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

PRIMAL-DUAL KERNEL MACHINES

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

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Kristiaan PELCKMANS

May 2005

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

PRIMAL-DUAL KERNEL MACHINES

Jury:
Prof. G. De Roeck, voorzitter
Prof. J. Suykens, promotor
Prof. B. De Moor, promotor
Prof. J. Vandewalle
Prof. P. Van Dooren (UCL)
Prof. J. Schoukens (VUB)
Prof. M. Hubert
Prof. M. Pontil (UC London)

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Kristiaan PELCKMANS

U.D.C. 681.3*G12

May 2005

©Katholieke Universiteit Leuven – Faculteit Ingenieurswetenschappen
Arenbergkasteel, B-3001 Heverlee (Belgium)

ISBN 90-5682-610-7

U.D.C. 681.3*I26

D/2005/7515/44

Voorwoord

Ruim vier jaar van onderzoek zijn uiteindelijk samengebald in het huidige werkstuk. Ik geloof dat ik met tevredenheid terug kan kijken op deze jaren van wetenschappelijke exploratie en persoonlijke evolutie. Deze periode heeft me in contact gebracht met vele nieuwe gezichten, en heeft academische zowel als industriële waarheden en waarden bijgebracht. Dit is dan ook een uitstekend moment om mijn wetenschappelijke wortels en persoonlijke ankerpunten te bedanken.

Vooreerst wil ik deze gelegenheid aangrijpen om de mensen te bedanken die mij de kansen gaven dit onderzoek te realizeren. Graag wil ik professor Bart De Moor en professor Joos Vandewalle bedanken voor de talrijke kansen die ze me hebben geboden. Bedankt Joos om mijn mogelijkheden in een zo vroeg stadium te erkennen en me binnen te loodsen in deze academische wereld van ideeën en uitvindingen. Bart, ik wil u graag bedanken voor de nadruk die je bent blijven leggen op de reële waarde van toepassingen en werkbaarheid van onderzoek.

Bovenal wil ik professor Johan Suykens bedanken die de missie heeft volbracht om mijn enthousiasme te stroomlijnen in de vorm van wetenschappelijke output. Johan, je toewijding en bezorgdheid voor je onderzoekers zou een voorrecht moeten zijn voor elke doctoraatsstudent.

De assessoren van het leescommité wil ik graag danken voor hun constructieve kritiek voor het verbeteren van de tekst. Professor Johan Schoukens ben ik erg erkentelijk voor de wetenschappelijke discussies tijdens de vele IUAP bijeenkomsten en conferenties. Verder kan ik zijn hulp betreffende de thesis tekst erg waarderen en kan ik zeggen dat zijn opmerkingen zeker mee gedragen hebben tot de "finishing touch" van dit werk. Professor Paul Van Dooren wil ik graag bedanken voor het grondig nalezen van het proefschrift.

Onderzoek zit vaak niet vervat in kant en klare antwoorden, maar in kruisbestuivingen tussen experten en andere praatjes aan de koffietafel. In die zin kan ik het belang van mijn bureaugenoten niet genoeg benadrukken. Luc, bedankt voor je lakonieke vriendschap, Jos, voor je geduldige meesterschap, Ivan, voor je relativerende en visionaire uitlatingen, Bart, voor je impulsieve idealisme, Tony, voor je nauwgezette berekeningen en inleiding in de praktijk van onderzoeker. Lieven, bedankt voor je stille aanwezigheid en vele suggesties. Marcello, Jairo and Nathalie, thanks for the cooperations! Maarten, Mustak, Sven, Dries, Oscar, Cynthia, Bert, Bert Raf en Tom wil ik graag bedanken voor hun suggesties en af en toe een frisse babbel. Steven en anderen, hoed af voor jullie vrijwillige investering in het ondersteunen van de SISTA frigo's.

Hoe kan ik eraan beginnen om mijn ouders hun steun en toeverlaat op een waardige manier te erkennen? Ik hoop dat ik ooit hetzelfde kan doen als jullie hebben gedaan. Simon, Sara, An, Werner, Bertje en Wardje, bedankt hé!

Graag wil ik deze thesis opdragen aan mijn vriendin: Boke, ik apprecieer van harte je geduld en bezorgdheid. Dit proefschrift moet gewoon gekleurd zijn door je frisse alternatieve kijk op de zaken!

Kristiaan Pelckmans
31 mei 2005

Abstract

This text presents a structured overview of recent advances in the research on machine learning and kernel machines. The general objective is the formulation and study of a broad methodology assisting the user in making decisions and predictions based on collections of observations in a number of complex tasks. The research issues are directly motivated by a number of questions of direct concern to the user. The proposed approaches are mainly studied in the context of convex optimization.

