogonal to both $\text{row}(X^{(1)}_n)$ and $\text{row}(X^{(2)}_n)$ (see (3.35)), the remaining n angles between the row spaces of $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ are equal to the n angles between $\text{row}(X^{(1)}_n)$ and $\text{row}(X^{(2)}_n)$. Hence,

$$\lim_{j\to\infty}\left[Y^{(1)}_{0|i-1} \sphericalangle Y^{(2)}_{0|i-1}\right] = \lim_{j\to\infty}\left[X^{(1)}_n \sphericalangle X^{(2)}_n\right], \underbrace{0,\ldots,0}_{i-n} \ .$$

The squared cosines of the principal angles between the row spaces of $X^{(1)}_n$ and $X^{(2)}_n$ are the eigenvalues of

$$\left(X^{(1)}_n X^{(1)^T}_n\right)^{-1} X^{(1)}_n X^{(2)^T}_n \left(X^{(2)}_n X^{(2)^T}_n\right)^{-1} X^{(2)}_n X^{(1)^T}_n \ .$$

From Remark 3.4 we know that

$$\lim_{j\to\infty} X_n^{(1)} X_n^{(1)^T} = \sigma^2 \mathcal{C}_\infty^{(1)} \mathcal{C}_\infty^{(1)^T} ,$$
$$\lim_{j\to\infty} X_n^{(2)} X_n^{(2)^T} = \sigma^2 \mathcal{C}_\infty^{(2)} \mathcal{C}_\infty^{(2)^T} .$$

Analogously, one can derive that $\lim_{j\to\infty} X_n^{(1)} X_n^{(2)^T} = \sigma^2 \mathcal{C}_\infty^{(1)} \mathcal{C}_\infty^{(2)^T}$. Consequently, the principal angles between the row spaces of $X_n^{(1)}$ and $X_n^{(2)}$ are equal to the principal angles between the row spaces of $\mathcal{C}_\infty^{(1)}$ and $\mathcal{C}_\infty^{(2)}$. □

Just as for the principal angles between the past and future input and output spaces of a SISO model (see Section 3.5.1), the principal angles between the row spaces of the output Hankel matrices $Y_{0|i-1}^{(1)}$ and $Y_{0|i-1}^{(2)}$ are independent of the input variance σ^2.

Remark 3.17. The canonical correlations of $y_f^{(1)}$ and $y_f^{(2)}$
The principal angles between the row spaces of the $i \times j$ Hankel matrices $Y_{0|i-1}^{(1)}$ and $Y_{0|i-1}^{(2)}$ are independent of i, if $j \to \infty$ and $i \geq n$. Hence, the smallest n canonical correlations of $y_f^{(1)}$ and $y_f^{(2)}$ are the cosines of the principal angles between the row spaces of $\mathcal{C}_\infty^{(1)}$ and $\mathcal{C}_\infty^{(2)}$ and the other canonical correlations are equal to 1. ○

Remark 3.18. The computation of the canonical correlations of two AR processes
We discuss the non-parametric estimation and a few parametric computation methods of the canonical correlations.

1. **Estimating the canonical correlations from measurements**
   If two AR processes $y^{(1)}(k)$ and $y^{(2)}(k)$ that are generated by the same white input process, are observed, then the canonical correlations of the future processes can be estimated by computing the cosines of the principal angles between the row spaces of the $i \times j$ Hankel matrices $Y_{0|i-1}^{(1)}$ and $Y_{0|i-1}^{(2)}$. In case the model order n is known, we can take $i = n$. Otherwise, we choose i large enough, i.e. equal to the maximum possible order. The quality of the estimates will improve with growing j. Increasing i, while j is constant, will not have any influence.
   However, starting from a small i, smaller than the model order, and computing the principal angles for increasing i, can be used to estimate the model order: from $i = n$ onwards, the principal angles will not change any more, provided j is large enough. See Example 3.10 below.

2. **Computing the canonical correlations from the model parameters**
   Although the canonical correlations are the square roots of the eigenval-

ues of

$$\left(\mathcal{C}_\infty^{(1)}\mathcal{C}_\infty^{(1)^T}\right)^{-1}\mathcal{C}_\infty^{(1)}\mathcal{C}_\infty^{(2)^T}\left(\mathcal{C}_\infty^{(2)}\mathcal{C}_\infty^{(2)^T}\right)^{-1}\mathcal{C}_\infty^{(2)}\mathcal{C}_\infty^{(1)^T},$$

we do not have to know the infinitely large controllability matrices of the models. The matrices $P_{kl} = \mathcal{C}_\infty^{(k)}\mathcal{C}_\infty^{(l)^T}$ $(k,l = 1,2)$ can be obtained by solving the controllability Lyapunov equation

$$\begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} A^{(1)} & 0 \\ 0 & A^{(2)} \end{pmatrix}\begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}\begin{pmatrix} A^{(1)^T} & 0 \\ 0 & A^{(2)^T} \end{pmatrix} + \begin{pmatrix} K^{(1)} \\ K^{(2)} \end{pmatrix}\begin{pmatrix} K^{(1)^T} & K^{(2)^T} \end{pmatrix}, \quad (3.108)$$

where $(A^{(1)}, K^{(1)}, C^{(1)})$ and $(A^{(2)}, K^{(2)}, C^{(2)})$ are the state space representations of the AR models. The solution of (3.108) exists and is unique due to the stability of the AR models.

Furthermore, we note that the principal angles between the row spaces of the controllability matrices are equal to the principal angles between the column spaces of the models' observability matrices[8], because for SISO models the row space of the controllability matrix and the column space of the observability matrix are equal (see Remark 3.9). Hence, we do not have to know the forward innovation structure (the Kalman gains $K^{(1)}$ and $K^{(2)}$) of the models. The smallest n canonical correlations of the future AR processes are also equal to the square roots of the eigenvalues of $Q_{11}^{-1}Q_{12}Q_{22}^{-1}Q_{21}$, where $Q_{kl} = \Gamma_\infty^{(k)^T}\Gamma_\infty^{(l)}(k,l = 1,2)$, which can be obtained by solving the observability Lyapunov equation (3.23) for $\left(\begin{pmatrix} A^{(1)} & 0 \\ 0 & A^{(2)} \end{pmatrix}, \begin{pmatrix} C^{(1)} & C^{(2)} \end{pmatrix}\right)$.

We can reduce the computation cost further by working only with the poles of the models, which is shown in the following proposition.

Proposition 3.9. Computation of the canonical correlations of two AR processes

Let the two nth order AR models have distinct poles $\alpha_1^{(1)}, \ldots, \alpha_n^{(1)}$ and $\alpha_1^{(2)}, \ldots, \alpha_n^{(2)}$, respectively. Then, the squared cosines of the principal angles between $\mathrm{row}(\mathcal{C}_\infty^{(1)})$ *and* $\mathrm{row}(\mathcal{C}_\infty^{(2)})$ *are equal to the eigenvalues of XY, where*

$$\begin{aligned} X(i,j) &= \prod_{k=1}^{n}\frac{(1-\bar{\alpha}_k^{(1)}\alpha_i^{(1)})}{(1-\bar{\alpha}_k^{(1)}\alpha_j^{(2)})}\prod_{k\neq i}^{n}\frac{(\alpha_j^{(2)}-\alpha_k^{(1)})}{(\alpha_i^{(1)}-\alpha_k^{(1)})}, \\ Y(i,j) &= \prod_{k=1}^{n}\frac{(1-\bar{\alpha}_k^{(2)}\alpha_i^{(2)})}{(1-\bar{\alpha}_k^{(2)}\alpha_j^{(1)})}\prod_{k\neq i}^{n}\frac{(\alpha_j^{(1)}-\alpha_k^{(2)})}{(\alpha_i^{(2)}-\alpha_k^{(2)})}, \end{aligned} \quad (3.109)$$

[8] In [36, 37, 38, 39] we have called these angles the subspace angles between the AR models.

in which $X(i,j)$ denotes the element of the matrix X on the ith row and the jth column and similarly for $Y(i,j)$.

Proof.
The row spaces of $\mathcal{C}_\infty^{(1)} = \mathcal{C}_\infty\left(\frac{z^n}{a^{(1)}(z)}\right)$ and $\mathcal{C}_\infty^{(2)} = \mathcal{C}_\infty\left(\frac{z^n}{a^{(2)}(z)}\right)$ are equal to the row spaces of $\mathcal{C}_\infty^*\left(\frac{z^n}{a^{(1)}(z)}\right)$ and $\mathcal{C}_\infty^*\left(\frac{z^n}{a^{(2)}(z)}\right)$ (see Proposition 3.2), where $\mathcal{C}_\infty^*(H(z))$ is defined as in (3.54). Hence, the squared cosines of the principal angles between $\mathrm{row}(\mathcal{C}_\infty^{(1)})$ and $\mathrm{row}(\mathcal{C}_\infty^{(2)})$ are equal to the eigenvalues of $P_{11}^{*^{-1}} P_{12}^* P_{22}^{*^{-1}} P_{21}^*$, where

$$P_{kl}^* = \mathcal{C}_\infty^*\left(\frac{z^n}{a^{(k)}(z)}\right) \mathcal{C}_\infty^{*^H}\left(\frac{z^n}{a^{(l)}(z)}\right) \quad (k,l=1,2)$$

(the superscript $\cdot^H$ denotes the complex conjugate transpose). The elements of these matrices can be expressed in terms of the poles of the models by applying

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \text{ for } |x|<1 \ .$$

Due to the stability of the models, P_{kl}^* $(k,l=1,2)$ are matrices with finite elements

$$P_{kl}^*(i,j) = \frac{1}{1-\alpha_i^{(k)}\bar{\alpha}_j^{(l)}} \ (k,l=1,2) \ , \tag{3.110}$$

where $\bar{c}$ denotes the complex conjugate of c.
Let $P_{11}^{*^{-1}} P_{12}^*$ be denoted by X and $P_{22}^{*^{-1}} P_{21}^*$ by Y, then the cosines of the principal angles between $\mathcal{C}_\infty^*\left(\frac{z^n}{a^{(1)}(z)}\right)$ and $\mathcal{C}_\infty^*\left(\frac{z^n}{a^{(2)}(z)}\right)$ are the square roots of the eigenvalues of XY. For the elements of the matrices X and Y the expressions in (3.109) can be derived, via the expressions of P_{kl}^* $(k,l=1,2)$ in (3.110). □

○

Example 3.10. The canonical correlations of two 15th order AR processes that are driven by the same input
We consider two AR models of order 15 of which the denominators are respectively

$$\begin{aligned}
a^{(1)}(z) \quad &= \quad z^{15}+0.23z^{14}-0.44z^{13}-0.23z^{12}-0.04z^{11}+0.20z^{10}+\\
&\quad\ 0.02z^9-0.09z^8-0.02z^7+0.36z^6+0.04z^5-0.24z^4-\\
&\quad\ 0.16z^3-0.14z^2-0.02z-0.09 \ ,\\
a^{(2)}(z) \quad &= \quad z^{15}+0.86z^{14}-0.85z^{13}-0.77z^{12}+0.42z^{11}+0.67z^{10}+\\
&\quad\ 0.40z^9+0.38z^8+0.42z^7-0.56z^6-0.85z^5+0.56z^4+\\
&\quad\ 0.96z^3+0.06z^2-0.33z^1-0.13
\end{aligned}$$

We have simulated the output processes by using a white input sequence of length 2000. The principal angles between the row spaces of the output Hankel matrices $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ are shown in Figure 3.19 for an increasing number of rows $i = 1, 2, \ldots, 30$. The left picture shows the non-parametric estimates of the principal angles where j is equal to 1971. From $i = 15$ onwards, the angles do not significantly change with increasing i. Hence, we could estimate the model order as $n = 15$.
In the right picture we give the principal angles between $\mathrm{row}(Y^{(1)}_{0|i-1})$ and $\mathrm{row}(Y^{(2)}_{0|i-1})$ for an infinite number of columns ($j = \infty$). As proven in Proposition 3.8, the angles remain constant as soon as the number of rows is equal to the model order, here $i = 15$. From $i = 15$ onwards, the largest 15 angles are equal to

$$\lim_{j\to\infty}\left[Y^{(1)}_{0|i-1} \sphericalangle Y^{(2)}_{0|i-1}\right] = \begin{array}{l}(2.01^\circ,\ 5.86^\circ,\ 9.54^\circ,\ 10.27^\circ,\ 13.40^\circ,\\ 29.32^\circ,\ 34.12^\circ,\ 46.86^\circ,\ 55.50^\circ,\\ 55.72^\circ,\ 63.66^\circ,\ 69.53^\circ,\ 73.40^\circ,\\ 73.97^\circ,\ 79.16^\circ)\ .\end{array} \tag{3.111}$$

These values are indeed equal to the principal angles between the row spaces of $\mathcal{C}_\infty\left(\frac{z^{15}}{a^{(1)}(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^{15}}{a^{(2)}(z)}\right)$. The other $i - 15$ angles are equal to 0. $\Diamond$

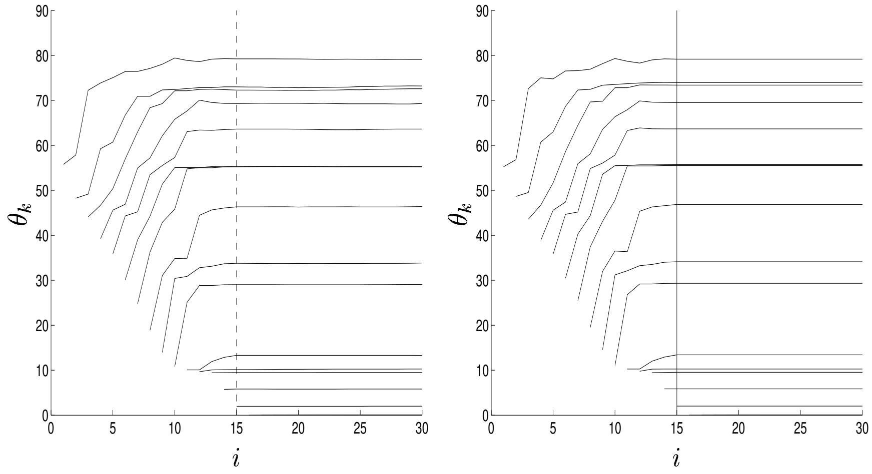


Figure 3.19: The principal angles (in degrees) between the row spaces of the $i \times j$ output Hankel matrices $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$, for $i = 1, 2, \ldots, 30$. The left picture shows the non-parametric estimates for $j = 1971$. From $i = 15$ onwards, the angles do not significantly change with increasing i. The right picture gives the angles for $j = \infty$. From $i = 15$ onwards, the angles remain constant and they are equal to the principal angles between the row spaces of the infinite controllability matrices of the models.

Although we will not need this result further on, we briefly describe how AR processes with a different model order should be treated.

Remark 3.19. The principal angles for AR models with a different order
Assume $\{y^{(1)}(k)\}$ is a stable AR process of order $n^{(1)}$ and $\{y^{(2)}(k)\}$ has order $n^{(2)}$, where $n^{(2)} > n^{(1)}$. One can show that if the number of rows i in the Hankel matrices $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ is larger than $n^{(2)}$, then the largest $n^{(2)}$ principal angles between $\text{row}(Y^{(1)}_{0|i-1})$ and $\text{row}(Y^{(2)}_{0|i-1})$ are equal to the principal angles between the row spaces of

$$\begin{pmatrix} & \mathcal{C}^{(1)}_{\infty} \\ I_{n^{(2)}-n^{(1)}} & 0_{n^{(2)}-n^{(1)},\infty} \end{pmatrix} \text{ and } \mathcal{C}^{(2)}_{\infty} \ ,$$

and the other $i - n^{(2)}$ principal angles are equal to 0, for $i \geq n^{(2)}$. ○

The input is white noise filtered by an AR model
We now show that Proposition 3.8 also applies if two AR models are driven by the same input that is white noise filtered by another AR model, which is depicted in the left picture of Figure 3.20. This result will be used for the computation of the canonical correlations between two ARMA processes (see below).

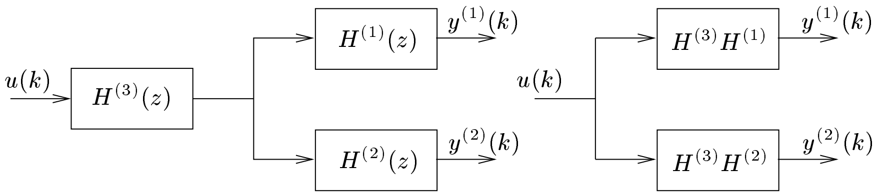


Figure 3.20: Setup for the computation of the canonical correlations between the future output of two AR models driven by the same AR filtered white noise.

Proposition 3.10. The principal angles between the row spaces of $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ when the input is an AR process
Let $H^{(1)}(z)$ and $H^{(2)}(z)$ be the transfer functions of two nth order AR models driven by the same input which is white noise filtered by an mth order AR model. Then, the largest n principal angles between the row spaces of the output Hankel matrices $Y^{(1)}_{0|i-1} \in \mathbb{R}^{i\times j}$ and $Y^{(2)}_{0|i-1} \in \mathbb{R}^{i\times j}$ are equal to the principal angles between $\mathcal{C}^{(1)}_{\infty}$ and $\mathcal{C}^{(2)}_{\infty}$ and the other principal angles are zero, provided $i \geq m+n$ and $j \to \infty$:

$$\lim_{j\to\infty} \left[Y^{(1)}_{0|i-1} \sphericalangle Y^{(2)}_{0|i-1} \right] = \left[\mathcal{C}^{(1)}_{\infty} \sphericalangle \mathcal{C}^{(2)}_{\infty} \right], \underbrace{0,\ldots,0}_{i-n} \qquad (i \geq m+n) \ .$$

Proof.
Let the mth order AR model by which the white input is filtered, be denoted by its transfer function $H^{(3)}(z)$ and the white input process by $\{u(k)\}$. The output processes $\{y^{(1)}(k)\}$ and $\{y^{(2)}(k)\}$ are then AR processes with transfer functions $H^{(3)}(z)H^{(1)}(z)$ and $H^{(3)}(z)H^{(2)}(z)$, respectively, driven by the same white noise input $\{u(k)\}$ (see the right picture of Figure 3.20). Hence, we can apply Proposition 3.8 to these AR processes of order $m+n$, provided that $i \geq m+n$:

$$\lim_{j\to\infty}\left[Y^{(1)}_{0|i-1} \triangleleft Y^{(2)}_{0|i-1}\right] = \left[\mathcal{C}_\infty\left(H^{(3)}H^{(1)}\right) \triangleleft \mathcal{C}_\infty\left(H^{(3)}H^{(2)}\right)\right] \underbrace{0,\ldots,0}_{i-(n+m)} \quad \text{for } i \geq m+n\ ,$$

where $\mathcal{C}_\infty(H^{(3)}H^{(1)})$ and $\mathcal{C}_\infty(H^{(3)}H^{(2)})$ are the infinite controllability matrices of the AR models with transfer function $H^{(3)}(z)H^{(1)}(z)$ and $H^{(3)}(z)H^{(2)}(z)$. It is easily seen from Proposition 3.2 that there are at least m principal angles between the row spaces of $\mathcal{C}_\infty(H^{(3)}(z)H^{(1)}(z))$ and $\mathcal{C}_\infty(H^{(3)}(z)H^{(2)}(z))$ equal to zero. By exploiting the structure of the row spaces of the controllability matrices (see (3.54)), one can prove that the other n principal angles are equal to the principal angles between the row spaces of $\mathcal{C}^{(1)}_\infty$ and $\mathcal{C}^{(2)}_\infty$. □

Example 3.11. The canonical correlations of two 15th order AR processes that are driven by a 5th order AR process
We consider the two AR models of Example 3.10, but now the input is white noise filtered by a 5th order AR model. The principal angles between the row spaces of the output Hankel matrices $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ are shown in Figure 3.21 for an increasing number of rows i and an infinite number of columns ($j=\infty$). As proven in Proposition 3.8, the angles remain constant as soon as the number of rows is equal to the sum of the model order, here $n=15$, and the order of the filter, 5. From $i=20$ onwards, the largest 15 angles are equal to the principal angles between $\mathcal{C}_\infty\left(\frac{z^{15}}{a^{(1)}(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^{15}}{a^{(2)}(z)}\right)$, which can be seen in Figure 3.19 and in (3.111). The other $i-20$ angles are equal to 0.
Although the filter does not change the canonical correlations of the two future output processes, it does influences the values when $i<20$. ◊

The canonical correlations between two ARMA processes

In this paragraph we compute the canonical correlations of two stable and minimum phase ARMA processes that are driven by the same white input. Figure 3.22 shows two equivalent schemes for this computation. The scheme on the right is obtained from the left side scheme as follows. We multiply both the transfer functions $H^{(1)}(z)=\frac{b^{(1)}(z)}{a^{(1)}(z)}$ and $H^{(2)}(z)=\frac{b^{(2)}(z)}{a^{(2)}(z)}$ by $\frac{z^{n^{(1)}+n^{(2)}}}{b^{(1)}(z)b^{(2)}(z)}$, which

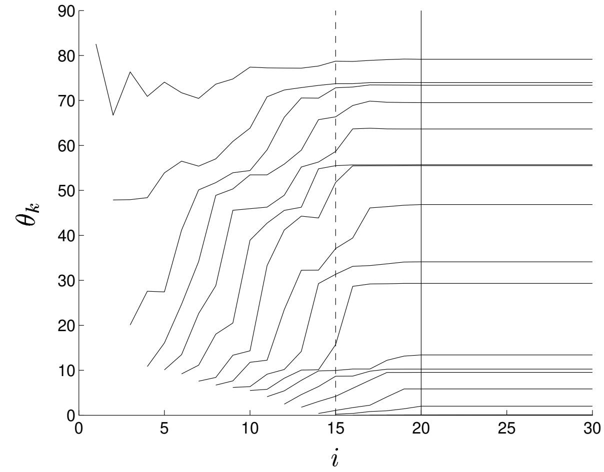


Figure 3.21: The principal angles (in degrees) between the row spaces of the $i \times j$ output Hankel matrices $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$, for $i = 1, 2, \ldots, 30$ and $j \to \infty$. From $i = 20$ onwards, the angles remain constant and they are equal to the principal angles between the row spaces of the infinite controllability matrices of the models.

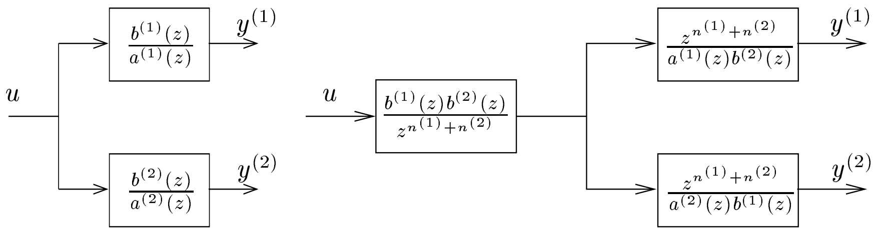


Figure 3.22: Setup for the computation of the canonical correlations between the future output of two ARMA models driven by the same white noise.

is then compensated by filtering the common input $\{u(k)\}$ with the inverse transfer function. The equivalence of both schemes and Proposition 3.10 form the basis of the following theorem.

Theorem 3.11. The canonical correlations of two ARMA processes driven by the same white noise input
Let $\{y^{(1)}(k)\}$ and $\{y^{(2)}(k)\}$ be the output processes of the stable and minimum phase ARMA models with transfer function $H^{(1)}(z) = \frac{b^{(1)}(z)}{a^{(1)}(z)}$ of order $n^{(1)}$ and $H^{(2)}(z) = \frac{b^{(2)}(z)}{a^{(2)}(z)}$ of order $n^{(2)}$, respectively that are driven by the same white input. The smallest $n^{(1)} + n^{(2)}$ canonical correlations of the future output pro-

cesses, $y_f^{(1)}$ and $y_f^{(2)}$, are equal to the cosines of the principal angles between the row spaces of $\mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{a^{(1)}(z)b^{(2)}(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{a^{(2)}(z)b^{(1)}(z)}\right)$. The other canonical correlations are equal to 1.

Proof.
From Figure 3.22 it follows that the canonical correlations between $y_f^{(1)}$ and $y_f^{(2)}$ can be computed as the cosines of the principal angles between the row spaces of the output Hankel matrices of two AR models of order $n^{(1)} + n^{(2)}$ that are driven by the same input. This input is the white noise $\{u(k)\}$ filtered by the minimum phase MA model with transfer function $\frac{b^{(1)}(z)b^{(2)}(z)}{z^{n^{(1)}+n^{(2)}}}$. It is well known (see for example [50, p. 70]) that a stable and minimum phase ARMA process driven by white noise, can be approximated arbitrarily well by a stable AR process of order m as $m \to \infty$. Doing this for the MA filter:

$$\frac{b^{(1)}(z)b^{(2)}(z)}{z^{n^{(1)}+n^{(2)}}} \approx \frac{z^m}{f(z)} \ (m \to \infty) \ ,$$

yields the same situation as in Proposition 3.10: there are two AR models that are driven by the same input, which is white noise filtered by a third AR model. Consequently, we can apply Proposition 3.10, where n is now equal to $n^{(1)} + n^{(2)}$ and $m \to \infty$. The condition that $i \geq m + n$ and $m \to \infty$ leads to $i \to \infty$. We hence obtain

$$\lim_{i\to\infty}\lim_{j\to\infty}\left[Y_{0|i-1}^{(1)} \sphericalangle Y_{0|i-1}^{(2)}\right] = \left[\mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{a^{(1)}(z)b^{(2)}(z)}\right) \sphericalangle \mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{a^{(2)}(z)b^{(1)}(z)}\right)\right], 0, 0, \ldots$$

□

Remark 3.20. From Proposition 3.2 it follows that the row space of the infinite controllability matrix $\mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{a^{(1)}(z)b^{(2)}(z)}\right)$ is equal to the row space of $\mathcal{C}_\infty\left(\frac{b^{(1)}(z)a^{(2)}(z)}{a^{(1)}(z)b^{(2)}(z)}\right)$, which is the row space of the infinite controllability matrix of $H^{(1)}(z)H^{(2)^{-1}}(z)$:

$$\operatorname{row}\left(\mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{a^{(1)}(z)b^{(2)}(z)}\right)\right) = \operatorname{row}\left(\mathcal{C}_\infty\left(H^{(1)}(z)H^{(2)^{-1}}(z)\right)\right) \ .$$

Similarly,

$$\operatorname{row}\left(\mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{a^{(2)}(z)b^{(1)}(z)}\right)\right) = \operatorname{row}\left(\mathcal{C}_\infty\left(H^{(2)}(z)H^{(1)^{-1}}(z)\right)\right) \ .$$

Consequently, the canonical correlations of the future ARMA processes $y_f^{(1)}$ and $y_f^{(2)}$ can be written as

$$\begin{aligned}\mathrm{cc}(y_f^{(1)}, y_f^{(2)}) &= \lim_{i\to\infty}\lim_{j\to\infty} \cos\left[Y_{0|i-1}^{(1)} \sphericalangle Y_{0|i-1}^{(2)}\right] \\ &= \cos\left[\mathcal{C}_\infty\left(H^{(1)}H^{(2)^{-1}}\right) \sphericalangle \mathcal{C}_\infty\left(H^{(2)}H^{(1)^{-1}}\right)\right], 1, 1, \ldots\end{aligned}$$

Note that the canonical correlations of the two processes are not a function of the difference of the transfer functions (or spectra), but of their ratio and hence of the difference of the log-spectra or of the difference of their cepstra. See Chapter 5 for the definition of the cepstrum of a stochastic process and Chapter 6 for the consequences of this property. ○

Remark 3.21. The computation of the canonical correlations of two ARMA processes

We first briefly describe the estimation of the canonical correlations from data and then we show how to compute the canonical correlations from the model parameters.

1. **Estimating the canonical correlations from measurements**
   If two ARMA processes $\{y^{(1)}(k)\}$ and $\{y^{(2)}(k)\}$ that are driven by the same white input process, are observed, then their canonical correlations can be estimated by computing the cosines of the principal angles between the row spaces of the $i \times j$ Hankel matrices $Y_{0|i-1}^{(1)}$ and $Y_{0|i-1}^{(2)}$. Contrary to the case of two AR processes, the estimates now improve with growing j and i.

2. **Computing the canonical correlations from the model parameters**
   If the model parameters are known, the exact values of the canonical correlations can be computed. Let $\{y^{(1)}(k)\}$ be an ARMA process of order $n^{(1)}$ with transfer function $\frac{b^{(1)}(z)}{a^{(1)}(z)}$ and $\{y^{(2)}(k)\}$ an ARMA process of order $n^{(2)}$ with transfer function $\frac{b^{(2)}(z)}{a^{(2)}(z)}$, driven by the same input. The canonical correlations of the future processes are equal to the cosines of the principal angles between the row spaces of the infinite controllability matrices of the AR processes with transfer function $\frac{z^{n^{(1)}+n^{(2)}}}{a^{(1)}(z)b^{(2)}(z)}$ and $\frac{z^{n^{(1)}+n^{(2)}}}{a^{(2)}(z)b^{(1)}(z)}$, respectively. Consequently, we can apply the parametric methods for the computation of the canonical correlations of AR processes, as described in Remark 3.18. ○

Example 3.12. The canonical correlations of two ARMA processes driven by the same white noise

In Figure 3.23 we show simulation results obtained as follows. We generate the output sequences (1 000 samples) of two stable and minimum phase ARMA models with transfer functions

$$H^{(1)}(z) = \frac{z^3 - 0.1377z^2 + 0.0166z - 0.0013}{z^3 - 2.3464z^2 + 2.2204z - 0.8268} ,$$
$$H^{(2)}(z) = \frac{z^2 + 1.1565z + 0.7750}{z^2 - 0.5797z + 0.0916} .$$

The input $u(k)$ is Gaussian white noise. In full line the cosines of the principal angles between the row spaces of the $i \times 951$ Hankel matrices $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ as a function of i are depicted. The results for $j = \infty$ are in dashed lines and the asymptotic values (for $i \to \infty$ and $j \to \infty$) are given by the dash-dot lines. The results show that the angles converge to the true ones by increasing the number of rows and columns of the output Hankel matrices. $\diamondsuit$

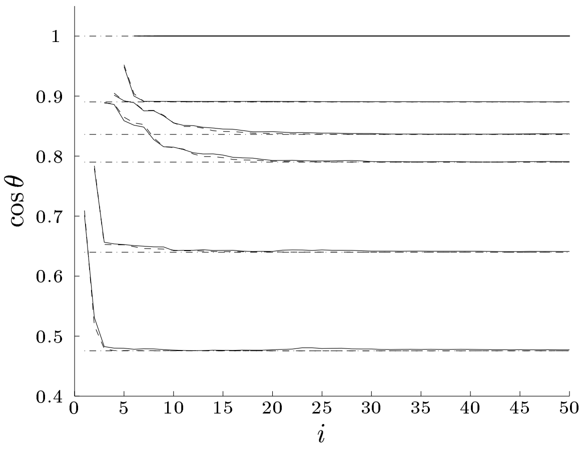


Figure 3.23: The cosines of the principal angles between the row spaces of the $i \times j$ output Hankel matrices $Y^{(1)}_{0|i-1}$ and $Y^{(2)}_{0|i-1}$ which contain the simulated output samples of a second order and a third order ARMA model. The number of rows, i, increases from 10 to 50. In full line we show the results for $j = 951$. The cosines for $j = \infty$ are given in dashed line and the canonical correlations, i.e. the cosines for $i, j = \infty$, in dash-dot line.

3.5.3 A new characterization of the canonical correlations of the future input and future output processes of a SISO ARMA model

Based on the results of the previous section, we can now prove a new characterization of the canonical correlations of the future input and output processes of a SISO ARMA model.

Theorem 3.12. The canonical correlations of the future input and future output of a SISO ARMA model
The smallest n canonical correlations of the future input and output processes of an nth order SISO ARMA model with transfer function $\frac{b(z)}{a(z)}$ are equal to the cosines of the principal angles between the row spaces of the infinite controllability matrices $\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)$. The other canonical correlations are equal to 1:

$$\mathrm{cc}(u_f, y_f) = \cos\left[\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right) \sphericalangle\, \mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)\right], 1, 1, \ldots$$

Proof.
The canonical correlations of the future input $\{u(k)\}$ and the future output $\{y(k)\}$ of an ARMA model with transfer function $\frac{b(z)}{a(z)}$ can be viewed as the canonical correlations between two ARMA process that are driven by the same input: the output of $H^{(1)}(z) = \frac{b(z)}{a(z)}$ and the output of the ARMA model with transfer function $H^{(2)}(z) = 1$, both driven by $\{u(k)\}$. Applying Theorem 3.11 to $H^{(1)}(z)$ and $H^{(2)}(z)$ immediately gives the statement in Theorem 3.12. □

A more direct and elegant proof, i.e. without having to rely on the results of Section 3.5.2, would follow from a property for Lyapunov equations, which is given below. However, since the proof of the latter is not complete yet, we present it as a conjecture and indicate its correspondence with the property above. A proof is given for the simple case that the matrices involved are scalars ($n = 1$).

Conjecture 3.13.
Let $A \in \mathbb{R}^{n\times n}$, $K \in \mathbb{R}^{n\times 1}$ and $C \in \mathbb{R}^{1\times n}$ be matrices so that there are no two eigenvalues λ_i and λ_j of $\begin{pmatrix} A & 0 \\ 0 & A-KC \end{pmatrix}$ for which $\lambda_i\lambda_j = 1$ $(i,j = 1,\ldots,2n)$. If the $n \times n$ matrices P, Q and R satisfy

$$\begin{pmatrix} P & R \\ R^T & Q \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & (A-KC)^T \end{pmatrix} \begin{pmatrix} P & R \\ R^T & Q \end{pmatrix} \begin{pmatrix} A^T & 0 \\ 0 & A-KC \end{pmatrix} + \begin{pmatrix} K \\ -C^T \end{pmatrix} \begin{pmatrix} K^T & -C \end{pmatrix},$$

and the matrices P, Q and $I_n + QP$ are invertible, then $P^{-1}RQ^{-1}R^T$ and $(I_n + QP)^{-1}$ have the same eigenvalues.

Relation of Conjecture 3.13 to Theorem 3.12 From Conjecture 3.13, Theorem 3.12 can be deduced as follows. Let A, K and C be the state space matrices of the nth order ARMA model with transfer function $\frac{b(z)}{a(z)}$. First note that since we deal with stable and minimum phase models, the condition

$\lambda_i\lambda_j \neq 1$ in Conjecture 3.13 is automatically fulfilled. Moreover, under fairly general conditions P, Q and $I_n + PQ$ are non-singular.
The matrix P is then the controllability Gramian of the model (see (3.22))

$$P = \mathcal{C}_\infty\left(\frac{b(z)}{a(z)}\right)\mathcal{C}_\infty^T\left(\frac{b(z)}{a(z)}\right) ,$$

and by (3.23) and (3.42), Q is the observability Gramian of the inverse model (denoted by Q_z above)

$$Q = \Gamma_\infty^T\left(\frac{a(z)}{b(z)}\right)\Gamma_\infty\left(\frac{a(z)}{b(z)}\right) .$$

The eigenvalues of $(I_n+QP)^{-1}$ are then equal to the smallest n squared canonical correlations of the future input and output process of the ARMA model (see (3.105)). Furthermore, the matrix R is equal to the product of the infinite controllability matrix of the ARMA model and the infinite observability matrix of the inverse model

$$R = \mathcal{C}_\infty\left(\frac{b(z)}{a(z)}\right)\Gamma_\infty\left(\frac{a(z)}{b(z)}\right) .$$

The eigenvalues of $P^{-1}RQ^{-1}R^T$ are thus the cosines squared of the principal angles between the row space of $\mathcal{C}_\infty\left(\frac{b(z)}{a(z)}\right)$ and the column space of $\Gamma_\infty\left(\frac{a(z)}{b(z)}\right)$. From Proposition 3.2 and Remark 3.9 it follows that these subspaces are equal to the row spaces of $\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)$, respectively.
Since we conjecture that the eigenvalues of $(I_n+QP)^{-1}$ are equal to the eigenvalues of $P^{-1}RQ^{-1}R^T$, proving the conjecture would prove Theorem 3.12.

Proof of Conjecture 3.13 for $n = 1$

Property 3.14.
Given $a \in \mathbb{R}$, $k, c \in \mathbb{R}_0$, so that $a \neq 1$, $a - kc \neq 1$ and $a(a-kc) \neq 1$. If $p, q, r \in \mathbb{R}$ satisfy

$$p = apa + k^2 , \tag{3.112a}$$
$$q = (a-kc)q(a-kc) + c^2 , \tag{3.112b}$$
$$r = ar(a-kc) - kc , \tag{3.112c}$$

and $(1+pq) \neq 0$, then

$$\frac{r^2}{pq} = \frac{1}{1+pq} .$$

Proof.
For simplicity of notation, we use $b = a - kc$. From the Equations (3.112a–3.112c) it follows immediately that p, q, r can be written as

$$p = \frac{k^2}{1-a^2}, \quad q = \frac{c^2}{1-b^2}, \quad r = \frac{-kc}{1-ab} .$$

We then have

$$\frac{r^2}{pq} = \frac{k^2c^2(1-a^2)(1-b^2)}{(1-ab)^2k^2c^2} = \frac{(1-a^2)(1-b^2)}{(1-ab)^2} \ ,$$

and

$$\frac{1}{1+pq} = \frac{(1-a^2)(1-b^2)}{(1-a^2)(1-b^2)+k^2c^2} = \frac{(1-a^2)(1-b^2)}{(1-a^2)(1-b^2)+(a-b)^2} \ .$$

Since $(1-a^2)(1-b^2)+(a-b)^2 = (1-ab)^2$, we obtain

$$\frac{r^2}{pq} = \frac{1}{1+pq} \ .$$

□

Example 3.13. The canonical correlations of the future input and output of a 15th order SISO ARMA process

Let us return to Example 3.10, where we computed the canonical correlations of two 15th order AR models with transfer functions $\frac{z^{15}}{a^{(1)}(z)}$ and $\frac{z^{15}}{a^{(2)}(z)}$. The polynomials $a^{(1)}(z)$ and $a^{(2)}(z)$ are given in Example 3.10. However, now we compute the canonical correlations of the future input and output process of the ARMA model with transfer function $\frac{a^{(1)}(z)}{a^{(2)}(z)}$. From Proposition 3.8 and Theorem 3.12 we know that the canonical correlations of the future AR processes in

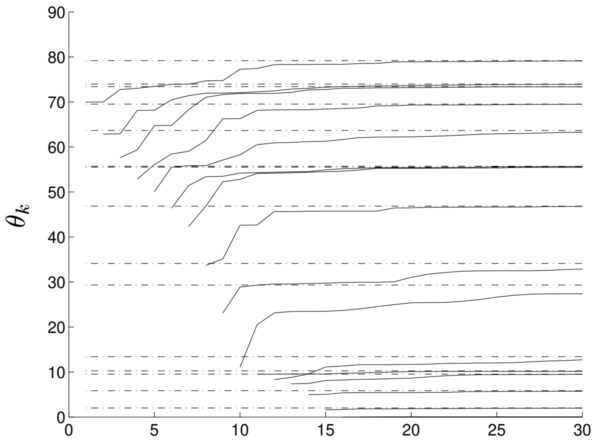


Figure 3.24: The principal angles (in degrees) between the row spaces of the $i \times j$ input and output Hankel matrices $U_{0|i-1}$, $Y_{0|i-1}$, for $i = 1, 2, \ldots, 30$ and $j \to \infty$. They converge to the principal angles between the row spaces of $\mathcal{C}_\infty\left(\frac{z^{15}}{a^{(1)}(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^{15}}{a^{(2)}(z)}\right)$, which are given in dash-dot line.

Example 3.10 and the canonical correlations of future input and future output of the ARMA process $\frac{a^{(1)}(z)}{a^{(2)}(z)}$ should be equal. For both cases the smallest n canonical correlations are equal to the cosines of the principal angles between $\text{row}\Big(\mathcal{C}_\infty\Big(\frac{z^{15}}{a^{(1)}(z)}\Big)\Big)$ and $\text{row}\Big(\mathcal{C}_\infty\Big(\frac{z^{15}}{a^{(2)}(z)}\Big)\Big)$ and the other canonical correlations are equal to 1.
The principal angles between the row spaces of the input and output Hankel matrices $U_{0|i-1}$ and $Y_{0|i-1}$ are shown in full line in Figure 3.24 for an increasing number of rows $i = 1, 2, \ldots, 30$ and an infinite number of columns ($j = \infty$). The dash-dot lines represent the principal angles between the row spaces of the controllability matrices $\mathcal{C}_\infty\Big(\frac{z^{15}}{a^{(1)}(z)}\Big)$ and $\mathcal{C}_\infty\Big(\frac{z^{15}}{a^{(2)}(z)}\Big)$, whose values are given in Example 3.10. The principal angles between $\text{row}(U_{0|i-1})$ and $\text{row}(Y_{0|i-1})$ indeed converge to these values, for $i \to \infty$. $\diamond$

From Theorem 3.12 and the complementarity property (Property 3.7), the following Corollary immediately follows.

Corollary 3.15. The canonical correlations of the past and the future output
The largest n canonical correlations of the past and future output processes of an nth order SISO ARMA model with transfer function $\frac{b(z)}{a(z)}$ are equal to the sines of the principal angles between the row spaces of the infinite controllability matrices $\mathcal{C}_\infty\Big(\frac{z^n}{a(z)}\Big)$ and $\mathcal{C}_\infty\Big(\frac{z^n}{b(z)}\Big)$. The other canonical correlations are equal to 0:

$$\text{cc}(y_p, y_f) = \sin\left[\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right) \sphericalangle \mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)\right], 0, 0, \ldots$$

It is also readily seen from Theorem 3.12 that the canonical correlations of the future input and future output of a SISO model are equal to the canonical correlations of the future input and future output of the inverse model. Similarly, by Corollary 3.15, the canonical correlations of the past and future output do not change when the model is inverted. This is formulated in the following corollary.

Corollary 3.16. The canonical correlations of the inverse model
Let $\{y(k)\}$ be the output process of a SISO linear model that is driven by white noise and $\{z(k)\}$ the output process of the inverse model when driven by white noise. Then, the canonical correlations of the past and the future process of $\{y(k)\}$ are equal to the canonical correlations of the past and the future of the process $\{z(k)\}$

$$\text{cc}(y_p, y_f) = \text{cc}(z_p, z_f) \; .$$

Note that in general Corollary 3.16 is *not* valid for MIMO models.

3.5.4 The canonical correlations of the past and future of an AR process

Property 3.17. The canonical correlations of the past and future of an AR process
The largest n squared canonical correlations of the past and the future of an nth order AR process with system matrix A and controllability Gramian P are the eigenvalues of $P^{-1}A^nPA^{n^T}$. The other canonical correlations are equal to 0.

Proof.
From Proposition 3.3 we know that the row spaces of Y_p and Y_f are equal to the row spaces of $\begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix}$ and $\begin{pmatrix} X_{i+n} \\ U_{i+n|2i-1} \end{pmatrix}$, respectively, for $j \to \infty$ and $i \geq n$. Since $U_{i+n|2i-1}$ is orthogonal to both $\begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix}$ and X_{i+n}, the principal angles between the row spaces of Y_p and Y_f are

$$[Y_p \sphericalangle Y_f] = \left[\begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix} \sphericalangle X_{i+n} \right], \underbrace{\frac{\pi}{2}, \ldots, \frac{\pi}{2}}_{i-n} \text{ for } j \to \infty \ .$$

By using (3.27), we obtain the following expressions

$$\begin{aligned} X_{i+n} X_{i+n}^T &= \Sigma \ , \\ X_{i+n} \begin{pmatrix} X_n^T & U_{n|i-1}^T \end{pmatrix} &= \begin{pmatrix} A^i \Sigma & A^n \Delta_{i-n} Q_{u_{i-n}} \end{pmatrix} \ , \\ \begin{pmatrix} X_n \\ U_{n|i-1} \end{pmatrix} \begin{pmatrix} X_n^T & U_{n|i-1}^T \end{pmatrix} &= \begin{pmatrix} \Sigma & 0 \\ 0 & Q_{u_{i-n}} \end{pmatrix} \ , \end{aligned}$$

where Σ is the state covariance matrix, $Q_{u_{i-n}}$ is the $(i-n) \times (i-n)$ input covariance matrix and $\Delta_{i-n} = \begin{pmatrix} A^{i-n-1}K & \cdots & AK & K \end{pmatrix}$. The squared cosines of the smallest n angles are then the eigenvalues of

$$\Sigma^{-1} A^n \left(A^{i-n} \Sigma A^{i-n^T} + \Delta_{i-n} Q_{u_{i-n}} \Delta_{i-n}^T \right) A^{n^T} \ .$$

Since we are working with a SISO model, the input covariance matrix $Q_{u_{i-n}}$ is equal to $\sigma^2 I_{i-n}$ and $\Sigma = \sigma^2 P$, where σ^2 is the input variance and P the controllability Gramian. Consequently, the squared cosines of the principal angles between the row spaces of Y_p and Y_f are, for $j \to \infty$ and $i \geq n$,

$$\begin{aligned} \cos^2 [Y_p \sphericalangle Y_f] &= \lambda \left(P^{-1} A^n (\underbrace{A^{i-n} P A^{i-n^T} + \Delta_{i-n} \Delta_{i-n}^T}_{P}) A^{n^T} \right), \underbrace{0, \ldots, 0}_{i-n} \ , \\ &= \lambda \left(P^{-1} A^n P A^{n^T} \right), \underbrace{0, \ldots, 0}_{i-n} \ , \end{aligned} \tag{3.113}$$

and the eigenvalues of $P^{-1}A^nPA^{n^T}$ are thus also the largest n squared canonical correlations of y_p and y_f. □

Example 3.14. The principal angles between the past and future of an AR process
From the expression (3.113) it is clear that the principal angles between the row spaces of the $i \times j$ Hankel matrices Y_p and Y_f are independent of i, for $j \to \infty$ and $i \geq n$. This is illustrated in Figure 3.25, where the cosines of the principal angles between row(Y_p) and row(Y_f) of a 15th order AR model are shown as a function of i ($j = \infty$). The transfer function of the AR model in the simulation is $\frac{z^{15}}{a^{(1)}(z)}$, where $a^{(1)}(z)$ is given in Example 3.10. The true values of the canonical correlations are equal to 0.7391, 0.7290, 0.6136, 0.5182, 0.4926, 0.4512, 0.3610, 0.3540, 0.2941, 0.2841, 0.1609, 0.0130, 0.0070, 0.0001, 0.0003, which are attained from $i = 15$ onwards in Figure 3.25. ◊

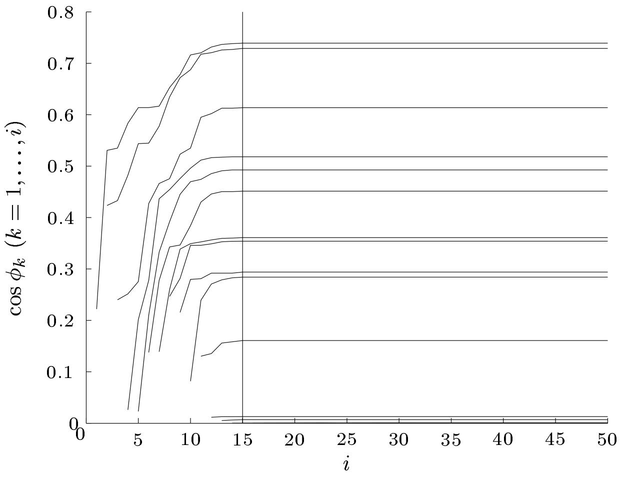


Figure 3.25: The cosines of the principal angles between the row spaces of the $i \times j$ Hankel matrices Y_p and Y_f of a 15th order AR process as a function of i. The number of columns is equal to ∞. From $i = 15$ onwards, the values remain constant.

3.6 Conclusions

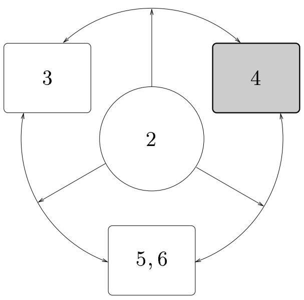
In this chapter, we have first described the model class that we work with: time-invariant linear stochastic models. We have deduced some useful properties for the single-input single-output (SISO) models. Based on the state space description of the model, we have derived parametric expressions for the principal angles and canonical correlations between different past and future input and output pairs. We have also shown that these quantities can be estimated from finite data block Hankel matrices. Furthermore, we have proven how the canonical correlations are related.

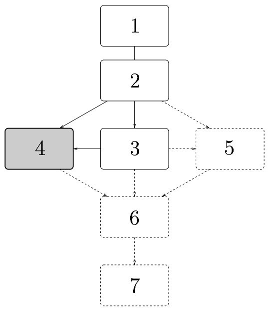
For SISO models, the parametric expressions have been simplified. We have first shown that they do not depend on the input variance. In addition, we have proven that the smallest n canonical correlations of the future input and output process of a SISO model are equal to the cosines of the principal angles between the row spaces of the controllability matrix of the model and the controllability matrix of the inverse model. This is derived from another new property that describes the canonical correlations of the future output processes of two ARMA models that are driven by the same input process, as the cosines of the principal angles between the row spaces of certain controllability matrices. Since the canonical correlations of two such ARMA processes depend on the ratio of the transfer functions, we already see that they are related to cepstral distance measures. This will be clarified in Chapter 6. Finally, we derived an expression for the principal angles between and the canonical correlations of the past and future of an AR process.

The expressions that we have obtained in this chapter will be used in Chapter 4 for the characterization of the entropy rate of the output process of a linear model, and the mutual information of its past and the future. In Chapter 6, we use the results of this chapter to derive different expressions for a cepstral norm and distance. Furthermore, the results obtained here will turn out to be useful in our proof that the cepstral norm in Chapter 6 is proportional to the mutual information of the past and the future of a scalar Gaussian process, which is discussed in Chapter 4.

Chapter 4

Some Concepts from Information Theory





The two preceding chapters both play a central role in this thesis. Chapter 2, which is the center in the left pictogram, introduced the mathematical tools and in Chapter 3, the center of the right picture, we derived expressions that will be applied in this and the following chapters.
This chapter, besides being an immersion into the world of information theory, forms a bridge between Chapter 3 and Chapter 6. For most examples in this chapter that are concerned with stochastic processes, we rely on the output processes of linear stochastic models, of which the assumptions, parameterizations and properties are given in Chapter 3. Furthermore, the canonical correlations of two Gaussian random variables are closely connected to their mutual information, and similarly for the canonical correlations and mutual information of two Gaussian stochastic processes. We can hence directly apply the results of Chapter 3 to obtain expressions for the mutual information of the past and future input and output processes of a linear stochastic model in terms of the model's parameters.

The contents of this chapter are based upon a number of books and papers. The definitions mainly come from the book of Cover and Thomas [29], which is a comprehensive and very understandable introduction to and survey of information theory. Some examples from Khinchin [87] are used to introduce the entropy concept. For the more mathematical details we have consulted the paper by Gel'fand and Yaglom [52] and the books of Pinsker [121] and Ihara [70]. Reza [123] brought clarity, partly because of the abundant historical notes. And of course the pioneering paper of Shannon [130] has been very helpful. This is a very readable paper and a good introduction to the concept of entropy. We also refer to the book of Pierce [120], which is a very easy introduction to the ideas of information theory, giving applications in e.g. language study, psychology and art.

Contributions In this chapter we apply the results of Chapter 3 to the information theoretic notions of entropy and mutual information. In particular for Gaussian random variables and processes, we can use the insights of Chapter 3. In the Gaussian case, the entropy is only a function of the covariance matrix and the mutual information of two Gaussian random variables or processes can be expressed in terms of their canonical correlations.
We focus on the output process of a linear stochastic model that is driven by Gaussian white noise. Based on the methods of Chapter 3, we give in **Section 4.2.2** a proof for the fact that the entropy rate of the Gaussian output process of a linear model only depends on the covariance matrix of the input process. It is not influenced by the other model parameters.
We compute in **Section 4.4.2** the mutual information of the past and the future of the Gaussian output of a linear stochastic model. Applying the expressions for the canonical correlations of the past and the future output in terms of the model parameters, as derived in Chapter 3, we obtain formulas for the mutual information in terms of the state space matrices (A, K, C) and the input covariance matrix S_u .
We show how the mutual information of the past and future can be estimated from measurements. We indicate that the principal angles between the row spaces of the past and future block Hankel matrices are useful for this purpose. For single input single output (SISO) models, we apply the new characterizations of Chapter 3. The results suggest that the mutual information of the past and the future can be viewed as a model norm, which is confirmed in Chapter 6.
From the characterization of the principal angles between the past and future output spaces of a SISO AR model, we derive a formula for the mutual information of the past and future of the corresponding AR process.

Chapter overview We briefly discuss two important notions from information theory: the entropy (in Sections 4.1 and 4.2) and the mutual information (in Sections 4.3 and 4.4). In Section 4.1 the entropy of a random variable is

discussed and in Section 4.2 the entropy of a stochastic process. In Section 4.3 we discuss the mutual information of random variables and Section 4.4 deals with the mutual information of stochastic processes.
Note that the emphasis will be placed on an intuitive understanding rather than on mathematical rigor.

4.1 The entropy of a random variable

We start this section with a motivating introduction to the entropy. Therefore we first concentrate on discrete random variables only. The entropy of a discrete random variable is known as a measure of uncertainty and of information. What is meant by uncertainty and information in this context, is explained in the Sections 4.1.1 and 4.1.2, respectively. We can then formulate some properties that we would like a measure for uncertainty to have, which leads to the definition of the entropy of a discrete random variable in Section 4.1.3. The properties of the entropy for discrete random variables are given in Section 4.1.4.
In Section 4.1.5 we move on to the entropy of a continuous random variable. The definition is now somewhat more cumbersome and the interpretation of the entropy as an uncertainty/information measure is not straightforward. However, as we will see in Section 4.3.3, these objections do not hold for the derived (and for this thesis far more important) notion of mutual information. In Section 4.1.6 we treat the continuous random variables with a Gaussian distribution, for which a simple expression of the entropy is given.

4.1.1 Entropy as a measure for uncertainty

The entropy of a random variable is a measure for its uncertainty, i.e., the uncertainty of the observer about the outcome of an experiment.
Let X be a random variable with a finite number of possible values (the alphabet) $x_1, x_2, \ldots, x_n$ and let $p(x_1), \ldots, p(x_n)$ be

$$\Pr\{X = x_k\} = p(x_k) \ (k = 1, \ldots, n) \ .$$

Example 4.1. Let X be a variable with only two possible values and consider the two alternatives for the probabilities:

$$\left\{ \begin{array}{rcl} p(x_1) & = & 0.5 \\ p(x_2) & = & 0.5 \end{array} \right. \text{and} \left\{ \begin{array}{rcl} p(x_1) & = & 0.99 \\ p(x_2) & = & 0.01 \end{array} \right. .$$

It is obvious that the first scheme represents more uncertainty than the second one since in the second case the result of the experiment will be x_1 in 99 out of 100 trials. As will turn out later, the entropy of the first scheme is larger than the entropy of the second scheme. In fact, the first scheme has the maximum entropy of a discrete random variable with two possible outcomes. $\diamondsuit$

4.1.2 Entropy as a measure of information

If we carry out an experiment, then we obtain information and the uncertainty about the outcome is completely eliminated. Thus, we can say that the information given by carrying out some experiment, consists in removing the uncertainty which existed before the experiment. The entropy of a random variable is consequently not only a measure for the uncertainty but also for the amount of information that can be gained.

4.1.3 Building the definition

Before giving the definition of the entropy of a random variable X, denoted by $H(X)$, we describe four properties that we might expect of a reasonable measure of uncertainty/information.

1. There is no uncertainty if one of the probabilities $p(x_k)$ is equal to 1 and all the others are 0:

$$H(X) = 0 \Leftrightarrow p(x_k) = 1 \text{ and } p(x_j) = 0 \text{ for } j = 1, \ldots, n \text{ and } j \neq k \ .$$

   Otherwise, the uncertainty is strictly positive.

2. For fixed n, the scheme with the highest uncertainty is the one with equally likely outcomes, i.e., $p(x_k) = \frac{1}{n}$ $(k = 1, \ldots, n)$

$$H(X) = H_{\max}(X) \Leftrightarrow p(x_1) = p(x_2) = \cdots = p(x_n) = \frac{1}{n} \ .$$

3. Suppose we have two random variables X with probabilities $p(x_k)$ $(k = 1, \ldots, n)$ and Y with probabilities $q(y_k)$ $(k = 1, \ldots, m)$ that are independent. Then the probability of the joint occurrence of the values x_k and y_j is $p(x_k)q(y_j)$. It is natural to require that the information given by the realization of X and Y is the sum of the two amounts of information given by X and Y. Consequently, we would like the entropy to be additive for independent variables:

$$H(X, Y) = H(X) + H(Y) \text{ for } X \text{ and } Y \text{ independent.}$$

4. It is also obvious that the following property must be met by a reasonable definition of entropy:

$$H(p(x_1), p(x_2), \ldots, p(x_n)) = H(p(x_1), p(x_2), \ldots, p(x_n), 0) \ ,$$

   since the uncertainty is not changed by adding an impossible event.

It can be shown that a function that satisfies the four preceding conditions is of the form

$$H(X) = -\lambda \sum_{k=1}^{n} p(x_k) \log p(x_k) \ ,$$

where λ is a positive constant.
The entropy of a random variable can therefore be defined as follows.

Definition 4.1. The entropy of a discrete random variable with finite alphabet
The entropy $H(X)$ of a discrete random variable X with finite alphabet $\mathcal{X}$ and probability function $p(x) = \Pr\{X = x\}$ for $x \in \mathcal{X}$ is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \ ,$$

with the convention that $0 \log 0 = 0$.

The quantity $-\log p(x_k)$ may be seen as the measure of uncertainty associated with the event $X = x_k$. Therefore, $H(X)$ represents the expected value of the uncertainty associated with p:

$$H(X) = E_p\{-\log p\} \ .$$

The definitions of the joint entropy and the conditional entropy follow then immediately.

Definition 4.2. The joint entropy of two discrete random variables
The joint entropy $H(X, Y)$ of a pair of discrete random variables (X, Y) with alphabet $\mathcal{X}$ and $\mathcal{Y}$, respectively, and joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \ .$$

The joint entropy of X and Y is a measure for the uncertainty about X and Y, or a measure for the information that is obtained by observing X and Y (see Example 4.4).

Definition 4.3. The conditional entropy
The conditional entropy of Y given X, denoted by $H(Y|X)$, of a pair of discrete random variables with alphabet $\mathcal{X}$ and $\mathcal{Y}$, respectively, and joint distribution $p(x, y)$ is defined as

$$H(Y|X) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \ .$$

The conditional entropy of Y given X is a measure for the uncertainty about Y when X is observed, or a measure for the information that is obtained by observing Y when X is already known.

Remark 4.1. The unit of entropy
The base of the logarithmic function determines the unit of the entropy. In communications theory one generally uses logarithms to the base 2, and the entropy is then expressed in bits. Indeed, when making binary codes for the outcomes of a random variable, the expected length (with respect to the probabilities) of the code words, expressed in bits, cannot be smaller than the entropy of the variable.

Example 4.2. Let X be a random variable with the following distribution and codeword assignment:

$$\begin{array}{lcll} p(x_1) & = & \frac{1}{2}, & \text{codeword } C_1 = 0, \\ p(x_2) & = & \frac{1}{4}, & \text{codeword } C_2 = 10, \\ p(x_3) & = & \frac{1}{8}, & \text{codeword } C_3 = 110, \\ p(x_4) & = & \frac{1}{8}, & \text{codeword } C_4 = 111. \end{array}$$

The entropy of X is $-(\frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{8}\log\frac{1}{8}+\frac{1}{8}\log\frac{1}{8}) = 1.75$ bits and the expected length of this code is $\frac{1}{2}+\frac{1}{4}2+\frac{1}{8}3+\frac{1}{8}3$, which is also 1.75 bits. There does not exist a code that has a shorter expected length.
Note that coding with 1 and 0 can also be interpreted as asking binary questions, i.e. questions with two possible answers, say 'yes' and 'no'. Given the probability distribution, the entropy is a lower bound for the average number of binary questions that should be asked to determine the outcome of an experiment. Our strategy for asking questions would for this example be as follows.

Question 1 Is the outcome equal to x_1? We have 50% chance to obtain a 'yes'.

Question 2 Is the outcome equal to x_2? Again the probability for an affirmative answer is $\frac{1}{2}$, since $p(x_2) = p(x_3)+p(x_4)$ (x_1 is not a possible outcome any more).

Question 3 Is the outcome equal to x_3?

If we always ask these questions in this order, the average number of questions that we have to ask before we obtain a 'yes' answer is equal to $1p(x_1)+2p(x_2)+3p(x_3)+3p(x_4) = 1.75$, which is exactly the entropy in bits of the distribution. In general, the strategy is the following. We partition the outcome set into two equiprobable subsets and ask to which subset the outcome belongs. The same procedure is repeated for each subset, and so on until each subset contains only one outcome. Note however, that this equiprobable partitioning might not be possible. In that case, we will need more questions on average than the entropy of the probability distribution, see Example 4.3. $\Diamond$

Example 4.3. The entropy of a loaded die
Assume we have a loaded die with the following probabilities for the six sides:

$$p(1) = \frac{2}{32},\ p(2) = \frac{2}{32},\ p(3) = \frac{8}{32},\ p(4) = \frac{8}{32},\ p(5) = \frac{5}{32},\ p(6) = \frac{7}{32}\ .$$

The entropy of the die is equal to 2.3981 bits. In order to determine the outcome of an experiment in as few binary questions as possible, we propose the strategy depicted in Figure 4.1. The first question is easy, since we can divide the set of possible outcomes in two equiprobable sets: the probability to obtain 3 or 4 is equal to the probability to obtain 1, 2, 5 or 6. But, assume the

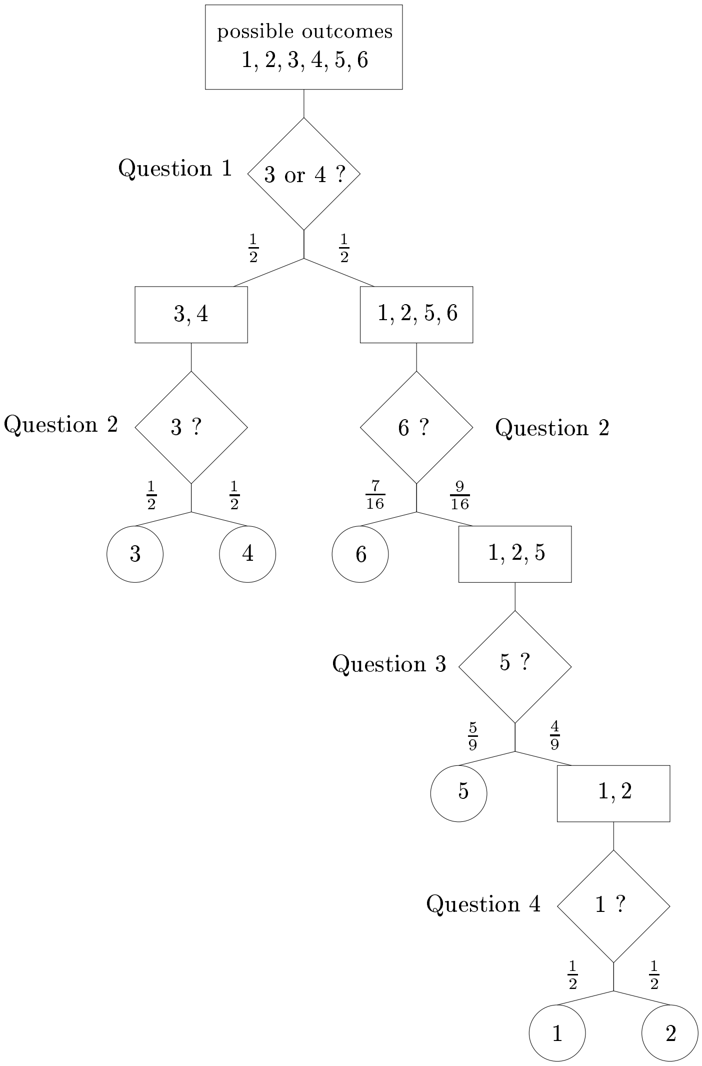


Figure 4.1: Flow chart of the strategy for asking binary questions to determine the outcome of the loaded die in Example 4.3.

answer to question 2 'Is the outcome 3 or 4?' is 'no', then we cannot divide the subset consisting of 1, 2, 5 and 6 into equiprobable subsets. The best we can get is $\frac{7}{16}$ and $\frac{9}{16}$. The average number of binary questions with the strategy of Figure 4.1 is $4\frac{2}{32} + 4\frac{2}{32} + 2\frac{8}{32} + 2\frac{8}{32} + 3\frac{5}{32} + 2\frac{7}{32} = 2.4063$, since we need four questions for the outcomes 1 and 2, which appear both with probability $\frac{2}{32}$. For 3, 4 and 6 we only need two questions and for 5 we need three. The average number of questions is larger than the entropy, but we manage to be close to the lower bound. ◊

One can as well take other bases for the logarithmic function. If the base is e, then the entropy is measured in nats, if the base is 10, then the unit of entropy is the digit. Unless otherwise mentioned, we will use base 2 and express the entropy in bits. ○

Example 4.4. The joint entropy of two discrete random variables
Assume we have two random variables X and Y that each can adopt 4 values. In order to make this example illustrative, we assume that the values are the letters A, C, G and T, which can be thought of as the nucleotides in DNA. The distributions of X and Y are different, which can be seen from the following strings of 30 realizations. Clearly, the probability that Y is equal to T is large.

X	A G T A C G C A G T G C G A A C G G G C A G C A T C C C A G
Y	C T T T T A T T T T T G A T T T T T T C G T G T A T C T T T

If a large number of realizations is given, then the probabilities for X, Y and (X, Y) can be estimated by counting the number of occurrences and dividing by the total number of samples. We use the values with which we have simulated the strings. The joint probabilities $p(x, y)$ are given in Table 4.1. The marginal

$x \setminus y$	A	C	G	T	$p_x(x)$
A	0.01	0.03	0.02	0.19	0.25
C	0.02	0.01	0.03	0.19	0.25
G	0.03	0.04	0.01	0.17	0.25
T	0.01	0.01	0.04	0.19	0.25
$p_y(y)$	0.07	0.09	0.1	0.74	1

Table 4.1: The joint probabilities of (X, Y).

probability functions of the variables X and Y, denoted by $p_x(x)$ and $p_y(y)$ can then be obtained by adding the values in each row and column, respectively (see Table 4.1).
The entropy of X is equal to $H(X) = 2$ bits, the entropy of Y is $H(Y) = 1.2349$ bits and their joint entropy is $H(X, Y) = 3.1850$ bits, which is smaller than the sum of the entropies. This indicates that X and Y are dependent (see (4.3) below and also Section 4.3). ◊

4.1.4 Properties

We now enumerate some useful properties, most of which are easily derived from the previous definitions and elementary rules of probability theory.

1. Non-negativity of the entropy:
$$H(X) \geq 0 \ .$$

2. Relation of entropy, conditional entropy and joint entropy:
$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \ . \tag{4.1}$$

3. Conditioning reduces the entropy:
$$H(X|Y) \leq H(X) \ , \tag{4.2}$$
with equality if and only if X and Y are independent. This property follows from Jensen's inequality [72].
The property says that, on the average, knowing a random variable Y can only reduce the uncertainty in another variable X.

4. Upper bound on joint entropy:
By applying (4.1) to (4.2) we obtain;
$$H(X,Y) \leq H(X) + H(Y) \ , \tag{4.3}$$
with equality if and only if X and Y are independent.

4.1.5 The entropy of a continuous random variable

The original definition of Shannon [130] for the entropy of a continuous random variable is based on the observation that for discrete random variables the entropy is the expected value of $-\log p(x)$ (see Definition 4.1). However, from a mathematical point of view, there are some shortcomings. If the density function $f(x)$ of the continuous random variable is approximated by a discrete scheme and the quantization is made finer and finer, the entropy of the discrete random variable keeps on increasing. In the limit, the entropy of the discretized scheme becomes infinitely large. The interpretation is that the average information in a continuous variable is infinite since there is an infinite (and uncountable) number of possible outcomes.

Remark 4.2. The infinite amount of uncertainty in a continuous random variable can also be interpreted in terms of the binary questions of Example 4.2 and Example 4.3. It would not be possible to determine the exact value of a real number with a finite number of binary questions. However, if we beforehand agree on the degree of accuracy, i.e. we make a finite discretization of the real axis and try to determine in which interval the outcome lies, then we will again only need a finite number of questions. Indeed, the continuous variable is replaced by a discrete approximation. ○

We clarify the above argument in the following example from [123, pp. 270–272].

Example 4.5. The entropy of a continuous random variable
Let X be a continuous random variable with density function $f(x)$ on the interval $[0, a]$. We divide the interval into N subintervals of equal length $\Delta = \frac{a}{N}$ (see Figure 4.2). We can then look at the discrete random variable $\hat{X}$ with values $x_1 = \frac{a}{N}, x_2 = \frac{2a}{N}, \ldots, x_N = a$ and approximate the corresponding probabilities as follows

$$p(x_k) = \Pr\{x_{k-1} < x \leq x_k\} = \int_{x_{k-1}}^{x_k} f(x)\, dx \approx f(x_{k-1})\Delta \ .$$

The entropy of the discrete random variable $\hat{X}$, which is finite, can then be approximated as

$$\begin{aligned} H(\hat{X}) &= -\sum_{k=1}^{N} p(x_k) \log p(x_k) \ , \\ &\approx -\sum_{k=1}^{N} f(x_{k-1})\Delta \log \left(f(x_{k-1})\Delta\right) \ . \end{aligned} \tag{4.4}$$

By making the intervals smaller, the entropy of the discrete random variable will increase. In the limit (for an infinite number of intervals of length 0) we

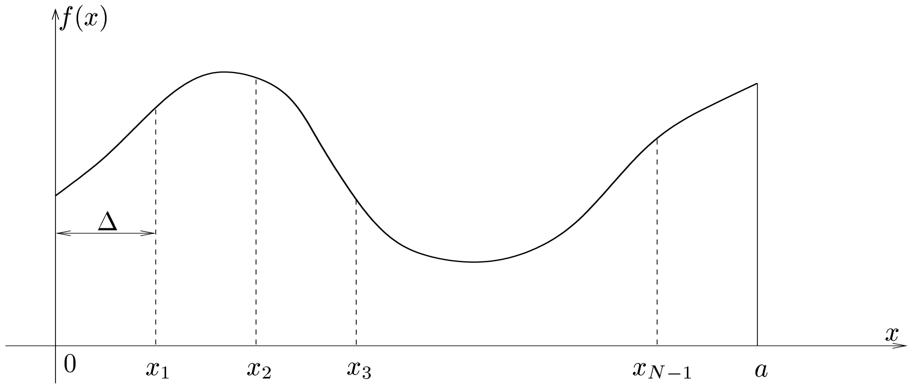


Figure 4.2: The discretization of the probability density function $f(x)$. The points $x_1, x_2, \ldots, x_N = a$ are chosen equidistantly, with distance Δ.

have

$$\lim_{\Delta \to 0} H(\hat{X}) = \lim_{\Delta \to 0} - \sum_{k=1}^{N} f(x_{k-1}) \Delta \log f(x_{k-1})$$
$$- \lim_{\Delta \to 0} \sum_{k=1}^{N} f(x_{k-1}) \Delta \cdot \lim_{\Delta \to 0} \log \Delta \ ,$$

which can further be written as

$$\lim_{\Delta \to 0} H(\hat{X}) = - \int_0^a f(x) \log f(x) \, dx - \int_0^a f(x) \, dx \cdot \lim_{\Delta \to 0} \log \Delta \ ,$$

and since $\int_0^a f(x) \, dx = 1$, we obtain

$$\lim_{\Delta \to 0} H(\hat{X}) = - \int_0^a f(x) \log f(x) \, dx - \lim_{\Delta \to 0} \log \Delta \ . \tag{4.5}$$

The first term in (4.5) is finite, but the second term in (4.5) is equal to ∞. ◊

The limiting process thus leads to an infinite entropy for continuous random variables. As said before, this can be understood by recognizing that the uncertainty about the value of a continuous random variable is infinitely large since an infinite number of values is possible. This would be true if we required an absolutely accurate measurement. In practice, however, only a certain degree of approximation is necessary and/or possible. This means that the entropy of a continuous random variable depends on the discretization level of the probability density, Δ, as in (4.4).

If one still wants an entropy quantity that is independent of the discretization level, the integral $-\int f(x) \log f(x) \, dx$ in the first term of (4.5) can be used. It is called the differential entropy of X and is denoted by $h(X)$. The differential entropy turns out to be useful for comparing the entropy of two continuous random variables. Indeed, let X and Y be continuous random variables with density function $f(x)$ and $g(y)$, respectively, then the derivation above leads to

$$\begin{aligned} \lim_{\Delta \to 0} \left(H(\hat{X}) - H(\hat{Y}) \right) &= - \int f(x) \log f(x) \, dx - \int g(y) \log g(y) \, dy \ , \\ &= h(X) - h(Y) \ , \end{aligned}$$

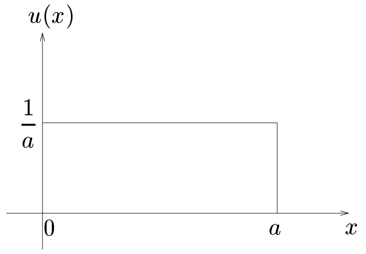
if the same discretization Δ is used for $\hat{Y}$ as for $\hat{X}$. Consequently, the difference between the entropies of X and Y, although both infinitely large, can be measured by the difference of their differential entropies.

Definition 4.4. The differential entropy
Let X be a random variable with probability density function $f(x)$. The differential entropy $h(X)$ is defined as

$$h(X) = - \int f(x) \log f(x) \, dx \ . \tag{4.6}$$

Note that unlike the entropy of a discrete random variable, the differential entropy can be negative.

Example 4.6. Consider, for example, a random variable that is distributed uniformly from 0 to a as in the figure below.



Its density is $\frac{1}{a}$ from 0 to a and 0 elsewhere. Then its differential entropy is

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a}\, dx = \log a \ ,$$

and for $a < 1$, the differential entropy is negative. $\Diamond$

Remark 4.3. The differential entropy of X as the divergence of $f(x)$ to the uniform distribution
The divergence of two probability distributions (also called the relative entropy or the Kullback–Leibler distance) is a term first introduced by Kullback for applications of information theory to statistics [91]. Intuitively speaking, it is a measure for the distance between two probability distributions. It is defined as follows. Let $f(x)$ and $g(x)$ be two distribution functions. Then, the divergence of f with respect to g is equal to

$$H(f;g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx \ .$$

In some books, see e.g. [91, 70], the divergence is used as a basis quantity and the entropy and mutual information are treated as special cases of the divergence.
If the continuous random variable X has a distribution $f(x)$ on a finite support $\mathcal{X} = [0, a]$, as e.g. in Example 4.5, then the divergence of $f(x)$ with respect to the uniform distribution $u(x) = \frac{1}{a}$ (Example 4.6), is equal to

$$\begin{aligned} H(f;u) &= \int_0^a f(x) \log f(x)\, dx - \int_0^a f(x) \log u(x)\, dx \ , \\ &= -h(X) + \log a \ . \end{aligned}$$

Thus we see that the differential entropy $h(X)$ can be interpreted as the relative entropy with respect to the uniform distribution on the same finite interval, except for the additive constant $\log a$ and the multiplicative constant -1. $\bigcirc$

The joint and conditional differential entropy are defined analogously to the discrete versions:

$$h(X,Y) = -\iint f(x,y)\log f(x,y)\,dx\,dy \ ,$$

and

$$h(X|Y) = -\iint f(x,y)\log f(x|y)\,dx\,dy \ .$$

In general, $f(x,y) = f(x|y)f(y) = f(y|x)f(x)$ and thus (4.1) holds also for continuous variables. Furthermore, the inequalities in (4.2) and (4.3) are also valid for the continuous case.
From Definition 4.4 it follows that translation does not change the differential entropy

$$h(X+c) = h(X) \ .$$

A linear transformation of the variable, i.e., a change of coordinates, however, has an influence on the differential entropy of a continuous variable. Let X be an m-component random variable and $A \in \mathbb{R}^{m\times m}$ be a non-singular matrix, then

$$h(AX) = h(X) + \log|\det A| \ .$$

The entropy concept remains important in the continuous case, however, since the derived concept of mutual information does not depend on the coordinate system (see Section 4.3).

4.1.6 The differential entropy of a Gaussian random variable

We now turn to an important example of a continuous variable: the Gaussian distributed random variable. Gaussian random variables are important because of their ubiquitousness in nature. The underlying reason why this distribution is so widespread in nature, is given in the central limit theorem (see e.g. [123, pp. 254–256]). It basically says that observations which are the result of a sum of a large number of random and independent influences have a cumulative distribution function closely approximated by that of a Gaussian random variable. Gaussian random variables are also very easy to work with because they are simply characterized. A complete characterization of Gaussian random variables is obtained by knowing their first and second order moments, viz. the mean and the covariance matrix. Another important property is that linear combinations of Gaussian random variables produce Gaussian random variables. As a consequence, stochastic processes that result from linear filtering of Gaussian stochastic processes are also Gaussian processes.
Due to the translation invariance of the differential entropy, the differential entropy of a Gaussian random variable can be expressed as a function of its covariance matrix only.

Property 4.1. The differential entropy of a Gaussian variable
Let X be an m-component Gaussian random variable with mean $\bar{X}$. If its covariance matrix $Q = E\{(X-\bar{X})(X-\bar{X})^T\}$ is non-singular, then the differential entropy of X is

$$h(X) = \frac{1}{2}\log\left((2\pi e)^m\right) + \frac{1}{2}\log\det Q \ . \tag{4.7}$$

This can be derived by computing the integral in (4.6) where $f(x)$ is the Gaussian distribution function:

$$f(x) = \frac{1}{\sqrt{(2\pi)^m \det Q}} \exp\left(-\frac{1}{2}(x-\bar{X})^T Q^{-1}(x-\bar{X})\right) \ .$$

Example 4.7. The differential entropy of Gaussian distributions
Let X and Y be two scalar ($m = 1$) and zero-mean Gaussian variables with variance $\sigma_x^2 = 0.01$ and $\sigma_y^2 = 1$, respectively. Their distributions, which are denoted by $f(x)$ and $g(y)$, are shown in Figure 4.3. Due to the smaller variance of X, its distribution function is more concentrated around 0 than the distribution of Y. The differential entropy of X is equal to $h(X) = -1.2748$ bits, while $h(Y)$ is equal to 2.0471 bits. Although these values can not be interpreted as information or uncertainty, they can be used to compare the uncertainty of the two variables (see Example 4.5). The uncertainty about the outcomes of X is 3.3219 bits smaller than about the outcomes of Y. Indeed, it is clear from Figure 4.3 that we have more confidence in where x will lie than in the position of y, even though they both have an infinite number of possible values. $\Diamond$

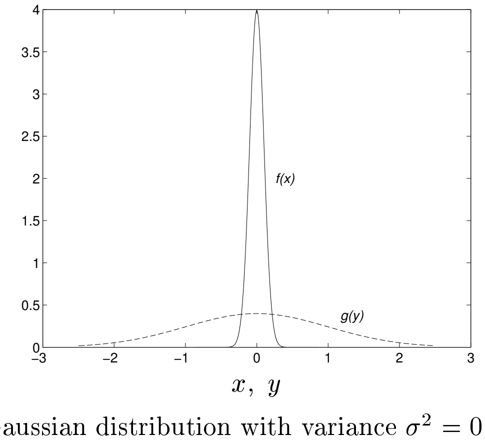


Figure 4.3: The Gaussian distribution with variance $\sigma^2 = 0.01$ in full line and with $\sigma^2 = 1$ in dashed line.

4.2 The entropy rate of a stationary stochastic process

In this section we discuss the entropy concept for stochastic processes. We deal with real-valued stationary stochastic processes that are discrete in time. We start in Section 4.2.1 with two examples from which the definition for the entropy rate of a stochastic process is deduced. In Section 4.2.2 a frequency domain interpretation for the entropy rate of a Gaussian stochastic process is derived.

4.2.1 Examples and definition

Example 4.8. The differential entropy of a Gaussian white stochastic process
Consider the input process $\{u(k)\}$ of Example 3.3 in Section 3.4.2. It is a zero mean stationary white Gaussian noise process with 2 components and covariance matrix

$$E\left\{u(k)u(k)^T\right\} = S_u = \begin{pmatrix} 0.3813 & -0.1512 \\ -0.1512 & 1.3302 \end{pmatrix} .$$

Consider the random vector $U_{0|i-1}$, which is defined as

$$U_{0|i-1} = \begin{pmatrix} U(0) \\ U(1) \\ \vdots \\ U(i-1) \end{pmatrix} .$$

It is a Gaussian random variable with $2i$ components. The differential entropy of $U_{0|i-1}$ can be computed by means of (4.7):

$$h(U_{0|i-1}) = \frac{1}{2}\log\left((2\pi e)^{2i}\right) + \frac{1}{2}\log\det Q_{u_i} ,$$

where

$$Q_{u_i} = E\left\{U_{0|i-1}U_{0|i-1}^T\right\} = \mathrm{diag}(S_u, S_u, \ldots, S_u) .$$

Hence,

$$h(U_{0|i-1}) = i\log(2\pi e) + \frac{1}{2}i\log\det S_u = i\log\left(2\pi e\sqrt{\det S_u}\right) . \tag{4.8}$$

The differential entropy of $U_{0|i-1}$ grows linearly with i.
In Figure 4.4 the evolution of the differential entropy $h(U_{0|i-1})$ as a function of i is shown in the dotted line. Assume that we obtain one new sample per second. Since $\det S_u = 0.4843$, the slope of the line is $\log(2\pi e\sqrt{\det S_u}) \approx 3.5712$ bits per second. $\Diamond$

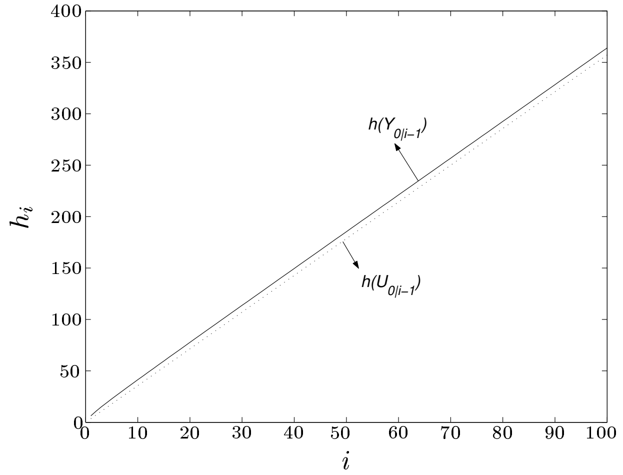


Figure 4.4: The differential entropy of the random vector $U_{0|i-1}$, composed of subsequent variables of the white noise process $\{u(k)\}$ of Example 4.8 is shown as a function of i in dotted line. The differential entropy of $Y_{0|i-1}$, derived from the output process $\{y(k)\}$ in Example 4.9 is given in full line. Note that they are not equal. However, after some time the lines have more or less the same slope (see also Figure 4.5).

Example 4.9. The differential entropy of a non-white Gaussian process

Let us now consider the 2-component output process $\{y(k)\}$ of the model in Example 3.3. Since the input $\{u(k)\}$ is a Gaussian process, $\{y(k)\}$ is Gaussian too. The differential entropy of the random variable $Y_{0|i-1}$, defined as

$$Y_{0|i-1} = \begin{pmatrix} Y(0) \\ Y(1) \\ \vdots \\ Y(i-1) \end{pmatrix} ,$$

can be computed as (see (4.7))

$$h(Y_{0|i-1}) = i \log(2\pi e) + \frac{1}{2} \log \det Q_{y_i} ,$$

where Q_{y_i} is the covariance matrix of $Y_{0|i-1}$. The full line in Figure 4.4 represents the evolution of $h(Y_{0|i-1})$ for increasing i. After some time the slope of $h(Y_{0|i-1})$ as a function of i becomes equal to the slope of $h(U_{0|i-1})$ (see further). $\diamond$

For stationary processes, as $\{u(k)\}$ and $\{y(k)\}$, the differential entropy of the variables $U_{0|i-1}$ and $Y_{0|i-1}$ is independent of the starting moment. Thus, we

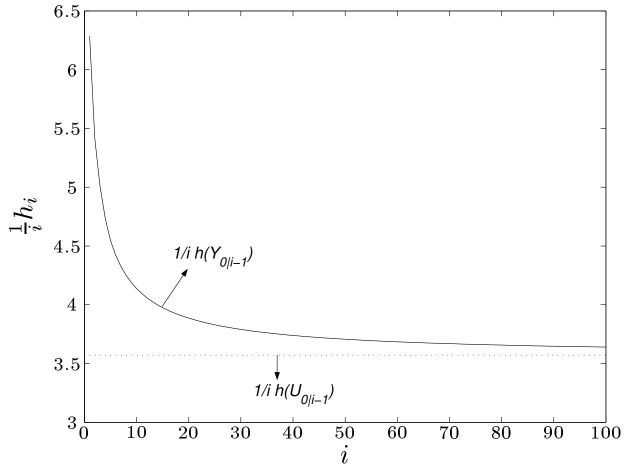


Figure 4.5: The differential entropy h_i divided by i as a function of i. The dotted line represents $\frac{1}{i}h(U_{0|i-1})$, and the full line shows $\frac{1}{i}h(Y_{0|i-1})$. For the white noise case, the value $\frac{1}{i}h(U_{0|i-1})$ is independent of i and it is equal to 3.5712 bits per second. The entropy per sample of the output process converges to this value.

can say that the sequence of i random variables gives a well defined amount of information h_i that depends only on i and on the process. On average, the amount of information per symbol emitted is $\frac{1}{i}h_i$. This quantity is shown in Figure 4.5 for the two processes $\{u(k)\}$ (dotted line) and $\{y(k)\}$ (full line) of Example 4.8 and Example 4.9, respectively. Due to the linear growth of the differential entropy of the white noise process $\{u(k)\}$ (see (4.8) and Figure 4.4), we obtain a constant curve for $\frac{1}{i}h(U_{0|i-1})$ at 3.5712 bits per second. We clearly see that the differential entropy of $\{y(k)\}$ does not increase linearly, but asymptotically, i.e. for $i \to \infty$, it does. The slope of $h(Y_{0|i-1})$ is equal to the slope of $h(U_{0|i-1})$ for $i \to \infty$.
It is consequently natural to call the quantity $\lim_{i\to\infty} \frac{1}{i}h_i$ the entropy rate of the stochastic process, i.e., the average amount of information conveyed by one emitted symbol.

Definition 4.5. The entropy rate of a stationary stochastic process
The entropy rate of a stationary stochastic process $\{x(k)\}$ is defined as

$$h(x) = \lim_{i\to\infty} \frac{1}{i} h(X_{0|i-1}) ,$$

if the limit exists.

It can be shown [87, p. 47–48] that the entropy rate exists for every stationary stochastic process.

4.2.2 The entropy rate of a stationary Gaussian process

The entropy rate of a zero mean m-component stationary Gaussian process $\{y(k)\}$ is equal to

$$h(y) = \frac{m}{2}\log(2\pi e) + \lim_{i\to\infty}\frac{1}{2i}\log\det Q_{y_i}\ , \tag{4.9}$$

where Q_{y_i} is the $mi \times mi$ covariance matrix of $Y_{0|i-1}$.
Let $\{y(k)\}$ be the output process of a linear system with Gaussian white noise input $\{u(k)\}$. In Example 4.9, we have seen that $\frac{1}{i}h(Y_{0|i-1})$ converges to $\frac{1}{i}h(U_{0|i-1})$ (see Figure 4.5). This is in general true: the entropy rate of the output process $\{y(k)\}$ is equal to the entropy rate of its white noise input $\{u(k)\}$. This is stated in the following proposition.

Proposition 4.2. The entropy rate of the output process of a linear stochastic model driven by zero-mean white Gaussian noise
Let $\{y(k)\}$ be the m-component output process of a linear stochastic model satisfying the conditions in Section 3.2.1, of which the input is a zero mean white Gaussian noise with covariance matrix S_u. The entropy rate of $\{y(k)\}$ is then equal to

$$h(y) = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\log\det S_u\ . \tag{4.10}$$

Proof.
Let the linear model be of order n with state space matrices (A, K, C). The covariance matrix $Q_{y_i} = E\left\{Y_{0|i-1}Y_{0|i-1}^T\right\}$ is equal to

$$Q_{y_i} = \Gamma_i\Sigma\Gamma_i^T + H_iQ_{u_i}H_i^T\ ,$$

where $\Gamma_i \in \mathbb{R}^{mi\times n}$ is the observability matrix (3.21), Σ the state covariance matrix (3.7), H_i the block Toeplitz matrix with the Markov parameters (3.30) and Q_{u_i} the input covariance matrix

$$Q_{u_i} = E\left\{U_{0|i-1}U_{0|i-1}^T\right\} = \mathrm{diag}(S_u, \ldots, S_u)\ . \tag{4.11}$$

The covariance matrix Q_{y_i} can be rewritten as

$$Q_{y_i} = (\Gamma_i\Sigma\Gamma_i^TH_i^{-T}Q_{u_i}^{-1}H_i^{-1} + I_i)H_iQ_{u_i}H_i^T\ ,$$

and its determinant is consequently equal to

$$\det Q_{y_i} = \det(\Gamma_i^TH_i^{-T}Q_{u_i}^{-1}H_i^{-1}\Gamma_i\Sigma + I_n)\det(H_iQ_{u_i}H_i^T)\ .$$

By including $\Gamma_{z_i} = -H_i^{-1}\Gamma_i$, the observability matrix of the inverse model, and applying $\det H_i = 1$ and (4.11), we obtain

$$\begin{aligned}\det Q_{y_i} &= \det(\Gamma_{z_i}^TQ_{u_i}^{-1}\Gamma_{z_i}\Sigma + I_n)(\det S_u)^i\ ,\\ &= \det(\mathcal{G}_{z_i}\Sigma + I_n)(\det S_u)^i\ ,\end{aligned}$$

where $\mathcal{G}_{z_i} = \Gamma_{z_i}^T Q_{u_i}^{-1} \Gamma_{z_i}$ (see also (3.70)). Since $\det(\mathcal{G}_{z_i}\Sigma + I_n)$ converges for $i \to \infty$ to the finite value $\det(\mathcal{G}_z\Sigma + I_n)$ (see (3.75)), we have

$$\lim_{i\to\infty} \frac{1}{i} \log\det Q_{y_i} = 0 + \log\det S_u \ .$$

Consequently, the entropy rate of $\{y(k)\}$ (4.9) is equal to

$$h(y) = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\log\det S_u \ .$$

□

Proposition 4.2 also follows from the following characterization of the entropy rate of a Gaussian process in terms of its spectral density (see e.g. [25, pp. 44–45], [59] and particularly interesting is [60]) Let $R(\tau)$ be the autocovariance function of $\{y(k)\}$, then Q_{y_i} is equal to the following symmetric block Toeplitz matrix (see also Section 2.6.3)

$$Q_{y_i} = \begin{pmatrix} R(0) & R(1)^T & \cdots & R(i-1)^T \\ R(1) & R(0) & \cdots & R(i-2)^T \\ \vdots & & \ddots & \\ R(i-1) & R(i-2) & \cdots & R(0) \end{pmatrix} \in \mathbb{R}^{mi\times mi} \ .$$

Let Φ be the spectral density matrix of the Gaussian process:

$$\Phi(e^{i\theta}) = \sum_{\tau=-\infty}^{\infty} R(\tau)e^{-i\tau\theta} \ .$$

Then, from the Szegő formula [63, p. 65], it follows that

$$\lim_{i\to\infty} \frac{1}{i}\log\det Q_{y_i} = \frac{1}{2\pi}\int_0^{2\pi} \log\det\Phi(e^{i\theta})\,d\theta \ .$$

Hence, the entropy rate of $\{y(k)\}$ is

$$h(y) = \frac{m}{2}\log(2\pi e) + \frac{1}{4\pi}\int_0^{2\pi} \log\det\Phi(e^{i\theta})\,d\theta \ . \tag{4.12}$$

If $\{y(k)\}$ is the output of a linear model with input covariance matrix S_u, then (4.12) reduces to (4.10).