The two main messages of the dissertation can be summarized as follows. At first the structure of the text reflects the observation that the problem of designing a good machine learning problem is intertwined with the question of regularization and kernel design. Those three different issues cannot be considered independently, and their relation can be studied consistently using tools of optimization theory. Furthermore, the problem of automatic model selection fused with model training is approached from an optimization point of view. It is argued that the joint problem can be written as an hierarchical programming problem which contrasts with other approaches of multi-objective programming problems. This viewpoint results in a number of formulations where one performs model training and model selection at the same time by solving a (convex) programming problem. We refer to such formulations as to fusion of training and model selection. Its relation to the use of appropriate regularization schemes is disccussed extensively.

Secondly, the thesis argues that the use of the primal-dual argument which originates from the theory on convex optimization constitutes a powerfull building block for designing appropriate kernel machines. This statement is largely motivated by the elaboration of new leaning machines incorporating prior knowlege known from the problem under study. Structure as additive models, semi-parameteric models, model symmetries and noise coloring schemes turn out to be related closely to the design of the kernel. Prior knowledge in the form of pointwise inequalities, occurence of known censoring mechanisms and a known noise level can be incorporated into an appriate learning machine easily using the primal-dual argument. This approach is related and contrasted to other commonly encountered techniques as smoothing splines, Guassian processes, wavelet methods and others. A related important step is the definition and study of the relevance of the measure of maximal variation which can be used to obtain an efficient way for detecting structure in the data and handling missing values.

The text is glued together to a consistent story by the addition of new results, including the formulation of new learning machines (e.g. the Support Vector Tube), study of new advanced regularization schemes (e.g. alternative least squares), investigation of the relation of the kernel design with model formulations and results in signal-processing and system identification (e.g. the relation of kernels with Fourier and wavelet decompositions). This results in a data-driven way to design an appropriate kernel for the learning machine based on the correlation measured in the data.

Korte Inhoud

Dit proefschrift presenteert een breed overzicht van nieuwe bijdragen in het onderzoek naar automatische leeralgoritmen. Het algemeen opzet is de formulering en de studie van een methodologie voor het assisteren van de expert in het maken van gefundeerde beslissingen of voorspellingen. Hoewel deze studie generiek van aard is en er academische problemen zullen bestudeerd worden, is de praktische relevantie van de gebruikte methode eerder aangetoond op verscheidene gevallenstudies. De kritische problemen die ervaren werden in dergelijke studies motiveerden de keuze van de onderzoeksonderwerpen. De aanpak is essentieel geworteld in een context van convexe optimalisatie.

Het proefschrift bestudeert en motiveert in hoofdzaak twee stellingen. Ten eerste wordt er geargumenteerd dat het probleem van het opstellen van een goed leeralgoritme, de vraag naar een goede maat van modelcomplexiteit en het ontwerp van een goede maat van similariteit in de vorm van een zogenaamde kernfunctie sterk gerelateerd zijn. De invalshoek van optimalisatie vormt een krachtig hupmiddel om de onderliggende relaties te bestuderen en constructief te gebruiken. Verder wordt het probleem van modelselectie dieper bestudeerd, eveneens vanuit een optimalisatieperspectief. Het modelselectieprobleem wordt geïnterpreteerd als een hiërarchisch programmeerprobleem. Dit laatste vormt een techniek voor het oplossen van optimalisatieproblemen waar meerdere kostfuncties moeten in rekening gebracht worden. Verschillende modelselectieproblemen worden dan geformuleerd als een optimalisatieprobleem en efficiënte manieren worden onderzocht om de taak van modelschatting en modelselectie tegelijkertijd op te lossen met betrekking tot verschillende deeltaken.

Ten tweede wordt er geargumenteerd dat het primair-duale raamwerk zoals bekend vanuit convexe optimalisatieproblemen een krachtige bouwblok vormt voor het formuleren van nieuwe leeralgoritmen. Deze bewering wordt gestaafd door het uitwerken van verschillende leermachines voor complexe taken. Het inbrengen van voorkennis met betrekking tot structuur en globale parameters in het leeralgoritme is in het bijzonder een sterkte van de methode. We bestuderen voornamelijk enerzijds de structuur van additieve modellen, gedeeltelijk parametrische kernfunctie methoden, het opleggen van modelsymmetrieën, en anderzijds de relatie van deze drie met het ontwerp van een goede kernfunctie. Andere bestudeerde vormen van opgelegde voorkennis omvatten puntsgewijze ongelijkheden, toegepaste vormen van censureringsmechanismen, het behandelen van onvolledige observaties en het

inbrengen van voorkennis met betrekking tot het ruisniveau. Dit centrale primair-duale argument wordt gerelateerd en gecontrasteerd met andere bekende methoden uit de literatuur. Verder werd een belangrijke stap gezet voor het detecteren van structuur uit de observaties door het uitwerken en bestuderen van de maat van maximale variatie van een functie.