Remark 4.4. The entropy rate of a stochastic process can thus not be used as a measure for the dynamic content of the process. It is not even a function of the model, it only depends on the variance of the input noise. We will see in Section 4.4.2 that in contrast to the entropy rate, the mutual information between the past and the future of a scalar stochastic process does not depend on the input variance and it does make a distinction between uncorrelated and correlated processes. ○

4.3 The mutual information of random variables

This section has the same structure as the section on the entropy of random variables. We start with discrete random variables, for which we give the definition of the mutual information in Section 4.3.1 and some properties in Section 4.3.2. Next, we deal with the mutual information of continuous random variables in Section 4.3.3. The important special case of Gaussian random variables is treated in Section 4.3.4. Their mutual information can be expressed in terms of the principal angles between certain linear subspaces, a relation that we derive in detail.

4.3.1 Definition

The definition of the mutual information of two discrete random variables follows very naturally from the inequality (4.2) in Section 4.1, $H(X|Y) \leq H(X)$, with equality if and only if X and Y are independent. If X and Y are mutually dependent, then observing Y reduces the uncertainty in X, or otherwise formulated, we gain information about X by observing Y. Thus, the variable Y must contain information about X, the amount of which is equal to $H(X) - H(X|Y)$. This is denoted by $I(X;Y)$. Since $H(X) - H(X|Y) = H(Y) - H(Y|X)$ (see Equation (4.1)), we have

$$I(X;Y) = I(Y;X) \; .$$

Hence, we can call $I(X;Y)$ the mutual information of X and Y.
Observe that by Equation (4.1), $H(X) - H(X|Y)$ is also equal to $H(X) + H(Y) - H(X,Y)$, in which we recognize the terms of the inequality in (4.3). Consequently,

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \; ,$$

and by using Definition 4.1 and Definition 4.2, we can express $I(X;Y)$ in terms of the probability functions as follows.

Definition 4.6. The mutual information of two discrete random variables
Consider the discrete random variables X and Y with joint probability function $p(x,y)$ and marginal probability function $p_x(x)$ and $p_y(y)$. The mutual information of X and Y is defined by

$$I(X;Y) = \sum_{\mathcal{X}} \sum_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p_x(x) p_y(y)} \; ,$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the alphabets of X, respectively Y.

4.3.2 Properties of the mutual information of two discrete random variables

A number of properties of the mutual information in relation to the entropy are given below. Some of the properties can be represented in the Venn diagram in Figure 4.6, which offers an intuitive feeling for their meaning.

1. Symmetry:

$$I(X;Y) = I(Y;X) \ .$$

   The variable X contains as much information about Y as Y about X.

2. Non-negativity:
   The mutual information of two random variables is always positive:

$$I(X;Y) \geq 0 \ ,$$

   with equality if and only if X and Y are independent.
   In contrast to the non-negativity of the entropy, this property also holds for continuous random variables.

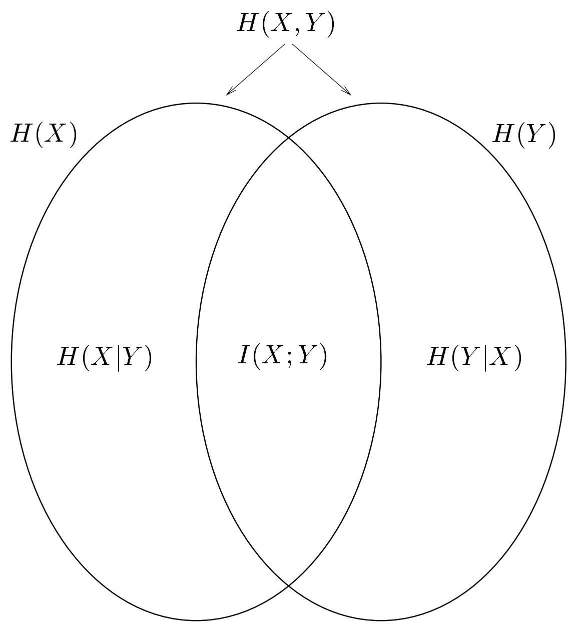


Figure 4.6: The relation of the mutual information and the entropy of random variables

3. Relation of the mutual information and the entropy:

$$I(X;Y) \leq I(X;X) = H(X) \ , \tag{4.13}$$

and also $I(X;Y) \leq H(Y)$. Since $I(X;X) = H(X)$, the entropy is sometimes called the self-information of a random variable.

4. Relation of the mutual information and the conditional entropy:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \ . \tag{4.14}$$

The mutual information of two variables is the reduction in the uncertainty of one of the variables due to the knowledge of the other one.

5. Relation of the mutual information and the joint entropy:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \ . \tag{4.15}$$

Note that the relation (4.15) and its representation in Figure 4.6 is completely analogous to the dimension theorem of Grassmann, which was given in Chapter 2, Theorem 2.2:

$$\dim(\text{range}(A) \cap \text{range}(B)) = \text{rank}(A) + \text{rank}(B) - \text{rank}(\begin{pmatrix} A & B \end{pmatrix}) \ ,$$

where A and B are matrices with the same number of rows.

4.3.3 The mutual information of continuous random variables

Although the differential entropy of a continuous variable, $h(X)$, does not have a direct interpretation as far as information content is concerned (see Section 4.1.5), the mutual information retains its information theoretic significance for continuous random variables.
Several definitions have been proposed for the mutual information of continuous random variables [130, 52, 121]. We will use the following.

Definition 4.7. The mutual information of continuous random variables
Let X and Y be random variables with joint probability density function $f(x,y)$ and marginal densities $f_x(x)$ and $f_y(y)$, respectively. Then, the mutual information of X and Y is defined as

$$I(X;Y) = \iint f(x,y) \log \frac{f(x,y)}{f_x(x) f_y(y)} \, dx \, dy \ ,$$

if the integral exists.

For most cases of interest, the integral exists. However, when computing the mutual information of jointly Gaussian variables (see Section 4.3.4), one could also deal with singular Gaussian distributions. A more general definition for the mutual information should then be used, e.g. the definition of Gel'fand and Yaglom [52], which is based on the discretization process that we have discussed in connection with the entropy of continuous random variables (Section 4.1.5).
We now mention some properties of the mutual information for continuous variables. The following properties, which were derived for discrete random variables, also hold for continuous random variables:

$$
\begin{aligned}
&I(X;Y) = I(Y;X) \ , \\
&I(X;Y) \geq 0 \ , \ \text{with equality if and only if } X \text{ and } Y \text{ are independent} \\
&I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \ , && (4.16) \\
&I(X;Y) = h(X) + h(Y) - h(X,Y) \ . && (4.17)
\end{aligned}
$$

Note that we have used in (4.16) and (4.17) the differential entropy from Definition 4.4. By applying the discretization process explained in Section 4.1.5, we also obtain these formulas because the infinities cancel out (see [123, p. 276]).

Remark 4.5. The self-information of a continuous random variable
As we have seen in (4.13), the mutual information of a discrete random variable and itself (its self-information), is equal to the entropy of the variable. This is also the case for continuous random variables. Note, however, that the entropy of a continuous random variable is infinitely large. For the continuous random variable X holds

$$I(X;X) = H(X) = \infty \ , \tag{4.18}$$

while the differential entropy of X, $h(X)$, can be finite. In contrast to the entropy of discrete random variables, the differential entropy can not be seen as a measure for the self-information. ○

Two additional properties should be mentioned. Apart from their general importance, they will be needed for the characterization of the mutual information of Gaussian random variables in terms of principal angles in Section 4.3.4. They both hold even if the integral in Definition 4.7 does not exist.
Contrary to the differential entropy, the mutual information is independent of the choice of bases for the coordinates of X and Y, more specifically, the following property holds:

Property 4.3. The influence of a linear transformation on the mutual information
Let X be an m-component random vector and $T \in \mathbb{R}^{m \times m}$ a linear transformation. Then,

$$
\begin{aligned}
I(X;Y) &= I(TX;Y) \quad \textit{if } T \textit{ is non-singular} \\
&> I(TX;Y) \quad \textit{if } T \textit{ is singular} \ .
\end{aligned}
$$

Property 4.3 also holds for more general transformations:

$$I(X;Y) \geq I(t(X);Y) \ ,$$

with equality if t is one-to-one [52, pp. 209, 214].
Property 4.4 says that the mutual information function is additive for independent pairs [121, p. 11].

Property 4.4. The joint mutual information of independent variables
Let X_1, X_2, Y_1 and Y_2 be continuous random variables so that the pair (X_1, Y_1) is independent of (X_2, Y_2). Then,

$$I(X_1, X_2; Y_1, Y_2) = I(X_1; Y_1) + I(X_2; Y_2) \ .$$

4.3.4 The mutual information of two Gaussian random variables

We now consider random variables that are jointly Gaussian distributed. For the Gaussian case, an explicit expression for $I(X;Y)$ can be obtained quite simply. Moreover, it can also be expressed in terms of the principal angles between certain subspaces (see Section 2.5).

Property 4.5. The mutual information of Gaussian variables
Let X be a p-component random variable and Y a q-component random variable that are zero mean and jointly Gaussian. The covariance matrix of X, Y and (X, Y) is denoted by Q_x, Q_y and $Q = \begin{pmatrix} Q_x & Q_{xy} \\ Q_{yx} & Q_y \end{pmatrix}$, respectively, which we assume to be non-singular. The mutual information of X and Y is given by

$$I(X;Y) = -\frac{1}{2} \log \frac{\det Q}{\det Q_x \det Q_y} \ . \tag{4.19}$$

This follows from the expression for the entropy of a Gaussian variable (4.7) and Equation (4.17).
We recall that the covariance matrices Q_x, Q_{xy}, Q_{yx} and Q_y can be viewed as matrices of inner products (see (2.34)). The following theorem relates the mutual information of the random variables X and Y to the principal angles between the row spaces of the matrices $X^{(j)}$ and $Y^{(j)}$ (for $j \to \infty$) that are defined in (2.33).

Theorem 4.6. The mutual information of Gaussian variables in terms of principal angles [52]
Let X be a p-component random variable and Y a q-component random variable that are zero mean and jointly Gaussian. The covariance matrices of X, Y and (X, Y) are denoted by Q_x, Q_y and $Q = \begin{pmatrix} Q_x & Q_{xy} \\ Q_{yx} & Q_y \end{pmatrix}$ respectively, which we

assume to be non-singular. The mutual information of X and Y is equal to

$$I(X;Y) = -\frac{1}{2}\log \prod_{k=1}^{\min(p,q)} \sin^2\phi_k = -\sum_{k=1}^{\min(p,q)} \log\sin\phi_k \ , \tag{4.20}$$

where the angles ϕ_k ($k = 1,\dots,\min(p,q)$) are the principal angles between the row spaces of $X^{(j)}$ and $Y^{(j)}$, for $j \to \infty$.

Proof.
The determinant of Q can be written in terms of its submatrices:

$$\det Q = \det Q_x \det(Q_y - Q_{yx}Q_x^{-1}Q_{xy}) \ .$$

Substituting this expression in (4.19), gives

$$I(X;Y) = -\frac{1}{2}\log\frac{\det(Q_y - Q_{yx}Q_x^{-1}Q_{xy})}{\det Q_y} \ ,$$

and this is also equal to

$$\begin{aligned} I(X;Y) &= -\frac{1}{2}\log\det\left(Q_y^{-1}(Q_y - Q_{yx}Q_x^{-1}Q_{xy})\right) \ , \\ &= -\frac{1}{2}\log\det(I_q - Q_y^{-1}Q_{yx}Q_x^{-1}Q_{xy}) \ . \end{aligned}$$

The largest $\min(p,q)$ eigenvalues of $Q_y^{-1}Q_{xy}Q_x^{-1}Q_{yx}$ are the squared canonical correlations of X and Y and the other eigenvalues are zero. As we have seen in Section 2.6.2, the canonical correlations are the squared cosines of the principal angles ϕ_k ($k = 1,\dots,\min(p,q)$) between the row spaces of $X^{(j)}$ and $Y^{(j)}$, for $j \to \infty$.
Hence,

$$\begin{aligned} I(X;Y) &= -\frac{1}{2}\log\prod_{k=1}^{\min(p,q)}(1-\cos^2\phi_k) \ , \\ &= -\frac{1}{2}\log\prod_{k=1}^{\min(p,q)}\sin^2\phi_k \ . \end{aligned}$$

□

Remark 4.6. The mutual information of singular Gaussian distributions
Based on Property 4.3 and Property 4.4, Gel'fand and Yaglom show in [52] that even for singular Gaussian distributions the relation (4.20) is valid. They start from a more general definition for the mutual information of continuous random variables. ○

Remark 4.7. The mutual information of X and Y as the sum of the mutual informationof their canonical variates
Since each principal angle ϕ_k between the row spaces of $X^{(j)}$ and $Y^{(j)}$ is the angle between the corresponding principal directions u_k and v_k, the expression in (4.20) can also be viewed as the sum of the mutual information of the different independent canonical variate pairs:

$$I(X;Y) = \sum_{k=1}^{\min(p,q)} I(U_k;V_k) \ , \tag{4.21}$$

where $I(U_k;V_k) = -\log \sin \phi_k$. ○

Let us now investigate the influence of the principal angles $\phi_1, \ldots, \phi_r$ (where $r = \min(p,q)$) on the mutual information of X and Y. If one of the angles is equal to zero, then the mutual information of X and Y will be infinitely large and vice versa:

$$I(X;Y) = \infty \Leftrightarrow \phi_1 = 0 \ ,$$

where we use the convention that the principal angles are ordered in non-decreasing order (see Definition 2.4). This can be understood as follows. If $\phi_1 = 0$, then $U_1 = V_1$. From (4.18) we know that $I(U_1;U_1) = \infty$. Hence, one of the terms in (4.21) is ∞, which explains that $I(X;Y) = \infty$. In this case the covariance matrix Q would not be of full rank, although Q_x and Q_y might be full rank matrices, see also Grassmann's dimension theorem (Theorem 2.2).
A small angle ϕ_k between two canonical variates U_k and V_k indicates that they are linearly closely related and their mutual information is therefore large. If the angle becomes larger, the mutual information $I(U_k;V_k) = -\log \sin \phi_k$ of the pair of variates decreases. If the principal angle ϕ_k becomes $\frac{\pi}{2}$, then U_k and V_k do not have any mutual information. The canonical variates are independent. Here we see that two scalar Gaussian random variables are independent if and only if they are uncorrelated (or orthogonal) while this is not the case in general (see also Section 2.6.1 where we discuss the relation of correlation and orthogonality). Furthermore, the necessary and sufficient condition for two multivariate Gaussian random variables X and Y to be independent is

$$I(X;Y) = 0 \Leftrightarrow \phi_1 = \phi_2 = \cdots = \phi_r = \frac{\pi}{2} \ ,$$

which is also equivalent to $\mathrm{row}(X^{(j)})$ and $\mathrm{row}(Y^{(j)})$ being orthogonal for $j \to \infty$:

$$\lim_{j\to\infty} X^{(j)} Y^{(j)^T} = 0 \text{ or } E\left\{XY^T\right\} = 0 \ .$$

4.4 The mutual information of stochastic processes

In analogy to the definition of the entropy rate of stochastic processes in Section 4.2, we can define the mutual information rate of stochastic processes. This is discussed in Section 4.4.1. In Section 4.4.2 the mutual information of the past and the future of a stochastic process is computed. We will encounter the notion of mutual information of past and future several times in this thesis. Therefore, we will treat it here in more detail than the mutual information rate.

4.4.1 The mutual information rate of two stochastic processes

We start with an example for which the mutual information of two vectors of random variables, derived from the stochastic processes, grows asymptotically linearly with the vector length. The definition of the mutual information rate of two stochastic processes follows then naturally. We will be short on this aspect of information theory, since we will not need it often throughout the thesis. More details can be found in [52], [121] and [123].

Example 4.10. We compute the mutual information of the Gaussian stochastic processes $\{y(k)\}$ of Examples 3.3–3.9 and Example 4.9, and $\{z(k)\}$, which is obtained by corrupting $\{y(k)\}$ with white noise. This leads to a definition of the mutual information rate of two stochastic processes. Consider the stochastic process $\{z(k)\}$ which is equal to the Gaussian output process $\{y(k)\}$ plus an independent white noise process $\{e(k)\}$ with covariance matrix $S_e = \begin{pmatrix} 3 & 0 \\ 0 & 0.5 \end{pmatrix}$:

$$z(k) = y(k) + e(k) \ .$$

In communication theory this is known as the memory-less Gaussian channel [29, p. 239]. We now compute the mutual information of

$$Y_{0|i-1} = \begin{pmatrix} Y(0) \\ Y(1) \\ \vdots \\ Y(i-1) \end{pmatrix} \text{ and } Z_{0|i-1} = \begin{pmatrix} Z(0) \\ Z(1) \\ \vdots \\ Z(i-1) \end{pmatrix} \ .$$

Because $Y_{0|i-1}$ and $Z_{0|i-1}$ are jointly Gaussian random variables, their mutual information is equal to (see Property 4.5)

$$I(Y_{0|i-1}; Z_{0|i-1}) = -\frac{1}{2} \left(\log \det(Q_i) - \log \det(Q_{y_i}) - \log \det(Q_{z_i}) \right) \ ,$$

where the covariance matrix Q_i is equal to

$$Q_i = \begin{pmatrix} Q_{y_i} & Q_{yz_i} \\ Q_{zy_i} & Q_{z_i} \end{pmatrix} = E\left\{ \begin{pmatrix} Y_{0|i-1} \\ Z_{0|i-1} \end{pmatrix} \begin{pmatrix} Y_{0|i-1}^T & Z_{0|i-1}^T \end{pmatrix} \right\} .$$

The mutual information of $Y_{0|i-1}$ and $Z_{0|i-1}$ is also equal to (Theorem 4.6)

$$I(Y_{0|i-1}; Z_{0|i-1}) = -\frac{1}{2} \log \prod_{k=1}^{2i} \sin^2 \phi_k ,$$

where ϕ_k $(k = 1, \dots, 2i)$ are the principal angles between the row spaces of the $2i \times j$ block Hankel matrices $Y_{0|i-1}^{(j)}$ and $Z_{0|i-1}^{(j)}$ for $j \to \infty$. There are indeed $2i$ principal angles between the subspaces, since $Y_{0|i-1}^{(j)}$ is of full row rank (see Proposition 3.1) and, as is easily shown, the rank of $Z_{0|i-1}^{(j)}$ is also equal to $2i$. In the top picture of Figure 4.7 the sines of the principal angles are shown

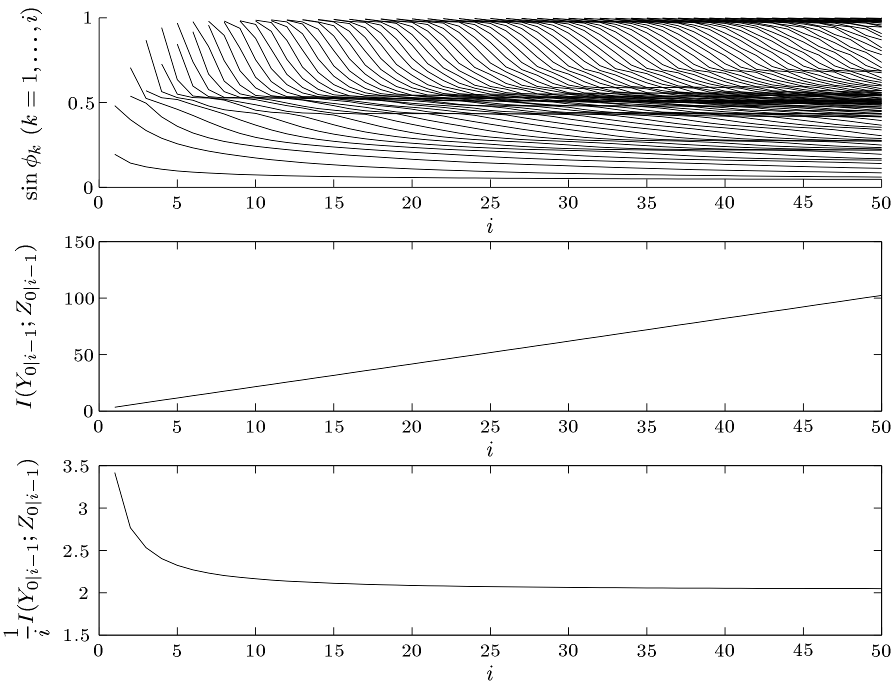


Figure 4.7: The top picture shows the sines of the principal angles between the row spaces of the $2i \times 10\,000$ block Hankel matrices $Y_{0|i-1}^{(j)}$ and $Z_{0|i-1}^{(j)}$ for increasing i. The approximated mutual information $I(Y_{0|i-1}; Z_{0|i-1})$ is given as a function of i in the middle. The bottom picture shows $\frac{1}{i} I(Y_{0|i-1}; Z_{0|i-1})$.

as a function of i. They are computed from the simulation data by means of the algorithm described in Section 2.5.6. The number of columns in the block Hankel matrices is equal to $j = 10\,000$. The picture in the middle gives the mutual information $I(Y_{0|i-1}; Z_{0|i-1})$, which is computed from the principal angles. This is an approximation, since j is finite. The mutual information of $Y_{0|i-1}$ and $Z_{0|i-1}$ grows for increasing i. Therefore, the fraction $\frac{1}{i}I(Y_{0|i-1}; Z_{0|i-1})$ is depicted in the bottom picture. This quantity is a measure for the average amount of information that we gain about y per symbol of z that is observed. We see that asymptotically, i.e. for $i \to \infty$, the mutual information of $Y_{0|i-1}$ and $Z_{0|i-1}$ increases linearly with i. $\Diamond$

Definition 4.8. The mutual information rate of two stochastic processes
Let $\{x(k)\}$ and $\{y(k)\}$ be discrete-time stochastic processes. The mutual information rate of $\{x(k)\}$ and $\{y(k)\}$ is defined by

$$I_r(x; y) = \lim_{i\to\infty} \frac{1}{i} I(X_{0|i-1}; Y_{0|i-1}) \ .$$

Several other definitions exist for mutual information rate of stochastic processes, see e.g. [121, p. 76].

Remark 4.8. The mutual information of two Gaussian processes cannot decrease with i
Let $\{y(k)\}$ and $\{z(k)\}$ be m-component Gaussian stationary processes. Then, the mutual information between $Y_{0|i-1}$ and $Z_{0|i-1}$ is equal to

$$I(Y_{0|i-1}; Z_{0|i-1}) = \frac{1}{2}\sum_{k=1}^{mi}(-\log \sin^2 \phi_k) \ ,$$

where $\phi_1, \ldots, \phi_{mi}$ are the principal angles between the row spaces of the block Hankel matrices Y and Z, provided these matrices are of full rank. Otherwise, less than mi angles are defined. Since $-\log \sin^2 \phi_k \geq 0 \ \forall k$, it is clear that

$$I(Y_{0|i-1}; Z_{0|i-1}) \geq I(Y_{0|i-2}; Z_{0|i-2}) \ .$$

In Example 4.10, the mutual information keeps increasing. There also exist, however, processes of which the mutual information remains constant (see Example 4.11) and processes of which the mutual information converges to a finite value for increasing i (Example 4.12). Furthermore, for some processes, the mutual information remains constant from a certain i onwards, which we will see in Example 4.14. $\bigcirc$

4.4.2 The mutual information of the past and the future of a stochastic process

The problem of characterizing the dependence of the future of a stationary process on its own past has received a great deal of attention because of its implications for the prediction and ergodic theories of Gaussian processes (see [69, 73, 74]). A model reduction technique based on the mutual information of the past and the future, which is also used in subspace identification algorithms [159], has been proposed by Desai and Pal [42], while Li and Xie have used this concept for model selection and order determination problems [95]. In this section, we first look at a few examples and give the definition of the mutual information of the past and the future of a stochastic process. Next, we give expressions for the mutual information of the past and the future for three more specified stochastic processes.

Example 4.11. The mutual information of the past and future of white noise
Let $\{u(k)\}$ be a scalar Gaussian white noise process with variance $\sigma^2 = 4$. The past and the future processes of $\{u(k)\}$ are jointly Gaussian. Hence, we can compute their mutual information as follows. Let $U_{0|i-1}$ and $U_{i|2i-1}$ be the past and the future random vectors, which are Gaussian random variables. Their mutual information is equal to (see Property 4.5)

$$I(U_{0|i-1}; U_{i|2i-1}) = -\frac{1}{2} \log \frac{\det Q_{2i}}{\det Q_{p_i} \det Q_{f_i}} ,$$

where Q_{2i} is the covariance matrix of $U_{0|2i-1}$, the matrix Q_{p_i} is the covariance matrix of the past random variable $U_{0|i-1}$ and Q_{f_i} is the covariance matrix of the future random variable $U_{i|2i-1}$. These matrices are all diagonal and their diagonal elements are all equal to σ^2. Consequently, the mutual information of the length-i past and future random variables is equal to 0 bits:

$$\begin{aligned} I(U_{0|i-1}; U_{i|2i-1}) &= -\tfrac{1}{2} \log \tfrac{\sigma^{2^{2i}}}{\sigma^{2^i} \sigma^{2^i}} \\ &= 0 \text{ bits.} \end{aligned}$$

This does not come as a surprise since the past and the future of white noise are completely independent. ◊

Example 4.12. The mutual information of the past and future of the output of a stochastic MIMO linear model
In this example, we consider the 2-component stochastic process $\{y(k)\}$ that was used in the Examples of Section 3.4.2. It is the output of the third order linear model with state space matrices (3.73) of which the input is white Gaussian noise with covariance matrix S_u, given in (3.74). There are 15 000 data points available. In order to get an idea how the mutual information of the past and the future of the output process evolves as a function of i, we compute the mutual information of $Y_{0|i-1}$ and $Y_{i|2i-1}$ for increasing i. Starting from the measured data, this is done as follows. Let Y_p be the past block Hankel matrix and Y_f the future block Hankel matrix (see (3.26)) with $j = 10\,000$ columns and i block rows and let $Y = \begin{pmatrix} Y_p \\ Y_f \end{pmatrix}$. The mutual information of the past and

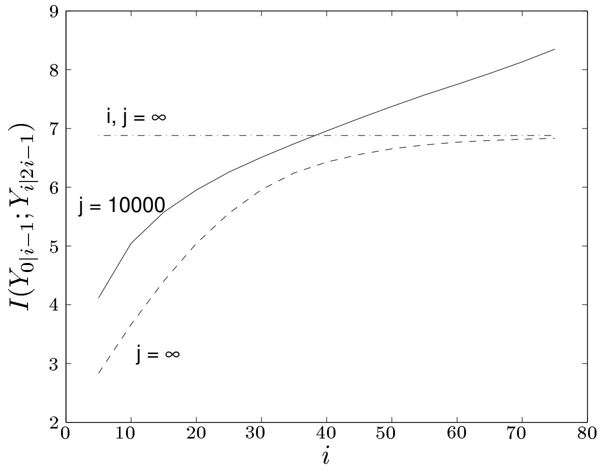


Figure 4.8: The mutual information $I(Y_{0|i-1};Y_{i|2i-1})$ as a function of i. In full line the estimates obtained via $2i \times 10\,000$ block Hankel matrices. The dashed line depicts the values for $j=\infty$ and the dash-dot line shows $I(y_p;y_f)$, the mutual information of the past and future processes, which is equal to 6.8808 bits.

the future random variables is then estimated as

$$I(Y_{0|i-1};Y_{i|2i-1}) \approx -\frac{1}{2}\log\prod_{k=1}^{2i}\sin^2\phi_k \ ,$$

where $\phi_1,\ldots,\phi_{2i}$ are the principal angles between the row spaces of Y_p and Y_f. By increasing the number of rows i from 5 till 75, with steps of 5, we obtain the curve in full line in Figure 4.8 (see below for further explanation). Using the exact covariance matrices, i.e. letting j go to infinity, gives the dashed curve, which converges neatly to the value that is obtained for infinite j and i, depicted in dash-dot line.

Remark 4.9. The dashed and dash-dot curves in Figure 4.8
The dashed curve ($j\to\infty$) is obtained as follows. The mutual information of $Y_{0|i-1}$ and $Y_{i|2i-1}$ is equal to

$$I(Y_{0|i-1};Y_{i|2i-1}) = -\frac{1}{2}\log\prod_{k=1}^{2i}\sin^2\phi_k \ ,$$

where $\phi_1,\ldots,\phi_{2i}$ are the principal angles between the row spaces of Y_p and Y_f (for $j\to\infty$). The smallest 3 principal angles can be computed from the system matrices, since their squared cosines are the eigenvalues of the matrix in (3.91), and the other $2i-3$ principal angles are equal to $\frac{\pi}{2}$. Let us denote the matrix in (3.91) by M_i. Then, the mutual information is also equal to

$$I(Y_{0|i-1}; Y_{i|2i-1}) = -\frac{1}{2} \log \det(I_3 - M_i) \ .$$

Hence, we do not need to compute the eigenvalues of M_i to know the mutual information of $Y_{0|i-1}$ and $Y_{i|2i-1}$.
When $i \to \infty$, the cosines of the principal angles between $\mathrm{row}(Y_p)$ and $\mathrm{row}(Y_f)$ are equal to the canonical correlations of y_p and y_f. The three canonical correlations of y_p and y_f that are not equal to 0, which we denote by ρ_1, ρ_2, ρ_3, can be obtained as the square roots of the eigenvalues of $\Sigma \left(\mathcal{G}_z^{-1} + \Sigma\right)^{-1}$ (see (3.100)). From Example 3.9, we know their values: $\rho_1 = 0.9990$, $\rho_2 = 0.9643$ and $\rho_3 = 0.6825$. The asymptotic value for the mutual information of the past and the future is consequently obtained as

$$I(y_p; y_f) = -\frac{1}{2} \log \prod_{k=1}^{3} (1 - \rho_k^2) = 6.8808 \text{ bits.}$$

We can also use the forward and backward state covariance matrices Σ and $\bar{\Sigma}$. The mutual information of the past and the future of $\{y(k)\}$ is then equal to

$$I(y_p; y_f) = -\frac{1}{2} \log \det(I - \Sigma\bar{\Sigma}) \ ,$$

see Property 3.5. ○

Returning to the estimation of the mutual information, starting from a finite amount of data and without knowing the model, we see in Figure 4.8, that we do not get a good estimate. The full line does not converge at all. The bad estimate can be explained by looking at the principal angles $\phi_1, \ldots, \phi_{2i}$ between the row spaces of Y_p and Y_f, of which the cosines are shown in Figure 3.17 for $i = 5, 10, \ldots, 50$. The $2i - 3$ angles that should be equal to $\frac{\pi}{2}$, are underestimated for finite j and since the mutual information of $Y_{0|i-1}$ and $Y_{i|2i-1}$ is approximately equal to $-\frac{1}{2} \log \prod_{k=1}^{i} \sin^2 \phi_k$, the estimate of the mutual information is too large. It is, however, clear from Figure 3.17 that only three angles should contribute to the mutual information. In other words, we can estimate the model order n from the principal angles and only use the smallest n principal angles in the computation of the mutual information. Determining how many principal angles between the past and future are significantly different from $\frac{\pi}{2}$, is exactly how the model order is estimated in a number of subspace identification methods [94, 156, 159, 10].
Thus, for the given data, we estimate the mutual information of the past and the future as

$$I(Y_{0|i-1}; Y_{i|2i-1}) \approx -\frac{1}{2} \log \prod_{k=1}^{3} \sin^2 \phi_k \ ,$$

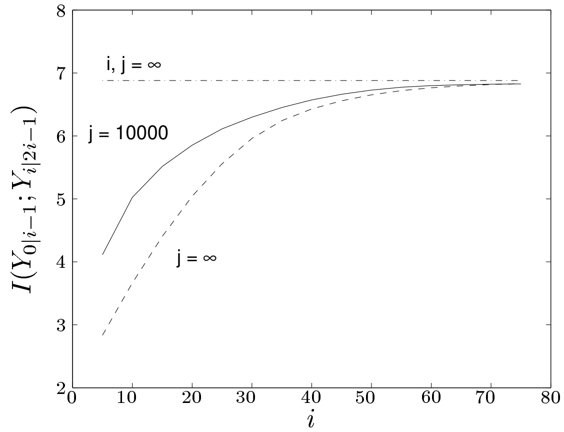


Figure 4.9: The mutual information $I(Y_{0|i-1}; Y_{i|2i-1})$ as a function of i. In full line we show the estimates obtained via the smallest 3 principal angles between the row spaces of the $2i \times 10\,000$ block Hankel matrices. The dashed line depicts the values for $j = \infty$ and the dash-dot line shows $I(y_p; y_f)$, the mutual information of the past and future processes.

where ϕ_1, ϕ_2, ϕ_3 are the smallest three principal angles between $\mathrm{row}(Y_p)$ and $\mathrm{row}(Y_f)$. This new estimate is shown in Figure 4.9 as the full line, which converges to the right value. $\Diamond$

Remark 4.10. The model generates mutual information
In the two examples above, we have seen that the mutual information of the past and future of a white noise process is equal to 0, while the output of a linear model that is driven by white noise, does have mutual information between its past and future. Since the model has generated correlations between past and future, they are not independent any more. This will further be discussed in Example 4.13 and Example 4.14. $\bigcirc$

As indicated in Example 4.12, we define the mutual information of the past and the future of a stochastic process $\{y(k)\}$ as follows.

Definition 4.9. The mutual information of the past and the future of a stochastic process
The mutual information of the past and the future of a stochastic process $\{y(k)\}$ is defined as

$$I(y_p; y_f) = \lim_{i \to \infty} I(Y_{0|i-1}; Y_{i|2i-1}) \ ,$$

if the limit exists.

We now treat three special cases, where the specification of the process becomes more restricted with each case: Gaussian stochastic processes, Gaussian processes that are generated by a linear model and Gaussian processes that are generated by a SISO linear model.

The mutual information of past and future of a Gaussian process

If $\{y(k)\}$ is an m-component Gaussian stationary stochastic process, then, $Y_{0|i-1}$ and $Y_{i|2i-1}$ are jointly Gaussian and hence the mutual information of the past and the future of $\{y(k)\}$ can be expressed in the following ways (as already applied in Example 4.11 and Example 4.12)

1. In terms of the covariance matrices:

$$I(y_p; y_f) = \lim_{i\to\infty} -\frac{1}{2}\log\frac{\det Q_{2i}}{\det Q_{p_i}\det Q_{f_i}} \ ,$$

   where Q_{2i} is the covariance matrix of $Y_{0|2i-1}$ and Q_{p_i} and Q_{f_i} are the covariance matrices of $Y_{0|i-1}$ and $Y_{i|2i-1}$, respectively, or, by using the output block Hankel matrices: $Q_{2i} = \lim_{j\to\infty} YY^T$, $Q_{p_i} = \lim_{j\to\infty} Y_pY_p^T$ and $Q_{f_i} = \lim_{j\to\infty} Y_fY_f^T$. Due to the stationarity of the stochastic process, $Q_{p_i} = Q_{f_i}$. Hence,

$$\begin{aligned} I(y_p; y_f) &= \lim_{i\to\infty}\left(\log\det Q_{p_i} - \frac{1}{2}\log\det Q_{2i}\right) \ , \\ &= \lim_{i\to\infty}\lim_{j\to\infty}\left(\log\det Y_pY_p^T - \frac{1}{2}\log\det YY^T\right) \ . \end{aligned}$$

2. In terms of the principal angles $\phi_1, \ldots, \phi_{mi}$ between the row spaces of the $mi \times j$ past and future block Hankel matrices Y_p and Y_f

$$I(y_p; y_f) = \lim_{i\to\infty}\lim_{j\to\infty} -\frac{1}{2}\log\prod_{k=1}^{mi}\sin^2\phi_k \ .$$

3. In terms of the canonical correlations $\rho_1, \rho_2, \ldots$ of the past and the future:

$$I(y_p; y_f) = -\frac{1}{2}\log\prod_{k=1}^{\infty}(1-\rho_k^2) \ .$$

The mutual information of past and future of a Gaussian process generated by a linear model

For Gaussian processes that are generated by a linear stochastic model of order n, only the largest n canonical correlations of the past and the future can be different from 0, which leads to the following corollary.

Corollary 4.7.
The mutual information of the past and future of the output of a linear model that is driven by white Gaussian noise is equal to

$$I(y_p; y_f) = -\frac{1}{2}\log\prod_{k=1}^{n}(1-\rho_k^2) = -\frac{1}{2}\log\prod_{k=1}^{n}\sin^2\phi_k \ ,$$

where $\rho_1,\ldots,\rho_n$ are the largest n canonical correlations of the past and the future process and ϕ_k $(k=1,\ldots,n)$ the corresponding principal angles.

The squares of the canonical correlations are equal to the eigenvalues of $\Sigma(\mathcal{G}_z^{-1}+\Sigma)^{-1}$ (see Equation (3.100), which are also the eigenvalues of $\Sigma\bar{\Sigma}$. The mutual information of the past and the future process is therefore equal to

$$\begin{aligned} I(y_p; y_f) &= -\frac{1}{2}\log\det(I_n - \Sigma(\mathcal{G}_z^{-1}+\Sigma)^{-1}) \ , \\ &= -\frac{1}{2}\log\det(I_n - \Sigma\bar{\Sigma}) \ , \end{aligned}$$

where Σ is the forward state covariance matrix, the solution of the Lyapunov equation (3.8), $\mathcal{G}_z$ is the solution of the Lyapunov equation in (3.99), and $\bar{\Sigma}$ is the backward state covariance matrix, which is the solution of the Riccati equation in (3.20).
Alternatively, by using the complementarity of the canonical correlations of the future input and the future output on the one hand and the canonical correlations of the past and future output on the other hand (see Property 3.7) we have

$$I(y_p; y_f) = -\frac{1}{2}\log\prod_{k=1}^{n}\tau_k^2 = -\frac{1}{2}\log\prod_{k=1}^{n}\cos^2\theta_k \ ,$$

where $\tau_1,\ldots,\tau_n$ are the smallest n canonical correlations of the future input and the future output process and θ_k $(k=1,\ldots,n)$ the corresponding principal angles. Together with Equation (3.98) , this leads to

$$I(y_p; y_f) = \frac{1}{2}\log\det(I_n + \mathcal{G}_z\Sigma) \ .$$

The mutual information of past and future of a Gaussian process generated by a SISO linear model

The canonical correlations of the past and the future of the output of a SISO linear model are independent of the variance of the driving noise, as explained in Section 3.5.1. The largest n squared canonical correlations of the past and the future are the eigenvalues of $P(Q_z^{-1}+P)^{-1}$ and the other canonical correlations are equal to 0. Consequently, if the input process is white Gaussian noise, then the mutual information of the past and the future is equal to

$$I(y_p;y_f) = -\frac{1}{2}\log\det(I_n - P(Q_z^{-1}+P)^{-1}) \ ,$$
$$= \frac{1}{2}\log\det(I_n + Q_zP) \ , \tag{4.22}$$

where P is the controllability Gramian of the model and Q_z is the observability Gramian of the inverse model, satisfying the Lyapunov equations in (3.22) and (3.46), respectively.
Applying Theorem 3.12 gives

$$I(y_p;y_f) = -\frac{1}{2}\log\det(P^{-1}RQ_z^{-1}R^T) \ ,$$
$$= -\frac{1}{2}\log\frac{(\det R)^2}{\det P\det Q_z} \ , \tag{4.23}$$

where R is the product of the infinite controllability matrix of the model and the infinite observability matrix of the inverse model: $R = \mathcal{C}_\infty\Gamma_{z_\infty}$. The matrices P, Q_z and R in (4.23) can be derived from the following Lyapunov equation, which we encountered in Conjecture 3.13:

$$\begin{pmatrix} P & R \\ R^T & Q_z \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & (A-KC)^T \end{pmatrix}\begin{pmatrix} P & R \\ R^T & Q_z \end{pmatrix}\begin{pmatrix} A^T & 0 \\ 0 & A-KC \end{pmatrix} + \begin{pmatrix} K \\ -C^T \end{pmatrix}\begin{pmatrix} K^T & -C \end{pmatrix} \ .$$

Remark 4.11. The mutual information of the past and the future output of the inverse model
As we mentioned in Corollary 3.16, the canonical correlations of the past and future output process of a SISO linear model are equal to the canonical correlations of the past and future output process of the inverse model (provided both models are driven by white noise). Let $\{y(k)\}$ be the output process of a SISO linear model that is driven by white Gaussian noise and $\{z(k)\}$ the output process of the inverse model when driven by white Gaussian noise. Since we can express the mutual information of the past and the future of these processes in terms of the canonical correlations of their past and future processes, we obtain

$$I(y_p;y_f) = I(z_p;z_f) \ .$$

○

Remark 4.12. The mutual information of past and future as a model norm
The canonical correlations of the future input and future output of a SISO ARMA model with transfer function $H(z) = \frac{b(z)}{a(z)}$ can also be seen as the canonical correlations of the future output processes of the AR models $\frac{z^n}{a(z)}$

and $\frac{z^n}{b(z)}$, respectively, when the models are driven by the same input (see Section 3.5.2 and Section 3.5.3). The n principal angles corresponding to the smallest n canonical correlations, which we have called in [36, 37, 38, 39] the subspace angles between the AR models, are indicators of the distance between the two models. Since the mutual information of the past and future of the output of the ARMA model is equal to $-\frac{1}{2}\log\prod_{k=1}^{n}\cos^2\theta_k$, where $\theta_1,\ldots,\theta_n$ are the subspace angles between $\frac{z^n}{a(z)}$ and $\frac{z^n}{b(z)}$, the mutual information seems to express a distance between the AR models. Alternatively, it indicates the distance between the numerator and denominator, or between the poles and zeros of the ARMA model. This also means that it is a measure for the amount of dynamics in a data sequence. If all poles are cancelled by zeros, then the transfer function is equal to 1, and the output process will be white Gaussian noise, and consequently the mutual information is equal to 0 bits (see also Example 4.11).
The following examples give further indications. ○

Example 4.13. Influence of relative position of poles and zeros on the mutual information of past and future
We consider the 5th order ARMA model with poles -0.5, $0.3496 \pm 0.3389i$, 0.3599 and 0.1336 and of which we change the zeros. We take the zeros equal to $k/0.5$ times the poles, where k varies from 0 to 0.9 in steps of 0.01. Hence, the zeros lie first all in the origin ($k = 0$), then move towards the poles, cancel the poles at $k = 0.5$ (this is the spectral radius of the poles) and go further to the unit circle. The poles and 10 of the 91 zero-realizations, namely for $k = 0, 0.1, 0.2, \ldots, 0.9$ are shown in Figure 4.10.
We compute for each ARMA model the mutual information of the past and future of its output process when the input is white Gaussian noise. The re-

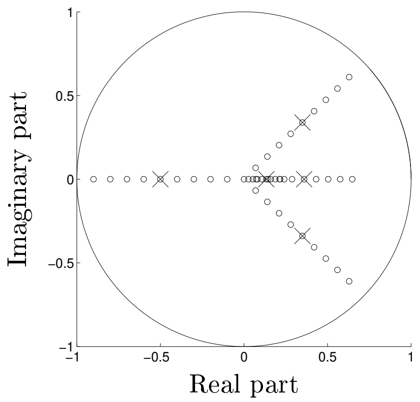


Figure 4.10: The poles (×) and 10 zero-realizations (∘) for a 5th order ARMA model. The zeros move from the origin, towards the poles, cancel the poles exactly and move on towards the unit circle.

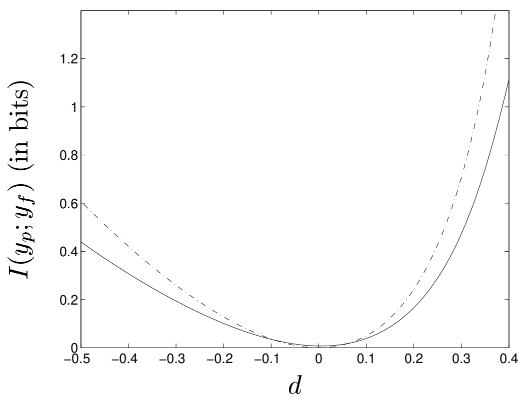


Figure 4.11: The mutual information of past and future of the output of the 91 different ARMA models as a function of the difference d between the pole and zero of largest absolute value. The estimated mutual information is given in full line. We use a data set of 10 000 data samples. The past and future Hankel matrices have $i = 10$ rows and $j = 9981$ columns. The true values of the mutual information (i.e. for $i, j \to \infty$) for the different ARMA models are given in dash-dot line. They are obtained via the parametric expression (4.22). When the zeros cancel the poles (at $d = 0$), then the mutual information is equal to 0 bits. The output process is then white noise. When the zeros move away from the poles, the mutual information becomes larger.

sult is shown in Figure 4.11, where we give the mutual information of the past and the future of the ARMA models as a function of the difference d between the pole in -0.5 (this is the pole with largest absolute value) and the zero with largest absolute value, which lies subsequently in $0,-0.01,-0.02,\ldots,-0.9$. The difference between the pole in -0.5 and these zeros consequently goes from $-0.5,-0.49,\ldots,0.4$. We estimate the mutual information of past and future from data obtained as follows. We simulate 10 000 samples of the output process of each ARMA model and compute the mutual information via the principal angles between the row spaces of the Hankel matrices Y_p and Y_f, which are 10×9981. The estimates are shown in Figure 4.11 in full line, while the true values (for $i, j \to \infty$) are given in dash-dot line. These asymptotic values are computed via the parametric expression (4.22).
As can be seen in Figure 4.11, the mutual information of the past and the future decreases as the zeros come closer to the poles. When they cancel the poles, which happens when $d = 0$, then the mutual information is equal to 0 bits. We knew this would be the case, since the output process is then white noise and the mutual information of past and future of white noise is equal to 0 bits. When the zeros move further to the unit circle, the mutual information again increases.
The principal angles between the past and future spaces of the different out-

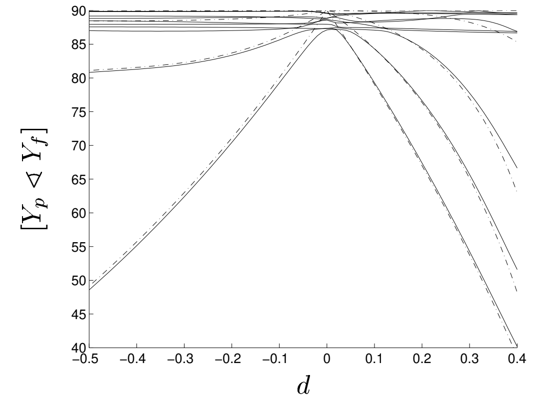


Figure 4.12: The principal angles (in degrees) between the past and future output spaces for the different simulated ARMA processes. The parameter d indicates the difference between the pole and zero of maximum absolute value. The full line gives the computed angles, while the dash-dot line represents the asymptotic values for $i, j \to \infty$, i.e. the canonical correlations of y_p and y_f.

put processes are shown in Figure 4.12. For $d = 0$, the asymptotic values in dash-dot line, which are computed via the parametric expression (3.100), are all equal to $\frac{\pi}{2}$. The past and future are consequently orthogonal when the zeros cancel the poles. The further the zeros lie from the poles, the smaller the angles between past and future. The dynamics of the system decrease the 'distance' between the past and future subspaces.
The principal angles between the future input and future output spaces in Fig-

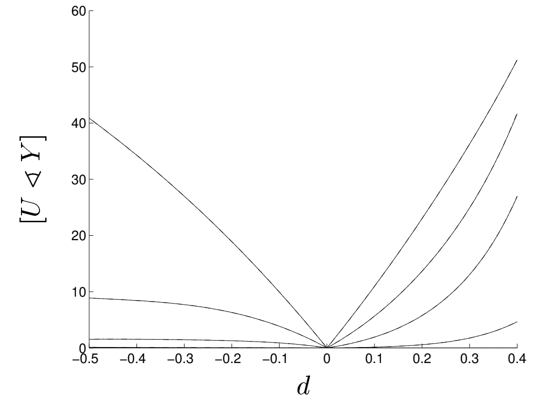


Figure 4.13: The principal angles (in degrees) between the input (innovations) and output spaces of the different ARMA models. The computed values coincide with the asymptotic values, which are computed according to (3.98).

ure 4.13 shed another light on this example. As noted in Remark 4.12, these angles are the so-called subspace angles between the numerator and denominator AR models. They give an indication for the distance between the poles and the zeros of the different models. Indeed, when the zeros cancel the poles, which happens for $d = 0$, then the principal angles between the future input and output are all equal to 0 and they become larger when the zeros move further from the poles. The asymptotic values were computed via the parametric expression (3.98). $\diamond$

Example 4.14. The mutual information of past and future of an AR process

Let $\{y^{(1)}(k)\}$ be the autoregressive (AR) process obtained as the output of the 10th order linear model with transfer function $G^{(1)}(z) = \frac{z^{10}}{\prod_{k=1}^{10}(z-\alpha_k^{(1)})}$, where $\alpha_k^{(1)}$ $(k = 1, \ldots, 10)$ are the poles of $G^{(1)}(z)$, which are given as $\times$-marks in Figure 4.14. The spectral radius of the poles is 0.8. The input of the model is white noise. A second AR process $\{y^{(2)}(k)\}$ is obtained as follows: we multiply the poles $p_k^{(1)}$ with $\frac{0.9}{0.8}$ to obtain the poles of the second model $G^{(2)}(z)$, so that the maximum radius now becomes 0.9. These poles are indicated in Figure 4.14 by the stars $(*)$. Since the poles of $G^{(2)}(z)$ lie closer to the unit circle than the poles of $G^{(1)}(z)$, the second output process resembles less white noise than the first output process. Indeed, since all zeros of both models lie in the origin, the output process would be white only if the poles also lie in the origin. The closer the poles are to the origin, the whiter the output process. This can e.g. also be seen by correlation analysis or in the power spectrum. If the mutual information of the past and the future is a measure for the distance to white noise, it should be larger for the second process.

We compute for both processes the mutual information of $Y_{0|i-1}$ and $Y_{i|2i-1}$.

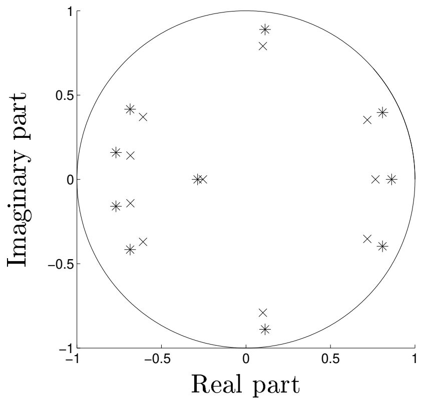


Figure 4.14: The poles of the two AR models of order 10. The $\times$-marks are the poles of $G_1(z)$, the stars $(*)$ are the poles of $G_2(z)$.

Therefore, we first estimate the model order n, based on the principal angles ϕ_k $(k = 1, \ldots, i)$ between the row spaces of the $i \times 10\,000$ Hankel matrices Y_p and Y_f, as in Example 4.12. For $\{y^{(1)}(k)\}$ the estimated order is 4 and for $\{y^{(2)}(k)\}$ it is 7. The mutual information $I(Y_{0|i-1}; Y_{i|2i-1})$ is then estimated as $-\frac{1}{2} \log \prod_{k=1}^{n} \sin^2 \phi_k$. The result as a function of the length i is given in Figure 4.15. The full lines gives the estimated mutual information and the dashed lines show the values for $j \to \infty$ and correct model order and model parameters. The true mutual information of the past and the future of the AR processes is equal to 1.3113, respectively 2.6610 bits. We observe that the mutual information of the past and the future of $\{y^{(2)}\}$ is larger than the mutual information of the past and the future of $\{y^{(1)}\}$. This is indeed what we have predicted from the location of the poles.
In Figure 4.15 we also notice that for $j \to \infty$, the mutual information of the past and the future of the 10th order AR processes remains constant from $i = 10$ onwards. This follows directly from Property 3.17, where we have shown that for an AR process $\{y(k)\}$, the principal angles between the row spaces of Y_p and Y_f remain constant as a function of i from $i = n$ onwards (for $j \to \infty$). Moreover, by using the expression for the canonical correlations of the past and the future of an AR process in Property 3.17, the mutual information of the past and the future is equal to

$$I(y_p; y_f) = -\frac{1}{2} \log \det(I_n - P^{-1} A^n P A^{n^T}) \ ,$$

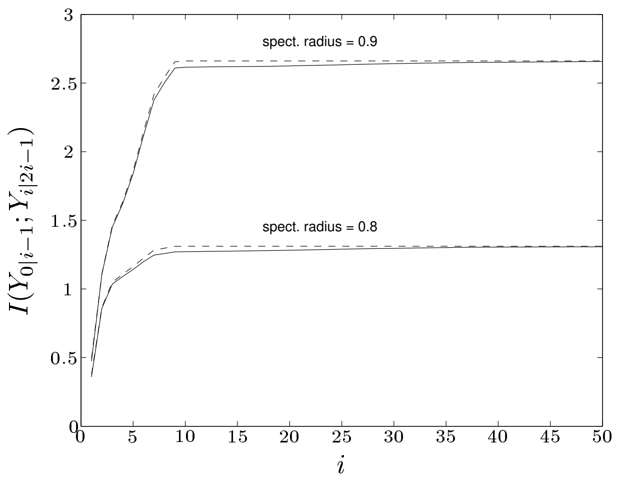


Figure 4.15: The mutual information $I_i(y_p; y_f)$ of the past and the future of the AR processes $\{y^{(1)}(k)\}$ (spectral radius is 0.8) and $\{y^{(2)}(k)\}$ (spectral radius 0.9) as a function of i. The dashed lines give the values for $j = \infty$, the full lines are estimates based on the principal angles between row(Y_p) and row(Y_f), with order estimation.

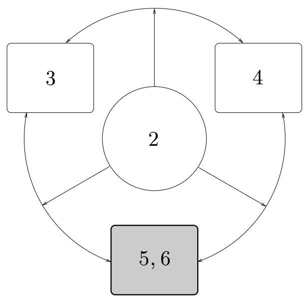
where P is the controllability Gramian of the AR model and A is its system matrix. $\diamond$

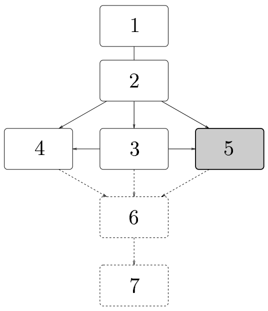
4.5 Conclusions

In this chapter, we have first given the necessary background information on the information theoretic notions of entropy and mutual information. We have treated the entropy of a random variable as well as the entropy rate of a stochastic process. Similarly, we have given definitions and properties for the mutual information of two random variables and of two stochastic processes.
We have seen that the mutual information of two Gaussian random variables can be expressed as a function of their canonical correlations. The same can be done for the mutual information of two jointly Gaussian stochastic processes. This property allowed us to apply the results of Chapter 3, i.e. the expressions for the canonical correlations of the different pairs of input and output processes, which resulted in parametric expressions for the mutual information of the past and future output process of a linear model.
Furthermore, we have shown how the mutual information of the past and future can be estimated from measurements. The principal angles between the row spaces of the past and future block Hankel matrices turn out to be useful for this, since by looking at the principal angles, we can first estimate the model order and then only take that number of principal angles into account.
From the characterization of the principal angles between the past and future output spaces of a SISO AR model, finally, we have derived a formula for the mutual information of the past and future of the corresponding AR process.
As we will see in Chapter 6, the mutual information of the past and future of a scalar Gaussian stochastic process, corresponds to a cepstral model norm, i.e. a norm based on the cepstrum of the model (or output process). We have already given indications in this chapter that make this predicted property plausible. However, first we have to define this cepstrum and explain its properties. This is done in the next chapter.

Chapter 5

The Cepstrum





There is only one missing link left in order to close the circle in the left picture above, which is the signal processing part, spread over Chapter 5 and Chapter 6. Indeed, in order to define the cepstral norm in the next chapter, we first need to define the cepstrum and to explain its properties. From Figure 1.10 and the right picture above, we see that this chapter forms a bit an excursion from the main path. As a matter of fact, this is the only chapter where we do not encounter principal angles between subspaces nor canonical correlations. However, it will become clear that we will need the cepstrum and its properties for the derivations in Chapter 6.

Contributions We try to give a clear and concise overview of the power cepstrum and the complex cepstrum, interlaced with examples to illustrate their applicability. Furthermore, we point out how the cepstrum can be used for the identification of linear models. There exists a simple relation between the Markov parameters of the model and its cepstrum. Consequently, one could obtain an estimate of the Markov parameters by first estimating a number of cepstrum coefficients from measurements. A state space description of the identified model can then be obtained by applying an algorithm for the deterministic realization problem (see Section 3.3.1).

Chapter overview We start with a historical introduction in Section 5.1. In Section 5.2, we define the power cepstrum. The complex cepstrum is treated in Section 5.3. The power cepstrum of an ARMA process can easily be expressed in terms of its poles and zeros. This is done in Section 5.4. In Section 5.5 we discuss the inverse problem, where an ARMA model is estimated from a series of cepstral coefficients. This is an identification problem, based on the cepstrum.

5.1 Introduction

In 1963, Bogert, Healy and Tukey published a paper with the unusual title 'The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking', in which they introduced the cepstrum and showed that it is a useful tool for the detection of echoes in seismological data [17]. As can be seen from the title, they proposed a whole new terminology by paraphrasing existing words. The terms that are still in use nowadays, are given in Table 5.1. We will only adopt the term 'cepstrum' for the inverse Fourier transform of the logarithm of the spectral density[1].

spectrum	cepstrum
filtering	liftering
frequency	quefrency
harmonics	rahmonics

Table 5.1: A few signal processing terms and their paraphrases as introduced in [17].

Meanwhile, Oppenheim had been developing his theory of homomorphic[2] systems [110], for which soon applications in signal processing were found, such as image enhancement [139], speech analysis [112] and blind deconvolution. A classic example of blind deconvolution is the restoration of the old phonograph recordings of Caruso [140], where the horn distortions are removed from the

[1] More precisely, we will call this the 'power cepstrum', in order to make a clear distinction with the complex cepstrum (see further).

[2] A homomorphism is a mapping that preserves all relevant structure. If A and B are groups with respective operations $\bigcirc$ and $\Box$, then f is a homomorphism of A into B if $f(a_1 \bigcirc a_2) = f(a_1) \Box f(a_2)$ for all $a_1, a_2 \in A$. We are only interested in the homomorphic system for the convolution of signals. By converting two convolved signals to the cepstral domain, the converted signals are added and can be treated with the usual filter techniques designed for added signals.

audio signal[3].
The cepstrum analysis as described by Bogert et al. turned out to be a special case of Oppenheim's homomorphic signal processing. Oppenheim takes the complex logarithm of the Z-transform of a signal, whereas Bogert et al. consider the logarithm of the power spectrum, which is a real function. Therefore, Oppenheim's version of the cepstrum is called the 'complex cepstrum', although the cepstrum itself is a real function (see Section 5.3).
An interesting document to get to know the circumstances and spirit of the times in which these concepts were developed, is the interview with Oppenheim, conducted for the IEEE History Center [54].
A good introduction to homomorphic signal processing is Chapter 10 in the book by Oppenheim and Schafer [113]. They start with an abstract description of homomorphic systems, consider systems for the multiplication and convolution of signals, discuss the complex cepstrum and end up with a number of interesting applications. Deller et al. [41, Chapter 6] treat cepstral analysis for speech signals, emphasizing the signal processing aspects, which we will not treat in detail.
Besides the above mentioned applications, the cepstrum is also used for machine diagnostics because of its ability to detect periodicities in the spectrum [161] and for measuring the distance between two signals, see [6] and references therein and [105]. The application of the cepstrum to distance measures will be treated in Chapter 6.
Oppenheim [111] already indicated that the cepstrum can also be used for the identification of a stochastic model (see Section 3.3 for an introduction to stochastic system identification). We will describe his approach in Section 5.5. Recently, Byrnes et al. [23, 22] and Enqvist [46, 47] showed that the cepstrum coefficients can be applied to find solutions for the rational covariance extension problem [23, 22, 46, 47].

5.2 The power cepstrum

As we have mentioned in the introduction of this chapter, the power cepstrum originated in the work of Bogert et al. [17] as a solution for the detection of echoes in seismic data. Before we give the formal definitions of the power cepstrum, we consider an example where it is introduced naturally and the need for two definitions becomes clear. We construct in Example 5.1 a signal with one echo and show how the cepstrum reveals this echo, whereas it is difficult to detect it in the autocovariance function. Next, we give two definitions for the power cepstrum, the first one defines the power cepstrum of a stochastic

[3]The result of this restoration can be heard via the video of the Research Laboratory of Electronics of the MIT, available on `http://rleweb.mit.edu/video/video.htm`. The audio fragment is contained in video clip 10 'Digital Signal Processing – continued'.

process and the second one the power cepstrum of an absolutely summable sequence.

Example 5.1. Echo detection
The detection (and cancellation) of echoes is an important problem in many areas. The original application in [17] dealt with seismic data. But also in most communication channels (telephony, wireless communication, ...), echoes appear as the result of e.g. an impedance mismatch along the telephone line, or due to leakage from the loudspeaker to the microphone in e.g. hands-free telephony. State-of-the-art signal processing techniques for echo cancelling in modern ADSL modems are described in [148, 149] and references herein. Acoustic echo cancelling methods can e.g. be found in [44, 45, 126] and references therein.
A signal with one echo can be modeled as $z(k) = y(k) + \gamma y(k - D)$, where γ is the attenuation coefficient ($|\gamma| < 1$) and D is the delay. The purpose of echo detection is to determine D, the time at which the echo appears. The signal $z(k)$ can also be viewed as the convolution of $y(k)$ and the signal $d(k)$ which consists of two pulses:

$$d(k) = \begin{cases} 1 & k = 0 \ , \\ \gamma & k = D \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{5.1}$$

Thus,

$$z(k) = d(k) * y(k) \ .$$

In this example, the time series $\{y(k)\}$ is obtained by sending white noise with variance 1 through a fourth order ARMA model with transfer function $H(z)$, see Figure 5.1. To make it concrete, the model can be viewed as the model of a transmission channel and the time series $\{y(k)\}$ is the signal that is sent. The

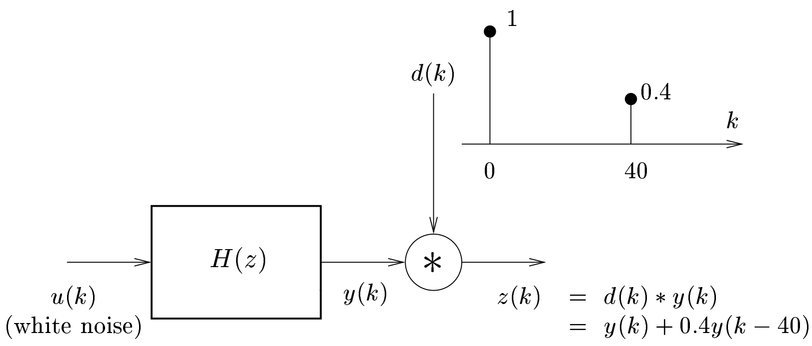


Figure 5.1: The construction of the signal $z(k)$, which is equal to $y(k)$ with an echo of $y(k)$ at $k = 40$. The attenuation coefficient is 0.4. It can be viewed as the convolution of $y(k)$ and $d(k)$ which consists of two pulses, one at $k = 0$ with magnitude 1 and the other at $k = 40$, valued 0.4.

poles of the ARMA(4, 4) model are $0.9095, -0.9289$ and $-0.4193 \pm 0.8525\,i$, and the zeros are $-0.6668 \pm 0.2835\,i$ and $0.6192 \pm 0.1494\,i$. The attenuation coefficient is chosen to be $\gamma = 0.4$ and the time at which the echo appears is $D = 40$.
The discrete time signals we work with, are shown in the top row of Figure 5.3. As will become clear, the power cepstrum is the obvious tool for detecting the echo in $\{z(k)\}$. The different steps that are taken to compute the power cepstrum of $\{z(k)\}$ are given in Figure 5.2. We discuss each step in detail.

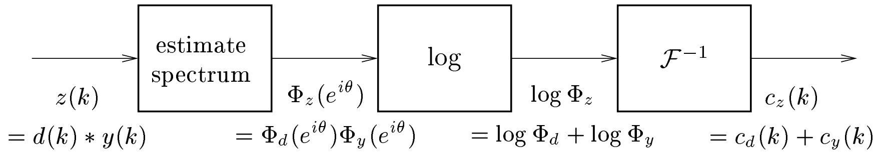


Figure 5.2: The computation of the power cepstrum of $\{z(k)\}$.

1. First, the power spectrum of $\{z(k)\}$, denoted by $\Phi_z(e^{i\theta})$, is estimated. In this example, we will use the theoretical values. Since we know the relation of $\{z(k)\}$ and $\{y(k)\}$, we obtain, via the autocovariance function $R_z(\tau)$, the exact expression for $\Phi_z(e^{i\theta})$ as a function of the power spectrum of $\{y(k)\}$, denoted by $\Phi_y(e^{i\theta})$:

$$\begin{aligned}\Phi_z(e^{i\theta}) &= \mathcal{F}(R_z) &(5.2)\\ &= \mathcal{F}\left((1+\gamma^2)R_y(\tau) + \gamma R_y(\tau + D) + \gamma R_y(\tau - D)\right) \\ &= \left(1 + \gamma^2 + 2\gamma\cos(D\theta)\right)\Phi_y(e^{i\theta}) \ . &(5.3)\end{aligned}$$

We now note that $d(k)$ can be viewed as the impulse response of the MA model with transfer function $\frac{z^D+\gamma}{z^D}$. The squared magnitude of its Fourier transform is therefore equal to the power spectral density of the model's output when driven by white noise with variance 1 (see Example 3.2). This output process of the MA model is denoted by $\{y_d(k)\}$ and its spectral density by $\Phi_d(e^{i\theta})$, although it is not the spectral density of $d(k)$

$$\begin{aligned}\Phi_d(e^{i\theta}) &= \mathcal{F}(R_{y_d}) &(5.4)\\ &= |\mathcal{F}(d)|^2 \ , \\ &= 1 + \gamma^2 + 2\gamma\cos(D\theta) \ . &(5.5)\end{aligned}$$

Using (5.5) in (5.3) gives

$$\Phi_z(e^{i\theta}) = \Phi_d(e^{i\theta})\Phi_y(e^{i\theta}) \ .$$

The three spectral densities for this example, Φ_z, Φ_d and Φ_y, are shown in the second row of pictures of Figure 5.3. In the first picture we see the

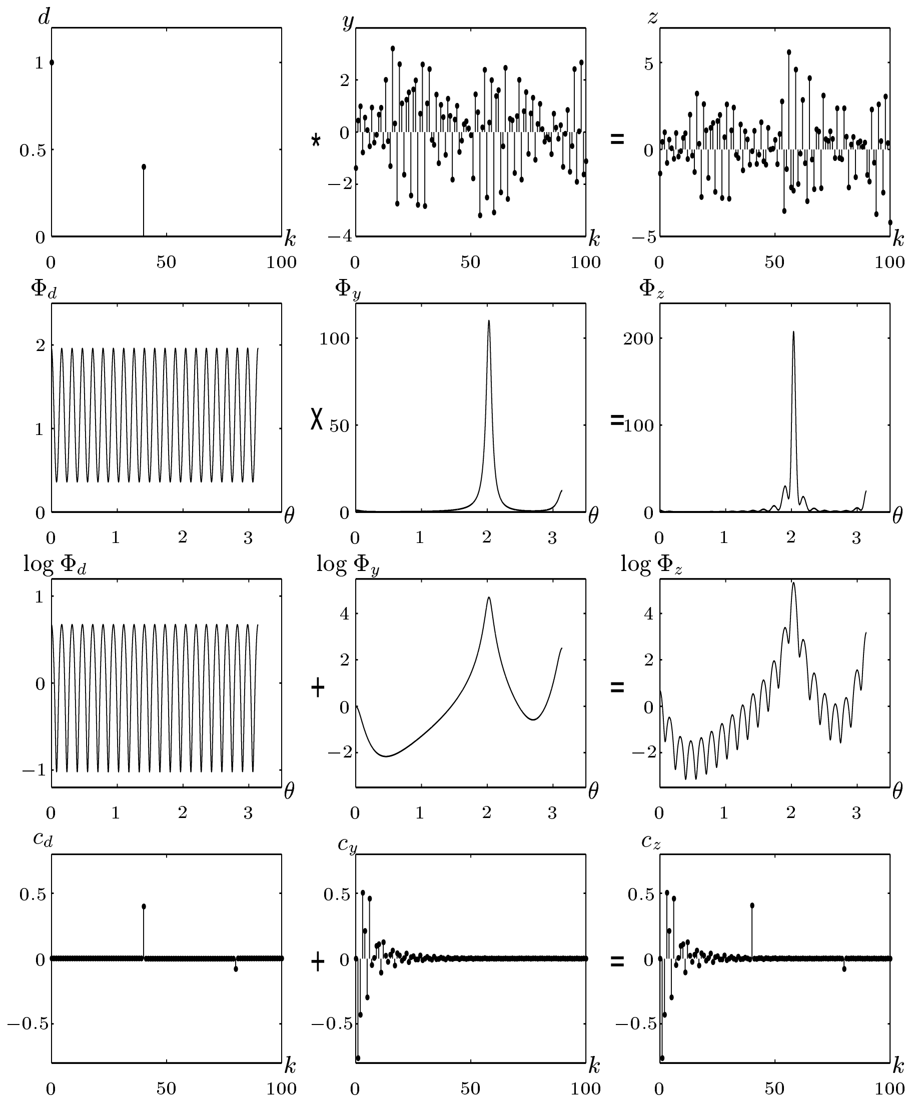


Figure 5.3: The different signals we work with in Example 5.1. In the third column, the different steps to obtain the power cepstrum of $\{z(k)\}$ are recognized. We start from the time signal z, then compute the power spectral density $\Phi_z(e^{i\theta})$, we take the logarithm and transform back to the time domain to obtain $c_z(k)$. In the horizontal direction we see how each z-signal is composed of the corresponding d- and y-signal. In the time domain $d(k)$ and $y(k)$ are convolved, the spectra Φ_d and Φ_y are multiplied, the log spectra and the cepstra are added.

cosine function with period $\frac{2\pi}{40} \approx 0.1571$. The second picture gives the spectral density of the ARMA process $\{y(k)\}$ with a resonance peak at $\theta \approx 2.0279$, i.e. the angle of the complex conjugated pair of poles. The spectral density of $\{z(k)\}$ is the product of these two functions, as shown in the third picture.
In practice, only Φ_z (and not Φ_d nor Φ_y) is known. The influence of the echo (γ and D) is not easily determined from Φ_z.

2. The second step in Figure 5.2 consists of taking the logarithm[4] of the power spectrum of $\{z(k)\}$, by which we obtain an additive combination:

$$\log \Phi_z(e^{i\theta}) = \log \Phi_d(e^{i\theta}) + \log \Phi_y(e^{i\theta}) \ .$$

The first term is equal to

$$\log \Phi_d(e^{i\theta}) = \log(1 + \gamma^2 + 2\gamma \cos(D\theta)) \ ,$$

which is still a periodic function with period $\frac{2\pi}{D}$, see the first picture on row 3 of Figure 5.3. It causes a periodic ripple in $\Phi_z(e^{i\theta})$, which can be seen in the third picture of row 3 in Figure 5.3. This ripple is more clearly seen than the one in the power spectrum (last picture on row 2). The parameter D could now be estimated by determining the period of $\log \Phi_z(e^{i\theta})$.
In practice, however, the ripple will be obscured by noise, stemming from the spectrum estimation and the measurements of $\{z(k)\}$. One of the techniques to determine periodic phenomena in a noisy signal is to look at the Fourier transform of the signal. Since we are already in the frequency domain, we can as well transform $\log \Phi_z(e^{i\theta})$ back to the time domain, which is the last step in Figure 5.2.

3. By taking the inverse Fourier transform of $\log \Phi_z$, we obtain a discrete time signal with spikes at $D, 2D, 3D, \ldots$, which can be seen in the third picture on the last row of Figure 5.3. The explanation is simple. Since $\log \Phi_z$ is equal to the sum of the logarithms of the spectra of $\{y_d(k)\}$ and $\{y(k)\}$, we have

$$\mathcal{F}^{-1}(\log \Phi_z) = \mathcal{F}^{-1}(\log \Phi_d) + \mathcal{F}^{-1}(\log \Phi_y) \ .$$

The inverse Fourier transform of the logarithm of the power spectrum is called the power cepstrum. Denoting the power cepstra of $d(k)$, $\{y(k)\}$ and $\{z(k)\}$, respectively by $c_d(k)$, $c_y(k)$ and $c_z(k)$, gives

$$c_z(k) = c_d(k) + c_y(k) \qquad k \in \mathbb{Z} \, .$$

Recall that $\Phi_d(e^{i\theta})$ is not the power spectrum of $d(k)$, but of the MA process $\{y_d(k)\}$, obtained by sending white noise with variance 1 through

[4] In this chapter we use the logarithm with base e.

the MA filter with impulse response $d(k)$. Consequently, $c_d(k)$ is strictly spoken not the power cepstrum of $d(k)$, but of the related MA process. It is easy to show (a more general assertion is proven in Proposition 5.1) that $c_d(k)$ is a sequence of pulses at multiples of D (the rahmonics):

$$c_d(k) = \begin{cases} (-1)^{|n|+1}\dfrac{\gamma^{|n|}}{|n|} & k = nD \quad (n \in \mathbb{Z}), \\ 0 & \text{otherwise,} \end{cases}$$

which is shown in the first picture of the last row in Figure 5.3. The sequence of pulses are also observed in the power cepstrum of $\{z(k)\}$, shown in the last picture of Figure 5.3, because the cepstrum of $\{y(k)\}$ dies out sufficiently fast. It will be harder to see the pulses when the decay of $c_y(k)$ is slower. In this example, we can estimate the echo parameter D from the power cepstrum of $\{z(k)\}$. Moreover, we get an indication of the attenuation coefficient's value by measuring the magnitude of the spikes.

The autocovariance function of $\{z(k)\}$ is equal to the inverse Fourier transform of its power spectral density (see (5.2)):

$$R_z = \mathcal{F}^{-1}(\Phi_z) \ .$$

Thus, the cepstrum differs from the autocovariance only by the logarithmic conversion of the spectrum (see also Table 5.2). The autocovariance by itself can also be used for echo detection. Since

$$R_z(\tau) = (1+\gamma^2)R_y(\tau) + \gamma R_y(\tau + D) + \gamma R_y(\tau - D) \ ,$$

we get in R_z an echo of R_y at $\tau = D$, which can be detected if $R_y(\tau)$ decreases fast enough with respect to D, or, if D is large enough with respect to the time

$$\begin{array}{ccc}
\text{autocovariance} & & \text{power spectrum} \\
R(\tau) & \underset{\mathcal{F}^{-1}}{\overset{\mathcal{F}}{\rightleftarrows}} & \Phi(e^{i\theta}) \\
\text{one to one} \Updownarrow & & \text{exp} \uparrow\downarrow \text{log} \\
c(k) & \underset{\mathcal{F}}{\overset{\mathcal{F}^{-1}}{\leftrightarrows}} & \log \Phi(e^{i\theta}) \\
\text{power cepstrum} & & \text{log spectrum}
\end{array}$$

Table 5.2: Relations of autocovariance, power spectrum and power cepstrum.

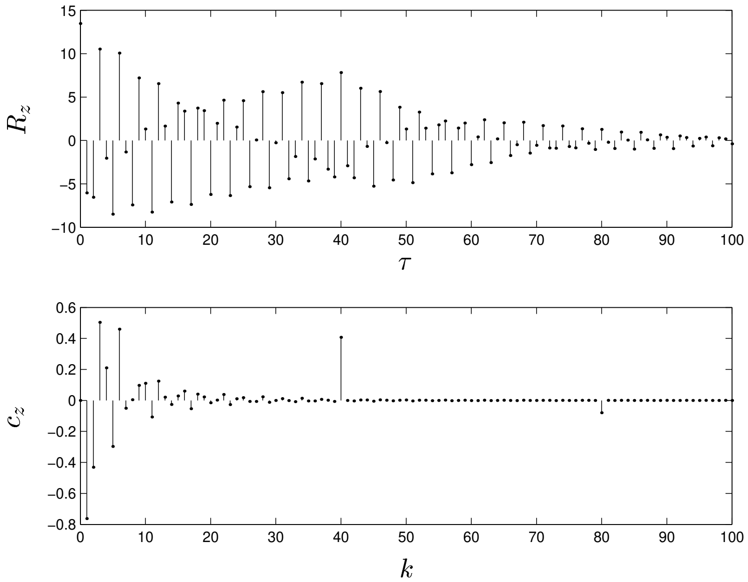


Figure 5.4: Comparison of the autocovariance function $R_z(\tau)$ and the power cepstrum $c_z(k)$ of the stochastic process $\{z(k)\}$. The echo is more clearly visible in the power cepstrum than in the autocovariance. This is partly due to the faster decay of the cepstrum.

constant of R_y. In our example, however, the influence of R_y is still too large at $\tau = D$ to detect the echo clearly. In Figure 5.4 we compare the autocovariance function R_z (top picture) and the power cepstrum c_z (bottom) of $\{z(k)\}$. It is not easy to observe the echo in the autocovariance, but, as we have seen in Figure 5.3, there is a clear peak at $k = 40$ in the cepstrum. Moreover, the 'rahmonic' at $k = 80$ is also an indication of the echo at $k = 40$.
The difference of autocovariance and cepstrum can be understood as follows: the autocovariance is dominated by the large peak in the spectrum Φ_z, while in $\log \Phi_z$ the resonance peak is attenuated and more details of the echo show up. From Figure 5.4 we also see that the cepstrum of $\{y(k)\}$ decays faster than its autocovariance (see also Section 5.4).
Finally, we remark that in more difficult echo detection problems, filtering the log spectrum signal by an appropriate filter (i.e. applying a lifter) can improve the result considerably [17]. $\Diamond$

We can now formally give the definition of the power cepstrum.

Definition 5.1. The power cepstrum of a stationary stochastic process
The power cepstrum of a stationary scalar stochastic process $\{y(k)\}$ with spectral density Φ_y is defined as the inverse Fourier transform of the logarithm of

its power spectral density:

$$c_y(k) = \frac{1}{2\pi}\int_0^{2\pi} \log(\Phi_y(e^{i\theta}))e^{ik\theta}\,d\theta \ .$$

This can also be written as

$$\log\Phi_y(e^{i\theta}) = \sum_{k=-\infty}^{\infty} c_y(k)e^{-ik\theta} \ ,$$

and hence the kth cepstral coefficient is the coordinate of $\log\Phi_y$ w.r.t. the basis signal $e^{-ik\theta}$.
Recall that the spectral density is the Fourier transform of the autocovariance function. For the stationary stochastic process $\{y(k)\}$ with autocovariance $R_y(\tau)$, an alternative expression for the power cepstrum is therefore

$$c_y = \mathcal{F}^{-1}\left(\log\mathcal{F}(R_y)\right) \ .$$

Note that we start from a signal in the time domain (R_y), transform it to the frequency domain to obtain the spectrum ($\Phi_y = \mathcal{F}(R_y)$), and after taking the logarithm, we transform it back to the time domain and obtain the cepstrum ($c_y = \mathcal{F}^{-1}(\log\Phi_y)$). As already mentioned in Example 5.1, the logarithm is the essential difference between the autocovariance and the cepstrum. From a homomorphic processing point of view (see footnote 2 on page 188) the logarithm is necessary to transform the convolved signals to cepstra that are added. In some applications, however, taking the logarithm of the spectrum can also be motivated more directly. For instance, decibels are used when measuring the loudness of sounds because our sense of hearing is roughly logarithmic. Another example is the Richter scale, which measures the energy released by an earth quake [124] (see also [136]). In controller design, but also in modal analysis, one most often displays the Bode plot, which shows the logarithm of the transfer function's magnitude and its phase as a function of the frequency. For stochastic processes this boils down to the logarithm of the spectrum, as was shown in Figure 3.4, where the logarithmic spectrum of the transmitter mast's data is given.
As we have noticed in Example 5.1 with $d(k)$, it is desirable to have a definition for the power cepstrum of sequences that have a Fourier transform themselves and for which one should not rely on the autocovariance. For a sequence $s(k)$ that is absolutely summable, the sum $\sum_{k=-\infty}^{\infty} s(k)e^{-ik\theta}$ converges and hence its Fourier transform exists. An example of an absolutely summable sequence is the impulse response of a linear stable model, of which the Fourier transform is the model's transfer function evaluated on the unit circle. The power cepstrum of absolutely summable sequences can be defined as follows.

Definition 5.2. The power cepstrum of an absolutely summable sequence
The power cepstrum of an absolutely summable sequence $s(k)$ whose Fourier

transform $S(e^{i\theta})$ *satisfies* $S(e^{i\theta}) \neq 0\ \forall\theta$, *is defined as the inverse Fourier transform of the logarithm of the magnitude squared of its Fourier transform:*

$$c_s(k) = \frac{1}{2\pi}\int_0^{2\pi} \log(|S(e^{i\theta})|^2)e^{ik\theta}\,d\theta\ ,$$

where $S(e^{i\theta}) = \sum_{k=-\infty}^{\infty} s(k)e^{-ik\theta}$.

Remark 5.1. We take the magnitude squared of $S(e^{i\theta})$ and not just its magnitude, as e.g. in [41], because we want Definition 5.2 to be consistent with Definition 5.1. In particular, if $s(k)$ is the impulse response of an ARMA model, then with Definitions 5.1 and 5.2 its power cepstrum is equal to the power cepstrum of the ARMA process that is the output of the model when it is driven by white noise with variance 1. ○

The power cepstrum is a real and even function:

$$c(k) = c(-k) \in \mathbb{R}\ . \tag{5.6}$$

The estimation of the power cepstrum of a stationary stochastic process

Parametric as well as non-parametric power spectral density estimates can be used for the estimation of the power cepstrum of a stationary stochastic process (see Remark 3.5). The most commonly used parametric power spectral density estimate is obtained from AR modeling of the signal. Once a model is identified, the cepstrum can be computed from the model parameters directly, as we will see in Section 5.4.

Examples of non-parametric power spectral density estimates used in cepstral representation are the periodogram and the smoothed periodogram. We now describe how the power cepstrum of a stationary process can be estimated from a finite sequence of observations via the periodogram (without smoothing). For a given sequence $y(0), \ldots, y(K-1)$ with discrete Fourier transform (DFT)

$$Y(n) = \sum_{k=0}^{K-1} y(k)e^{-i\frac{2\pi}{N}nk} \quad (n = 0, \ldots, K-1)\ ,$$

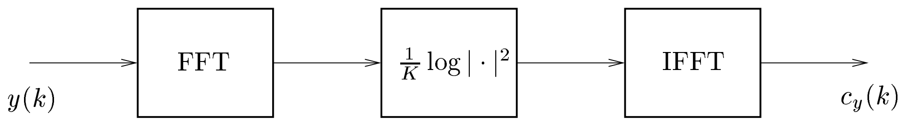


Figure 5.5: Estimation of the power cepstrum.

the periodogram estimate of the spectrum is $S_y(n) = \frac{1}{K}|Y(n)|^2$, and the cepstral components are obtained as the inverse DFT of the logarithm of the periodogram:

$$c_y(k) = \frac{1}{K}\sum_{n=0}^{K-1} \log\left(\frac{1}{K}|Y(n)|^2\right) e^{i\frac{2\pi}{N}nk} \quad (k = 0, \ldots, K-1)\ .$$

By using the fast Fourier transform algorithm (FFT) [28], we obtain a computationally efficient method, depicted in Figure 5.5.

Example 5.2. The cepstrum of the mast data
As an example, we compute the power cepstrum of the data used in Example 3.1, which are acceleration measurements on a steel transmitter mast. In Figure 5.6 a non-parametric estimate is compared to a parametric estimate. The non-parametric estimate is based on the periodogram of the 3839 measurements and it is obtained by the method in Figure 5.5. For the parametric estimate, we use the 12th order ARMA model, identified in Example 3.1 and of which the autocovariance function and the spectrum are shown in Figure 3.4. From the model, the power cepstrum is obtained by applying the formulas in Proposition 5.1 below. ◊

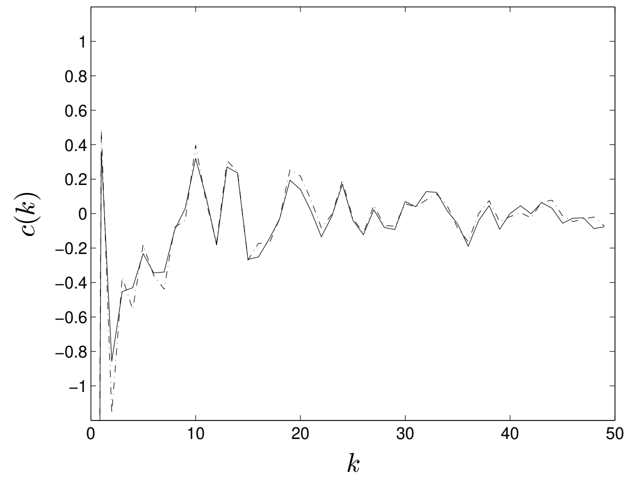


Figure 5.6: A non-parametric (full line) and a parametric (dash-dot line) estimation of the power cepstrum of the mast measurements, introduced in Example 3.1. The model used for the parametric estimation is the 12th order ARMA model that we identified in Example 3.1.

5.3 The complex cepstrum

Sometimes, it is useful not only to detect the echoes, but also to remove them and recover the original signal without echoes. Or, in general, to deconvolve two signals. Using the power cepstrum, however, one discards the phase information and an inversion of the operations is thus not possible. The complex cepstrum does not have this disadvantage. Again, we start with a motivational example and give the definition of the complex cepstrum afterwards.

Example 5.3. Echo detection and removal
Let the signal $s(k)$ be equal to $h(k) + 0.4h(k-40)$, where $h(k)$ is the impulse response of the ARMA model with poles and zeros as in Example 5.1. So, instead of sending white noise through the ARMA model of Example 5.1, we consider its impulse response. The signal $s(k)$ can be viewed as the convolution of $d(k)$, defined in (5.1), and $h(k)$. It contains an echo of $h(k)$ at $k = 40$ with intensity 0.4. Assume that we want to restore the original signal without echo, i.e. $h(k)$.
We know how to detect the echo in the power cepstrum $c_s(k)$ and by filtering the cepstrum with a comb filter, which boils down to leaving out the rahmonics produced by the echo, it can be removed from the cepstrum. Transforming the adapted cepstrum back to the time domain is, however, not possible because the magnitude of the Fourier transform of $s(k)$ was taken to obtain $c_s(k)$ (see Definition 5.2) and the phase information has been lost. Yet, instead of taking the squared magnitude, one could as well work with the Fourier transform $S(e^{i\theta})$ (or the more general Z-transform) directly, which then results in the complex cepstrum $\bar{c}_s(k)$.
A diagram of the procedure to obtain the complex cepstrum of $s(k)$ is given in Figure 5.7. Note that the logarithm of the complex function $S(z)$ is taken. Hence its name 'complex' cepstrum. The complex logarithm should be defined properly, as will be discussed further.

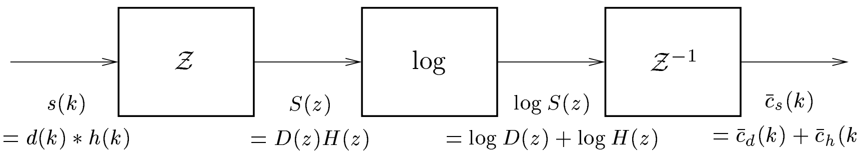


Figure 5.7: Schematic presentation of the computation of the complex cepstrum of $s(k)$.

As will be shown in Corollary 5.2, the complex cepstra of $d(k)$ and $h(k)$ are equal to the power cepstra of $d(k)$, respectively $\{y(k)\}$ in Example 5.1 for $k \geq 0$. Consequently, $\bar{c}_s(k)$ is equal to $c_z(k)$ for $k \geq 0$, which can be seen in Figure 5.4. After removing the spikes at multiples of $D = 40$, we transform the adapted complex cepstrum, denoted by $\bar{c}_s^*(k)$ back to the time domain as in Figure 5.8.

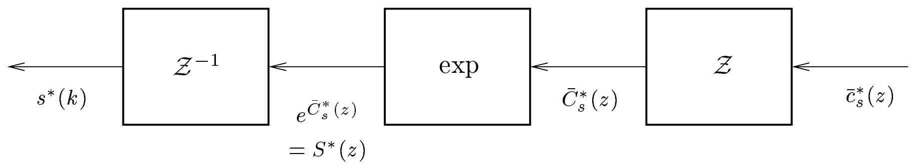


Figure 5.8: Schematic presentation of the inverse operation: going from the cepstral domain back to the time domain.

The result of the echo-removal can be seen in Figure 5.9, where $h(k)$, $s(k)$ and $s^*(k)$ are shown in the top picture and the difference between $s^*(k)$ and $h(k)$ in the bottom picture. We managed to deprive $s(k)$ of its echo pretty well. $\diamond$

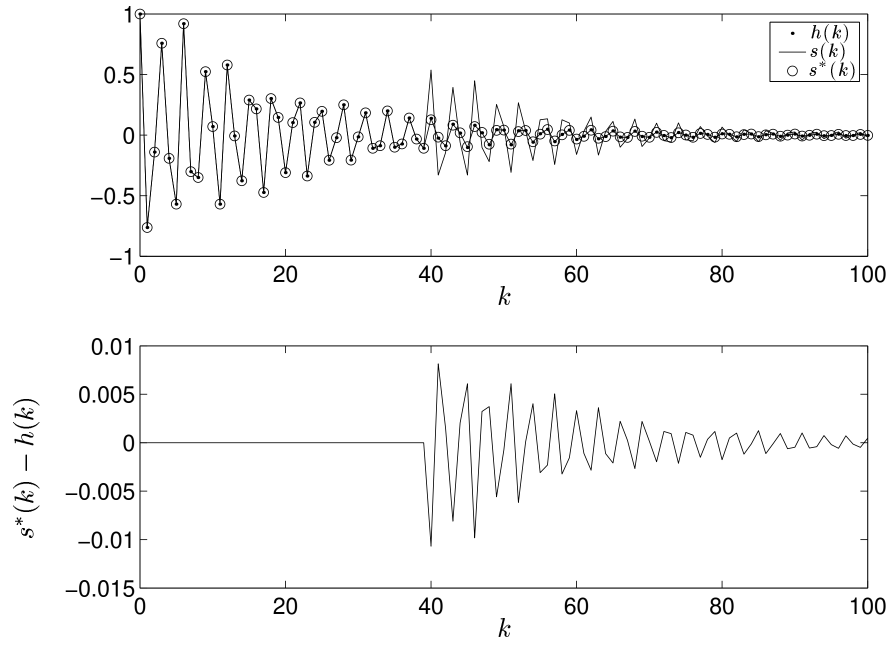


Figure 5.9: The result of the echo-removal from $s(k)$. In the top picture we show three signals: the signal without echo (h), represented by the dots, the signal with an echo at $k = 40$ (s) in full line and the signal from which the echo is removed (s^*) by the circles. The second picture gives the difference of s^* and h, i.e. the error of the restored signal compared to the original echo-free signal. Note that the scales in the two pictures is different!

The example was given as a motivation for the following definition of the complex cepstrum (see also Figure 5.7).

Definition 5.3. The complex cepstrum
Let $s(k)$ be an absolutely summable sequence with Z-transform $S(z)$. The complex cepstrum $\bar{c}_s(k)$ of $s(k)$ is defined as the inverse Z-transform of the complex logarithm of $S(z)$:

$$\bar{c}_s(k) = \frac{1}{2\pi i} \oint_C \log(S(z)) z^{k-1} \, dz \ ,$$

where the complex logarithm of $S(z)$ is appropriately defined and the contour C is the unit circle.