Het verhaal is samengebracht tot een consistent geheel door het toevoegen van een scala van nieuwe resultaten zoals het uitwerken van nieuwe leeralgoritmen, bijvoorbeeld voor het schatten van onzekerheden (Support Vector Tubes), de studie van nieuwe mechanismen voor complexiteitscontrole of regularisatie (zoals bijvoorbeeld de formulering van het alternatieve kleinste kwadraten probleem), en de verdere studie van de relatie tussen modelcomplexiteit, het ontwerp van de kernfunctie en resultaten vanuit de theorie van systeemidentificatie. In het bijzonder wordt er een methode voorgesteld voor het schatten van een goede kernfunctie uit de observaties gebaseerd op de berekende correlatie geschat op de gegeven dataset.

Primair-duale Kernfunctie Methoden

Vele problemen kunnen herleid worden tot het zoeken van geschikte mathematische modellen op basis van een verzameling observaties en het maken van voorspellingen op basis van deze modellen. Dit sleutelidee vormt een belangrijk ingrediënt van verschillende wetenschappelijke deelgebieden zoals statistiek, systeemidentificatie en artificiële intelligentie, en vindt een directe toepassing in een breed spectrum van praktische problemen gaande van medische overlevingsanalyse tot het regelen van complexe chemische processen. In het kielzog van de zogenaamde Support Vector Machines (SVMs) (Cortes and Vapnik, 1995; Vapnik, 1998) is een nieuwe sterke impuls gegeven aan het wetenschappelijk onderzoek naar algoritmen voor het automatisch leren met behulp van leermachines ("Machine Learning"). Deze nederlandstalige samenvatting van het proefschrift bevat twee delen. Het eerste bespreekt de algemene methodologie van SVMs en kernfunctie methoden op een inleidend niveau. Het tweede deel geeft hierop gesteund een overzicht van de bijdrage van het proefschrift.

Dit onderzoek richt zich vooral op het ontwerp en de analyse van leersystemen voor de automatische classificatie en het benaderen van functionele verbanden gegeven een eindige verzameling observaties. Deze klasse van problemen werd bekeken vanuit een nieuwe theoretische invalshoek bekend als de theorie van statistische leeralgoritmen (Vapnik, 1998; Bousquet *et al.*, 2004). Door de recente beschikbaarheid van mogelijkheden om grote berekeningen op een automatische manier uit te voeren en door de formulering van efficiënte numerieke algoritmen mag men spreken van een doorbraak van de kernel methoden zowel op theoretisch vlak als in de praktijk. De huidige tendens is om de klasse van kernelmethoden als een volwaardige aanvulling te zien op de klassieke statistische methodologie (Hastie *et al.*, 2001). De onderzoeksgroep SCD-SISTA en ondergetekende richtten zich de voorbije jaren op het bestuderen en toepassen van een variant, de kleinste kwadraten SVMs (LS-SVMs) (Suykens *et al.*, 2002*b*). Dit onderzoek onderscheidt zich voornamelijk van andere kernelgebaseerde methoden door het uitbuiten van expliciete verbanden met de theorie van convexe optimalisatie (Boyd en Vandenberge, 2004). Belangrijke elementen van de LS-SVMs zijn de resulterende algoritmen die eenvoudiger en sneller zijn dan de doorsnee SVM methoden, en de expliciete verbanden met methoden als neurale en regularizatie netwerken, wavelets en splines (voor de laatste zie b.v. (Wahba, 1990)). De praktische werkbaarheid van de algoritmen was de voorbije jaren bewezen onder meer in het veld van medische signaalverwerking, bioinformatica, econometrie en regeltoepassingen, zie (Suykens *et al.*, 2002*b*).

A. Introductie tot Machine Leeralgoritmen en Kernfuncties

A.1 Machine Leeralgoritmen

Het onderzoeksgebied van machine leeralgoritmen bevat het onderzoek naar hoe programma's te ontwerpen die verbeteren met de gegevens die ze opdoen (Mitchell, 1997). Zodoende is men geïnteresseerd in een automatisch formalisme of algoritme Alg dat gegevens $\mathscr{D}$ - bijvoorbeeld in de vorm van observaties van een bepaald fenomeen - en voorkennis van het probleem $\mathscr{A}$ (bijvoorbeeld in de vorm van assumpties over het bestudeerde fenomeen) omzetten in een expertsysteem in de vorm van wiskundige vergelijkingen. In het algemeen behoort het bekomen expert systeem tot een voorgedefinieerde klasse $\mathscr{F}$ van potentiële beschrijvingen die gedetermineerd zijn op enkele onbekende parameters na. Een leeralgoritme kan aldus formeel beschreven worden als een optimale afbeelding als

$$\mathsf{Alg} : \mathscr{D} \times \mathscr{A} \to \mathscr{F}.$$

Men refereert naar deze mapping ook als *inferentie*, *schatter* (in een statistische context), *leeralgoritme* (in een context van artificiële intelligentie). Hier beperken we ons tot de taak waarbij de observaties uiteenvallen in twee klassen, namelijk de bekende *invoer variabelen* en de overeenkomde *uitvoer onbekenden* of *uitvoer etiketten*. Het doel van het geleerde resultaat is dan om voorspellingen te doen van de uitvoer overeenkomende met nieuwe observaties van de invoer. In dit geval kan de klasse $\mathscr{F}$ van potentiële beschrijvingen f nauwkeuriger beschreven worden in termen van een aantal onbekende parameters $\theta \in \Theta$ als volgt

$$\mathscr{F} = \left\{ f : \mathbb{R}^D \to \mathbb{D} \;\middle|\; f(x,\theta) = y \right\},$$

waar $x \in \mathbb{R}^D$ een mogelijke invoer en $y \in \mathbb{D}$ een mogelijke uitvoer representeert. Details van de mapping Alg bepalen in grote mate de specificaties van het leeralgoritme in kwestie:

Afbeelding: Alg Door een leeralgoritme te beschrijven als een welgedefinieerde afbeelding van een set van observaties en een verzameling aannames op een klasse van mogelijke modellen wordt impliciet aangenomen dat het resultaat uniek is en worden globale optimalisatiemethoden (zoals dikwijls gebruikt in artificiële neurale netwerken) uitgesloten. Deze definitie maakt het mogelijk om begrippen als gevoeligheid van het algoritme aan kleine perturbaties op de observaties formeel te definiëren.

Optimaliteit: Het begrip optimaliteit staat centraal in deze definitie: elke gegeven dataset en verzameling veronderstellingen impliceert een resultaat dat het beste is onder alle mogelijke hypothesen. De gebruikte vorm van optimaliteit is in belangrijke mate bepaald door het uiteindelijke doel van het leeralgoritme (e.g. verklaring en inzicht, voorspelling, de observaties ontdoen van ruis,...). Optimaliteit wordt uitgedrukt in wiskundige symboliek die eigen is aan de exacte context van het leerprobleem (klassiek statistisch, Bayesiaans, deterministische benadering,...).

Gegevens $\mathscr{D}$**:** De observaties worden vaak verschaft in de volgende vorm

$$\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}, \tag{0.1}$$

met $x_i \in \mathbb{D}^D$ de input observaties en $y_i \in \mathbb{D}$ de overeenkomstige uitvoer observaties. De exacte vorm van het domein $\mathbb{D}$ van de variabelen bepaalt in grote mate de probleemstelling. Men maakt vaak een onderscheid tussen $\mathbb{D} = \mathbb{R}$ (continue onbekenden), $\mathbb{D} = \{-1, 1\}$ (binaire observaties), nominale variabelen (bv. $\mathbb{D} = \{Jazz, pop, classic\}$) of ordinale variabelen (bv. $\mathbb{D} = \{slecht, goed, super\}$). Bovendien kunnen observaties ontbreken (“missen”) of fout zijn omwille van verscheidene redenen.

Aannames $\mathscr{A}$**:** Veronderstellingen komen voor in verschillende vormen: kwalitatief (bijvoorbeeld het functioneel verband is strict stijgend), kwantitatief (bijvoorbeeld er is een signaal-ruis verhouding), een a-priori bekend probabilistisch model (bv. de ruis is normaal verdeeld) of in de vorm van latente kennis. In de laatste zitten alle eigenschappen en resultaten bevat met betrekking tot de probleemstelling zelf.

Klasse $\mathscr{F}$**:** Een belangrijke vorm van voorkennis met betrekking tot de probleemstelling wordt verwerkt in de preciese klasse van modellen (bijvoorbeeld welke gemeten variabelen zijn relevant voor het model). Bovendien legt de klasse van hypothesen dikwijls een inherente structuur op het leerproces. Men maakt bijvoorbeeld een onderscheid tussen oorzakelijke modellen (met een inherente tijdscomponent), of beslissingsbomen met een hiërarchische structuur. Verder is de klasse $\mathscr{F}$ van modellen vaak bepaald door de specifieke vorm van de uitvoervariabelen (bijvoorbeeld regressie voor continue uitvoer en classificatie voor binaire uitkomsten).

Analyse**:** Een uiteindelijke analyse van de resulterende modellen van het leeralgoritme zoekt een antwoord op de vraag of het geleerde verband inderdaad bruikbaar is. Hiervoor bestaan verschillende mogelijkheden. In eerste instantie kan men de veralgemeningsperformantie (“generalisatie performantie”) van de schatting evalueren met een toepasselijk model selectie criterium. Een voorbeeld hiervan is om het geleerde model te gebruiken voor het voorspellen van de uitvoer van nieuwe observaties in een validatiefase. Een meer theoretische aanpak kan gebasseerd worden op een mate van gevoeligheid van het leeralgoritme aan kleine perturbaties in de data of de aannames.

A.2 Support Vector Machines en Kernfuncties

We beschouwen op dit ogenblik het specifieke geval waar de uitgang een binaire waarde (-1 of 1) aanneemt. Dit geval van classificatie wordt dikwijls beschouwd als een van de minst complexe maar meest generieke taken en verdiende zodoende een groot deel van de interesse in het wetenschappelijk onderzoek van leertechnieken.