Remark 5.2. Since a stationary stochastic process $\{y(k)\}$ does not have a Z-transform, its complex cepstrum cannot be defined. The complex cepstrum of the autocovariance function $R_y(\tau)$ of the process does exist and it is equal to the power cepstrum of $\{y(k)\}$ because the Fourier transform of $R_y(\tau)$ is the power spectrum of $\{y(k)\}$

$$\bar{c}_{R_y}(k) = c_y(k) \ .$$

○

Remark 5.3. The complex logarithm
Let z be a complex number with magnitude r and phase φ, thus $z = re^{i\varphi}$. The logarithm of z is then equal to

$$\log(re^{i\varphi}) = \log r + i(\varphi + k2\pi) \qquad k \in \mathbb{Z} \ .$$

While there is no ambiguity in defining the real part of the complex logarithm ($\log r$), the imaginary part is multi-valued.
In defining the complex cepstrum of a sequence $s(k)$, the complex logarithm of its Z-transform $S(z)$ is taken, which gives

$$\log S(z) = \log|S(z)| + i \log\left(\arg(S(z)) + k2\pi\right) \ ,$$

where $\arg(S(z))$ is the phase of $S(z)$. Hence, a problem of uniqueness arises in the definition of the complex cepstrum of a sequence $s(k)$ because any multiple of 2π can be added to any point of $\arg(S(z))$ without changing the original signal $s(k)$, but the complex cepstrum is altered by such phase changes.
The problem is overcome by demanding that on the unit circle the phase of $S(z)$ be an odd, continuous function of θ with period[5] 2π (for details see [113, pp. 495–497]). This implies that the complex cepstrum $\bar{c}_s(k)$ is a real (not necessarily even) function. The procedure ensuring that $\arg(S(e^{i\theta}))$ has the demanded properties is called 'phase unwrapping', for which algorithms are discussed in [147]. The inverse transformation from cepstral domain to time domain as illustrated in Figure 5.8, does not cause any problems since the (complex) exponential function is single-valued. ○

[5] The period of $\arg(S(e^{i\theta}))$ can also be smaller than 2π, as long as $\arg(S(e^{i(\theta+2\pi)}))$ is equal to $\arg(S(e^{i(\theta)}))$.

It can be shown (see e.g. [113, p. 497–498]) that an absolutely summable sequence $s(k)$ and its complex cepstrum are implicitly related as follows:

$$s(k) = \frac{1}{k} \sum_{n=-\infty}^{\infty} n\bar{c}_s(n) s(k-n), \;\; k \neq 0 \; . \tag{5.7}$$

Based on this expression, the cepstral coefficients of an ARMA process can be obtained from the Markov parameters of the corresponding model and vice versa (see Section 5.5).

5.4 The cepstrum of an ARMA process

Throughout this thesis at different occasions, we have drawn the attention to the properties of SISO ARMA processes (Section 3.2.5, Section 3.5.1, Example 4.14). The examples in this Chapter were also based on them. In the next chapter we will define a norm (and associated metric) for these processes which is based on the cepstrum. Therefore, we discuss in this section the properties of the power cepstrum of ARMA processes. As we will see in Proposition 5.1, the power cepstrum of an ARMA process can be expressed in terms of the input noise variance σ^2 and the poles and zeros of the model. The same is possible for the complex cepstrum of the impulse response of the ARMA model. It does not come as a surprise that these two cepstra are strongly related. How they are related, is given in Corollary 5.2.
Let us first recall that we have defined ARMA models without gain in the transfer function. The gain in the process is modeled by the input variance σ^2 (see also Remark 3.6).

Proposition 5.1. The power cepstrum of a stable and minimum phase ARMA process [113, p. 502]
Let $\{y(k)\}$ be the output of the stable and minimum phase ARMA(p,q) model with poles $\alpha_1, \ldots, \alpha_p$ and zeros $\beta_1, \ldots, \beta_q$ that is driven by white noise with variance σ^2. Then its power cepstrum is equal to

$$c_y(k) = \begin{cases} \log \sigma^2 & k = 0 \; , \\ \displaystyle\sum_{i=1}^{p} \frac{\alpha_i^{|k|}}{|k|} - \sum_{i=1}^{q} \frac{\beta_i^{|k|}}{|k|} & k \neq 0 \; . \end{cases} \tag{5.8}$$

Proof.
The spectral density of the ARMA model can be written as

$$\Phi(e^{i\theta}) = \sigma^2 \frac{\prod_{i=1}^{q} |e^{i\theta} - \beta_i|^2}{\prod_{i=1}^{p} |e^{i\theta} - \alpha_i|^2} = \frac{\prod_{i=1}^{q} |1 - \beta_i e^{-i\theta}|^2}{\prod_{i=1}^{p} |1 - \alpha_i e^{-i\theta}|^2} \; .$$

The logarithm of $\Phi(e^{i\theta})$ is then

$$\log\Phi(e^{i\theta}) = \log(\sigma^2) + \sum_{i=1}^{q}\left(\log(1-\beta_i e^{-i\theta}) + \log(1-\bar{\beta}_i e^{i\theta})\right) - \sum_{i=1}^{p}\left(\log(1-\alpha_i e^{-i\theta}) + \log(1-\bar{\alpha}_i e^{i\theta})\right) \ .$$

Since the poles and zeros lie inside the unit circle, we can apply the power-series expansion $\log(1-x) = -\sum_{k=1}^{\infty}\frac{x^k}{k}$ for $|x|<1$:

$$\log\Phi(e^{i\theta}) = \log(\sigma^2) + \sum_{i=1}^{p}\left(\sum_{k=1}^{\infty}\frac{\alpha_i^k}{k}e^{-ik\theta} + \sum_{k=1}^{\infty}\frac{\bar{\alpha}_i{}^k}{k}e^{ik\theta}\right) - \sum_{i=1}^{q}\left(\sum_{k=1}^{\infty}\frac{\beta_i^k}{k}e^{-ik\theta} + \sum_{k=1}^{\infty}\frac{\bar{\beta}_i^k}{k}e^{ik\theta}\right) \ . \tag{5.9}$$

On the other hand, the logarithm of the spectrum is also the Fourier transform of the power cepstrum of $\{y(k)\}$:

$$\log\Phi(e^{i\theta}) = \sum_{k=-\infty}^{\infty} c_y(k)e^{-ik\theta} \ . \tag{5.10}$$

By equating the terms in (5.9) and (5.10), we obtain

$$c_y(k) = \begin{cases} \log\sigma^2 & k=0 \ , \\ \displaystyle\sum_{i=1}^{p}\frac{\alpha_i^k}{k} - \sum_{i=1}^{q}\frac{\beta_i^k}{k} & k>0 \ , \\ \displaystyle\sum_{i=1}^{q}\frac{\bar{\beta}_i^{-k}}{k} - \sum_{i=1}^{p}\frac{\bar{\alpha}_i^{-k}}{k} & k<0 \ , \end{cases}$$

and since the poles and zeros occur in complex conjugate pairs,

$$c_y(k) = \begin{cases} \log\sigma^2 & k=0 \ , \\ \displaystyle\sum_{i=1}^{p}\frac{\alpha_i^{|k|}}{|k|} - \sum_{i=1}^{q}\frac{\beta_i^{|k|}}{|k|} & k\neq 0 \ . \end{cases}$$

□

Note that $c_y(0)=0$ if the input variance σ^2 is equal to 1.
In Example 5.1 we have seen that the power cepstrum of the ARMA process $\{y(k)\}$ decays faster than its autocovariance function (Figure 5.4). This is partly explained by formula (5.8). While the kth point in the autocovariance

of an ARMA process is a linear combination of the poles to the power $k-1$ (this can be deduced from Equation (3.11)), the kth cepstral coefficient is attenuated by an extra factor $\frac{1}{k}$. That the decay of the cepstrum is not always faster than the decay of the autocovariance function is shown in the following example.

Example 5.4. Autocovariance function and cepstrum of an MA model
We consider the 4th order MA model with transfer function

$$\frac{z^4 + 0.3586z^3 - 0.1994z^2 - 0.8650z - 0.0171}{z^4} .$$

In Figure 5.10 the autocovariance function $R(\tau)$ is given in full line, while the cepstrum of the model $c(k)$ is depicted in dash-dot line. The cepstrum keeps oscillating, while the autocovariance function is equal to zero from $\tau = 5$ onwards. ◊

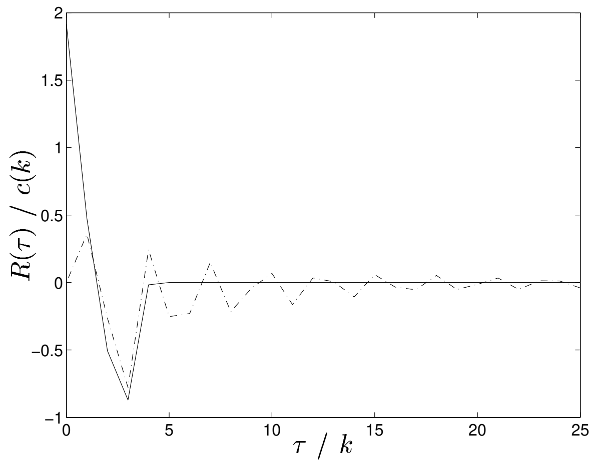


Figure 5.10: The autocovariance (full) and cepstrum (dash-dot) of a 4th order moving average model.

The power cepstrum of the ARMA process $\{y(k)\}$ is related to the complex cepstrum of the impulse response. This result will be needed in Chapter 6.

Corollary 5.2. The power cepstrum of an ARMA process and the complex cepstrum of the model's impulse response
For $k > 0$, the impulse response's complex cepstrum, $\bar{c}_h(k)$, is equal to the power cepstrum $c_y(k)$ of the ARMA process. The coefficients of $\bar{c}_h(k)$ for $k \leq 0$ are zero:

$$\bar{c}_h(k) = \begin{cases} c_y(k) & k > 0 , \\ 0 & k \leq 0 . \end{cases} \tag{5.11}$$

Proof.
The Z-transform of the impulse response $h(k)$ is the transfer function of the ARMA model $H(z) = \frac{\prod_{i=1}^{q}(1-\beta_i z^{-1})}{\prod_{i=1}^{p}(1-\alpha_i z^{-1})}$. In an analogous way as the proof of Proposition 5.1 one obtains

$$\bar{c}_h(k) = \begin{cases} \sum_{i=1}^{p} \frac{\alpha_i^k}{k} - \sum_{i=1}^{q} \frac{\beta_i^k}{k} & k > 0 \ , \\ 0 & k \leq 0 \ . \end{cases}$$

□

Remark 5.4. The complex cepstrum of an ARMA model
Henceforth we will call the complex cepstrum of the impulse response of an ARMA model briefly the complex cepstrum of the model. ○

Remark 5.5. The complex cepstrum of the inverse model
Let $\bar{c}_H(k)$ be the complex cepstrum of the ARMA model with transfer function $H(z)$. Then, the complex cepstrum of the inverse model, denoted by $\bar{c}_{H^{-1}}(k)$ satisfies

$$\bar{c}_{H^{-1}}(k) = -\bar{c}_H(k) \ .$$

This is easy to see as the poles become zeros and vice versa.
The same holds true for the power cepstra of the output processes of the models, if the noise variance of the input of both models is equal. ○

5.5 System identification based on the cepstrum

Until now, we have worked with the cepstrum as a tool for signal processing, because this is the area where it originated. In this way we could indicate how these ideas emerged. The cepstrum has, however, also made its entry in the system identification community. In particular, methods have been developed for the identification of a linear model of a scalar stationary stochastic process, based on (an estimate of) the power cepstrum of that process. In this section we give a brief overview of two possible methods. In Section 5.5.1 we show how the Markov parameters (i.e. the impulse response sequence) can be computed from the cepstral coefficients. The Markov parameters can then be used to obtain a linear model, as explained in Section 3.3.1. Another approach, called cepstral prediction [111], is explained in Section 5.5.2. Using this method, one obtains the poles and zeros of the ARMA model directly from the cepstrum, without the detour via the Markov parameters.
Byrnes et al. [23, 22] and Enqvist [46, 47] take another approach to the incorporation of cepstral coefficients in system identification. They start from the rational covariance extension problem and bring in the cepstral coefficients in two ways. The first approach is inspired by the maximum entropy model and the second consists in simultaneously interpolating a window of covariance estimates and a window of cepstral coefficients.

5.5.1 Identification of an ARMA model from the cepstral coefficients via the Markov parameters

Proposition 5.1 shows how the power cepstrum of an ARMA process can be computed if the input variance and the poles and zeros of the model are known. It can also be obtained from the input variance and the model's Markov parameters (the impulse response). Conversely, the Markov parameters and the input variance can be derived from the power cepstrum. Indeed, applying (5.7) to the impulse response $h(k)$ of a stable and minimum phase ARMA model, leads to the following recursion:

$$h(k) = \frac{1}{k}\sum_{n=1}^{k} n\bar{c}_h(n)h(k-n) \text{ for } k > 0 \ . \tag{5.12}$$

Since $h(0) = 1$ for the models that we consider (see Section 3.2.5), the expression in (5.12) gives a way to compute the Markov parameters $h(1), \ldots, h(N)$ from the complex cepstral coefficients $\bar{c}_h(1), \ldots, \bar{c}_h(N)$. Note that in an identification problem we do not have the complex cepstrum $\bar{c}_h(k)$ of the impulse response at our disposal, but an estimate of the power cepstrum $c_y(k)$ of the observed ARMA process. However, from Corollary 5.2 we know that $\bar{c}_h(1) = c_y(1), \ldots, \bar{c}_h(N) = c_y(N)$. Consequently, the Markov parameters can be determined from the power cepstrum. Furthermore, the input variance is immediately obtained as $\sigma^2 = e^{c_y(0)}$ (see Proposition 5.1).
Having obtained the Markov parameters, the next step is given by the algorithms designed for the deterministic identification problem, where the model parameters are computed from the Markov parameters, e.g. the algorithm described in Section 3.3.

Example 5.5. Identification of an ARMA model from the power cepstrum estimate
We have generated a time series $y(k)$ by sending a sequence of 50 000 samples of white noise with variance 1 through the ARMA$(4, 4)$ model of Example 5.1. The power cepstrum is estimated via the periodogram, as shown in Figure 5.5 from which the first 100 Markov parameters are computed as in (5.12). Via the deterministic realization algorithm of Kung, which is described in Section 3.3, we obtain an ARMA$(4, 4)$ model. This procedure is repeated 100 times and the location of the resulting poles and zeros of the models is shown in Figure 5.11. $\diamondsuit$

5.5.2 Cepstral prediction

Let the transfer function of an ARMA(n, n) model with poles α_i $(i = 1, \ldots, n)$ and zeros β_i $(i = 1, \ldots, n)$ be denoted by $H(z) = \frac{b(z)}{a(z)}$ The impulse response

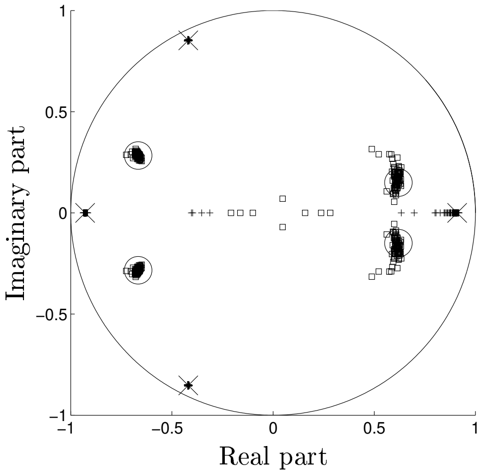


Figure 5.11: The poles ($\times$) and zeros (in the center of the big circles) of the ARMA model used to generate the process $\{y(k)\}$ and the poles (+) and zeros ($\square$) of the identified ARMA(4, 4) models.

$h(k)$ of the model can then be written as

$$h(k) = \sum_{i=1}^{n} c_i \alpha_i^k \qquad k > 0 \ ,$$

and also its autocovariance function is a weighted sum of damped exponentials, determined by the poles:

$$R(\tau) = \sum_{i=1}^{n} d_i \alpha_i^\tau \qquad \tau > 0 \ ,$$

where c_i and d_i, $i = 1, \ldots, n$ are weighting constants. From Proposition 5.1 we know that the power cepstrum of the model is equal to

$$c_y(k) = \frac{1}{k} \sum_{i=1}^{n} (\alpha_i^k - \beta_i^k) \text{ for } k > 0 \ ,$$

and thus can the weighted sequence $kc_y(k)$ $(k > 0)$ be viewed as the impulse response or the autocovariance function of an ARMA$(2n, 2n)$ model with poles that correspond to the poles and zeros of the original nth order model. This can also be seen from the Z-transform of $kc_y(k) = k\bar{c}_h(k)$ $(k > 0)$, namely $-z\frac{d}{dz}\log H(z)$, which is a rational function with denominator $a(z)b(z)$.
Given a finite time series $y(k)$ from which the power cepstrum $c_y(k)$ is estimated, a deterministic identification algorithm, or an efficient covariance-based AR identification algorithm applied to $kc_y(k)$ produces a model with poles that

are estimates for the poles and zeros of the original model. However, classifying which pole of the model is a pole and which is a zero of the original model is not always easy. Two methods for this classification problem are given in [111]. Even for data originating from models with an irrational transfer function of the form

$$H(z) = \prod_{i=1}^{p} \frac{1}{(1-\alpha_i z^{-1})^{r_i}} ,$$

where $r_i \in \mathbb{R}$, which are sometimes called fractional models, the sequence $kc_y(k) = k\bar{c}_h(k)$ $(k > 0)$ has a rational Z-transform:

$$-z\frac{d}{dz}\log H(z) = \sum_{i=1}^{p} \frac{r_i \alpha_i z^{-1}}{1-\alpha_i z^{-1}} .$$

Consequently, the same identification technique as described above could be used to find fractional models from cepstral data.

Remark 5.6. Multiplying with k

Although the variance on the estimates of the cepstral coefficients decreases with k, which is shown in [78], one should take care not to amplify numerical and estimation errors when multiplying $c(k)$ by k for large k's. This is illustrated in Figure 5.12, where we consider the spectrum, the cepstrum and the weighted cepstrum of the ARMA model used in the Examples 5.1, 5.3 and 5.5. The power cepstrum of the ARMA process (input is white noise with variance 1) is estimated in two ways. A parametric estimation of the spectrum is obtained by identifying a 10th order AR model. The true spectrum (full line) of the ARMA process and the parametric estimate (dash-dot line) are given in picture A of Figure 5.12. A non-parametric estimate, namely the periodogram without smoothing, which we have computed with the FFT algorithm, is shown in picture B. In picture C, the true power cepstrum (full line), its parametric estimate (dash-dot) and non-parametric estimate (dotted) are depicted. It can be seen that both estimation methods give a good power cepstrum. However, by linearly weighing the estimated cepstra, small differences are amplified, as can be seen in picture D. The parametric estimate (dash-dot) still behaves well, while the non-parametric estimate (dotted) becomes unstable. ○

5.6 Conclusions

Via the motivating example of echo detection and echo removal, we have introduced the cepstrum. The power cepstrum as well as the complex cepstrum were defined, together with their properties that are useful for the derivations in Chapter 6. Furthermore, we have briefly indicated how the cepstrum can be used to identify a linear model. Especially the fact that the zeros of the model can be determined more easily than from the autocovariances, seems promising.

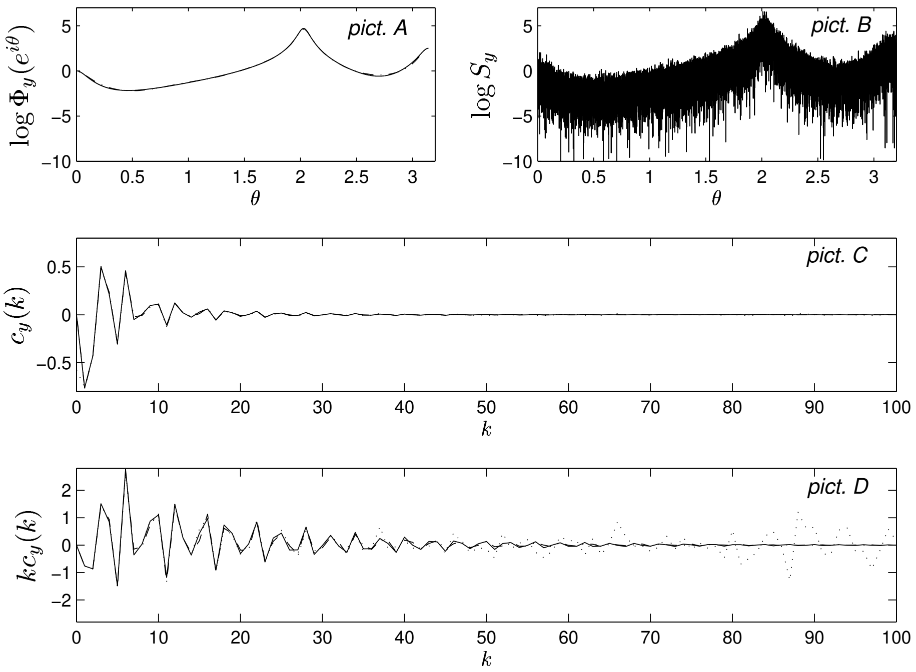
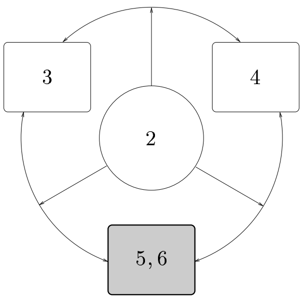
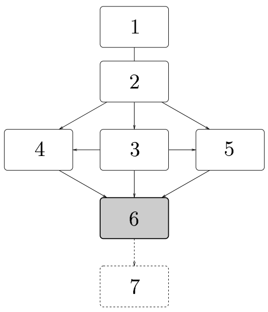


Figure 5.12: The logarithm of the spectrum, power cepstrum and weighted power cepstrum of an ARMA(4, 4) process, estimated in a parametric (via a 10th order AR model) and non-parametric (via the periodogram) way. In picture A the true spectrum (full line) and its parametric estimate (dash-dot) are given. The periodogram is shown in picture B. The true power cepstrum (full line) and its two estimates (parametric in dash-dot, non-parametric in dotted line) are depicted in picture C and in picture D their weighted versions are drawn.

Chapter 6

A Cepstral Norm for ARMA Models





As illustrated in the right pictogram above, the results and insights of Chapter 3, Chapter 4 and Chapter 5 are brought together in this chapter. The cepstrum, principal angles and canonical correlations, the mutual information of two stochastic processes, controllability and observability matrices, Gramians, Lyapunov equations, ... in this chapter we apply all these concepts to derive different interpretations of a cepstral norm for stable and minimum-phase ARMA models. Some relations have already been announced in the preceding chapters. In Remark 3.20, e.g., we stated that the canonical correlations of two ARMA processes with the same input depend on the ratio of the transfer functions, which is also the case for cepstral distance measures. Remark 4.12, Example 4.13 and Example 4.14 gave indications that the mutual information of the past and future of a scalar Gaussian stochastic process is a measure for the distance between the numerator and denominator AR models. In this chapter, these hypotheses are confirmed.

Contributions Martin proposed in [105] a new cepstral metric for ARMA models. We study first the associated cepstral norm of a model, of which the transfer function is denoted by $H(z) = \frac{b(z)}{a(z)}$ and its order is n. The cepstral norm of $\frac{b(z)}{a(z)}$ can be interpreted as the distance of the output process to white noise, or as the cepstral distance between the AR models $\frac{z^n}{a(z)}$ and $\frac{z^n}{b(z)}$. Starting from the expression of the cepstral norm in terms of the poles and zeros of the model, we prove a number of different representations for the norm.

- The cepstral norm is a function of the principal angles between the row spaces of the infinite controllability matrix of the model and the infinite controllability matrix of the inverse model.
- The cepstral norm as a function of canonical correlations:
  - the canonical correlations of the input and output process of the model,
  - the canonical correlations of the output processes of $\frac{z^n}{a(z)}$ and $\frac{z^n}{b(z)}$, if they have the same input,
  - the canonical correlations of the past and the future of the output process.
- Expressions for the cepstral norm in terms of Gramians, derived from the model parameters.

We prove that the cepstral norm of an ARMA model is equal to two times the mutual information of the past and the future output process, if the input is white Gaussian noise[1]. Furthermore, from the characterization of the cepstral norm as the Hilbert–Schmidt norm of an infinite Hankel matrix with cepstral coefficients, we derive a graphical interpretation. The squared cepstral norm of the ARMA model with transfer function $H(z)$ is proportional to the area enclosed by the polar plot of $\log H(z)$.
All these results for the cepstral norm of an ARMA model are translated to the cepstral distance between two models.
We also show how the cepstral norm and cepstral distance can be estimated from measurements. Estimating the distance between a given model and a measured signal can be done by filtering the signal by the inverse model, i.e. without explicitly modeling the observed data. We obtain elegant formulas for this distance between a model and a signal, which can easily be updated.

Chapter overview In Section 6.1 we introduce cepstral distance measures for signals and models. Throughout this chapter we will be dealing with one

[1] Note that we use the logarithm with base e throughout this chapter. The mutual information of the past and the future is then expressed in nats.

particular cepstral metric and its associated norm. They are defined in Section 6.2. We also give the formulation of the norm of an autoregressive moving average (ARMA) model in terms of the poles and zeros of the model. In Section 6.3 we relate the cepstral norm of an ARMA model to the principal angles between the row spaces of the controllability matrix of the model and the controllability matrix of the inverse model. Via the results of Chapter 3, this leads in Section 6.4 to the characterization of the norm as a function of the canonical correlations of the input and output process of the model and also of the canonical correlations of the past and future output. Before relating this to the mutual information of the past and the future in Section 6.6, we derive in Section 6.5 a number of expressions for the norm in terms of Gramians. In Section 6.6 we then show that the cepstral norm of an ARMA model is proportional to the mutual information of the past and the future of the corresponding ARMA process, provided it is Gaussian. How the cepstral norm of an ARMA model can be estimated from measurements of the output process, is explained in Section 6.7. A graphical interpretation of the norm is given in Section 6.8. We conclude our discussion of the cepstral norm in Section 6.9 with an overview of expressions for the cepstral norm of an ARMA model.
We then return in Section 6.10 to the cepstral distance between two models and characterize it by means of the derived interpretations for the norm, from which we obtain methods to compute the distance. Estimating the distance between a given model and measurements, is particularly interesting for monitoring applications, which is illustrated in one of the examples of Section 6.11.

6.1 Introduction

Distance measures between two signals, between a signal and a model, or between two stochastic models are widely used concepts in signal processing for solving various problems such as detection, automatic segmentation, classification, pattern recognition, coding [6]. By looking at distance measures that depend on the ratio between the spectra of the signals or models, i.e. on the difference between the log-spectra and hence on the difference between the cepstra, we obtain measures that are independent of the measurement scale.

Example 6.1. Cepstral distance measures are independent of the measurement scale
Let $\Phi^{(1)}(e^{i\theta})$ and $\Phi^{(2)}(e^{i\theta})$ be the power spectra of two processes, measured with the same scale. When measuring them with another scale, for example one of which the unit is 10 times smaller than the unit of the first scale, we obtain the values $10\Phi^{(1)}(e^{i\theta})$ and $10\Phi^{(2)}(e^{i\theta})$. Cepstral distance measures depend on the difference between the log-spectra, which gives

$$\log\left(10\Phi^{(1)}(e^{i\theta})\right) - \log\left(10\Phi^{(2)}(e^{i\theta})\right) = \log\left(\Phi^{(1)}(e^{i\theta})\right) - \left(\Phi^{(2)}(e^{i\theta})\right) \; .$$

The two measurement scales will thus give the same distance. A cepstral distance between two ARMA processes that are generated by white input processes with the same variance will consequently always be the same, irrespective of the common input variance:

$$\log\left(\sigma^2 \frac{|b^{(1)}(e^{i\theta})|^2}{|a^{(1)}(e^{i\theta})|^2}\right) - \log\left(\sigma^2 \frac{|b^{(2)}(e^{i\theta})|^2}{|a^{(2)}(e^{i\theta})|^2}\right) = \log\left(\eta^2 \frac{|b^{(1)}(e^{i\theta})|^2}{|a^{(1)}(e^{i\theta})|^2}\right) - \log\left(\eta^2 \frac{|b^{(2)}(e^{i\theta})|^2}{|a^{(2)}(e^{i\theta})|^2}\right) .$$

◊

In speech processing, in particular, these cepstral distance measures are intensively used and studied, see [41, pp. 377–379], [135, 145, 78, 61, 62]. But also in modal analysis, one often displays the logarithm of the spectrum, see for example Figure 3.4. Furthermore, when comparing two Bode diagrams one is also comparing two models on a logarithmic scale.
One possible cepstral distance measure for stationary stochastic processes is the Euclidean distance between two cepstral sequences,

$$\sqrt{\sum_{k=0}^{\infty} \left(c^{(1)}(k) - c^{(2)}(k)\right)^2} .$$

Note that since the power cepstra are even functions (see (5.6)), we only need to sum over $k \geq 0$. In addition to the Euclidean distance, various weighting schemes of the form

$$\sum_{k=0}^{\infty} w_k \left(c^{(1)}(k) - c^{(2)}(k)\right)^2 , \tag{6.1}$$

where w_k $(k = 0, 1, \ldots)$ are weighting coefficients, have been proposed, which might lead to performance improvements, depending on the application. We pick out one of the weighting schemes, namely the one proposed by Martin in [105] ($w_k = k$), and give several new interpretations for this distance measure for ARMA models.

6.2 Definition and properties

The definition of Martin [105] for the cepstral distance between two ARMA processes is a weighted version of the Euclidean distance between two power cepstrum sequences, where the weights w_k are equal to k. Let $\{y^{(1)}(k)\}$ and $\{y^{(2)}(k)\}$ be ARMA processes with power cepstrum $c^{(1)}(k)$ and $c^{(2)}(k)$ respectively. Then, the squared cepstral distance between the two processes, which we will temporarily denote by $d(y^{(1)}, y^{(2)})^2$, is

$$d(y^{(1)}, y^{(2)})^2 = \sum_{k=0}^{\infty} k \left(c^{(1)}(k) - c^{(2)}(k) \right)^2 \ . \tag{6.2}$$

In practice, of course, only finite sequences of estimated cepstral coefficients are used in the distance computation.
In addition to the scale invariance property of cepstrum-based distance measures (see Example 6.1), the distance measure of (6.2) is independent of the zeroth cepstral coefficient of each process, which is for an ARMA process equal to the logarithm of the input variance (see Proposition 5.1). Since the other cepstral coefficients are independent of the input variance, the distance measure is independent of the input variance of the individual processes. It only depends on the dynamic characteristics, i.e. the poles and the zeros, of both models.

Example 6.2. A cepstral distance between two ARMA processes
Consider two ARMA models with transfer function $H^{(1)}(z)$ and $H^{(2)}(z)$, respectively. Let $\{y_a^{(1)}(k)\}$ and $\{y_b^{(1)}(k)\}$ be the output processes of $H^{(1)}$ driven by white noise with variance σ_a^2 and σ_b^2 respectively, while the output processes of $H^{(2)}$ driven by white noise with variance η_a^2 and η_b^2, respectively are denoted by $\{y_a^{(2)}(k)\}$ and $\{y_b^{(2)}(k)\}$. The power cepstra of $\{y_a^{(1)}\}$ and $\{y_b^{(1)}\}$ are equal, except for the zeroth coefficient, which is $\log \sigma_a^2$ and $\log \sigma_b^2$ and similarly for the power cepstra of $\{y_a^{(2)}\}$ and $\{y_b^{(2)}\}$. Consequently, the cepstral distance (6.2) between $\{y_a^{(1)}(k)\}$ and $\{y_a^{(2)}(k)\}$ is equal to the distance between $\{y_b^{(1)}(k)\}$ and $\{y_b^{(2)}(k)\}$

$$d(y_a^{(1)}, y_a^{(2)}) = d(y_b^{(1)}, y_b^{(2)}) \ ,$$

and also

$$d(y_a^{(1)}, y_a^{(2)}) = d(y_a^{(1)}, y_b^{(2)}) = d(y_b^{(1)}, y_a^{(2)}) \ .$$

Furthermore, $\{y_a^{(1)}\}$ and $\{y_b^{(1)}\}$ lie at a distance 0 from each other, which also holds for $\{y_a^{(2)}\}$ and $\{y_b^{(2)}\}$

$$d(y_a^{(1)}, y_b^{(1)}) = d(y_a^{(2)}, y_b^{(2)}) = 0 \ .$$

◊

This implies that the distance measure in (6.2) is not a metric for ARMA processes, since a metric must obey the condition $d(y^{(1)}, y^{(2)}) = 0 \Leftrightarrow y^{(1)} = y^{(2)}$. Nevertheless, it is a metric for the set of ARMA processes that are generated by sending white noise with variance 1 through an ARMA model with transfer function $H(z) = \frac{b(z)}{a(z)}$, of which the polynomial coefficients b_0 and a_0 are equal to 1 (see also Section 3.2.5). In other words, it is a metric for a set of equivalence classes in which all ARMA processes that differ only in the zeroth cepstral coefficient (or equivalently, the input variance) are considered to be equivalent.

Alternatively, one can solve the ambiguity by defining the metric on the set of stable and minimum phase ARMA *models* and use the complex cepstrum of the impulse response of the models instead of the power cepstrum of the output processes. Since the zeroth coefficient of the complex cepstrum of an ARMA model is equal to 0 (see Corollary 5.2), it does not matter that the zeroth weight w_0 is equal to 0. We will use this definition.

Definition 6.1. A weighted cepstral distance measure for ARMA models
Given two stable and minimum phase ARMA models with transfer functions $H^{(1)}(z)$ and $H^{(2)}(z)$ and complex cepstra $\bar{c}^{(1)}(k)$ and $\bar{c}^{(2)}(k)$ respectively, the squared distance between the models is defined as

$$d(\log H^{(1)}, \log H^{(2)})^2 = \sum_{k=1}^{\infty} k \left(\bar{c}^{(1)}(k) - \bar{c}^{(2)}(k)\right)^2 . \tag{6.3}$$

Remark 6.1. Falling back on the transfer functions and the complex cepstrum instead of the ARMA processes and their power cepstrum does not mean that we cannot compute the distance between two observed processes any more. For stable and minimum phase models the complex cepstrum coefficients $\bar{c}(k)$ for $k \geq 0$ are equal to the power cepstrum coefficients of the output processes (see Corollary 5.2). By estimating the power cepstrum of the observed sequences and applying (6.2), we consequently obtain an estimate of the distance between the generating ARMA models, which can be used to quantify the distance between the processes, bearing in mind that it is independent of the input variances. ○

The cepstral distance between the models with transfer function $H^{(1)}(z)$ and $H^{(2)}(z)$ is denoted by $d(\log H^{(1)}, \log H^{(2)})$ (and not $d(H^{(1)}, H^{(2)})$ for example) in order to have a metric on a linear space. In this way, we are able to associate the following norm with the metric.

Definition 6.2. A weighted cepstral norm for ARMA models
Given a stable and minimum phase ARMA model with transfer function $H(z)$ and complex cepstrum $\bar{c}(k)$, the cepstral norm of this model is defined as

$$\|\log H\|^2 = \sum_{k=1}^{\infty} k \bar{c}(k)^2 . \tag{6.4}$$

Remark 6.2. Cepstral norm and cepstral distance
Consider the ARMA model with transfer function $\frac{H^{(1)}(z)}{H^{(2)}(z)}$. Its complex cepstrum is $\bar{c}^{(1)}(k) - \bar{c}^{(2)}(k)$, where $\bar{c}^{(1)}(k)$ and $\bar{c}^{(2)}(k)$ are the complex cepstra of $H^{(1)}(z)$ and $H^{(2)}(z)$ respectively. The squared norm of this model is thus equal

to $\sum_{k=1}^{\infty} k(\bar{c}^{(1)}(k) - \bar{c}^{(2)}(k))^2$, which is the cepstral distance of Definition 6.1 between the models with transfer function $H^{(1)}(z)$ and $H^{(2)}(z)$. Hence, we can write

$$\| \log H^{(1)} - \log H^{(2)} \| = d(\log H^{(1)}, \log H^{(2)}) \ , \tag{6.5}$$

which partly explains why we keep the logarithm in the notation of the norm and the distance. It also shows that by first studying the properties of the norm, we will be able to translate the results to the distance measure.

On the other hand, we note that the cepstral norm of a model with transfer function $H^{(1)}(z)$ can be seen as the cepstral distance between the model and the model with transfer function $H^{(2)} = 1$:

$$d(\log H^{(1)}, \log H^{(2)}) = \| \log H^{(1)} - \log H^{(2)} \| = \| \log H^{(1)} \| \ .$$

Hence, the cepstral norm is a measure for the non-whiteness of the output process of the model. The smaller the norm, the whiter the process.

From Equation (6.5) immediately follows another interpretation of the cepstral norm as a cepstral distance. The cepstral norm of an ARMA model with transfer function $H(z) = \frac{b(z)}{a(z)}$ is equal to the cepstral distance between the AR models with transfer function $\frac{z^n}{a(z)}$ and $\frac{z^n}{b(z)}$, respectively, and also to the cepstral distance between the MA models with transfer function $\frac{a(z)}{z^n}$ and $\frac{b(z)}{z^n}$:

$$\left\| \log \frac{b(z)}{a(z)} \right\| = d\left(\log \frac{z^n}{b(z)}, \log \frac{z^n}{a(z)} \right) = d\left(\log \frac{b(z)}{z^n}, \log \frac{a(z)}{z^n} \right) \ , \tag{6.6}$$

○

Remark 6.3. The cepstral norm of the inverse model
The cepstral norm of the inverse model is equal to the cepstral norm of the model, which is easily seen as follows:

$$\| \log H^{-1}(z) \| = \| - \log H(z) \| = \| \log H(z) \| \ .$$

○

Following the derivation of Martin [105], the cepstral norm of an ARMA model can be expressed in terms of its poles and zeros, which will be useful in Section 6.3.

Proposition 6.1. The cepstral norm of an ARMA model in terms of the poles and zeros
The squared cepstral norm of a stable and minimum phase ARMA(p,q) model with poles $\alpha_1, \ldots, \alpha_p$ and zeros $\beta_1, \ldots, \beta_q$ is equal to

$$\| \log H \|^2 = \log \frac{\prod_{i=1}^{p} \prod_{j=1}^{q} \left| 1 - \alpha_i \bar{\beta}_j \right|^2}{\prod_{i,j=1}^{p} (1 - \alpha_i \bar{\alpha}_j) \prod_{i,j=1}^{q} (1 - \beta_i \bar{\beta}_j)} \ . \tag{6.7}$$

This equality follows from the expression that relates the cepstrum coefficients of an ARMA model to its poles and zeros (see Proposition 5.1) and the identity $\sum_{k=1}^{\infty} \frac{x^k}{k} = -\log(1-x)$ for $|x| < 1$.
Formula (6.7) was the main inspiration source for our derivation of the cepstral norm in terms of principal angles, as will be shown in the next section.

6.3 The cepstral norm as a function of principal angles

In this section we derive a relationship between the cepstral norm of an nth order ARMA model with transfer function $\frac{b(z)}{a(z)}$ and the principal angles between the row spaces of the infinite controllability matrices of the AR models $\frac{z^n}{a(z)}$ and $\frac{z^n}{b(z)}$. This result will then be used in Section 6.6 to show that the cepstral norm of an ARMA model and the mutual information of the past and the future of the output process of the model are closely related.
We start this section with a simple example, which reveals the connection between the norm of Definition 6.2 and certain principal angles.

Example 6.3. The cepstral norm of an ARMA model of order 1
Let $H(z) = \frac{b(z)}{a(z)}$ be the transfer function of an ARMA model of order 1 with pole α and zero β. Then, by Proposition 6.1, the squared cepstral norm of this model is equal to

$$\|\log H\|^2 = \log \frac{(1-\alpha\beta)^2}{(1-\alpha^2)(1-\beta^2)} .$$

Consider the vectors

$$u = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots \end{pmatrix} \in \mathbb{R}^{1\times\infty} \text{ and } v = \begin{pmatrix} 1 & \beta & \beta^2 & \cdots \end{pmatrix} \in \mathbb{R}^{1\times\infty} ,$$

which can be seen as the infinite controllability matrices of the first order AR models $\frac{z}{a(z)}$ and $\frac{z}{b(z)}$. Since $|\alpha|$ and $|\beta|$ are smaller than 1 ($H(z)$ is stable and minimum phase), the squared cosine of the angle θ between u and v is equal to

$$\begin{aligned} \cos^2\theta &= \frac{|uv^T|^2}{\|u\|^2\|v\|^2} , \\ &= \frac{(1-\alpha^2)(1-\beta^2)}{(1-\alpha\beta)^2} , \end{aligned}$$

which is obtained by applying $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ for $|x| < 1$.
Consequently, the squared cepstral norm of the ARMA model of order 1 is equal to $\|\log H\|^2 = -\log\cos^2\theta$. ◊

For higher order models, this can be generalized by using the principal angles between the row spaces of the corresponding controllability matrices.

Theorem 6.2. The cepstral norm of an nth order ARMA model in terms of principal angles
For the stable and minimum phase ARMA model of order n with transfer function $H(z) = \frac{b(z)}{a(z)}$, the weighted cepstral norm of Definition 6.2 can be expressed in terms of the principal angles $\theta_1, \ldots, \theta_n$ between the row spaces of the infinite controllability matrices $\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)$ as follows

$$\|\log H\|^2 = -\log \prod_{k=1}^{n} \cos^2 \theta_k \ .$$

Proof.
The first part of the proof is similar to the proof of Proposition 3.9.
Assume the ARMA model has distinct poles $\alpha_1, \ldots, \alpha_n$ and distinct zeros $\beta_1, \ldots, \beta_n$. Then, by Proposition 3.2, the row spaces of the matrices $\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)$ are equal to the row spaces of $\mathcal{C}^*_\infty\left(\frac{z^n}{a(z)}\right)$ and $\mathcal{C}^*_\infty\left(\frac{z^n}{b(z)}\right)$ respectively, where

$$\mathcal{C}^*_\infty\left(\frac{z^n}{a(z)}\right) = \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots \\ 1 & \alpha_2 & \alpha_2^2 & \cdots \\ \vdots & \vdots & \vdots & \ldots \\ 1 & \alpha_n & \alpha_n^2 & \cdots \end{pmatrix} , \ \mathcal{C}^*_\infty\left(\frac{z^n}{b(z)}\right) = \begin{pmatrix} 1 & \beta_1 & \beta_1^2 & \cdots \\ 1 & \beta_2 & \beta_2^2 & \cdots \\ \vdots & \vdots & \vdots & \ldots \\ 1 & \beta_n & \beta_n^2 & \cdots \end{pmatrix} .$$

The squared cosines of the principal angles between the row spaces of $\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)$ can then be computed as the n eigenvalues of ${P^*_{aa}}^{-1} P^*_{ab} {P^*_{bb}}^{-1} P^*_{ba}$, where $P^*_{ab} = \mathcal{C}^*_\infty\left(\frac{z^n}{a(z)}\right) {\mathcal{C}^*_\infty}^H\left(\frac{z^n}{b(z)}\right)$ and analogously for P^*_{aa}, P^*_{bb} and $P^*_{ba} = {P^*_{ab}}^H$ (the superscript $\cdot^H$ denotes the complex conjugate transpose).
The product of the squared cosines of the principal angles between the row spaces of $\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)$ is therefore equal to

$$\prod_{k=1}^{n} \cos^2 \theta_k = \det\left({P^*_{aa}}^{-1} P^*_{ab} {P^*_{bb}}^{-1} P^*_{ba}\right) = \frac{|\det P^*_{ab}|^2}{\det P^*_{aa} \det P^*_{bb}} \ . \tag{6.8}$$

The elements of the matrices P^*_{aa}, P^*_{ab} and P^*_{bb} can be computed from the poles and zeros of the ARMA model by applying

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \text{ for } |x| < 1 \ .$$

For P^*_{ab} e.g. one obtains:

$$P^*_{ab}(i,j) = \frac{1}{1-\alpha_i \bar{\beta}_j} \ (i,j = 1, \ldots, n) \ , \tag{6.9}$$

where $P^*_{ab}(i,j)$ is the element of the matrix P^*_{ab} on the ith row and the jth column. The determinant of P^*_{ab} is equal to

$$\det P^*_{ab} = \frac{\prod_{i<j}^{n}(\alpha_i - \alpha_j)(\bar{\beta}_i - \bar{\beta}_j)}{\prod_{i,j}^{n}(1-\alpha_i\bar{\beta}_j)} \ ,$$

which can be proven by induction.
Similarly,

$$\det P^*_{aa} = \frac{\prod_{i<j}^{n}(\alpha_i - \alpha_j)(\bar{\alpha}_i - \bar{\alpha}_j)}{\prod_{i,j}^{n}(1-\alpha_i\bar{\alpha}_j)} \quad \text{and} \quad \det P^*_{bb} = \frac{\prod_{i<j}^{n}(\beta_i - \beta_j)(\bar{\beta}_i - \bar{\beta}_j)}{\prod_{i,j}^{n}(1-\beta_i\bar{\beta}_j)} \ .$$

Consequently,

$$\log \prod_{k=1}^{n} \cos^2\theta_k = \log \frac{\prod_{i,j=1}^{n}(1-\alpha_i\bar{\alpha}_j)(1-\beta_i\bar{\beta}_j)}{\prod_{i,j=1}^{n}\left|1-\alpha_i\bar{\beta}_j\right|^2} \ . \tag{6.10}$$

By Proposition 6.1, the right hand side of (6.10) is equal to $-\|\log H\|^2$. Consequently,

$$\|\log H\|^2 = -\log \prod_{k=1}^{n} \cos^2\theta_k \ . \tag{6.11}$$

If the ARMA model has poles or zeros with multiplicity larger than 1, Equation (6.10) remains valid and consequently, (6.11) holds for this situation too. □

6.4 The cepstral norm in terms of canonical correlations

By combining the results of Chapter 3 with Theorem 6.2, we obtain relations between the cepstral norm and canonical correlations between certain processes. This is shown in the following corollaries.

Corollary 6.3. The cepstral norm in terms of the canonical correlations of the future input and future output
Let $\tau_1, \ldots, \tau_n$ be the smallest n canonical correlations of the future input and future output processes of the nth order ARMA model with transfer function $H(z)$. Then, the cepstral norm of the model is equal to

$$\|\log H\|^2 = -\log \prod_{k=1}^{n} \tau_k^2 \ . \tag{6.12}$$

Proof.
As shown in Theorem 3.12, the smallest n canonical correlations $\tau_1, \ldots, \tau_n$ of the future input and output processes of the ARMA model are equal to the cosines of the principal angles between the row spaces of $\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)$. The other canonical correlations of the processes are equal to 1. By Theorem 6.2, the cepstral norm can consequently be expressed as

$$\|\log H\|^2 = -\log \prod_{k=1}^{n} \tau_k^2 \ .$$

□

Corollary 6.4. The cepstral norm in terms of the canonical correlations of the past and future output process
Let the largest n canonical correlations of the past and the future output process of the ARMA model be denoted by ρ_k $(k = 1, \ldots, n)$. Then

$$\|\log H\|^2 = -\log \prod_{k=1}^{n} (1 - \rho_k^2) \ . \tag{6.13}$$

Proof.
This follows immediately from the complementarity of the canonical correlations of the future input and the future output on the one hand and the canonical correlations of the past and the future output on the other hand (see Property 3.7). Together with Corollary 6.3, this gives (6.13). □

The relation between $\sum_{k=1}^{\infty} kc(k)^2$ and the canonical correlations of the past and the future in (6.13) can also be found in [74, Proposition 2], where it is derived in a different way and is used to obtain a bound on the maximum canonical correlation.
Note that, since only the smallest n canonical correlations of u_f and y_f are different from 1 and only the largest n canonical correlations of y_p and y_f are different from 0, we can express the cepstral norm also as follows:

$$\begin{aligned} \|\log H\|^2 &= -\log \prod \tau_k^2 \ , \\ &= -\log \prod (1 - \rho_k^2) \ , \end{aligned} \tag{6.14}$$

where all canonical correlations are taken into account.