Probleemstelling

De methode van Support Vector Machines (SVMs) stamt uit het onderzoek naar het induceren van een goede binaire classificatie regel uit een eindige verzameling observaties. Concreet zoekt men een regel $c : \mathbb{R}^D \rightarrow \{-1, 1\}$ die het verwachte etiket behorende bij toekomstige datapunten voorspelt. Laat de observaties samples zijn van de random variabele X en Y overeenkomstig de in- en uitvoer variabelen. Gegeven een vaste maar onbekende distributie P_{XY} over de random variabele X en Y, de optimale classificatie regel c met minimaal risico op verkeerde voorspellingen kan geformaliseerd worden als

$$\hat{c} = \operatorname*{arg\,min}_{c:\mathbb{R}^D \rightarrow \{-1,1\}} \int I(y \neq c(x)) dP_{XY}(xy),$$

waar de indicator functie $I(x \neq y)$ gelijk is aan 1 als $x \neq y$ en aan nul in het andere geval.

Support Vector Machines

We beschouwen classificatieregels van de volgende vorm

$$\operatorname{sign}\left[w^T \varphi(x) + b\right].$$

Hierbij is $\varphi : \mathbb{R}^D \rightarrow \mathbb{R}^{D_\varphi}$ een afbeelding van de gegevens met dimensie $D \in \mathbb{N}$ naar een kenmerkruimte D_φ met mogelijk oneindige dimensie ($D_\varphi = +\infty$), $w \in \mathbb{R}^{D_\varphi}$ is een parameter vector en $b \in \mathbb{R}$ een constante. Anders gesteld, men voorspelt een positief of een negatief etiket bij een nieuwe invoer $x_* \in \mathbb{R}^D$ afhankelijk aan welke kant dit punt zich bevindt ten opzichte van het hypervlak Hp gegeven als volgt

$$\operatorname{Hp}(w, b) = \left\{x_0 \in \mathbb{R}^D \mid w^T \varphi(x_0) + b = 0\right\}.$$

Het is een klassiek resultaat dat de afstand van een punt x_i tot het hypervlak $\operatorname{Hp}(w, b)$ begrensd wordt als volgt

$$d_i = \frac{\left|w^T \varphi(x_i) + b\right|}{w^T w} \geq \frac{y_i \left(w^T \varphi(x_i) + b\right)}{w^T w}, \;\; \forall i = 1, \ldots, N.$$

Resultaten in het domein van statistische machine leeralgoritmen geven dan garanties dat het hypervlak Hp goede resultaten levert indien de observaties op maximale afstand liggen van het hypervlak. Het optimale hypervlak wordt gegeven als de oplossing van het volgende optimalisatieprobleem

$$\max_{w,b,d} d \quad \text{s.t.} \quad \frac{y_i \left(w^T \varphi(x_i) + b\right)}{w^T w} \geq d, \ \ \forall i = 1, \ldots, N.$$

Dit probleem kan herschreven worden door d te vervangen door $1/w^T w$ wat altijd kan gedaan worden (de locatie van het hypervlak is niet afhankelijk van zijn norm)

$$\min_{w,b} \mathscr{J}(w) = w^T w \quad \text{s.t.} \quad y_i \left(w^T \varphi(x_i) + b\right) \geq 1, \ \ \forall i = 1, \ldots, N.$$

Dit probleem is convex en heeft zodoende slechts één globaal minimum. Indien de afbeelding φ bekend is kan bovenstaand optimalisatieprobleem efficiënt opgelost worden.

We bekijken nu het geval dat de afbeelding φ niet bekend is maar enkel de overeenkomstige kernfunctie gedefinieerd als

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \quad \forall x_i, x_j \in \mathbb{R}^D.$$

Het Mercer theorema stelt dan dat onder bepaalde voorwaarden op K (K is een positief definiete functie) er een unieke overeenkomstige afbeelding φ bestaat. Vaak kan het schattingsprobleem herschreven worden in functie van de kernel zodat de afbeelding φ impliciet kan blijven in de berekening. Dit biedt concrete voordelen indien enkel iets geweten is over het globale verloop van de functie (bijvoorbeeld “de functie is traag variërend”) en men niet zozeer de expliciete parametrische vorm kan neerschrijven.

Een mogelijk pad om dergelijke problemen te herschrijven in functie van de kernfunctie K is gegeven door resultaten in de theorie van convexe optimalisatie (Boyd en Vandenberge, 2004). Beschouw de zadelpuntbeschrijving van het probleem die bekomen wordt door het opstellen van de Lagrangiaan met Lagrange vermenigvuldigers α_i voor $i = 1, \ldots, N$

$$\max_{\alpha} \min_{w,b} \mathscr{L}(w, b; \alpha) = w^T w - \sum_{i=1}^{N} \alpha_i \left(y_i \left(w^T \varphi(x_i)\right) - 1\right),$$

met beperking dat $\alpha_i \geq 0$ voor alle $i = 1, \ldots, N$. Het minimum met betrekking tot de zogenoemde primaire variabelen w en b wordt gegeven door de volgende voorwaarden:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial w} = 0 \rightarrow & w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i) \\ \dfrac{\partial \mathscr{L}}{\partial b} = 0 \rightarrow & \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}$$