6.5 The cepstral norm in terms of Gramians

Starting from the formula $\|\log H\|^2 = -\log \prod_{k=1}^{n} \cos^2\theta_k$ (see Theorem 6.2), where $H(z) = \frac{b(z)}{a(z)}$ is of order n and

$$(\theta_1, \theta_2, \ldots, \theta_n) = \left[\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right) \sphericalangle \mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)\right] , \tag{6.15}$$

we can express the norm in terms of controllability and observability Gramians. First note that since the row space of a controllability matrix of a SISO ARMA model only depends on its poles (Proposition 3.2), we can also write

$$\mathrm{row}\left(\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)\right) = \mathrm{row}(\mathcal{C}_\infty(H(z))) \ ,$$

and

$$\mathrm{row}\left(\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)\right) = \mathrm{row}\big(\mathcal{C}_\infty(H^{-1}(z))\big) \ .$$

We denote $\mathcal{C}_\infty(H(z))$ by $\mathcal{C}_\infty$ and $\mathcal{C}_\infty(H^{-1}(z))$ by $\mathcal{C}_{z_\infty}$. We also know that $\mathrm{row}(\mathcal{C}_\infty) = \mathrm{range}(\Gamma_\infty)$ and $\mathrm{row}(\mathcal{C}_{z_\infty}) = \mathrm{range}(\Gamma_{z_\infty})$ (see Remark 3.9). A consequence of these observations is that the principal angles $\theta_1, \ldots, \theta_n$ in (6.15) can be described in the following ways:

$$\begin{aligned}(\theta_1, \theta_2, \ldots, \theta_n) &= [\mathcal{C}_\infty \sphericalangle \mathcal{C}_{z_\infty}] = [\mathcal{C}_\infty \sphericalangle \Gamma^T_{z_\infty}] \\ &= [\Gamma^T \sphericalangle \mathcal{C}_{z_\infty}] = [\Gamma^T \sphericalangle \Gamma^T_{z_\infty}] \ .\end{aligned} \tag{6.16}$$

Consequently, the cepstral norm can be written as a function of the Gramians, corresponding to the different pairs of spaces in (6.16).

1. $(\theta_1, \ldots, \theta_n) = [\mathcal{C}_\infty \sphericalangle \mathcal{C}_{z_\infty}]$
   The squared cosines of the principal angles $(\theta_1, \theta_2, \ldots, \theta_n)$ are the eigenvalues of $P^{-1}P_{ab}P_z^{-1}P_{ba}$, where

   $$\begin{pmatrix} P & P_{ab} \\ P_{ba} & P_z \end{pmatrix} = \begin{pmatrix} \mathcal{C}_\infty \\ \mathcal{C}_{z_\infty} \end{pmatrix} \begin{pmatrix} \mathcal{C}^T_\infty & \mathcal{C}^T_{z_\infty} \end{pmatrix}$$

   is the controllability Gramian of $\left(\begin{pmatrix} A & 0 \\ 0 & A-KC \end{pmatrix}, \begin{pmatrix} K \\ K \end{pmatrix}\right)$, which can be computed via the controllability Lyapunov equation (3.22). The squared cepstral norm is thus equal to

   $$\|\log H\|^2 = -\log \frac{(\det P_{ab})^2}{\det P \det P_z} \ .$$

2. $(\theta_1, \ldots, \theta_n) = \left[\Gamma_\infty^T \sphericalangle \Gamma_{z_\infty}^T\right]$
The squared cosines of the principal angles $(\theta_1, \theta_2, \ldots, \theta_n)$ are the eigenvalues of $Q^{-1} Q_{ab} Q_z^{-1} Q_{ba}$, where
$$\begin{pmatrix} Q & Q_{ab} \\ Q_{ba} & Q_z \end{pmatrix} = \begin{pmatrix} \Gamma_\infty^T \\ \Gamma_{z_\infty}^T \end{pmatrix} \begin{pmatrix} \Gamma_\infty & \Gamma_{z_\infty} \end{pmatrix}$$
is the observability Gramian of $\left(\begin{pmatrix} A & 0 \\ 0 & A - KC \end{pmatrix}, \begin{pmatrix} C & -C \end{pmatrix}\right)$, which can be computed via the observability Lyapunov equation (3.23). Consequently, the squared cepstral norm can be written as
$$\| \log H \|^2 = -\log \frac{(\det Q_{ab})^2}{\det Q \det Q_z} \ .$$

3. $(\theta_1, \ldots, \theta_n) = \left[\mathcal{C}_\infty \sphericalangle \Gamma_{z_\infty}^T\right]$
The squared cosines of the principal angles $(\theta_1, \theta_2, \ldots, \theta_n)$ are the eigenvalues of $P^{-1} R Q_z^{-1} R^T$, where
$$\begin{pmatrix} P & R \\ R^T & Q_z \end{pmatrix} = \begin{pmatrix} \mathcal{C}_\infty \\ \Gamma_{z_\infty}^T \end{pmatrix} \begin{pmatrix} \mathcal{C}_\infty^T & \Gamma_{z_\infty} \end{pmatrix}$$
is the controllability Gramian of $\left(\begin{pmatrix} A & 0 \\ 0 & (A - KC)^T \end{pmatrix}, \begin{pmatrix} K \\ -C^T \end{pmatrix}\right)$. This leads to
$$\| \log H \|^2 = -\log \frac{(\det R)^2}{\det P \det Q_z} \ .$$

4. $(\theta_1, \ldots, \theta_n) = \left[\Gamma_\infty^T \sphericalangle \mathcal{C}_{z_\infty}\right]$
The squared cosines of the principal angles $(\theta_1, \theta_2, \ldots, \theta_n)$ are the eigenvalues of $Q^{-1} S P_z^{-1} S^T$, where
$$\begin{pmatrix} Q & S \\ S^T & P_z \end{pmatrix} = \begin{pmatrix} \Gamma_\infty^T \\ \mathcal{C}_{z_\infty} \end{pmatrix} \begin{pmatrix} \Gamma_\infty & \mathcal{C}_{z_\infty}^T \end{pmatrix}$$
is the observability Gramian of $\left(\begin{pmatrix} A & 0 \\ 0 & (A - KC)^T \end{pmatrix}, \begin{pmatrix} C & K^T \end{pmatrix}\right)$. This leads to
$$\| \log H \|^2 = -\log \frac{(\det S)^2}{\det Q \det P_z} \ .$$

Based on the expression of the cepstral norm in terms of the canonical correlations of the future input and future output processes in (6.12), we obtain yet another description. The smallest n squared canonical correlations of the future input and output are the eigenvalues of $(I_n + Q_z P)^{-1}$, see (3.105), and the other canonical correlations are equal to 1. As above, the matrix Q_z is the observability Gramian of the inverse model and P is the controllability Gramian of the model. The squared cepstral norm is consequently also equal to
$$\| \log H \|^2 = \log \det (I_n + Q_z P) \ .$$

6.6 The cepstral norm and the mutual information of the past and the future output

We now derive the relation of the information theoretic notion of mutual information and the cepstral norm of an ARMA model. Although the proof is short, the complete derivation includes many results of the thesis. In Figure 6.1 we describe the crucial steps that lead to the theorem. Based on [74], this characterization of the mutual information of past and future in terms of the cepstral coefficients is also derived in [95], where it is applied to model selection and order determination.

Theorem 6.5. The cepstral norm and the mutual information of past and future
The squared cepstral norm of an ARMA model is equal to twice the mutual information of the past and the future output process of the model, when driven by white Gaussian noise:

$$\|\log H\|^2 = \sum_{k=1}^{\infty} k\bar{c}^2(k) = 2I(y_p; y_f) \ . \tag{6.17}$$

Proof.
The mutual information of the past and the future of a Gaussian process that is generated by an nth order linear model is equal to (see Corollary 4.7)

$$I(y_p; y_f) = -\frac{1}{2}\log \prod_{k=1}^{n} (1-\rho_k^2) \ ,$$

where $\rho_1, \rho_2, \ldots, \rho_n$ are the largest n canonical correlations of the past and the future output. Comparing this expression with (6.13), immediately leads to Equation (6.17). □

Remark 6.4. Estimation of the mutual information of the past and the future of a scalar Gaussian process
Theorem 6.5 gives us a new way to estimate the mutual information of the past and the future of a scalar Gaussian process. First estimate a sequence of power cepstral coefficients $\hat{c}(1), \hat{c}(2), \ldots, \hat{c}(K)$. The mutual information can then be approximated by $\frac{1}{2}\sum_{k=1}^{K} k\hat{c}^2(k)$.
Since the cepstral norm is independent of $\bar{c}(0)$ and hence of the input variance σ^2, we get here yet another proof for the fact that the mutual information of the past and the future of the output of a linear SISO model is independent of σ^2. This was also derived in Section 4.4.2, based on the results of Section 3.5.1. ○

cepstral norm

$$\|\log H\|^2 = \sum k\bar{c}^2(k)$$

Proposition 6.1

poles and zeros
$H(z) = \prod \frac{(z-\beta_k)}{(z-\alpha_k)}$

$$\|\log H\|^2 = \log \prod \frac{\left|1-\alpha_i\bar{\beta}_j\right|^2}{(1-\alpha_i\bar{\alpha}_j)(1-\beta_i\bar{\beta}_j)}$$

Theorem 6.2

$(\theta_1, \ldots, \theta_n) = [\mathcal{C}_\infty \sphericalangle \mathcal{C}_{z_\infty}]$

$$\|\log H\|^2 = -\log \prod \cos^2\theta_k$$

Theorem 3.12

$\mathrm{cc}(u_f, y_f) = \cos[\mathcal{C}_\infty \sphericalangle \mathcal{C}_{z_\infty}], 1, \ldots, 1$

$$\|\log H\|^2 = -\log \prod \tau_k^2$$

Property 3.7
(based on Proposition 3.8
and Theorem 3.11)

$\mathrm{cc}(y_p, y_f) = \sin[\mathcal{C}_\infty \sphericalangle \mathcal{C}_{z_\infty}], 1, \ldots, 1$

$$\|\log H\|^2 = -\log \prod (1-\rho_k^2)$$

Corollary 4.7

$\{y(k)\}$ Gaussian
$I(y_p; y_f) = -\frac{1}{2}\log \prod_{k=1}^{n}(1-\rho_k^2)$

$$\|\log H\|^2 = 2I(y_p; y_f)$$

Figure 6.1: The different steps that lead to Theorem 6.5.

Remark 6.5. The mutual information of past and future as a non-whiteness measure
Since the cepstral norm of an ARMA model is a measure for the distance of the output process to white noise (see Remark 6.2), we here see that the mutual information of past and future of the Gaussian output process of a linear model is also a measure for the non-whiteness of the process. We already indicated this property at the end of Chapter 4, see Remark 4.12 and Example 4.13. ○

6.7 Estimating the cepstral norm from measurements

Until now we have given a number of parametric methods for the computation of the cepstral norm. It can, however, also be estimated from measured output data. Obviously, one could estimate the power cepstrum of the data and apply Definition 6.2. But we could work even more directly on the data, without estimating the cepstrum.
A first approach is via the principal angles between the row spaces of the past and future output Hankel matrices. If a finite amount of data is available, then the canonical correlations of the past and future output process, $\rho_1, \rho_2, \ldots$, can be approximated by the cosines of the principal angles between the row spaces of the $i \times j$ data Hankel matrices. Let the principal angles between the row space of Y_p and Y_f be denoted by $\phi_1, \ldots, \phi_i$, then $\rho_k \approx \cos \phi_k$. By (6.14), the following expression can be used as an approximation of the cepstral norm of the model:

$$\|\log H\|^2 \approx -\log \prod_{k=1}^{i} \sin^2 \phi_k \ . \tag{6.18}$$

Or, as in Example 4.12, we could first determine the model order from the principal angles and then take only the smallest n principal angles into account.
The computation can be made simpler since we do not need to know the principal angles separately, but only the product of the sines. In the following corollary we show how the cepstral norm can be estimated by computing determinants of output covariance matrices. A similar characterization of the mutual information of the past and the future of a Gaussian stationary stochastic process can be found in [95].

Corollary 6.6. Computing the cepstral norm from the output Hankel matrix
Let a finite amount of measurements of an ARMA process $\{y(k)\}$ be given. Then, the cepstral norm of the underlying ARMA model can be estimated from the $2i \times j$ data Hankel matrix $Y = \begin{pmatrix} Y_p \\ Y_f \end{pmatrix}$ as

$$\|\log H\|^2 \approx 2 \log \det(Y_p Y_p^T) - \log \det(Y Y^T) \ . \tag{6.19}$$

Proof.
Throughout the proof, we use the following notation:

$$Q_{2i} = YY^T = \begin{pmatrix} Y_p \\ Y_f \end{pmatrix} \begin{pmatrix} Y_p^T & Y_f^T \end{pmatrix} = \begin{pmatrix} Q_{p_i} & Q_{pf_i} \\ Q_{fp_i} & Q_{f_i} \end{pmatrix} .$$

The squared cosines of the principal angles $\phi_1, \ldots, \phi_i$ between row(Y_p) and row(Y_f) are the eigenvalues of $Q_{p_i}^{-1} Q_{pf_i} Q_{f_i}^{-1} Q_{fp_i}$. The product of the squared sines of the angles, as in (6.18), is then equal to

$$\begin{aligned} \prod_{k=1}^{i} \sin^2 \phi_k &= \det(I_i - Q_{p_i}^{-1} Q_{pf_i} Q_{f_i}^{-1} Q_{fp_i}) , \\ &= \det\left(Q_{p_i}^{-1}(Q_{p_i} - Q_{pf_i} Q_{f_i}^{-1} Q_{pf_i}) \right) , \\ &= \frac{\det(Q_{p_i} - Q_{pf_i} Q_{f_i}^{-1} Q_{pf_i})}{\det Q_{p_i}} . \end{aligned}$$

Since the determinant of Q_{2i} is equal to

$$\det Q_{2i} = \det Q_{p_i} \det \left(Q_{p_i} - Q_{pf_i} Q_{f_i}^{-1} Q_{fp_i} \right) ,$$

we can write

$$\prod_{k=1}^{i} \sin^2 \phi_k = \frac{\det Q_{2i}}{(\det Q_{p_i})^2} ,$$

and by (6.18) we consequently obtain

$$\| \log H \|^2 \approx - \log \frac{\det Q_{2i}}{(\det Q_{p_i})^2} = 2 \log \det(Y_p Y_p^T) - \log \det(YY^T) .$$

□

Remark 6.6. Using the LQ decomposition to compute the determinants
Let the LQ decomposition of the $2i \times j$ data Hankel matrix Y be equal to LQ^T, where $L \in \mathbb{R}^{2i \times j}$ is lower triangular and $Q \in \mathbb{R}^{j \times j}$ is orthogonal (see Definition 2.5). We only have to compute the triangular matrix L. The $2i$ diagonal elements of L are denoted by $l_1, l_2, \ldots, l_{2i}$. Then, the determinants of YY^T and $Y_p Y_p^T$ are respectively equal to

$$\det(YY^T) = \prod_{k=1}^{2i} l_k^2 \quad \text{and} \quad \det(Y_p Y_p^T) = \prod_{k=1}^{i} l_k^2 ,$$

The approximation of the cepstral norm in (6.19) can consequently be rewritten as

$$\| \log H \|^2 \approx \sum_{k=1}^{i} \log l_k^2 - \sum_{k=i+1}^{2i} \log l_k^2 .$$

In this way, the cepstral norm can be updated when new data become available, since updating the diagonal elements of L can easily be done by Givens rotations, see e.g. [57, pp. 610–611]. We will see in Example 6.6 that updating the cepstral norm is very useful when one wants to track the distance between a nominal model of a system and measurements of the actual system. ○

6.8 A graphical interpretation of the cepstral norm

In this section we obtain the relation between the system norm $\|\log H\|$, defined in (6.4), and the area enclosed by the curve $\delta(\theta) = \log H(e^{i\theta})$ where θ is running from 0 to 2π. This curve is indeed closed in the complex plane due to the appropriate choice of the complex logarithm (see Remark 5.3). By the area enclosed by $\delta(\theta)$ we mean the following integral:

$$A(\delta) = \int_{x+iy\in\delta} y\,dx \ . \tag{6.20}$$

Note that areas that are encircled twice should be counted twice. For example, the area enclosed by $\delta(\theta)$ in Figure 6.2 is equal to $A_1 + 2A_2 + 2A_3$, where A_1, A_2 and A_3 are indicated in Figure 6.2.
In Section 6.8.1 we first show that the cepstral norm $\|\log H\|$ is equal to the Hilbert–Schmidt norm of the Hankel matrix with the cepstrum coefficients. In Section 6.8.2 we give the relation between $\|\log H\|$ and the area enclosed by $\delta(\theta)$.

6.8.1 The Hilbert–Schmidt–Hankel norm for cepstra

Let $\bar{c}(k)$ denote the complex cepstrum coefficients of the model with transfer function $H(z)$ and consider the double infinite Hankel matrix $\mathcal{H}$

$$\mathcal{H} = \begin{pmatrix} \bar{c}(1) & \bar{c}(2) & \bar{c}(3) & \cdots \\ \bar{c}(2) & \bar{c}(3) & \bar{c}(4) & \cdots \\ \bar{c}(3) & \bar{c}(4) & \bar{c}(5) & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \in \mathbb{R}^{\infty\times\infty} \ .$$

The Hilbert–Schmidt norm of $\mathcal{H}$ is given by

$$\|\mathcal{H}\|_{HS}^2 = \mathrm{tr}\mathcal{H}\mathcal{H}^T = \sum_{k=1}^{\infty} k\bar{c}(k)^2 \ .$$

Because the right hand side is equal to the squared cepstral system norm of Definition 6.2, we obtain

$$\|\log H\|^2 = \|\mathcal{H}\|_{HS}^2 \ .$$

6.8.2 Relation to the polar plot of $\log H(z)$

In [65] Hanzon considers the Hankel matrix with the Markov parameters $H(k)$ of a model without direct feed-through term (i.e. $H(0) = 0$). The Markov parameters form a sequence defined on $\mathbb{Z}^+$, whose Z-transform is equal to $H(z)$. The Hilbert–Schmidt norm of that Hankel matrix, i.e. the Hilbert–Schmidt–Hankel norm of $H(z)$, denoted by $\|H(z)\|_{HSH}$, is related to the Nyquist diagram of the transfer function $H(z)$ in the following way:

$$\|H(z)\|_{HSH}^2 = \sum_{k=1}^{\infty} kH(k)^2 = \frac{1}{\pi}A(\gamma) \ ,$$

where $A(\gamma)$ is the area enclosed by the curve $\gamma(\theta) = H(e^{i\theta})$, $\theta \in [0, 2\pi)$ running from 0 to 2π. An analogous result holds for the cepstral norm of Definition 6.2.

Proposition 6.7. The cepstral norm as the area enclosed by the curve $\log H(e^{i\theta})$
Let $H(z)$ be the transfer function of a stable and minimum phase ARMA model. The squared cepstral norm of $H(z)$ is then equal to

$$\|\log H\|^2 = \frac{1}{\pi}A(\delta) \ , \tag{6.21}$$

where $A(\delta)$ is the area enclosed by $\delta(\theta) = \log H(e^{i\theta})$, $\theta \in [0, 2\pi)$ running from 0 to 2π.

Proof.
Due to the fact that $\bar{c}(k) = 0$ for $k \leq 0$ (see Corollary 5.2), the cepstral norm of $H(z)$ that is defined in (6.4), is equal to

$$\|\log H\|^2 = \sum_{k=-\infty}^{\infty} k\bar{c}(k)^2 \ .$$

Applying Parseval's theorem leads to the following line integral over the unit circle

$$\begin{aligned} \|\log H\|^2 &= -\frac{1}{2\pi i}\oint \tfrac{d}{dz}(\log H(z))\overline{\log H(z)}\, dz \\ &= -\frac{1}{2\pi i}\int_\delta \overline{\log H(z)}\, d(\log H(z)) \ , \end{aligned}$$

where $\delta(\theta)$ is the closed curve $\log H(e^{i\theta})$, and θ is running from 0 to 2π. This can further be written as

$$\|\log H\|^2 = -\frac{1}{2\pi i}\int_{x+iy\in\delta} (x - iy)\, d(x + iy) \ .$$

By Stokes' theorem this is equal to [2]

$$\|\log H\|^2 = \frac{1}{2\pi i}\iint d\big((x-iy)\,d(x+iy)\big) \ ,$$

which leads to

$$\begin{aligned}\|\log H\|^2 &= \frac{1}{2\pi i}\iint d(x-iy)\,d(x+iy)\\ &= \frac{1}{2\pi i}\iint 2i\,dx\,dy\\ &= \frac{1}{\pi}\iint dx\,dy \ ,\end{aligned}$$

and this is $\frac{1}{\pi}A(\delta)$. □

Example 6.4. Graphical interpretation of the cepstral norm
As an example we take the fifth order ARMA model with poles $0.9 \pm 0.1i$, $0.2 \pm 0.8i$ and -0.95 and zeros $-0.5 \pm 0.82i$, $-0.1 \pm 0.7i$ and 0.92. Its transfer function

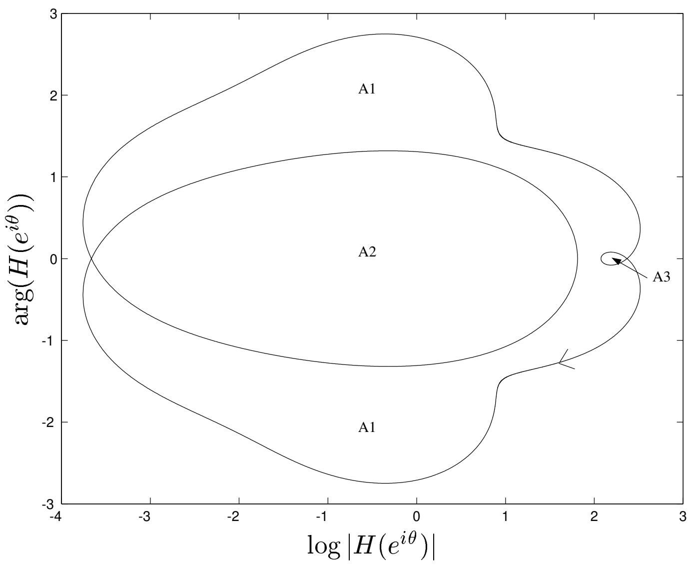


Figure 6.2: The curve $\delta(\theta) = \log H(e^{i\theta})$ where θ is running from 0 to 2π. The starting point is $(2.0746, 0)$ and the orientation is indicated by the arrow.The area enclosed by δ is $A_1 + 2A_2 + 2A_3$ and this is equal to 11.1925π.

is denoted by $H(z)$. In Figure 6.2 the curve $\delta(\theta) = \log H(e^{i\theta})$ corresponding

[2] The change of sign is due to the clockwise orientation of δ.

to the ARMA model is drawn. The curve starts and ends in $(2.0746, 0)$ and its orientation is indicated by the arrow. The area enclosed by δ is the sum

$$A(\delta) = A_1 + 2A_2 + 2A_3 \ ,$$

where A_1, A_2 and A_3 are indicated in Figure 6.2. The area is obtained by computing the integral in (6.20) numerically in Matlab, see Table 6.1 for the program. Multiplying $A(\delta)$ by $\frac{1}{\pi}$ results in 11.1925, and the formula for the cepstral norm as a function of the poles and zeros in (6.7), gives the same result. Note that we do *not* need to identify A_1, A_2, A_3 in order to calculate the area. ◊

```
function A = area(num,den,w)
% AREA(NUM,DEN) computes the area enclosed by
% log(NUM(z)/DEN(z))
% AREA(NUM,DEN,W) uses the user-supplied vector W
% of frequencies (in radian/second from 0 to π)

if nargin < 3
  [mag,phase,w] = dbode(num,den,1);
else
  [mag,phase,w] = dbode(num,den,1,w);
end

phase_rad = pi/180*phase;
x = log(mag);
A = 2*trapz(x,phase_rad);
```

Table 6.1: The Matlab program `area.m` for the computation of the area enclosed by $\log H(e^{i\theta})$, $\theta \in [0, 2\pi)$.

6.9 Expressions for the cepstral norm

In the previous sections, we have seen that the cepstral norm of Definition 6.2 is related to the principal angles between controllability matrices, the canonical correlations of the past and future output, the mutual information of the past and the future, the surface of a polar plot. We now give an overview of expressions for the norm of an nth order ARMA model. If the model is known, the norm can be computed exactly. This is the parametric approach. We discuss a number of possibilities in Section 6.9.1. The norm can also be estimated from observations of the model's output process (non-parametric approach), which is dealt with in Section 6.9.2.

6.9.1 The exact computation of the cepstral norm

Consider the nth order stable and minimum phase ARMA model with transfer function $H(z) = \frac{b(z)}{a(z)}$, whose poles are denoted by $\alpha_1, \ldots, \alpha_n$ and the zeros by $\beta_1, \ldots, \beta_n$. Its state space matrices are A, K and C. The forward and backward state covariance matrices Σ and $\bar{\Sigma}$ can be determined by the Lyapunov equations (3.8) and (3.19). Furthermore, the controllability Gramian of the model is P and the observability Gramian of the inverse model is Q_z, which satisfy respectively the Lyapunov equations (3.22) and (3.46). We also consider the model $\left(\begin{pmatrix} A & 0 \\ 0 & A-KC \end{pmatrix}, \begin{pmatrix} K \\ K \end{pmatrix}, \begin{pmatrix} C & -C \end{pmatrix}\right)$, with controllability and observability Gramian $\begin{pmatrix} P & P_{ab} \\ P_{ba} & P_z \end{pmatrix}$ and $\begin{pmatrix} Q & Q_{ab} \\ Q_{ba} & Q_z \end{pmatrix}$ respectively. Then, the cepstral norm of the model can be expressed as follows.

1. In terms of the poles and zeros of the model [105]:

$$\|\log H\|^2 = \log \prod_{i,j=1}^{n} \frac{\left|1-\alpha_i\bar{\beta}_j\right|^2}{(1-\alpha_i\bar{\alpha}_j)(1-\beta_i\bar{\beta}_j)} \ . \tag{6.22}$$

2. In terms of the coefficients of the transfer function polynomials $a(z)$ and $b(z)$ [105]:

$$\|\log H\|^2 = \log \frac{\left(\operatorname{res}\left(a(z), z^n b(z^{-1})\right)\right)^2}{\operatorname{res}\left(a(z), z^n a(z^{-1})\right)\operatorname{res}\left(b(z), z^n b(z^{-1})\right)} \ , \tag{6.23}$$

   where $\operatorname{res}(f(z), g(z))$ is the resultant of the polynomials $f(z)$ and $g(z)$, which is defined as follows. For the polynomials $f(z) = f_0 z^p + f_1 z^{p-1} + \cdots + f_p$ and $g(z) = g_0 z^q + g_1 z^{q-1} + \cdots + g_q$, the resultant of f and g is the determinant of the $(q+p) \times (q+p)$ Sylvester matrix[3]

$$\begin{pmatrix} f_0 & f_1 & \cdots & f_p & 0 & \cdots & 0 \\ 0 & f_0 & f_1 & \cdots & f_p & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \ddots & \\ 0 & \cdots & 0 & f_0 & f_1 & \cdots & f_p \\ g_0 & g_1 & \cdots & g_q & 0 & \cdots & 0 \\ 0 & g_0 & g_1 & \cdots & g_q & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \ddots & \\ 0 & \cdots & 0 & g_0 & g_1 & \cdots & g_q \end{pmatrix} .$$

[3] The Sylvester matrix is the key element in the elimination theory of Sylvester for the solution of sets of polynomial equations. Moreover, note that the resultant of two polynomials is zero if and only if they have a common root.

3. In terms of the Gramians (Section 6.5):

$$\begin{aligned}\|\log H\|^2 &= -\log \frac{(\det P_{ab})^2}{\det P \det P_z} \ , \\ &= -\log \frac{(\det Q_{ab})^2}{\det Q \det Q_z} \ , \\ &= \log \det(I_n + Q_z P) \ .\end{aligned}$$

4. In terms of the state covariance matrices:

$$\|\log H\|^2 = -\log \det(I_n - \Sigma\bar{\Sigma}) \ .$$

   This immediately follows from (6.13) and (3.100).

5. For an AR model, we can express the cepstral norm as

$$\|\log H\|^2 = \log \det(I_n - P^{-1}A^n P A^{n^T}) \ ,$$

   which follows from (6.13) and the expression for the canonical correlations of the past and future of an AR process in Property 3.17.

6.9.2 Estimating the norm from observations

Assume that a sequence of K samples of an ARMA process is observed $y(1)$, ..., $y(K-1)$. Then, the norm of the underlying ARMA model can be estimated in several ways.

1. Based on the estimated cepstrum.
   (a) If one opts for a parametric estimation of the cepstrum, i.e. computing the cepstrum of an identified model, one should not compute the cepstrum, since the norm can be estimated from the model directly, as discussed in Section 6.9.1.
   (b) If the cepstrum is estimated by a non-parametric method, e.g. via the periodogram (see Figure 5.5), and assume that K cepstrum coefficients $\hat{c}(0), \ldots, \hat{c}(K-1)$ are estimated, then, the norm can approximately be obtained as

$$\|\log H\|^2 \approx \sum_{k=1}^{K-1} k\hat{c}(k)^2 \ .$$

2. In terms of the principal angles $\phi_1, \ldots, \phi_i$ between row(Y_p) and row(Y_f).

$$\|\log H\|^2 \approx -\log \prod_{k=1}^{i} \sin^2 \phi_k \ . \tag{6.24}$$

By looking at the principal angles, one could also first estimate the order n of the model, i.e. the number of angles that differ significantly from $\frac{\pi}{2}$, and then take into account only the smallest n angles:

$$\|\log H\|^2 \approx -\log \prod_{k=1}^{n} \sin^2 \phi_k \ ,$$

see also Example 4.12.

3. Directly from the data Hankel matrices (Section 6.7).

$$\|\log H\|^2 \approx 2 \log \det(Y_p Y_p^T) - \log \det(Y Y^T) \ ,$$

or with $l_1, \ldots, l_{2i}$ the diagonal elements of the triangular factor L in the LQ factorization of Y:

$$\|\log H\|^2 \approx \sum_{k=1}^{i} \log l_k^2 - \sum_{k=i+1}^{2i} \log l_k^2 \ .$$

6.10 The cepstral distance between two ARMA models

Let $H^{(1)}(z) = \frac{b^{(1)}(z)}{a^{(1)}(z)}$ and $H^{(2)}(z) = \frac{b^{(2)}(z)}{a^{(2)}(z)}$ be the transfer functions of two stable and minimum phase ARMA models of order $n^{(1)}$ and $n^{(2)}$ respectively. As we have seen in Section 6.2, the cepstral distance as defined in Definition 6.1 between $H^{(1)}(z)$ and $H^{(2)}(z)$ is equal to

$$d(\log H^{(1)}, \log H^{(2)}) = \|\log H^{(1)} - \log H^{(2)}\| \ .$$

The distance between $H^{(1)}(z)$ and $H^{(2)}(z)$ is hence equal to the cepstral norm of the ARMA model of order $n^{(1)} + n^{(2)}$ with transfer function $\frac{H^{(1)}(z)}{H^{(2)}(z)} = \frac{b^{(1)}(z)a^{(2)}(z)}{a^{(1)}(z)b^{(2)}(z)}$:

$$d(\log H^{(1)}, \log H^{(2)}) = \left\| \log \frac{b^{(1)} a^{(2)}}{a^{(1)} b^{(2)}} \right\| \ . \tag{6.25}$$

Therefore, the results of the previous sections concerning the cepstral norm, can readily be translated to the cepstral distance.
Two examples:

1. The squared distance between the two ARMA models as the mutual information of the past and the future:

$$d(\log H^{(1)}, \log H^{(2)})^2 = 2I(y_p; y_f) \ ,$$

where $\{y(k)\}$ is the output process of the ARMA model with transfer function $\frac{H^{(1)}(z)}{H^{(2)}(z)}$ of which the input is white Gaussian noise.

2. The squared distance between the ARMA models as an area in the complex plane:
$$d(\log H^{(1)}, \log H^{(2)})^2 = \frac{1}{\pi} A(\delta_{1,2}) ,$$
where $A(\delta_{1,2})$ is the area enclosed by $\delta_{1,2}(\theta) = \log H^{(1)}(e^{i\theta}) - \log H^{(2)}(e^{i\theta})$ and $\theta \in [0, 2\pi)$, running from 0 to 2π.

We pay somewhat more attention to two other descriptions, namely the distance in terms of the principal angles between the row spaces of infinite controllability matrices, and the distance in terms of the controllability and observability Gramians. Next, we discuss how the cepstral distance between two observed ARMA processes can be estimated, and we conclude this section with the computation of the distance between an observed process and a given model.

6.10.1 The cepstral distance between two ARMA models in terms of principal angles

Let $\theta_1, \ldots, \theta_{n^{(1)}+n^{(2)}}$ be the principal angles between the row spaces of the matrices $\mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{a^{(1)}(z)b^{(2)}(z)}\right)$ and $\mathcal{C}_\infty\left(\frac{z^{n^{(1)}+n^{(2)}}}{b^{(1)}(z)a^{(2)}(z)}\right)$. Then, it follows from Equation (6.25) and Theorem 6.2 that the squared distance between the models with transfer function $H^{(1)}(z) = \frac{b^{(1)}(z)}{a^{(1)}(z)}$ and $H^{(2)}(z) = \frac{b^{(2)}(z)}{a^{(2)}(z)}$ is equal to
$$d(\log H^{(1)}, \log H^{(2)})^2 = -\log \prod_{k=1}^{n^{(1)}+n^{(2)}} \cos^2 \theta_k .$$

Interestingly enough, the cosines of these angles are the smallest $n^{(1)} + n^{(2)}$ canonical correlations of the output processes of the models $H^{(1)}(z)$ and $H^{(2)}(z)$ when they have the same input, which is white noise possibly filtered by a stable, minimum phase model. The other canonical correlations of these output processes are equal to 1. This was shown in Theorem 3.11. The distance can thus be expressed as
$$d(\log H^{(1)}, \log H^{(2)})^2 = -\log \prod \mathrm{cc}^2(y^{(1)}, y^{(2)}) ,$$
where $\{y^{(1)}(k)\}$ and $\{y^{(2)}(k)\}$ are the output processes of $H^{(1)}(z)$ and $H^{(2)}(z)$, respectively.
An overview of the expressions of the norm of a model and distance between two models in terms of canonical correlations is given in Table 6.2

squared cepstral norm of ARMA model	$-\log\prod\left(1-\mathrm{cc}^2(y_p,y_f)\right)$ $=-\log\prod\sin^2\left[Y_p \sphericalangle Y_f\right]$	$i,j\to\infty$
squared cepstral distance between two AR models	$-\log\prod \mathrm{cc}^2(y^{(1)},y^{(2)})$ $=-\log\prod\cos^2\left[Y^{(1)}\sphericalangle Y^{(2)}\right]$	$i\geq n, j\to\infty$
squared cepstral distance between two ARMA models	$-\log\prod \mathrm{cc}^2(y^{(1)},y^{(2)})$ $=-\log\prod\cos^2\left[Y^{(1)}\sphericalangle Y^{(2)}\right]$	$i,j\to\infty$

Table 6.2: The cepstral norm of an ARMA model and the cepstral distance between two AR models and between two ARMA models in terms of canonical correlations of output processes. The norm is expressed as a function of the canonical correlations of the past and future output process, whereas the model distances are a function of the canonical correlations of the output processes, provided the two models are driven by the same white input process. The third column gives for each case the specifications for the number of rows (i) and the number of columns (j) in the Hankel matrices.

6.10.2 The cepstral distance between two ARMA models in terms of the Gramians

In order to give the distance between $H^{(1)}(z)$ and $H^{(2)}(z)$ in terms of the controllability Gramian of $\frac{H^{(1)}(z)}{H^{(2)}(z)}$ and the observability Gramian of $\frac{H^{(2)}(z)}{H^{(1)}(z)}$, we first derive the state space matrices of the model $\frac{H^{(1)}(z)}{H^{(2)}(z)}$ as a function of the state space matrices of the two models.
Let $A^{(1)}\in\mathbb{R}^{n^{(1)}\times n^{(1)}}$, $K^{(1)}\in\mathbb{R}^{n^{(1)}}$ and $C^{(1)}\in\mathbb{R}^{1\times n^{(1)}}$ be the state space matrices of the ARMA model with transfer function $H^{(1)}(z)$ and $A^{(2)}\in\mathbb{R}^{n^{(2)}\times n^{(2)}}$, $K^{(2)}\in\mathbb{R}^{n^{(2)}}$ and $C^{(2)}\in\mathbb{R}^{1\times n^{(2)}}$ those of $H^{(2)}(z)$. Then, the state space matrices of the ARMA model with transfer function $\frac{H^{(1)}(z)}{H^{(2)}(z)}$ can be chosen to be

$$
\begin{aligned}
A^{(12)} &= \begin{pmatrix} A^{(1)} & 0 \\ K^{(2)}C^{(1)} & A^{(2)}-K^{(2)}C^{(2)} \end{pmatrix} \in \mathbb{R}^{(n^{(1)}+n^{(2)})\times(n^{(1)}+n^{(2)})} , \\
K^{(12)} &= \begin{pmatrix} K^{(1)} \\ K^{(2)} \end{pmatrix} \in \mathbb{R}^{(n^{(1)}+n^{(2)})\times 1} , \\
C^{(12)} &= \begin{pmatrix} C^{(1)} & -C^{(2)} \end{pmatrix} \in \mathbb{R}^{1\times(n^{(1)}+n^{(2)})} .
\end{aligned}
$$

The controllability Gramian $P^{(12)}$ of this model is then the solution of the Lyapunov equation

$$
P^{(12)} = A^{(12)}P^{(12)}A^{(12)^T} + K^{(12)}K^{(12)^T} ,
$$

and the observability Gramian $Q_z^{(12)}$ of the inverse model is the solution of

$$Q_z^{(12)} = \left(A^{(12)} - K^{(12)}C^{(12)}\right)^T Q_z^{(12)} \left(A^{(12)} - K^{(12)}C^{(12)}\right) + C^{(12)^T} C^{(12)} \ .$$

The squared distance between $H^{(1)}(z)$ and $H^{(2)}$ is then equal to

$$d(\log H^{(1)}, \log H^{(2)})^2 = \log\det\left(I_{n^{(1)}+n^{(2)}} + Q_z^{(12)} P^{(12)}\right) \ . \tag{6.26}$$

6.10.3 Estimating the distance between two ARMA processes from observations

We give two methods, the first one is based on an estimation of the cepstra of the two processes and the second method goes via the identification of two models.

1. Assume that we use a non-parametric method for the estimation of the cepstrum (as e.g. in Figure 5.5) of the two processes and we obtain K estimated cepstral coefficients for both, denoted by $c^{(1)}(0),\ldots,c^{(1)}(K-1)$ and $c^{(2)}(0),\ldots,c^{(2)}(K-1)$. Then, the squared cepstral distance between the underlying ARMA models is approximately equal to

$$d(\log H^{(1)}, \log H^{(2)})^2 \approx \sum_{k=1}^{K-1} k(c^{(1)}(k) - c^{(2)}(k))^2 \ .$$

   Using a parametric estimation method for the cepstra boils down to the identification of two ARMA models, from which the distance can be computed as discussed in the next item.

2. If we identify two ARMA (or AR) models, based on the two observed sequences, we can apply e.g. (6.26) to obtain the distance between the models. Furthermore, the expressions in (6.22) and (6.23) for the cepstral norm can also be applied to the ARMA model that is obtained by dividing the transfer functions of the identified models.

6.10.4 Estimating the distance between an observed ARMA process and a given ARMA model

In a signal classification problem, one typically has a number of models and the aim is to assign an observed signal to the model that gives the best description of the signal. To this end, the distance between the observation and the different models is computed and the model that is closest to the signal is chosen. This is

one of the applications for which one needs to be able to compute the distance between a model and an observed process, which can be done in several ways, as we will show. A relevant application is monitoring and fault detection. By monitoring the distance between measurements and a nominal model, one can detect changes and possibly faults in the system. For an illustration, see Example 6.6 at the end of the chapter.
Assume we have an ARMA model with transfer function $H^{(1)}(z)$ and K data samples $y(0),\ldots,y(K-1)$ of an ARMA process $\{y(k)\}$, originating from the model with transfer function $H^{(2)}(z)$. The distance between the ARMA process $\{y(k)\}$ and the ARMA model $H^{(1)}(z)$ can be estimated as follows.

1. By estimating the cepstrum of the process.
   Let the complex cepstrum of $H^{(1)}(z)$ be denoted by $\bar{c}^{(1)}(k)$. One could estimate the power cepstrum of $\{y(k)\}$ by a non-parametric method as described in Section 5.2, which gives $\hat{c}^{(2)}(0)$, $\hat{c}^{(2)}(1)$, …, $\hat{c}^{(2)}(K-1)$. The squared distance between $\{y(k)\}$ and the given ARMA model can then be approximated as
$$\sum_{k=1}^{K-1} k(\bar{c}^{(1)}(k) - \hat{c}^{(2)}(k))^2 \ .$$

2. By identifying a model for the process.
   Let $\hat{H}^{(2)}$ be a model identified from the observations $y(k)$ $(k=0,\ldots,K-1)$, which is hence an approximation of the real underlying model $H^{(2)}(z)$. Since the distance between $H^{(1)}(z)$ and $\hat{H}^{(2)}(z)$ is equal to the norm of $\frac{H^{(1)}(z)}{\hat{H}^{(2)}(z)}$, we can use the methods described in Section 6.9.1 to determine the distance between the given model and the identified model.

3. By filtering the data.
   Filtering the given data $y(0),\ldots,y(K-1)$ by the inverse model ${H^{(1)}}^{-1}(z)$, can be seen as a series connection of the model $H^{(2)}(z)$ and ${H^{(1)}}^{-1}(z)$. Let us denote the output of this filtering process, by $z(0),\ldots,z(K-1)$. These samples thus can be viewed to originate from the ARMA model with transfer function $\frac{H^{(2)}(z)}{H^{(1)}(z)}$. Since the distance between $H^{(1)}$ and $H^{(2)}$ is equal to the norm of $\frac{H^{(2)}(z)}{H^{(1)}(z)}$, we can estimate the distance from the samples $z(0),\ldots,z(K-1)$ with the techniques in Section 6.9.2.
   For example, construct the $2i \times j$ Hankel matrix Z with the samples $z(0),\ldots,z(K-1)$. Let $l_1, l_2, \ldots, l_{2i}$ be the diagonal elements of the triangular factor L in the LQ factorization of Z. The squared distance between $H^{(1)}(z)$ and $H^{(2)}(z)$ is then approximately equal to
$$d(\log H^{(1)}, \log H^{(2)})^2 \approx \sum_{k=1}^{i} \log l_k^2 - \sum_{k=i+1}^{2i} \log l_k^2 \ . \qquad (6.27)$$

Or, compute the principal angles between the row spaces of Z_p and Z_f and apply (6.24).
Note that filtering the data with the inverse of the given model, can be seen as applying the Kalman filter of the model to the measurements by which the residuals of the data are determined.

The third method can now be interpreted as follows. The norm of the output process $\{z(k)\}$ indicates how colored $\{z(k)\}$ is. If the norm is very small, then $\{z(k)\}$ is almost white, which suggests that the model by which $\{z(k)\}$ is generated, has a transfer function that is close to 1, in particular, $\frac{H^{(2)}(z)}{H^{(1)}(z)} \approx 1$ and hence $H^{(2)}(z) \approx H^{(1)}(z)$. Filtering an ARMA process by the inverse of a model that resembles the original model, gives a process that is close to white noise.

6.11 Some illustrative examples

In this section we describe three examples of possible applications of the concepts in this chapter: the stabilization diagram, fault detection and the clustering of models. They mainly serve as an illustration of the theoretical results in this chapter.

6.11.1 Tools for the stabilization diagram

The stabilization diagram is a graphical tool in modal analysis for determining the important physical modes from measurements of a mechanical or structural system. For increasing model order, the frequency of the poles of the identified models are shown. Poles and corresponding modal forms that remain around the same frequency when the order increases, are chosen as physical modes. These poles are said to 'stabilize' around the frequency, hence the name 'stabilization diagram'. Poles whose frequency varies considerably for the different orders are due to overmodeling. They do not have a physical meaning.
The analysis of the stabilization diagram is not completely automated yet. The determination of the model order and distinction between physical and non-physical poles is still being done manually. We think that the notions introduced in this chapter might help to perform these tasks. This is indicated in the following example.

Example 6.5. In this example we return to the data of the transmitter mast, denoted by $y(k)$, which were introduced in Example 3.1 and also used in Example 5.2. We illustrate the concepts of this chapter by subsequently identifying ARMA models of order $1, 2, \ldots, 40$ for this data sequence. We therefore use the `armax` function of the Matlab Identification Toolbox [100]. The transfer functions of these models are denoted by $H^{(n)}(z)$, where n is the model order.

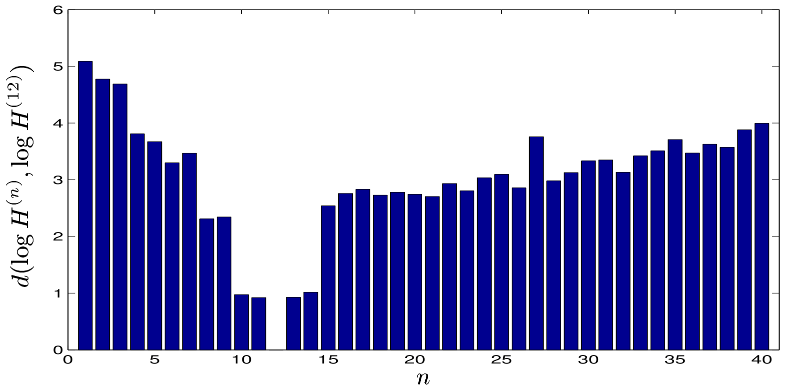


Figure 6.3: The distance between the models of order $n = 1, 2, \ldots, 40$ and the 12th order model that was identified in Example 3.1.

To illustrate the cepstral distance between two ARMA models, we show in Figure 6.3 the cepstral distance of each model to the 12th order model that

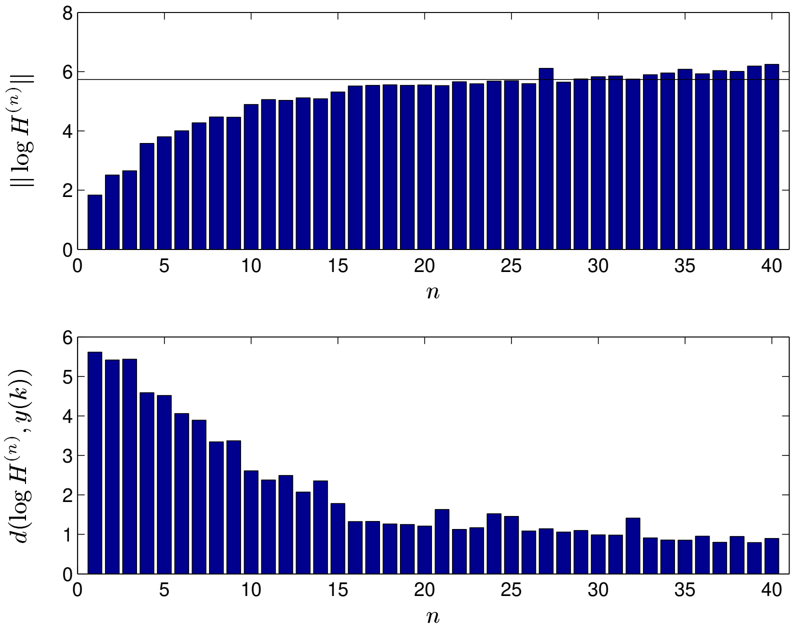


Figure 6.4: In the top picture the cepstral norm of the different models is shown, together with the norm of the measurements (the horizontal line). The bottom picture gives the distance of the measurement signal to the different models.

was identified in Example 3.1. The distance between the model of order 1 and the model of order 12 is quite large. For increasing model order, the distance becomes smaller until it is zero for $n = 12$. Not surprisingly, the models of order $10, 11, 13, 14$ lie closest to the model of order 12.
Whereas Figure 6.3 was only an illustration, the pictures in Figure 6.4 might be useful for the analysis of the stabilization diagram. The first picture shows the cepstral norm of each model, computed from the model parameters. The norm of the model increases for growing order, indicating that more dynamics are incorporated, since the norm is a measure for the distance to whiteness. However, the rate by which the norm increases, diminishes with increasing order. This is caused by the fact that mathematical (not physical) poles and zeros are added, which tend to cancel each other's influence. The norm of the imaginary model that generated the data, is the horizontal line at 5.7369 in the first picture of Figure 6.4. It is estimated from the data by

$$\| \log H \|^2 \approx -\log \prod_{k=1}^{i} \sin^2 \phi_k \ ,$$

where $\phi_1, \ldots, \phi_i$ are the principal angles between the row spaces of the past and future Hankel matrices Y_p and Y_f and the parameter i (the number of rows in Y_p) is equal to 50. The principal angles $\phi_1, \ldots, \phi_i$ are shown in Figure 6.5. As we know from the theoretical derivations in Chapter 3, only n principal angles between row(Y_p) and row(Y_f) of an nth order model are smaller than $\frac{\pi}{2}$ (for the number of columns $j \to \infty$), see (3.87) and (3.91). In Figure 6.5 we have more than 12 acute angles. This is partly due to the finite number of columns in the data Hankel matrices (also see Example 3.8), but also because the data are not coming from a 12th order linear model. The principal angles could however be used to estimate a good model order for a low order approximation. Zooming in to Figure 6.5, $n = 14$ seems to be a better choice than 12,

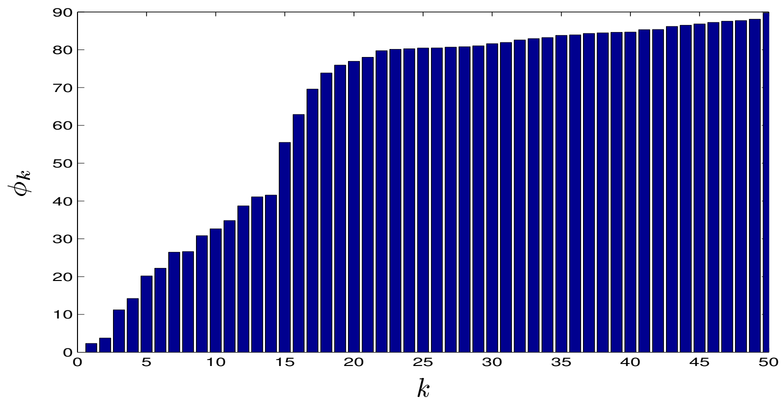


Figure 6.5: The principal angles between the row spaces of the 50×3740 past and future Hankel matrices of the given mast measurements (in degrees).

which was estimated from the singular values in Figure 3.3.
The second picture in Figure 6.4 shows the distance of the signal $y(k)$ to the respective models. This is computed by filtering the measurements with the inverse models, as explained in Section 6.10.4. The distance of the measurements to the linear system behind it, would of course be equal to 0.
Both representations can be used to analyze the influence of taking larger model orders. They could also be used to estimate the order of the model. In the top picture the model norms can be compared to the measurements norm, while the distances in the bottom picture should be compared to 0.
Finally, we give in Figure 6.6 the graphical interpretation of the norm of the models of order $1, 2, 3$ and 4. The squared norm is equal to the area enclosed by the curve, divided by π. Taking into account that area that is encircled twice should be counted twice, it is clear that the norm increases for increasing order, as we already have seen in the second plot of Figure 6.4. ◊

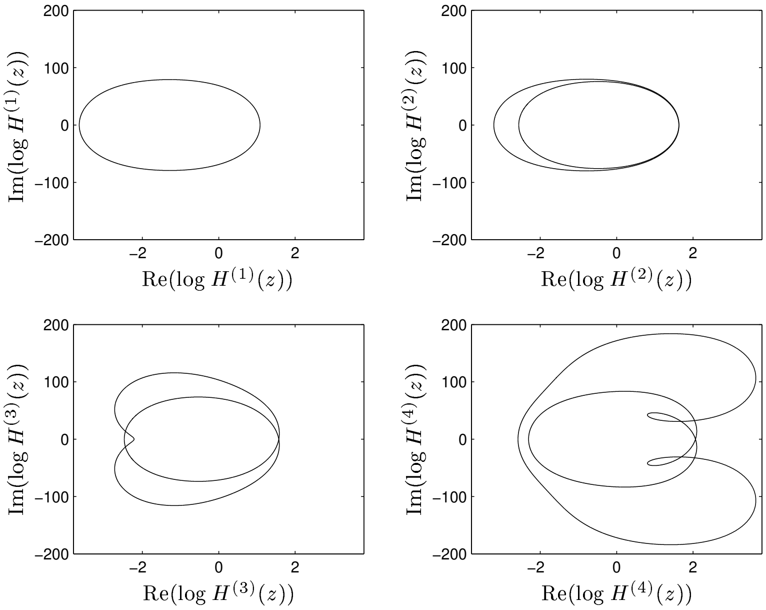


Figure 6.6: The polar plots of $\log H^{(n)}$ for $n = 1, 2, 3, 4$ (left to right and top to bottom). The area that is enclosed by the curves is proportional to the squared cepstral norm of the models.

6.11.2 Monitoring and fault detection

If at different moments in time a model is identified for the same system, there will be small or larger differences between the models. It is important to understand the causes of the differences. They can be uncertainties innate to

the estimation method, a dysfunction of a sensor, or real structural changes of the system. Little cracks in a bridge, e.g., will change the modal structure, but also the environmental temperature influences the modes [3, 125, 116, 117]. A good fault detection method should be robust against such disturbances, but sensitive enough for real defects.
One possible approach to detect changes in the dynamics of a system is to identify models at different time instants and compare the models, e.g. by means of the cepstral distance or related principal angles. In the following example, we show that we do not have to identify complete models all the time. As explained in Section 6.10.4, we can compute the distance between a given model and measurements. Hence, if we have a model of the system without damage, then we can continuously compare new measurements with this model using the techniques of Section 6.10.4. As soon as the difference between the models is significantly too large, one should intervene.

Example 6.6. Detection of the change of dynamics in measurements
We consider two fourth order ARMA models of which the poles and zeros are depicted in Figure 6.7. The poles and zeros of $M^{(2)}$ (stars ($*$) and squares ($\Box$)) are shifted versions of the poles and zeros of $M^{(1)}$ (crosses ($\times$) and circles ($\circ$)). They are shifted towards the unit circle. The first model is a model for the undamaged system and is assumed to be known, e.g. by identification, while the second model represents a damaged version of the original system.
We simulate the output of these models with a white noise input sequence. The first 5 000 samples originate from the first model and then we jump to the output samples of $M^{(2)}$. The goal is to detect this sudden change. Three methods of Section 6.10.4 are applied.

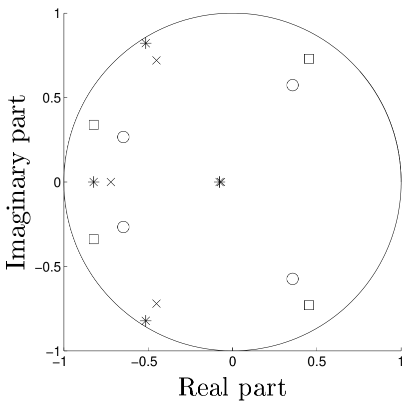


Figure 6.7: The poles and zeros of the models used for the simulation of the data. The poles of $M^{(1)}$ are the crosses ($\times$), those of $M^{(2)}$ are indicated by stars ($*$). The zeros of $M^{(1)}$ are the circles ($\circ$) and the zeros of $M^{(2)}$ are the squares ($\Box$).

1. Estimating the cepstrum from the measurements.
   We use the first 1039 data points to get an estimate of the cepstrum. When a new measurement point is made, namely $y(1040)$, the cepstrum is estimated from the data sequence $y(2), y(3), \ldots, y(1040)$. The estimated cepstrum is denoted by $\hat{c}(j)_k$, where k is the time index of the measurements, and j denotes the different cepstral coefficients. For example, $\hat{c}(5)_{1042}$ is the estimate of the fifth cepstral coefficient, based on the data $y(4), y(5), \ldots, y(1042)$. We only use the first 20 coefficients in the computation of the distance between the model and the measurements to avoid problems described in Remark 5.6. The cepstrum estimate based on the measurements $y(1), y(2), \ldots, y(1039)$ is consequently $\hat{c}_{1039}(0)$, $\hat{c}_{1039}(1)$, ...,$\hat{c}_{1039}(19)$. The cepstrum of nominal model $M^{(1)}$ is indexed as $c^{(1)}(j)$ and is computed by applying (5.8).
   The distance at each time instant k is then computed as

$$d(c^{(1)}, \hat{c}_k)^2 = \sum_{j=1}^{19} j(c^{(1)}(j) - \hat{c}_k(j))^2 \ .$$

   The result of this monitoring method is shown in picture A of Figure 6.8. The dashed line gives the true distance between the models $M^{(1)}$ and $M^{(2)}$. We can recognize the change in dynamics, although the distance estimate is rather noisy.

2. By filtering the data with the inverse model and updating determinants.
   We filter the measurements $y(k)$ by the inverse model ${M^{(1)}}^{-1}$. The output of this operation is denoted by $z(k)$. We again use 1039 data points at a time to form the $40 \times 1\,000$ Hankel matrix Z_k, where k denotes the time index of the last data point. The distance between the measurements $y(k)$ and the model $M^{(1)}$ is then estimated as

$$d_Z^2(k) \approx \sum_{j=1}^{20} \log l_{j_k}^2 - \sum_{j=21}^{40} \log l_{j_k}^2 \ .$$

   $l_{1_k}, l_{2_k}, \ldots, l_{40_k}$ are the diagonal elements of L_k, the L factor in the LQ factorization of Z_k. If a new measurement is made, the first column of Z_k is deleted and a new column is appended in order to get a new estimate for the distance. Updating the diagonal elements of L_k to those of L_{k+1} when a column in Z_k is deleted and another one is appended, can easily be done by Givens rotations, see e.g. [57, pp. 610–611].
   The result of this method is shown in picture B of Figure 6.8. Again, the dashed line gives the true distance between the models $M^{(1)}$ and $M^{(2)}$.

3. By filtering the data and computing the principal angles between past and future.
   We start from the same filtered data $z(k)$ as in the previous method, but

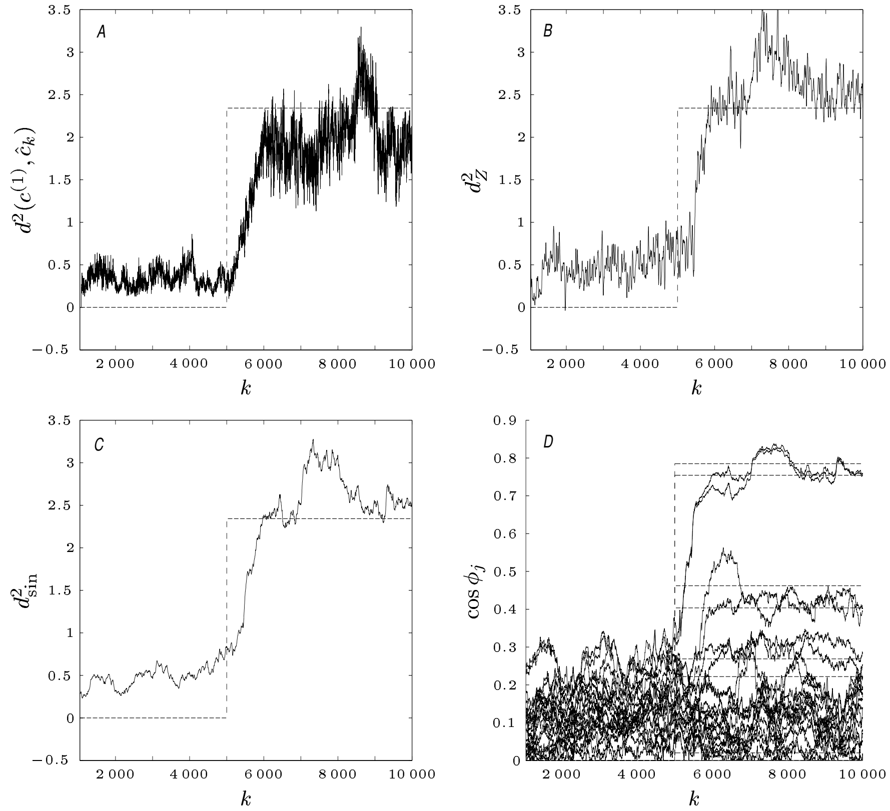


Figure 6.8: The results of three monitoring methods for the fault detection problem in Example 6.6. At time instant $k = 5\,000$, the dynamics of the measured data change from model $M^{(1)}$ to model $M^{(2)}$. Picture A shows the squared distance between the cepstrum of model $M^{(1)}$ and the cepstrum estimated from the measurements. Picture B gives the distance estimated by filtering the data by the inverse model and applying (6.27). In picture C the distance is based on the principal angles between the past and the future of the filtered data. The cosines of these angles are given in picture D. The dashed lines always represent the true values, based on the parameters of $M^{(1)}$ and $M^{(2)}$.

now we estimate the distance between model $M^{(1)}$ and the data as

$$d_{\sin}^2(k) = -\log \prod_{j=1}^{20} \sin^2 \phi_{j_k} \ ,$$

where $\phi_{1_k}, \phi_{2_k}, \ldots, \phi_{20_k}$ are the principal angles between the row spaces of Z_{p_k} and Z_{f_k}.

The distance based on this estimate, as a function of the time index k

is shown in picture C of Figure 6.8. In the first part ($k < 5\,000$), the theoretic distance is equal to 0 (see dashed line). However, due to the finite number of data, we get estimation errors. These are always positive since we are estimating a distance, which explains the bias in picture C of Figure 6.8.

The cosines of the past and the future are depicted in picture D. The true values of the cosines, based on the parameters of $M^{(1)}$ and $M^{(2)}$ are given in dashed lines. Although for $k < 5\,000$, all cosines should theoretically be equal to 0, it is not the case for our computations. This is due to the small amount of data that we use to estimate the angles. The bias is again due to the fact that the estimated cosines are always positive. The distance estimate based on the principal angles seems to average better than the other ones. It could of course be improved by taking more data samples for each estimate of the principal angles, but then the transition from the no-error zone to the zone where a fault is detected, would become larger.

Note that in order to decide when a distance between a model and measurements is too large, we need the statistical properties of the principal angles and the derived distance. Knowing the uncertainty can then be applied in hypothesis tests. $\Diamond$

6.11.3 Clustering and classification of models and signals

In the introduction to the previous example is mentioned that temperature differences can also cause changes in the modal parameters of a construction, such as a bridge. This is corroborated by experimental evidence [3, 125, 116, 117] as well as theoretical arguments. In order to determine whether changes that have been measured, originate from environmental changes or from damage in the structure, it might be useful to determine clusters of models or data sets. Each cluster corresponds to e.g. a certain temperature range. When new measurements are made, the distance between the measurements and the different clusters can then indicate to which cluster the measurements belong, or it could show that the measurements cannot be explained by one of the clusters and hence, the system is damaged. Clustering and classification is always based on a certain distance measure. In the following example, we use the cepstral distance that was described in this chapter.

Remark 6.7. Applications in computer vision
Inspired by our results in [37, 36], Saisan et al. [127] and Soatto et al. [133] have applied the ideas described above to the clustering and classification of dynamic textures, which are sequences of images that exhibit some form of temporal stationarity. Another problem that has been tackled in computer vision, based on our results, is described in [15]. $\bigcirc$

Example 6.7. The clustering of mast models
We use 9 different measurements sets of the transmitter mast that was discussed in Example 3.1, Example 5.2 and Example 6.5. They come from 9 different sensors on the mast. To reduce the amount of data and to zoom in on the frequency part of interest, the data are decimated. In order to illustrate the clustering possibilities, we use two decimation factors. First, we decimate with a factor 8 and the second time with a factor 6. The non-parametric spectra of the data sets are shown in Figure 6.9 on a logarithmic scale. In the top picture we give the spectra of the first 9 data sets, those of the other 9 data sets are depicted in the bottom picture. Note that due to the different filter, there are clear differences between the two groups of spectra, but one can also distinguish groups of similar spectra within each picture.
For each data sequence, an AR model is identified with the `ar` function of the Matlab Identification Toolbox [100]. The models are denoted by $M^{(1)}$,

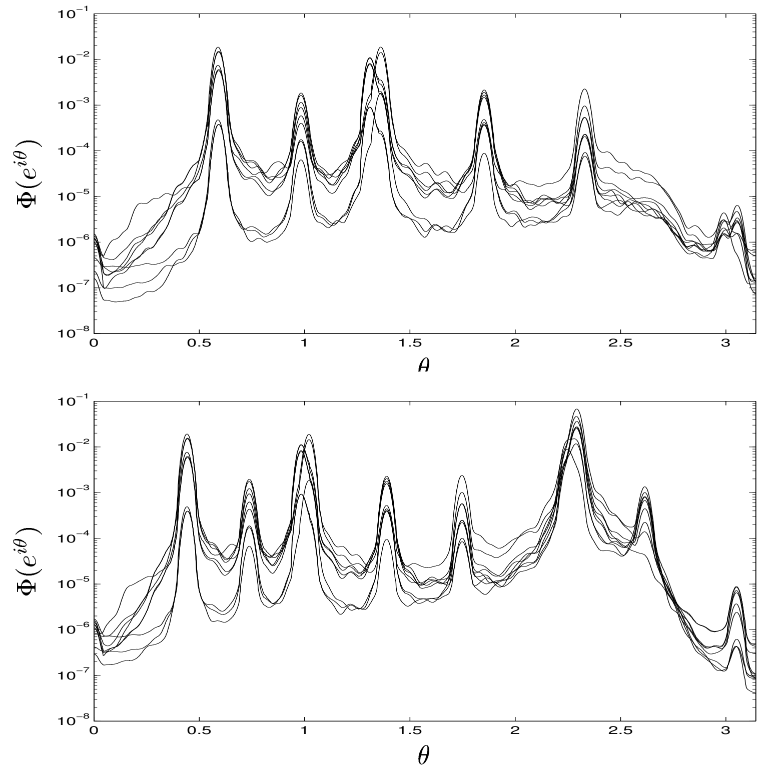


Figure 6.9: The spectra of the 18 data sets on a logarithmic scale. In the top picture we give the spectra of the first data sets, obtained by decimating the original data by a factor 8, those of the other 9 data sets, decimated by a factor 6, are depicted in the bottom picture.

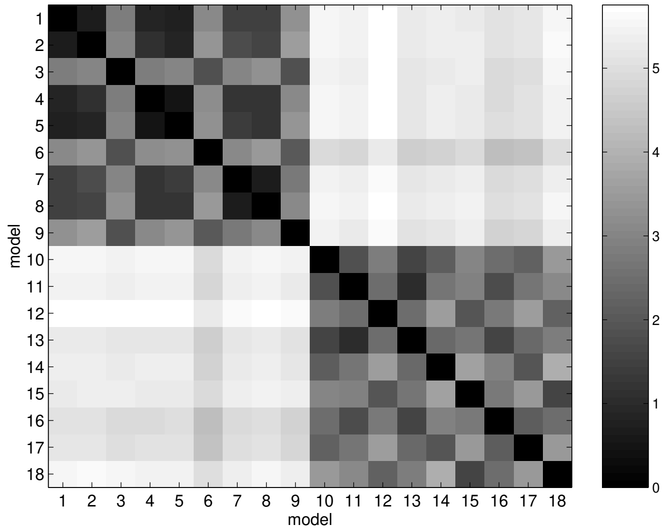


Figure 6.10: Graphical representation of the distance matrix. The element (i, j) in the matrix contains the distance between the ith and jth AR model. A black square indicates that the distance between the corresponding models is equal to 0. All diagonal squares are consequently black. A white square, on the other hand represents a large cepstral distance between the models. This (and the distance corresponding to all intermediate hues) can also be read from the colorbar at the right.

$M^{(2)}, \ldots, M^{(18)}$. Next, we compute the cepstral distances between all pairs of models via the formula based on the poles of the models in (6.7). The resulting distance matrix is depicted in Figure 6.10. The values corresponding to the different grey values are explained in the colorbar. The darker a square (i, j), the smaller the distance between the models $M^{(i)}$ and $M^{(j)}$. The diagonal elements are of course black and it is obvious that the matrix is symmetric. But it is also immediately clear that two groups of models can be distinguished: the models $M^{(1)},\ldots,M^{(9)}$ form one group and the models $M^{(10)},\ldots,M^{(18)}$ the other group. Applying a nearest neighbor algorithm for the clustering of the models (see e.g. [71]), where two clusters are to be found, gives indeed the following result:

	models								
cluster 1	1	2	3	4	5	6	7	8	9
cluster 2	10	11	12	13	14	15	16	17	18

If we take a closer look at the left upper 9×9 square, we can obtain more detailed information on the first cluster. Model $M^{(1)}$ lies close to model $M^{(2)}$, $M^{(4)}$ and

$M^{(5)}$, and to a smaller extent close to $M^{(7)}$ and $M^{(8)}$. Similar observations can be made for the other models. Clustering with 3 clusters indeed puts these models together:

	models								
cluster 1	1	2	4	5	7	8			
cluster 2	3	6	9						
cluster 3	10	11	12	13	14	15	16	17	18

With four clusters, we also obtain subgroups in the second group of models:

	models					
cluster 1	1	2	4	5	7	8
cluster 2	3	6	9			
cluster 3	10	11	13	14	16	17
cluster 4	12	15	18			

$\Diamond$

6.12 Conclusions

In this chapter, we have studied a cepstral norm and the related cepstral distance for stable and minimum phase SISO ARMA models. By means of the concepts, insights and results of the preceding chapters, we have shown that this norm and distance can be interpreted in different ways. We have derived expressions for the norm in terms of principal angles between subspaces derived from the model parameters, in terms of canonical correlations, Gramians. Furthermore, we have shown that the cepstral norm of a model is proportional to the mutual information of the past and the future of its output process, provided it is Gaussian. By recognizing in the cepstral norm the Hilbert–Schmidt norm of the Hankel matrix with the cepstral coefficients, we have given a graphical interpretation for the cepstral norm. We have proven that the area enclosed by the polar plot of the logarithm of the transfer function, is equal to π times the cepstral norm squared.
Via the expressions for the cepstral norm, we have derived system theoretic and information theoretic interpretations for the cepstral distance between ARMA-models. Furthermore, we have given an overview of how the norm and the distance can be computed in both parametric (via the model parameters) and non-parametric (directly from given measurements) ways. Also computing the distance between a model and a signal has been treated. In the last section, finally, we have given three examples of how these concepts can be applied.

Chapter 7

Conclusions and further research

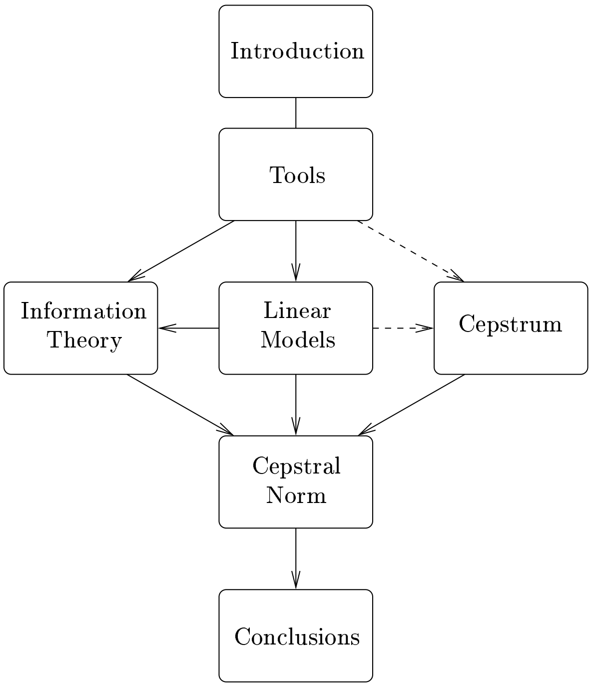


Figure 7.1: Overview of the thesis with the keywords that are used in this conclusive chapter to refer to the different chapters.

7.1 Conclusions

In this thesis we have established connections between system theory, information theory and signal processing for which we rely on geometrical insights obtained by computing the principal angles between pairs of linear subspaces. We give an overview of the different chapters, referring to the key words in Figure 7.1.

Tools In Chapter 2 we have given the definition of the principal angles between and corresponding principal vectors in two subspaces, we have shown how they can be computed (via the singular value decomposition, the eigenvalue decomposition and the LQ-factorization) and how they are related to orthogonal projections.
Furthermore, we have described how the canonical correlations of two random variables and of two stochastic processes can be seen as the cosines of the principal angles between two subspaces.

Linear Models Most attention has been paid to stationary, ergodic processes that can be modeled as the output of a linear time-invariant stochastic model. That is why we have described these models and their different representations in Chapter 3. For the single-input single-output (SISO) models we have deduced some useful properties.
Next, we have derived expressions for the principal angles and canonical correlations between the different past and future input and output pairs. As we have shown in the examples of Chapter 3, the angles can be estimated from observations of the stochastic processes, without knowing a model for the processes. These angles are the principal angles between the row spaces of the $mi \times j$ block Hankel matrices with the observations. That is the non-parametric approach. However, by expressing the principal angles in terms of the model parameters, one obtains the exact values for finite i and $j \to \infty$, from which the formulas for the canonical correlations of the different processes have been deduced, by letting i go to ∞. We have proven the relations of the different sets of angles and canonical correlations via their parametric expressions.
For SISO-models the parametric expressions have been simplified.

- We have shown that the principal angles and canonical correlations of the different combinations of past and future input and output processes are independent of the input variance.
- We have proven that the smallest n canonical correlations of the input and the output of an nth order linear time-invariant stochastic SISO-model (which can be estimated in a non-parametric way) are equal to the cosines of the principal angles between the row space of the infinite controllability

matrix of the model and the row space of the infinite controllability matrix of the inverse model. The other canonical correlations are equal to 1.

- Based on the last property, we have formulated a conjecture concerning the eigenvalues of matrices obtained by solving a Lyapunov equation.
- We have shown that the principal angles between the row spaces of the past and future $i \times j$ Hankel matrices of an autoregressive process are independent of i if $j \to \infty$ and $i \geq n$.

Information Theory In Chapter 4 we have first given the necessary background information on the information theoretic notions of entropy and mutual information. The mutual information of two Gaussian random variables can be expressed as a function of their canonical correlations. The same can be done for the mutual information of two jointly Gaussian stochastic processes.
We have mainly dealt with stationary stochastic processes that can be modeled as the output of a linear, time-invariant stochastic model. The results of Chapter 3, i.e. the expressions for the canonical correlations of the different pairs of input and output processes, are then used to write the mutual information of the processes as a function of the model parameters.

Cepstrum Via the motivating example of echo detection and echo removal, we have introduced the cepstrum in Chapter 5. The power cepstrum as well as the complex cepstrum were defined, together with their properties that are useful for the derivations in Chapter 6. Next, we have briefly indicated how the cepstrum can be used to identify a linear model. Especially the fact that the zeros of the model can be determined more easily than from the autocovariances, is promising.

Cepstral Norm The cepstrum, principal angles and canonical correlations, the mutual information of two stochastic processes, controllability and observability matrices, Gramians, Lyapunov equations and the Hilbert–Schmidt–Hankel norm, we have applied all these concepts to derive different interpretations of a cepstral norm for stable and minimum-phase ARMA models.

- We have shown that the cepstral norm can be expressed as a function of the principal angles between the row space of the infinite controllability matrix of the model and the row space of the infinite controllability matrix of the inverse model.
- In Chapter 3 we have proven that the cosines of these angles are equal to the smallest n canonical correlations of the input and output process of the model. The cepstral norm can consequently also be seen as a function of these canonical correlations. The expressions for the canonical

correlations as a function of the model parameters in Chapter 3 have been applied to the cepstral norm.

- Via the geometric relations deduced in Chapter 3 (the principal angles between the input and output are the complements of the principal angles between the past and future output), we have shown that the cepstral norm squared is equal to twice the mutual information of the past and the future output process.
- By recognizing in the cepstral norm the Hilbert–Schmidt norm of the Hankel matrix with the cepstral coefficients, we have given a graphical interpretation for the cepstral norm. We have proven that the area enclosed by the polar plot of the logarithm of the transfer function, is equal to π times the cepstral norm squared.

Via the expressions for the cepstral norm, we have derived a system theoretic and information theoretic interpretation for the cepstral distance between ARMA-models. Furthermore, we have given an overview of how the norm and the distance can be computed in both parametric (via the model parameters) and non-parametric (directly from given measurements) ways. Also computing the distance between a model and a signal has been treated.

7.2 Further research

We suggest possible directions for further research and mention some open questions.

The integration of information theory in system theory, in particular system identification and control theory, is far from complete. Stoorvogel and van Schuppen have started this new framework in [142], but there are still a lot of research opportunities. For example, because of the connection of subspace identification methods and canonical correlation analysis, we believe that it might be possible to determine the cost function that is minimized by subspace identification algorithms. Information theory could be helpful in this pursuit, as shown by the attempts in [143, 144].

In this thesis we have mainly worked with Gaussian stochastic processes, for which dependence and correlation are equivalent. However, for non-Gaussian processes, the mutual information is a much more general measure for dependence than canonical correlations. Of course, higher order statistics will then be taken into account. Maybe for these cases too, geometric interpretations could be made.

Subspace techniques have very important advantages in vibration analysis and

fault detection of mechanical structures. In the future, we want to study the application possibilities of subspace identification and subspace angles and related distance measures in these areas. In the illustrative examples of Chapter 6, we have indicated how this could be done. However, a thorough statistical analysis of the principal angles that we use and the related cepstral distance measure is necessary for the application of these concepts in e.g. fault detection and model order determination.
For fault detection applications it is also important to develop algorithms for on-line monitoring. The updating techniques based on the LQ decomposition that were mentioned in Section 6.11.2, will be useful. However, one can come up with smarter updating methods than the ones suggested, e.g. by exploiting the Hankel structure of the data matrices.
Other applications that we want to focus on are situated in bio-informatics (clustering of genes) and biomedical signal processing (detection of dynamic similarities in two signals), which we briefly discussed in Chapter 1.

In Chapter 6, we have given different representations and computation methods for the cepstral norm and distance. A comparative study could point out which method is the most reliable and which method is the fastest.

The cepstral distance that is studied in Chapter 6, is only defined for SISO stochastic models. Is it possible to define a similar distance measure, with the same interesting interpretations, for deterministic models and MIMO (multiple input multiple output) models?

As we have briefly shown in Chapter 5, the cepstrum can be used to identify the poles and zeros of a fractional model. This idea should be worked out to obtain a good working algorithm that can be applied to realistic problems. Furthermore, the use of the cepstrum for identification problems in general is an interesting research topic.

In the thesis we have mainly dealt with the principal angles between two subspaces, while the corresponding principal directions are hardly treated. Analogously to the angles, which can be computed between subspaces of data, but for which we also have analytical expressions, it should be possible to derive parametric expressions for the principal directions.

Of course we are still looking for a proof of Conjecture 3.13. An interpretation of this result and possible consequences should also be studied.

We have not spent any attention to the concepts in Appendix B: compound matrices, Grassmann vectors, volume of a matrix. The relations with the topics in the thesis are however apparent. We would like to understand the deeper connections.

Appendix A

Generalization of geometric properties in the plane

In this appendix we give the generalization of the observations made in Section 2.5.4 concerning the relations between the principal angles and principal directions in the plane spanned by u_i and v_i, between which the principal angle θ_i is acute. We generalize the properties in the plane to all the principal angles between and principal directions in the row spaces of the matrices A and B. Let $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{q\times n}$ be of rank r_a and r_b respectively. Without loss of generality we assume that

$$r_a \leq r_b \leq n - r_b \leq n - r_a \ .$$

Then, there are r_a principal angles between row(A) and row(B), denoted by

$$[A \sphericalangle B] = (\theta_1, \theta_2, \ldots, \theta_{r_a}) \ ,$$

r_a principal angles between row(A) and row$(B)^\perp$:

$$\left[A \sphericalangle B^\perp\right] = (\eta_1, \eta_2, \ldots, \eta_{r_a}) \ ,$$

r_b principal angles between row$(A)^\perp$ and row(B):

$$\left[A^\perp \sphericalangle B\right] = (\zeta_1, \zeta_2, \ldots, \zeta_{r_b}) \ ,$$

and $n - r_b$ principal angles between row$(A)^\perp$ and row$(B)^\perp$:

$$\left[A^\perp \sphericalangle B^\perp\right] = (\mu_1, \mu_2, \ldots, \mu_{n-r_b}) \ .$$

We can now prove the following theorem.

Theorem A.1.
For $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{q\times n}$ of rank r_a and r_b respectively and $r_a \leq r_b \leq n-r_b \leq n-r_a$, the following relations between the principal angles and principal directions in the different spaces hold:

$$\eta_i = \frac{\pi}{2} - \theta_{r_a-i+1} \qquad \textit{for } i = 1,\ldots,r_a, \tag{A.1}$$
$$\zeta_i = 0 \qquad \textit{for } i = 1,\ldots,r_b - r_a, \tag{A.2}$$
$$\zeta_{r_b-r_a+i} = \frac{\pi}{2} - \theta_{r_a-i+1} \qquad \textit{for } i = 1,\ldots,r_a, \tag{A.3}$$
$$\mu_i = 0 \qquad \textit{for } i = 1,\ldots,n-(r_a+r_b), \tag{A.4}$$
$$\mu_{n-(r_a+r_b)+i} = \theta_{r_a-i+1} \qquad \textit{for } i = 1,\ldots,r_a, \tag{A.5}$$
$$u_i(\left[A \sphericalangle B^{\perp}\right]) = u_{r_a-i+1}([A \sphericalangle B]) \qquad \textit{for } i = 1,\ldots,r_a, \tag{A.6}$$
$$v_{r_b-r_a+i}(\left[A^{\perp} \sphericalangle B\right]) = v_{r_a-i+1}([A \sphericalangle B]) \qquad \textit{for } i = 1,\ldots,r_a, \tag{A.7}$$
$$x_{r_b-r_a+i}(\left[A^{\perp} \sphericalangle B\right]) = x_{m-(r_a+r_b)+i}(\left[A^{\perp} \sphericalangle B^{\perp}\right]) \qquad \textit{for } i = 1,\ldots,r_a, \tag{A.8}$$
$$y_i(\left[A \sphericalangle B^{\perp}\right]) = y_{m-(r_a+r_b)+i}(\left[A^{\perp} \sphericalangle B^{\perp}\right]) \qquad \textit{for } i = 1,\ldots,r_a, \tag{A.9}$$

In words this is,

- The principal angles between row(A) and row$(B)^{\perp}$ are complementary to the principal angles between row(A) and row(B).
- The largest r_a principal angles between row$(A)^{\perp}$ and row(B) are complementary to the principal angles between row(A) and row(B) and the other $r_b - r_a$ principal angles are zero.
- The largest r_a principal angles between row$(A)^{\perp}$ and row$(B)^{\perp}$ are equal to the principal angles between row(A) and row(B) and the other $n - (r_a + r_b)$ principal angles are zero.

For the principal directions holds:

- The r_a principal directions in row(A) corresponding to $[A \sphericalangle B]$ are also the principal directions in row(A) corresponding to $\left[A \sphericalangle B^{\perp}\right]$.
- The r_a principal directions in row(B) corresponding to $[A \sphericalangle B]$ are also the principal directions in row(B) corresponding to the largest r_a principal angles $\left[A^{\perp} \sphericalangle B\right]$.
- The r_a principal directions in row$(A)^{\perp}$ that correspond to the largest r_a principal angles $\left[A^{\perp} \sphericalangle B^{\perp}\right]$ are also the r_a principal directions in row$(A)^{\perp}$ corresponding to the largest r_a $\left[A^{\perp} \sphericalangle B\right]$.

- The r_a principal directions in $\text{row}(B)^\perp$ that correspond to the largest r_a principal angles $[A^\perp \sphericalangle B^\perp]$ are also the r_a principal directions in $\text{row}(B)^\perp$ corresponding to the $[A \sphericalangle B^\perp]$.

Proof.
Let the right singular matrices of A and B be partitioned as

$$\begin{aligned} V_a &= \begin{pmatrix} V_{a_1} & V_{a_2} \end{pmatrix} \quad \text{where } V_{a_1} \in \mathbb{R}^{n \times r_a} \ , \\ V_b &= \begin{pmatrix} V_{b_1} & V_{b_2} \end{pmatrix} \quad \text{where } V_{b_1} \in \mathbb{R}^{n \times r_b} \ . \end{aligned}$$

Consider the four full SVDs that will be used to compute the principal angles and principal directions (see Property 2.8):

$$V_{a_i}^T V_{b_j} = P_{ij} S_{ij} Q_{ij}^T \qquad (i, j = 1, 2) \ . \tag{A.10}$$

First, we investigate the relations between $[A \sphericalangle B]$ and $[A \sphericalangle B^\perp]$. It follows from the orthogonality of V_b and V_a that

$$V_{a_1}^T V_{b_2} V_{b_2}^T V_{a_1} = I_{r_a} - V_{a_1}^T V_{b_1} V_{b_1}^T V_{a_1} \ .$$

Using the SVDs in (A.10) gives

$$P_{12} S_{12} S_{12}^T P_{12}^T = P_{11} (I_{r_a} - S_{11} S_{11}^T) P_{11}^T \ .$$

From the uniqueness properties of the SVD (see Remark 2.1) it follows that

$$I_{r_a} - S_{11} S_{11}^T = S_{12} S_{12}^T \ , \tag{A.11}$$

and we can take

$$P_{12} = P_{11} \ . \tag{A.12}$$

In the case of non-distinct singular values, the equality in Equation (A.12) should be interpreted in the sense that it is always possible to choose the vectors in the associated singular subspaces in such a way, that it is satisfied.
After ordering the singular values in non-ascending order and the singular vectors in the same way, (A.11) and (A.12) lead to (A.1) and (A.6), respectively.
In a similar way we obtain

- for $[B \sphericalangle A^\perp]$

$$\begin{aligned} S_{21}^T S_{21} &= I_{r_b} - S_{11}^T S_{11} \ , && \text{(A.13)} \\ Q_{21} &= Q_{11} \ , && \text{(A.14)} \end{aligned}$$

  which gives (A.2), (A.3) and (A.7),

- for $[A^\perp \sphericalangle B^\perp]$

$$\begin{aligned} S_{22} S_{22}^T &= I_{m-r_a} - S_{21} S_{21}^T \ , \\ P_{22} &= P_{21} \ , \end{aligned}$$

  which, together with (A.13) and (A.14) leads to (A.4), (A.5) and (A.8),

- for $\left[B^{\perp} \lessdot A^{\perp}\right]$

$$\begin{aligned} S_{12}^T S_{12} &= I_{m-r_b} - S_{22}^T S_{22} \ , \\ Q_{12} &= Q_{22} \ , \end{aligned}$$

which again leads to (A.4), (A.5) and also gives (A.9).

$\square$

Appendix B

Compound matrices, Grassmann vectors and the volume of a matrix

In this appendix we introduce a number of interrelated concepts: compound matrices (Section B.1), Grassmann coordinates (Section B.2) and the volume of a matrix (Section B.3). We also show how they are related to the SVD, orthogonal projectors and the principal angles between two subspaces. Throughout the appendix, we will derive a few more expressions for the cepstral norm of Chapter 6, which are briefly formulated in the different remarks.

B.1 Compound matrices

We start in Section B.1.1 with the definition of the kth compound matrix and an example. Then we proceed by giving some properties of compound matrices in Section B.1.2.

B.1.1 Definition and example

Definition B.1. The kth compound matrix
The kth compound matrix of a matrix $A \in \mathbb{R}^{m\times n}$, denoted by $C_k(A)$, is the $\binom{m}{k} \times \binom{n}{k}$ matrix whose element (i,j) is the determinant of the $k \times k$ submatrix $A(\alpha_i, \beta_j)$ of A, where $\alpha_i \subseteq \{1, 2, \ldots, m\}$ and $\beta_j \subseteq \{1, 2, \ldots, n\}$ are index sets with k elements, ordered lexicographically.

What is meant by the lexicographic order of the index sets, will be clear from the following example.

Example B.1. For

$$A = \begin{pmatrix} 1 & -4 & -7 & 10 \\ -2 & 5 & 8 & -11 \\ 3 & -6 & 9 & 12 \end{pmatrix} \tag{B.1}$$

the first compound matrix is equal to A itself:

$$C_1(A) = A \ ,$$

the second compound matrix is

$$\begin{aligned} C_2(A) &= \begin{pmatrix} \det\begin{pmatrix} 1 & -4 \\ -2 & 5 \end{pmatrix} & \det\begin{pmatrix} 1 & -7 \\ -2 & 8 \end{pmatrix} & \det\begin{pmatrix} 1 & 10 \\ -2 & -11 \end{pmatrix} & \det\begin{pmatrix} -4 & -7 \\ 5 & 8 \end{pmatrix} & \det\begin{pmatrix} -4 & 10 \\ 5 & -11 \end{pmatrix} & \det\begin{pmatrix} -7 & 10 \\ 8 & -11 \end{pmatrix} \\ \det\begin{pmatrix} 1 & -4 \\ 3 & -6 \end{pmatrix} & \det\begin{pmatrix} 1 & -7 \\ 3 & 9 \end{pmatrix} & \det\begin{pmatrix} 1 & 10 \\ 3 & 12 \end{pmatrix} & \det\begin{pmatrix} -4 & 10 \\ -6 & 12 \end{pmatrix} & \det\begin{pmatrix} -7 & 10 \\ 9 & 12 \end{pmatrix} & \det\begin{pmatrix} -4 & -7 \\ -6 & 9 \end{pmatrix} \\ \det\begin{pmatrix} -2 & 5 \\ 3 & -6 \end{pmatrix} & \det\begin{pmatrix} -2 & 8 \\ 3 & 9 \end{pmatrix} & \det\begin{pmatrix} -2 & -11 \\ 3 & 12 \end{pmatrix} & \det\begin{pmatrix} 5 & -11 \\ -6 & 12 \end{pmatrix} & \det\begin{pmatrix} 8 & -11 \\ 9 & 12 \end{pmatrix} & \det\begin{pmatrix} 5 & 8 \\ -6 & 9 \end{pmatrix} \end{pmatrix} , \\ &= \begin{pmatrix} -3 & -6 & 9 & 3 & -6 & -3 \\ 6 & 30 & -18 & -78 & 12 & 174 \\ -3 & -42 & 9 & 93 & -6 & 195 \end{pmatrix} , \end{aligned}$$

and the third compound matrix of A is

$$\begin{aligned} C_3(A) &= \begin{pmatrix} \det\begin{pmatrix} 1 & -4 & -7 \\ -2 & 5 & 8 \\ 3 & -6 & 9 \end{pmatrix} & \det\begin{pmatrix} 1 & -4 & 10 \\ -2 & 5 & -11 \\ 3 & -6 & 12 \end{pmatrix} & \det\begin{pmatrix} 1 & -7 & 10 \\ -2 & 8 & -11 \\ 3 & 9 & 12 \end{pmatrix} & \det\begin{pmatrix} -4 & -7 & 10 \\ 5 & 8 & -11 \\ -6 & 9 & 12 \end{pmatrix} \end{pmatrix} , \\ &= \begin{pmatrix} -54 & 0 & -162 & 108 \end{pmatrix} . \end{aligned}$$

The index sets used for $C_2(A)$ are $\alpha_1 = \{1,2\}$, $\alpha_2 = \{1,3\}$, $\alpha_3 = \{2,3\}$ and $\beta_1 = \{1,2\}$, $\beta_2 = \{1,3\}$, $\beta_3 = \{1,4\}$, $\beta_4 = \{2,3\}$, $\beta_5 = \{2,4\}$, $\beta_6 = \{3,4\}$. For $C_3(A)$ the index sets are $\alpha_1 = \{1,2,3\}$ and $\beta_1 = \{1,2,3\}$, $\beta_2 = \{1,2,4\}$, $\beta_3 = \{1,3,4\}$, $\beta_4 = \{2,3,4\}$.