Laat de vector $Y \in \mathbb{R}^N$ gedefinieerd zijn als volgt $Y = (y_1, \ldots, y_N)^T$ en laat de matrix $\Omega_Y \in \mathbb{R}^{N \times N}$ voldaan zijn aan $\Omega_{Y,ij} = y_i y_j K(x_i, x_j)$ voor alle $i, j = 1, \ldots, N$. Laat 1_N

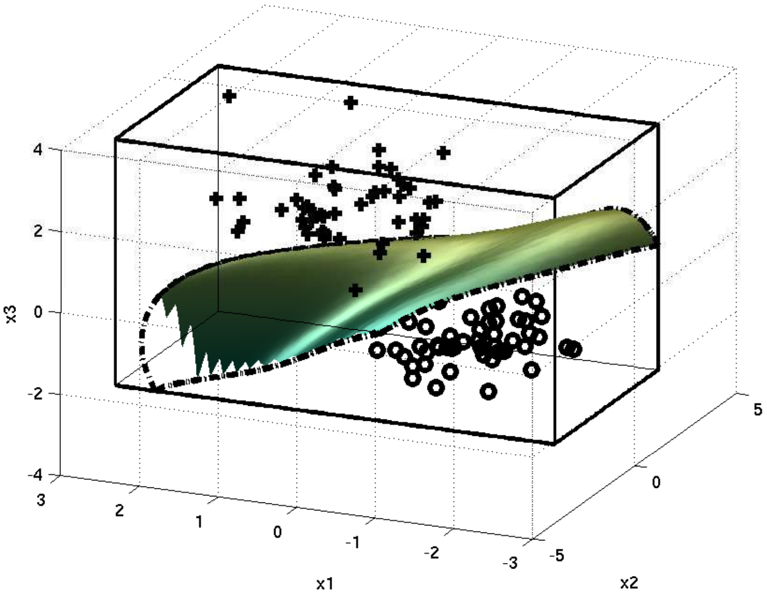


Figure 0.1: *Voorbeeld van een classificatieprobleem en het model bekomen door toepassing van een SVM. Positieve ("+") en negatieve ("o") observaties zijn gegroepeerd in twee verschillende klassen. De Support Vector Machine genereert een model (voorgesteld als het hellende vlak) dat de beslissing maakt of een nieuw datapunt meest waarschijnlijk een voorbeeld is van de klasse van positieve (boven het vlak) of negatieve samples (onder het vlak).*

gedefinieerd zijn als de vector $1_N = (1,\ldots,1)^T \in \mathbb{R}^N$. Gebruik makende van deze voorwaarden om dan de primaire variabelen te elimineren uit de zadelpuntformulering resulteert in het volgende duale probleem

$$\max_{\alpha} \mathscr{J}^D(\alpha) = \frac{-1}{2}\alpha^T\Omega_y\alpha + 1_N^T\alpha \quad \text{s.t.} \quad \begin{cases} Y^T\alpha = 0 \\ \alpha_i \geq 0 \qquad \forall i = 1,\ldots,N, \end{cases}$$

dat uitgedrukt wordt in termen van de duale vermenigvuldigers $\alpha = (\alpha_1,\ldots,\alpha_N)^T \in \mathbb{R}^N$. Door een verdere technische ingreep (het uitbuiten van de zogenaamde complimentariteitsvoorwaarden in de Karush-Kuhn-Tucker condities voor optimaliteit) kan uit het beschreven duale probleem niet alleen de vector α geschat worden, maar ook de impliciet overeenkomstige schatting van b kan gevonden worden. Eens zowel $\hat{\alpha}$ als $\hat{b}$ berekend is, kan het impliciet geschatte model geëvalueerd worden in een nieuw datapunt $x_* \in \mathbb{R}^D$ als volgt

$$\text{sign}\left[\sum_{i=1}^{N} \hat{\alpha}_i y_i K(x_i, x_*) + \hat{b}\right].$$

Afgeleide resultaten relaxeren dan de maximale marge door toe te laten dat de gevonden marge geschonden wordt door enkele observaties. Verdere uitbreidingen bestuderen gelijkaardige formuleringen waar de uitvoer continue of ordinale waarden kan aannemen.

Uitbreidingen

Deze aanpak heeft zijn kracht bewezen zowel op theoretisch als op praktisch vlak (see e.g. (Schölkopf and Smola, 2002)). Er resteren echter nog een verzameling pijnpunten waaronder de volgende: “Welke afweging tussen fit en modelcomplexiteit moet er gemaakt worden?”, “Wat is de specifieke vorm en tunings parameter van de kernfunctie die optimaal is voor de taak?”, of “Hoe kan men uit de observaties afleiden welke invoervariabelen relevant zijn voor de taak?”. Deze vragen zijn allen een specifieke vorm van het probleem van modelselectie. Op deze vraagstukken zal een antwoord worden geformuleerd in het tweede en derde deel van het proefschrift.