B.1.2 Properties of the compound matrices

1. The kth compound matrix of the identity matrix:

$$C_k(I_n) = I_{\binom{n}{k}} \ .$$

2. The kth compound matrix of a scalar multiple of a matrix:
$$C_k(tA) = t^k C_k(A),\ t \in \mathbb{R} .$$
3. The kth compound matrix of the inverse matrix:
$$C_k(A^{-1}) = C_k(A)^{-1},\ \text{for nonsingular } A .$$
4. The kth compound matrix of the transpose:
$$C_k(A^T) = C_k(A)^T . \tag{B.2}$$
5. The kth compound matrix of a product of matrices:
$$C_k(AB) = C_k(A)C_k(B),\ \text{for } A \in \mathbb{R}^{m\times n},\ B \in \mathbb{R}^{n\times p} \text{ and } k \le \min(m,n,p) . \tag{B.3}$$

Property B.1. The singular values of the kth compound matrix
The singular values of the kth compound matrix of A, $C_k(A)$, are all products $\sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_k}$ of singular values of A.

For the matrix A in (B.1) with singular spectrum
$$\sigma(A) = (\sigma_1, \sigma_2, \sigma_3) = (21.5310, 13.6361, 0.6882) ,$$
we obtain
$$\begin{array}{lclcl} \sigma(C_1(A)) & = & (\sigma_1, \sigma_2, \sigma_3) & = & (21.5310, 13.6361, 0.6882) , \\ \sigma(C_2(A)) & = & (\sigma_1\sigma_2, \sigma_1\sigma_3, \sigma_2\sigma_3) & = & (293.5990, 14.8172, 9.3841) , \\ \sigma(C_3(A)) & = & \sigma_1\sigma_2\sigma_3 & = & 202.0495 . \end{array}$$

B.2 Grassmann coordinates

If a matrix $A \in \mathbb{R}^{m\times n}$ is of full row rank, then its mth compound matrix is a non-zero vector. This is the Grassmann vector, associated with the row space of A.

Definition B.2. Grassmann vector and Grassmann coordinates
Let $A \in \mathbb{R}^{m\times n}$ be a matrix of full row rank. Then, the Grassmann vector $g(A)$ associated with row(A) *is the mth compound matrix of A. The components of $g(A)$ are called the Grassmann coordinates (also the Plücker coordinates).*

Note that with this definition the Grassmann vector associated with a subspace depends on the matrix that represents the subspace. Therefore we will denote the Grassmann vector associated with the row space of a matrix A by $g(A)$ (and not $g(\text{row}(A))$).
However, one can prove the following:

Property B.2.
Let the rows of $A_1 \in \mathbb{R}^{m \times n}$ and $A_2 \in \mathbb{R}^{m \times n}$ span the same m-dimensional space. Then, the Grassmann vectors obtained from both matrices differ only by a scalar multiple:

$$g(A_1) = cg(A_2), \ c \in \mathbb{R} \ .$$

Consequently, the Grassmann coordinates associated with an m-dimensional subspace of $\mathbb{R}^n$ span a line in $\mathbb{R}^{\binom{n}{m}}$. Thus, there is a one-to-one correspondence between m-dimensional subspaces in $\mathbb{R}^n$ and 1-dimensional subspaces in $\mathbb{R}^{\binom{n}{m}}$, see e.g. [132, Theorem 4.9].

Example B.2. As an example, take the matrix A in (B.1). The Grassmann vector associated with its row space has already been computed in Example B.1. It is equal to $C_3(A)$. $\Diamond$

The following theorem is proven in [102]. It establishes a nice relation between the principal angles between two subspaces and the Grassmann vectors associated to the subspaces.

Theorem B.3. Relation of principal angles and Grassmann vectors
Let $A, B \in \mathbb{R}^{m \times n}$ be full row rank matrices. Then the cosine of the angle between the Grassmann vectors $g(A)$ and $g(B)$ corresponding to the row spaces of A and B, is equal to the product of the cosines of the principal angles between the row spaces:

$$\cos\left[g(A) \sphericalangle g(B)\right] = \prod_{i=1}^{m} \cos\theta_i \ ,$$

where $(\theta_1, \ldots, \theta_m) = [A \sphericalangle B]$.

Remark B.1. The cepstral norm in terms of the angle between Grassmann vectors
Let $H(z) = \frac{b(z)}{a(z)}$ be the transfer function of a minimal, stable and minimum phase ARMA model of order n. As was shown in Section 6.3 and Section 6.4, the cepstral norm of this model that was defined in Definition 6.2, can be expressed in terms of principal angles between subspaces. Since the norm depends on the product of the cosines of the angles, we see now that it can also be written as a function of the angle between the corresponding Grassmann vectors.
First, the cepstral norm is a function of the principal angles between the row spaces of two controllability matrices (see Theorem 6.2):

$$\|\log H\|^2 = -\log \prod_{k=1}^{n} \cos^2\theta_k \ ,$$

where

$$(\theta_1, \ldots, \theta_n) = \left[\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right) \sphericalangle \mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)\right] \ .$$

Consequently,

$$\|\log H\|^2 = -\log\cos^2\left[g\left(\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)\right) \sphericalangle g\left(\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)\right)\right] .$$

Secondly, since the cosines of $\theta_1, \ldots, \theta_n$ are the smallest n canonical correlations of the future input and output of the model, and the other canonical correlations are equal to 1, we can write the cepstral norm as

$$\|\log H\|^2 = \lim_{i,j\to\infty} -\log\prod_{k=1}^{i}\cos^2\left[U_f \sphericalangle Y_f\right] ,$$

where U_f and Y_f are the $i \times j$ future input and output Hankel matrices. By Theorem B.3, this leads to

$$\|\log H\|^2 = -\log\cos^2\left[g(U_f) \sphericalangle g(Y_f)\right], \quad \text{for } i, j \to \infty .$$

○

B.3 The volume of a matrix

It is well known that the volume of the parallelepiped spanned by the rows or columns of a square matrix is equal to the absolute value of its determinant:

$$\mathrm{vol}(\mathrm{parallelepiped}(A)) = |\det(A)|, \text{ for } A \in \mathbb{R}^{n\times n} .$$

For a non-square matrix $A \in \mathbb{R}^{m\times n}$ with $m < n$, the volume of the parallelepiped spanned by its rows is

$$\begin{aligned} \mathrm{vol}(\mathrm{parallelepiped}(A)) &= \sqrt{\det(AA^T)} , \\ &= \prod_{i=1}^{m}\sigma_i , \end{aligned} \tag{B.4}$$

where $\sigma_1, \ldots, \sigma_m$ are the singular values of A. Note that if A is not of full row rank, then $\mathrm{vol}(\mathrm{parallelepiped}(A)) = 0$. The volume of a matrix, however, is defined in such a way that it is strictly positive unless its rank is zero.

Definition B.3. The volume of a matrix [13]
The volume of a matrix $A \in \mathbb{R}^{m\times n}$ of rank r is equal to the product of its r non-zero singular values.

$$\mathrm{vol}(A) = \begin{cases} 0 & \text{if } r = 0 , \\ \prod_{i=1}^{r}\sigma_i & \text{if } r > 0 . \end{cases}$$

Observe that if A is not of full row rank, then $\mathrm{vol}(A)$ is not the volume of the parallelepiped spanned by its rows, which is equal to zero.

B.3.1 Relation of the volume and the rth compound matrix of a matrix

Due to Proposition B.1, the largest singular value of $C_r(A)$, where r is the rank of A, is vol(A):

$$\mathrm{vol}(A) = \sigma_{\max}(C_r(A)) \ .$$

Since the maximum singular value of a matrix is equal to its spectral norm, we also have

$$\mathrm{vol}(A) = \|C_r(A)\|_2 \ . \tag{B.5}$$

B.3.2 Relation of the volume and the Grassmann coordinates of a matrix

The mth compound matrix of an $m \times n$ matrix A of full row rank is equal to its Grassmann vector (Definition B.2). By (B.5) we thus obtain

$$(\mathrm{vol}(A))^2 = \|g(A)\|^2 = \sum_{i=1}^{\binom{n}{m}} (g_i(A))^2 \ \ , \tag{B.6}$$

where $g_i(A)$ is the ith component of the Grassmann vector $g(A)$.
For the full row rank matrices $A, B \in \mathbb{R}^{m\times n}$ the cosine of the angle between their Grassmann vectors $g(A)$ and $g(B)$ can be written as

$$\cos\left[g(A) \sphericalangle g(B)\right] = \frac{|g(A)g(B)^T|}{\|g(A)\|\|g(B)\|} \ .$$

Due to (B.2) and (B.3) we have

$$\cos\left[g(A) \sphericalangle g(B)\right] = \frac{|g(AB^T)|}{\|g(A)\|\|g(B)\|} \ ,$$

provided AB^T is of full rank. Note that AB^T is an $m \times m$ matrix and since $g(AB^T)$ is the mth compound matrix of AB^T, it is a scalar (see Definition B.1). Hence,

$$\begin{aligned} |g(AB^T)| &= \sqrt{\|g(AB^T)\|^2} \ , \\ &= \mathrm{vol}(AB^T) \qquad\qquad \text{by (B.6)} \ . \end{aligned}$$

Consequently,

$$\cos\left[g(A) \sphericalangle g(B)\right] = \frac{\mathrm{vol}(AB^T)}{\mathrm{vol}(A)\mathrm{vol}(B)} \ . \tag{B.7}$$

B.3.3 Relation of the matrix volume and the principal angles between two subspaces

By applying Theorem B.3 to (B.7) we obtain for $A, B \in \mathbb{R}^{m\times n}$ of full row rank and for which AB^T is full rank,

$$\prod_{i=1}^{m} \cos\theta_i = \frac{\mathrm{vol}(AB^T)}{\mathrm{vol}(A)\mathrm{vol}(B)} \ , \tag{B.8}$$

where $\theta_1, \ldots, \theta_m$ are the m principal angles between $\mathrm{row}(A)$ and $\mathrm{row}(B)$. Observe that for $m = 1$, the expression in (B.8) reduces to the cosine of the angle between two n-dimensional vectors:

$$\cos\theta_1 = \frac{|AB^T|}{\|A\|\|B\|} \ .$$

Remark B.2. The cepstral norm in terms of matrix volumes
We can again derive new expressions for the cepstral norm of a SISO ARMA model. Let the transfer function of the model be $H(z) = \frac{b(z)}{a(z)}$. Then, its cepstral norm is equal to

$$\begin{aligned}\|\log H\|^2 &= -\log \left(\frac{\mathrm{vol}\left(\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)^T\right)}{\mathrm{vol}\left(\mathcal{C}_\infty\left(\frac{z^n}{a(z)}\right)\right)\mathrm{vol}\left(\mathcal{C}_\infty\left(\frac{z^n}{b(z)}\right)\right)} \right)^2 \ , \\ &= -\log\left(\frac{\mathrm{vol}(U_f Y_f^T)}{\mathrm{vol}(U_f)\mathrm{vol}(Y_f)}\right)^2 \ , \quad \text{for } i, j \to \infty \ .\end{aligned}$$

○

The equality in (B.8) can be generalized for matrices that are not of full rank as in the following theorem, the proof of which can be found in [5].

Theorem B.4. The Cauchy–Schwarz inequality for volumes
Let $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{q\times n}$ be matrices of equal rank r and let the rank of AB^T be denoted by r_{ab}. If $r_{ab} = r$, then

$$\mathrm{vol}(AB^T) = \mathrm{vol}(A)\mathrm{vol}(B)\prod_{i=1}^{r}\cos\theta_i \ ,$$

where $\theta_1, \ldots, \theta_r$ are the r principal angles between $\mathrm{row}(A)$ *and* $\mathrm{row}(B)$.

Note that if $r_{ab} < r$, then there is at least one principal angle equal to $\frac{\pi}{2}$ (see Section 2.5.3) and hence $\prod_{i=1}^{r}\cos\theta_i = 0$. The volume of AB^T, however, is not equal to 0, unless $r_{ab} = 0$.

For matrices of full row rank (B.8) can also be obtained from the characterization of the cosines of the principal angles as eigenvalues (see Section 2.5.5). Let $A, B \in \mathbb{R}^{m\times n}$ be full row rank matrices. Since the squared cosines of the principal angles between the row spaces of A and B are equal to the eigenvalues of $(AA^T)^{-1}AB^T(BB^T)^{-1}BA^T$, their product is

$$\prod_{i=1}^{r}\cos^2\theta_i = \det\left((AA^T)^{-1}AB^T(BB^T)^{-1}BA^T\right) ,$$

which is also equal to

$$\begin{aligned}\prod_{i=1}^{r}\cos^2\theta_i &= \frac{\det(BA^T)\det(AB^T)}{\det(AA^T)\det(BB^T)} ,\\ &= \frac{\det((BA^T)(BA^T)^T)}{\det(AA^T)\det(BB^T)} ,\end{aligned}$$

and due to (B.4), we obtain (B.8).
From Property 2.8 and Property 2.9 we obtain two other representations of the product of the non-zero cosines of the principal angles as the volume of a matrix:

$$\prod_{i=1}^{r_{ab}}\cos\theta_i = \mathrm{vol}(V_{a_1}^T V_{b_1}) ,$$

and

$$\prod_{i=1}^{r_{ab}}\cos\theta_i = \mathrm{vol}(\Pi_{\mathrm{row}(A)}\Pi_{\mathrm{row}(B)}) ,$$

where $r_{ab} = \mathrm{rank}(AB^T)$, the number of angles smaller than $\frac{\pi}{2}$, V_{a_1} and V_{b_1} are the right singular matrices of the reduced SVD of A and B respectively and $\Pi_{\mathrm{row}(A)}$ and $\Pi_{\mathrm{row}(B)}$ are the orthogonal projectors onto the row spaces of A and B.

Theorem B.5. The volume of the matrix whose columns are the principal directions in row(A) and row(B)
Let $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{q\times n}$ be of rank r_a and r_b respectively, where $r_a \le r_b$ and $n \ge 2r_a$ and let the columns of $U \in \mathbb{R}^{n\times r_a}$ and $V \in \mathbb{R}^{n\times r_a}$ be the principal directions in row(A) *and* row(B) *respectively, corresponding to the principal angles between* row(A) *and* row(B).
Then,

$$\mathrm{vol}\left(\begin{pmatrix} U & V\end{pmatrix}\right) = 2^{\frac{t}{2}}\prod_{i=t+1}^{r_a}\sin\theta_i ,$$

where $\theta_1,\dots,\theta_{r_a}$ are the principal angles between row(A) *and* row(B) *and t is the number of zero angles.*

Proof.
In order to apply Definition B.3, we first determine the singular values of $W = \begin{pmatrix} U & V \end{pmatrix}$. The squares of the singular values of W are the eigenvalues of

$$W^T W = \begin{pmatrix} U^T \\ V^T \end{pmatrix} \begin{pmatrix} U & V \end{pmatrix} \text{ (see Section 2.2.3)} .$$

Due to the orthonormality of the principal directions,

$$W^T W = \begin{pmatrix} I_{r_a} & U^T V \\ V^T U & I_{r_a} \end{pmatrix} = \begin{pmatrix} I_{r_a} & S \\ S & I_{r_a} \end{pmatrix} ,$$

where S is the diagonal matrix with the cosines of the principal angles:

$$S = \begin{pmatrix} \cos\theta_1 & 0 & \cdots & 0 \\ 0 & \cos\theta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cos\theta_{r_a} \end{pmatrix} .$$

Consequently, the squared singular values of W are

$$\begin{aligned} \sigma_1^2 &= 1 + \cos\theta_1 , \\ &\vdots \\ \sigma_{r_a}^2 &= 1 + \cos\theta_{r_a} , \\ \sigma_{r_a+1}^2 &= 1 - \cos\theta_{r_a} , \\ &\vdots \\ \sigma_{2r_a}^2 &= 1 - \cos\theta_1 . \end{aligned}$$

If t principal angles between row(A) and row(B) are equal to zero, then we know that $\sigma_1^2 = \cdots = \sigma_t^2 = 2$ and $\sigma_{2r_a} = \cdots = \sigma_{2r_a-t+1} = 0$. Hence, the product of the non-zero squared singular values of $\begin{pmatrix} U & V \end{pmatrix}$ is

$$\begin{aligned} \prod_{i=1}^{2r_a-t} \sigma_i^2 &= 2^t \prod_{i=t+1}^{r_a} (1+\cos\theta_i)(1-\cos\theta_i) , \\ &= 2^t \prod_{i=t+1}^{r_a} (1-\cos^2\theta_i) , \\ &= 2^t \prod_{i=t+1}^{r_a} \sin^2\theta_i . \end{aligned}$$

Consequently,

$$\operatorname{vol}\left(\begin{pmatrix} U & V \end{pmatrix}\right) = \prod_{i=1}^{2r_a-t} \sigma_i = 2^{\frac{t}{2}} \prod_{i=t+1}^{r_a} \sin\theta_i .$$

□

Observe that if there are no zero principal angles between row(A) and row(B), then we can compute the volume of the parallelepiped spanned by the principal directions as

$$\mathrm{vol}(\text{parallelepiped}) = \prod_{i=1}^{r_a} \sin\theta_i \ . \tag{B.9}$$

Remark B.3. The cepstral norm as a function of the volume of the principal directions in row(Y_p) and row(Y_f)
The cepstral norm of the nth order SISO ARMA model with transfer function $H(z)$ can be written as (see Equation 6.13)

$$\|\log H\|^2 = -\log \prod_{k=1}^{n} \sin^2\phi_k \ ,$$

where $\phi_1, \ldots, \phi_n$ are the smallest n principal angles between row(Y_p) and row(Y_f) (if the number of rows, i, and the number of columns, j, in the Hankel matrices go to infinity). The other principal angles are equal to $\frac{\pi}{2}$. Let the rows of the matrices $U \in \mathbb{R}^{n\times j}$ and $V \in \mathbb{R}^{n\times j}$ be the principal directions in row(Y_p) and row(Y_f) respectively, corresponding to the principal angles $\phi_1, \ldots, \phi_n$. By Theorem B.5, we can consequently express the cepstral norm as

$$\|\log H\|^2 = -\log \mathrm{vol}^2 \begin{pmatrix} U \\ V \end{pmatrix} \ .$$

Note that the future forward innovation state sequence X_f and the past backward innovation state sequence $\bar{X}_p$ can be chosen to be equal to $S^{\frac{1}{2}}U$ and $S^{\frac{1}{2}}V$ respectively, where S is the diagonal matrix with the largest n canonical correlations of past and future, i.e. $\cos\phi_1, \ldots, \cos\phi_n$, on the diagonal [159, pp. 80–81] [156].[1] The cepstral norm can consequently also be written in terms of the volume of the state sequence matrices:

$$\|\log H\|^2 = -\log \mathrm{vol}^2 \begin{pmatrix} X_f \\ \bar{X}_p \end{pmatrix} + \log \prod_{k=1}^{n} \cos^2\phi_k \ .$$

○

[1] This choice leads to a stochastically balanced model (see Remark 3.14).

Bibliography

[1] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.

[2] H. Akaike. Stochastic theory of minimal realization. *IEEE Transactions on Automatic Control*, 19:667–674, 1974.

[3] S. Alampalli. Influence of in-service environment on modal parameters. In *Proceedings of the International Modal Analysis Conference (IMAC 16)*, pages 111–116, Santa Barbara, USA, February 1998.

[4] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 1984.

[5] R. B. Bapat and A. Ben-Israel. Singular values and maximum rank minors of generalized inverses. *Linear & Multilinear Algebra*, 40:153–161, 1995. Available on `http://rutcor.rutgers.edu/~bisrael`.

[6] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, December 1989.

[7] M. Basseville. An invariance property of some subspace-based detection algorithms. *IEEE Transactions on Signal Processing*, 47(12):3398–3400, December 1999.

[8] M. Basseville, M. Abdelghani, and A. Benveniste. Subspace-based fault detection algorithms for vibration monitoring. *Automatica*, 36:101–109, 2000.

[9] D. Bauer. *Some Asymptotic Theory for the Estimation of Linear Systems Using Maximum Likelihood Methods or Subspace Algorithms*. PhD thesis, Technische Universität Wien, Vienna, Austria, 1998. Available on `http://www.eos.tuwien.ac.at/Oeko/papers/ES174.zip`.

[10] D. Bauer. Order estimation for subspace methods. *Automatica*, 37:1561–1573, 2001.

[11] D. Bauer, M. Deistler, and W. Scherrer. Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs. *Automatica*, 35:1243–1254, 1999.

[12] E. Beltrami. Sulle funzioni bilineari. *Giornale di Matematiche ad Uso degli Studenti delle Universita*, 11:98–106, 1873. An English translation by D. Boley is available as University of Minnesota, Department of Computer Science, Technical Report 90–37, 1990.

[13] A. Ben-Israel. A volume associated with $m \times n$ matrices. *Linear Algebra and its Applications*, 167:87–111, 1992. Available on `http://rutcor.rutgers.edu/~bisrael/VOLUME.dvi`.

[14] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. Wiley, New York, 1977.

[15] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto. Recognition of human gaits. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 01)*, volume II, pages 52–58, Kauai, Hawaii, December 2001. Available on `black1.csl.uiuc.edu/~yima/psfile/gait_rcgntn.ps.gz`.

[16] Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27:579–594, 1973.

[17] B. P. Bogert, M. J. R. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In M. Rosenblat, editor, *Proceedings of the Symposium on Time Series Analysis*, pages 209–243. Wiley, New York, 1963.

[18] M. Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Linköping University, Linköping, Sweden, 1998. Available on `http://people.imt.liu.se/~magnus`.

[19] D. R. Brillinger. John Wilder Tukey (1915 – 2000). *Notices of the AMS*, pages 193–201, February 2002.

[20] P. M. T. Broersen. Facts and fiction in spectral analysis of stationary stochastic processes. In S. Theodoridis, I. Pitas, A. Stouraitis, and N. Kalouptsidis, editors, *Proceedings of the IX European Signal Processing Conference (EUSIPCO 98)*, pages 61–64, Rhodes, Greece, September 1998. Available on `http://www.tn.tudelft.nl/mmr/`.

[21] J. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, Stanford University, Stanford, USA, 1975. Available on `http://sepwww.stanford.edu/theses/sep06`.

[22] C. I. Byrnes, P. Enqvist, and A. Lindquist. Identifiability and well-posedness of shaping-filter parameterizations: a global analysis approach. Technical report, Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 2000. Available on `http://www.math.kth.se/optsyst/alq/Publications.html` as `BEL1.pdf`.

[23] C. I. Byrnes, P. Enqvist, and A. Lindquist. Cepstral coefficients, covariance lags, and pole-zero models for finite data strings. *IEEE Transactions on Signal Processing*, 49(4):677–693, April 2001.

[24] C. I. Byrnes and A. Lindquist. On the partial stochastic realization problem. *IEEE Transactions on Automatic Control*, 42:1049–1070, 1997. Available on `http://www.math.kth.se/optsys/alq/Publications.html` as `BL97.pdf`.

[25] P. E. Caines. *Linear Stochastic Systems.* Wiley, New York, 1988.

[26] H. Che and S. Q. Li. Fast algorithms for measurement-based traffic modeling. *IEEE Journal on Selected Areas in Communications*, 16(5), June 1998. Available on `http://www.ece.utexas.edu/~sanqi/papers.html`.

[27] D. G. Childers, D. P. Skinner, and R. C. Kemerait. The cepstrum: a guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443, October 1977.

[28] J. W. Cooley and J. W. Tukey. An algorithm for the machine computation of complex fourier series. *Mathematics of Computation*, 19:297–301, April 1965.

[29] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. Wiley, New York, 1991.

[30] A. Dahlén. *Identification of Stochastic Systems: Subspace Methods and Covariance Extension.* PhD thesis, Royal Institute of technology, Stockholm, Sweden, 2001.

[31] A. Dahlén, A. Lindquist, and J. Marí. Experimental evidence showing that stochastic subspace identification methods may fail. *Systems & Control Letters*, 34:303–312, 1998. Available on `http://www.math.kth.se/optsys/alq/Publications.html` as `DLM.pdf`.

[32] A. Dahlén and W. Scherrer. The relation of the CCA subspace method to a balanced reduction of an autoregressive model. Technical report, Institute for Econometrics, Operations Research and System Theory, Technische Universität Wien, Vienna, Austria, September 2000.

[33] K. De Cock and B. De Moor. Identification of the first order parameters of a circulant modulated Poisson process. In *Proceedings of the International Conference on Telecommunications (ICT 98)*, volume II, pages 420–424, Porto Carras, Greece, June 1998. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/decock/reports` as `97-104.ps.gz`.

[34] K. De Cock and B. De Moor. Stochastic system identification for ATM network traffic models. Technical Report 98-57, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 1998. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/decock/reports` as `98-57.ps.gz`.

[35] K. De Cock and B. De Moor. Stochastic system identification for ATM network traffic models: a time domain approach. In *Proceedings of the IX European Signal Processing Conference (EUSIPCO 98)*, pages 2137–2140, Rhodes, Greece, September 1998. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/decock/reports` as `98-22.ps.gz`.

[36] K. De Cock and B. De Moor. Subspace angles and distances between ARMA models. In *Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems (MTNS 2000)*, Perpignan, France, June 2000. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/decock/reports` as `00-39.ps.gz`.

[37] K. De Cock and B. De Moor. Subspace angles between linear stochastic models. In *Proceedings of the 39th IEEE Conference on Decision and Control (CDC 2000)*, pages 1561–1566, Sydney, Australia, December 2000. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/decock/reports` as `99-101a.ps.gz`.

[38] K. De Cock and B. De Moor. Subspace angles between AR models. Technical Report 00-44, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, January 2001. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/decock/reports` as `00-44.ps.gz`.

[39] K. De Cock and B. De Moor. Subspace angles between ARMA models. Technical Report 00-44a, ESAT-SISTA, K.U.Leuven, accepted for publication in *Systems & Control Letters*, Leuven, Belgium, January 2002. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/decock/reports` as `00-44a.ps.gz`.

[40] K. De Cock, T. Van Gestel, and B. De Moor. Identification of the circulant modulated Poisson process: a time domain approach. In *Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems (MTNS 98)*, pages 739–742, Padova, Italy, July 1998. Available

on ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/decock/reports as 98-31.ps.gz.

[41] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-time Processing of Speech Signals*. MacMillan, New York, 1993.

[42] U. B. Desai and D. Pal. A realization approach to stochastic model reduction and balanced stochastic realization. In *Proceedings of the 21st IEEE Conference on Decision and Control (CDC 82)*, pages 1105–1112, Orlando, USA, December 1982.

[43] U. B. Desai and D. Pal. A transformation approach to stochastic model reduction. *IEEE Transactions on Automatic Control*, AC–29(12):1097–1100, December 1984.

[44] S. Doclo, E. De Clippel, and M. Moonen. Combined acoustic echo and noise reduction using GSVD-based optimal filtering. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000)*, volume 2, pages 1061–1064, Istanbul, Turkey, June 2000. Available on ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/doclo/reports as 99-85.ps.gz.

[45] K. Eneman and M. Moonen. The iterated partitioned block frequency-domain adaptive filter for acoustic echo cancellation. In *Proceedings of the 2001 IEEE/EURASIP International Workshop on Acoustic Echo and Noise Control (IWAENC 01)*, pages 135–138, Darmstadt, Germany, September 2001. Available on ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/eneman/reports as 01-65.ps.gz.

[46] P. Enqvist. A convex optimization approach to ARMA(n, m) model design from covariance and cepstrum data. Technical report, Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 2001.

[47] P. Enqvist. *Spectral Estimation by Geometric, Topological and Optimization Methods*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 2001.

[48] M. I. Friswell, J. E. Mottershead, and H. Ahmadian. Combining subset selection and parameter constraints in model updating. *Journal of Vibration and Acoustics*, 120:854–859, 1998.

[49] M. I. Friswell, J. E. T. Penny, and S. D. Garvey. Parameter subset selection in damage location. *Inverse Problems in Engineering*, 5:189–215, 1997. Available on http://www.swan.ac.uk/mecheng/staff/mfriswell/PDF_Files/ as J33.pdf.

[50] W. A. Fuller. *Introduction to Statistical Time Series*. Wiley, New York, 1976.

[51] C. F. Gauss. Theoria combinationis observationum erroribus minimis obnoxiae, pars posterior. In *Werke*, volume IV, pages 27–53. Königlichen Gesellshaft der Wissenschaften zu Göttingen, 1823.

[52] I. M. Gel'fand and A. M. Yaglom. Calculation of the amount of information about a random function contained in another such function. *American Mathematical Society Translations, Series (2)*, 12:199–236, 1959.

[53] R. Gittins. *Canonical Analysis: a Review with Applications in Ecology*. Springer, Berlin, 1985.

[54] A. Goldstein. Alan V. Oppenheim, electrical engineer. An oral history conducted by Andrew Goldstein, IEEE History Center, Rutgers University, New Brunswick, NJ, USA, 1997. Available on `http://www.ieee.org/organizations/history_center/oral_histories/transcripts/oppenheim.html`.

[55] G. H. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *SIAM Journal on Numerical Analysis*, 2:205–224, 1965.

[56] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solution. *Numerische Mathematik*, 14:403–420, 1970.

[57] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1996.

[58] G. H. Golub and H. Zha. The canonical correlations of matrix pairs and their numerical computation. In A. Bojanczyk and G. Cybenko, editors, *Linear Algebra for Signal Processing*, pages 59–82. Springer, New York, 1995.

[59] R. M. Gray. On the asymptotic eigenvalue distribution of Toeplitz matrices. *IEEE Transactions on Information Theory*, IT–18(6):725–730, November 1972.

[60] R. M. Gray. Toeplitz and circulant matrices: A review. Technical report, Electrical Engineering Department, Stanford University, Stanford, USA, 2001. Available on `http://www-isl.stanford.edu/~gray/toeplitz.pdf`.

[61] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–28(4):367–376, August 1980.

[62] A. H. Gray, Jr. and J. D. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–24(5):380–391, October 1976.

[63] U. Grenander and G. Szegő. *Toeplitz Forms and their Applications.* Chelsea, New York, 1958.

[64] E. J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems.* Wiley, New York, 1988.

[65] B. Hanzon. The area enclosed by the (oriented) Nyquist diagram and the Hilbert-Schmidt-Hankel norm of a linear system. *IEEE Transactions on Automatic Control*, 37(6):835–839, June 1992.

[66] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, SAC–4(6):856–868, September 1986.

[67] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1990.

[68] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–372, 1936.

[69] I. A. Ibragimov and Y. A. Rozanov. *Gaussian Random Processes.* Springer, New York, 1978.

[70] S. Ihara. *Information Theory for Continuous Systems.* World Scientific, 1993.

[71] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[72] J. L. Jensen. Sur les fonctions convexes et les inégalitées entre les valeurs moyennes. *Acta Mathematica*, 30:175–193, 1906.

[73] N. P. Jewell and P. Bloomfield. Canonical correlations of past and future for time series: definitions and theory. *The Annals of Statistics*, 11(3):837–847, 1983.

[74] N. P. Jewell, P. Bloomfield, and F. C. Bartmann. Canonical correlations of past and future for time series: bounds and computation. *The Annals of Statistics*, 11(3):848–855, 1983.

[75] C. Jordan. Mémoire sur les formes bilinéaires. *Journal de Mathématiques Pures et Appliquées, Deuxième Série*, 19:35–54, 1874.

[76] C. Jordan. Sur la réduction des formes bilinéaires. *Comptes Rendus de l'Académie des Sciences, Paris*, 78:614–617, 1874.

[77] C. Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société Mathématique*, 3:103–174, 1875.

[78] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–35(7):947–954, July 1987.

[79] T. Kailath. A view of three decades of linear filtering theory. *IEEE Transactions on Information Theory*, 20:146–181, 1974.

[80] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, Upper Saddle River, NJ, 2000.

[81] *IEEE Transactions on Automatic Control, Special Issue on Applications of Kalman Filtering*, AC–28(3), 1983.

[82] R. E. Kalman. Contributions to the theory of optimal control. *Boletin de la Sociedad Matematica Mexicana*, 5:102–119, 1960.

[83] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers, Journal of Basic Engineering*, 83(1):35–45, March 1960.

[84] R. E. Kalman. On the general theory of control systems. In *Proceedings of the first IFAC Congress*, volume 1, pages 481–491, London, UK, 1960.

[85] R. E. Kalman. Mathematical description of linear dynamical systems. *SIAM Journal on Control*, 1:152–192, 1963.

[86] J. L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926, 1956.

[87] A. I. Khinchin. The entropy concept in probability theory. *Uspekhi Matematicheskikh Nauk*, VIII(3):3–20, 1953. English translation by R. A. Silverman and M. D. Friedman in *Mathematical Foundations of Information Theory*, Dover Publications, New York, 1957.

[88] A. N. Kolmogorov. Sur l'interpolation et extrapolation des suites stationnaires. *Comptes Rendus de l'Académie des Sciences*, 208:2043, 1939.

[89] A. N. Kolmogorov. Interpolation and extrapolation of stationary random processes. *Bull. Acad Sci. USSR*, 5, 1941.

[90] A. N. Kolmogorov. Stationary sequences in Hilbert space (in Russian). *Bull. Math. Univ. Moscow*, 2, 1941.

[91] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[92] S. Y. Kung. A new identification method and model reduction algorithm via singular value decomposition. In *Proceedings of the 12th Asilomar Conference on Circuits, Systems and Computations*, pages 705–714, Asilomar, USA, 1978.

[93] C. Lanczos. *Linear Differential Operators.* SIAM, Philadelphia, 1996.

[94] W. E. Larimore. System identification, reduced ordered filtering and modeling via canonical variate analysis. In *Proceedings of the American Control Conference (ACC 83)*, volume 2, pages 445–451, San Francisco, USA, 1983.

[95] L. Li and Z. Xie. Model selection and order determination for time series by information between the past and the future. *Journal of time series analysis*, 17(1):65–84, 1996. Available on `http://stat.fsu.edu/~lilei`.

[96] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer, New York, 1993.

[97] S. Q. Li, S. Park, and D. Arifler. SMAQ: A measurement-based tool for traffic modeling and queuing analysis part I: Design methodologies and software architecture. *IEEE Communications Magazine*, 36(8), August 1998. Available on `http://www.ece.utexas.edu/~sanqi/papers.html`.

[98] S. Q. Li, S. Park, and D. Arifler. SMAQ: A measurement-based tool for traffic modeling and queuing analysis part II: Network applications. *IEEE Communications Magazine*, 36(8), August 1998. Available on `http://www.ece.utexas.edu/~sanqi/papers.html`.

[99] A. Lindquist and G. Picci. Canonical correlation analysis, approximate covariance extension, and identification of stationary time series. *Automatica*, 32:709–733, 1996. Available on `http://www.math.kth.se/optsys/alq/Publications.html` as `LP96.pdf`.

[100] Matlab identification toolbox, developed by Lennart Ljung. `http://www.mathworks.com/products/sysid/`.

[101] L. Ljung. *System Identification: Theory for the User.* Prentice Hall, Englewood Cliffs, NJ, 1999.

[102] C. S. Macinnes. The solution to a structured matrix approximation problem using Grassman coordinates. *SIAM Journal on Matrix Analysis and Applications*, 21(2):446–453, 1999.

[103] The MacTutor history of mathematics archive. `http://www-groups.dcs.st-and.ac.uk/~history/`.

[104] J. Marí, A. Dahlén, and A. Lindquist. A covariance extension approach to identification of time series. *Automatica*, 36:379–398, 2000. Available on `http://www.math.kth.se/optsys/alq/Publications.html` as `MDL.pdf`.

[105] R. J. Martin. A metric for ARMA processes. *IEEE Transactions on Signal Processing*, 48(4):1164–1170, April 2000.

[106] B. P. McGinnie. *A Balanced View of System Identification*. PhD thesis, Cambridge University, Cambridge, UK, 1994. Available on `http://www-control.eng.cam.ac.uk/bpm/thesis/THESIS.ps`.

[107] B. C. Moore. Principal component analysis in linear systems: controllability, observability and model reduction. *IEEE Transactions on Automatic Control*, AC–26(1):17–32, 1981.

[108] G. Morren, P. Lemmerling, S. Van Huffel, G. Naulaers, H. Devlieger, and P. Casaer. Detection of autoregulation in the brain of premature infants using a novel subspace-based technique. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2001)*, Istanbull, Turkey, October 2001. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/morren/reports` as `01-53.ps.gz`.

[109] G. Morren, G. Naulaers, P. Lemmerling, S. Van Huffel, P. Casaer, and H. Devlieger. Quantitation of the concordance between cerebral intravascular oxygenation and mean arterial blood pressure for the detection of impaired autoregulation. In *Proceedings of the 29th Annual Meeting of the International Society on Oxygen Transport to Tissue (ISOTT 2001)*, University of Pennsylvania, USA, August 2001. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/morren/reports` as `01-129.ps.gz`.

[110] A. V. Oppenheim. Superposition in a class of nonlinear systems. Technical Report 432, Research Laboratory of Electronics, MIT, Cambridge, USA, March 1965.

[111] A. V. Oppenheim, G. E. Kopec, and J. M. Tribolet. Signal analysis by homomorphic prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, AASP–24(4):327–332, August 1976.

[112] A. V. Oppenheim and R. W. Schafer. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, AU–16(2):221–226, June 1968.

[113] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice Hall International, London, 1975.

[114] D. Pal. Balanced stochastic realization and model reduction. Master's thesis, Washington State University, Electrical Engineering, 1982.

[115] B. Peeters. Steel mast website. http://www.kuleuven.ac.be/bwm/macec/mast.htm, March 1999.

[116] B. Peeters. *System Identification and Damage Detection in Civil Engineering*. PhD thesis, K.U.Leuven, Leuven, Belgium, 2000. Available on http://www.kuleuven.ac.be/bwm/papersBartPeeters_PhD.pdf.

[117] B. Peeters, J. Maeck, and G. De Roeck. Vibration-based damage detection in civil engineering: excitation sources and temperature effects. *Smart Materials and Structures*, 30:149–171, 2001. Available on http://www.kuleuven.ac.be/bwm/paperspeetij01b.pdf.

[118] B. Peeters and G. De Roeck. Stochastic subspace system identification of a steel transmitter mast. In *Proceedings of the International Modal Analysis Conference (IMAC 16)*, pages 130–136, Santa Barbara, USA, February 1998.

[119] R. Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51:406–413, 1955.

[120] J. R. Pierce. *An Introduction to Information Theory – Symbols, Signals and Noise*. Dover, New York, 1980.

[121] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden–Day, San Francisco, 1964. Originally published in Russian in 1960.

[122] M. B. Priestley. *Spectral Analysis and Time Series*. London Academic Press, London, 1981.

[123] F. M. Reza. *An Introduction to Information Theory*. Mac Graw-Hill, New York, 1994.

[124] C. F. Richter. An instrumental earthquake magnitude scale. *Bulletin of the Seismological Society of Armica*, 25:1–32, 1935.

[125] G. P. Roberts and A. J. Pearson. Health monitoring of structures – towards a stethoscope for bridges. In *Proceedings of the International Conference on Noise and Vibration Engineering (ISMA 23)*, Leuven, Belgium, September 1998.

[126] G. Rombouts and M. Moonen. Regularized affine projection algorithms for multichannel acoustic echo cancellation. In *Proceedings of the Signal Processing Symposium 2000 (SPS 2000)*, Hilvarenbeek, The Netherlands, March 2000. Available on ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/rombouts/reports as RegMulti.ps.Z.

[127] P. Saisan, G. Doretto, Y. N. Wu, and S. Soatto. Dynamic texture recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 01)*, volume II, pages 58–63, Kauai, Hawaii, December 2001. Available on `http://vision.ucla.edu/papers/487_doreetto.pdf`.

[128] W. Scherrer. Local optimality of minimum phase balanced truncation. Technical report, Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, Vienna, Austria, 2001. To appear in *Proceedings of the 15th IFAC World Congress on Automatic Control*, Barcelona, Spain.

[129] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188:415–431, 1986. Available on `http://www.lecb.ncifcrf.gov/~toms/paper/`.

[130] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948. Available on `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html`.

[131] C. E. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656–715, 1949.

[132] G. C. Shephard. *Vector Spaces of Finite Dimension*. University mathematical texts 32. Oliver and Boyd, Edinburgh, 1966.

[133] S. Soatto, G. Doretto, and Y. N. Wu. Dynamic textures. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 01)*, volume 2, pages 439–446, Vancouver, Canada, July 2001. Available on `http://www.cs.ucla.edu/~doretto/papers/soattoDW01ICCV.ps.gz`.

[134] T. Söderström and P. Stoica. *System Identification*. Prentice Hall International, London, 1989.

[135] F. K. Soong and M. M. Sondhi. A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1):41–48, January 1988.

[136] H. Spall. Charles F. Richter – an interview. Earthquake Information Bulletin, January–February 1980. Abridged version available on `http://wwwneic.cr.usgs.gov/neis/seismology/people/int_richter.html`.

[137] G. W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35:551–566, 1993. Available on `ftp://thales.cs.umd.edu/pub/reports` as `ehsvd.ps`.

[138] G. W. Stewart. *Theory of the Combination of Observations Least Subject to Errors.* Classics in Applied mathematics. SIAM, PA, 1995. Translation of original works by C. F. Gauss under the title *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae.*

[139] T. G. Stockham. The application of generalized linearity to automatic gain control. *IEEE Transactions on Audio and Electroacoustics*, AU–16:267–270, June 1968.

[140] T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen. Blind deconvolution through digital signal processing. *Proceedings of the IEEE*, 63(4):678–692, April 1975.

[141] P. Stoica, T. McKelvey, and J. Marí. MA estimation in polynomial time. *IEEE Transactions on Signal Processing*, 48(7):1999–2012, July 2000.

[142] A. A. Stoorvogel and J. H. van Schuppen. System identification with information theoretic criteria. In S. Bittanti and G. Picci, editors, *Identification, Adaptation, Learning*, pages 289–338. Springer, Berlin, 1996. Available on `http://www.cwi.nl/ftp/CWIreports/BS/` as file `BS-R9513.ps.Z`.

[143] A. A. Stoorvogel and J. H. van Schuppen. Approximation problems with the divergence criterion for Gaussian variables and Gaussian processes. *Systems & Control Letters*, 35:207–218, 1998.

[144] A. A. Stoorvogel and J. H. van Schuppen. Divergence rate approximation of a stationary Gaussian process by the output of a Gaussian system. In A. Beghi, L. Finesso, and G. Picci, editors, *Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems (MTNS 98)*, pages 879–882, Padova, Italy, July 1998.

[145] Y. Tohkura. A weighted cepstral distance measure for speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–35(10):1414–1422, October 1987.

[146] L. N. Trefethen and D. Bau. *Numerical Linear Algebra.* SIAM, Philadelphia, 1997.

[147] J. M. Tribolet. A new phase unwrapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–25(2):170–177, April 1977.

[148] K. Van Acker. *Equalization and Echo Cancellation for DMT-based DSL Modems.* PhD thesis, K.U.Leuven, Leuven, Belgium, January 2001. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vanacker/reports` as `01-02.ps.gz`.

[149] K. Van Acker, M. Moonen, and T. Pollet. Per tone echo cancellation for DMT-based systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*, 2001. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vanacker/reports` as `00-101.ps.gz`.

[150] T. Van Gestel, M. Van Barel, and B. De Moor. On an alternative parameterization of the solutions of the partial realization problem. *Acta Applicandae Mathematicae (Special Issue on ICRA 99)*, 61:317–331, 2000. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vangestel/reports` as `TR99-05.ps.gz`.

[151] T. Van Gestel, K. De Cock, R. Jans, B. De Schutter, Z. Degraeve, and B. De Moor. Discrete stochastic modelling of ATM-traffic with circulant transition matrices. In *Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems (MTNS 98)*, pages 891–894, Padova, Italy, July 1998. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vangestel/reports` as `97-109.ps.gz`.

[152] T. Van Gestel, J. Suykens, P. Van Dooren, and B. De Moor. Imposing stability in subspace identification by regularization. In *Proceedings of the 39th IEEE Conference on Decision and Control (CDC 2000)*, pages 1555–1560, Sydney, Australia, December 2000. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vangestel/reports` as `rep00-30.ps.gz`.

[153] T. Van Gestel, J. Suykens, P. Van Dooren, and B. De Moor. Identification of stable models in subspace identification by using regularization. *IEEE Transactions on Automatic Control*, 46(9):1416–1420, September 2001.

[154] T. Van Gestel, J. A. K. Suykens, J. De Brabanter, B. De Moor, and J. Vandewalle. Kernel canonical correlation analysis and least squares support vector machines. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2001)*, Vienna, Austria, August 2001. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/suykens/reports` as `lssvm_01_24.ps.gz`.

[155] P. Van Mieghem. The mapping of traffic parameters into a MMP(N). Technical report, Alcatel Corporate Research Center, Antwerpen, Belgium, 1995.

[156] P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vanoverschee/reports` as `stoch_auto2.ps.Z`.

[157] P. Van Overschee and B. De Moor. N4SID – Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–94, January 1994. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vanoverschee/reports` as `comb_ext2.ps.Z`.

[158] P. Van Overschee and B. De Moor. A unifying theorem for three subspace system identification algorithms. *Automatica, Special Issue on Trends in System Identification*, 31(12):1853–1864, December 1995. Available on `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vanoverschee/reports` as `ifac_paper2.ps.Z`.

[159] P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems: Theory – Implementation – Applications.* Kluwer Academic Publishers, Boston, 1996.

[160] P. Å. Wedin. On angles between subspaces of finite dimensional inner product space. In A. Ruhe B. Kågström, editor, *Matrix Pencils*, pages 263–285. Springer, Berlin, 1983.

[161] J. Wismer. *Application Note: Gearbox Analysis Using Cepstrum Analysis and Comb Liftering.* Brüel & Kjær, Denmark. Available from the Literature Search on `http://www.bksv.com`.

[162] H. Wold. *A Study in the Analysis of Stationary Time Series.* Almqvist & Wiksell, Stockholm, 1954.

[163] C. Yi and B. De Moor. Traffic identification of ATM networks with optimization algorithms. In *Proceedings of the 35th IEEE Conference on Decision and Control (CDC 96)*, pages 277–282, Kobe, Japan, December 1996.

[164] H. Zeiger and A. McEwen. Approximate linear realization of given dimension via Ho's algorithm. *IEEE Transactions on Automatic Control*, 19:153, 1974.

Index