Een uitgebreid deel van het onderzoek naar kernfunctie gebaseerde leeralgoritmen richt zich op het formuleren van leermethodes voor het automatisch bouwen van modellen voor complexere taken. Niet alleen classificatie, maar ook het schatten van continue functionele verbanden uit de gegevens is een belangrijke taak voor leeralgoritmen. In geval de data expliciete tijdsafhankelijkheden vertoont verschuift de focus meer naar het onderzoeksgebied van systeemidentificatie. Dit blijkt een vruchtbaar gebied te zijn voor het gebruik van leermachines die structurele vereisten kunnen incalculeren. In het algemeen is het inbrengen van extra voorkennis in het leeralgoritme zelf niet alleen een belangrijk desideratum, maar worden ook verkeerde schattingen vermeden op die manier.

Andere vragen gerelateerd aan de formulering van SVMs en primair-duale kernfunctie methoden hebben betrekking tot hoe men efficiënt de optimale oplossing kan berekenen bijvoorbeeld voor grote datasets. Een andere tak van het onderzoek naar kernfunctie gebaseerde leeralgoritmen richt de focus op het iteratief bijwerken van het geschatte model overeenkomend met nieuwe observaties die men toekrijgt. Een veelbelovend onderzoek richt zich dan op het ontwikkelen van snelle hardware implementaties van het schattingsprobleem.

B. Bijdragen van het Doctoraatswerk

Het huidige doctoraatswerk beschrijft een verzameling nieuwe resultaten in het onderzoek naar automatische leeralgoritmen en kernfunctie methoden. Dit biedt een uniforme kijk op het onderzoek door volgende regels centraal te stellen:

Convexe Optimalisatie: Dit onderzoek in het verlengde van de methode van SVMs vertoont enkele grote verschillen met het klassiekere onderzoek naar artificiële neurale netwerken. Naast de stevige theoretische fundering springt vooral de eigenschap van globale optimaliteit in het oog. De eigenschap dat de optimale schattingen uniek is heeft als resultaat dat herhaling van een experiment gegarandeerd tot dezelfde oplossing zal leiden. Dit resulteert in de mogelijkheid om stevige theoretische analyzes te binden aan de optimale schattingen. De uitdaging om nieuwe formuleringen van niet-lineaire technieken te herformuleren als een standaard convex programmeringsprobleem vormt een rode draad doorheen het onderzoek.

Opleggen van voorkennis: In vele toepassingen bezit men niet alleen observaties om een model te bouwen maar heeft men ook voorkennis betreffende het bestudeerde fenomeen ter beschikking. Een goed leeralgoritme moet zo mogelijk rekening houden met die voorkennis zodat het resulteert in modellen die voldoen aan die voorkennis. Een belangrijke vorm om voorkennis op te leggen aan het leeralgoritme is om een specifieke model structuur voorop te stellen.

Modelselectie: Dikwijls is het resultaat van het leeralgoritme bepaald op enkele ontwerpparameters na. Een veel voorkomende parameter kwantificeert het ruisniveau van de observaties. Indien de exacte waarde van deze ontwerpparameter niet expliciet bekend is, kan men specifieke methoden gebruiken om deze waarden te leren uit de observaties. Ondanks het uitgebreide onderzoek naar mogelijke criteria die de kwaliteit bepalen van een specifieke ontwerpparameter, is de automatisatie van dit metaprobleem in vele gevallen een open probleem. Deze thesis bestudeert een dergelijk formalisme voor het automatisch uitvoeren van modelselectietaken door het formuleren van hiëarchische programmeringsproblemen.

Dit overzicht volgt in grote trekken de structuur van de tekst en legt de kernpunten van de vier delen bloot.

Hoofdstuk 1: Problemen en Doelstellingen

Dit hoofdstuk legt op een formele manier de achtergrond van het onderzoek vast zoals gegeven in Hoofdstuk A.1. Verder wordt de techniek van SVMs en LS-SVMs gerelateerd aan klassieke methoden als bekend vanuit statistiek en andere wetenschappelijke domeinen. Een groot deel van het eerste hoofdstuk is gewijd aan een overzicht van de verschillende onderzoeksdisciplines binnen het onderzoek van automatische leeralgoritmen en kernfunctie modellen.

Hoofdstuk 2: Overzicht van de Theorie van Convexe Optimalisatie

Zoals reeds geargumenteerd wordt de theorie en praktijk van convexe optimalisatie centraal gesteld in dit onderzoek: het primair-duale argument dat de hoeksteen vormt van vele uitgewerkte resultaten heeft een duidelijke afkomst in optimalisatietheorie. Daartoe is er ruime aandacht besteed om een overzicht te geven van deze theorie voor zover relevant voor dit onderzoek. Een convex programmeringsprobleem heeft de volgende vorm.

Definition 0.1. [Convex Programmeringsprobleem] *Laat $m,p \in \mathbb{N}$ en laat $b_i \in \mathbb{R}$ voor alle $i = 1,\ldots,m,\ldots,m+p$. Een wiskundig optimalisatieprobleem heeft in het algemeen de volgende vorm*

$$p^* = \min_{x\in\mathbb{R}^D} f_0(x) \quad s.t. \quad \begin{cases} f_i(x) \leq b_i & \forall i = 1,\ldots,m \\ f_j(x) = b_j & \forall j = m+1,\ldots,m+p, \end{cases} \tag{0.2}$$

waar $f_k : R^D \to \mathbb{R}$ functies voorstellen voor alle $k = 0,\ldots,m+p$. Men refereert naar f_0 als de objectieffunctie die geminimaliseerd dient te worden, f_i voor alle $i = 1,\ldots,m$ en f_j voor alle $j = m+1,\ldots,m+p$ stellen dan de functies van de ongelijkheids- en de gelijkheidsbeperkingen voor. De vector $(b_1,\ldots,b_m,\ldots,b_{m+p})^T \in \mathbb{R}^{m+p}$ representeert de begrenzingen van de beperkingen. Een optimalisatieprobleem is convex *indien de punten die voldoen aan de beperkingen convex zijn (i.e. elke lineaire interpolatie van twee oplossingen is opnieuw een oplossing) en de objectieffunctie convex is (i.e. elke lineaire interpollatie van twee punten behorende tot de objectieffunctie is groter dan of gelijk aan het overeenkomstige punt op de objectieffunctie).*

Optimalisatieproblemen met verschillende kostenfuncties worden traditioneel aangepakt door de verschillende objectieffuncties om te vormen tot één enkele globale kostenfunctie en deze dan te optimaliseren. In verschillende gevallen is een dergelijke aanpak niet direct toepasbaar, bijvoorbeeld omdat de verschillende objectieffuncties op een verschillend niveau staan. Dit proefschrift bestudeert een andere techniek om dergelijke problemen te beschrijven via hiärchisch programmeren.

Definition 0.2. [Hiërarchische Programmeringsproblemen] *Beschouw twee objectieffuncties f_0^1 en f_0^2 en bijbehorende beperkingen f_i^1 en f_j^2 allen gedefinieerd op dezelfde onbekende van gelijke dimensie ($\mathbb{R}^D$). Indien $\Gamma \subset \mathbb{R}^D$ de globale oplossingsruimte is van het eerste probleem f_0^1 en f_i^1 op enkele parameters na waarvan de waarden vast gehouden worden (ontwerpparameters), dan bekomt men een hiërarchische aanpak indien men op een tweede niveau het tweede probleem f_0^2 en f_j^2 beperkt tot de oplossingsruimte Γ.*

Dit wordt schematisch geïlusstreerd in Figuur 0.2.

Deel α

Dit hoofdstuk is in grote mate gewijd aan de afleiding van de resultaten die reeds in het kort beschreven zijn in Subsectie A.2. In aanvulling hiertoe wordt het primair-duale argument gebruikt om gelijkaardige leermachines te formuleren. Vooreerst wordt een eenvoudig geval bestudeerd. Stel dat de data de vorm aannemen $\mathscr{D} = \{(x_i,y_i)\}_{i=1}^N$ met

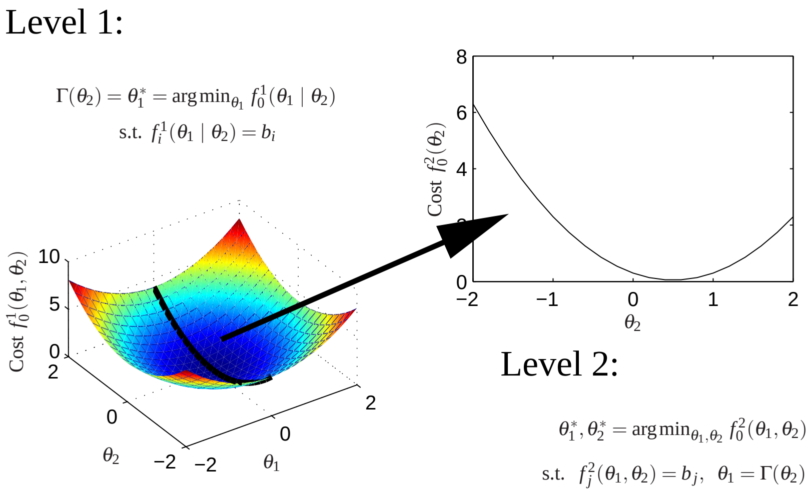

Level 1:
$\Gamma(\theta_2) = \theta_1^* = \arg\min_{\theta_1} f_0^1(\theta_1 \mid \theta_2)$
s.t. $f_i^1(\theta_1 \mid \theta_2) = b_i$
Cost $f_0^1(\theta_1, \theta_2)$
10
5
0
2
0
θ_2
−2
−2
0
2
θ_1
8
6
4
0
Cost $f_0^2(\theta_2)$
−2
−1
0
1
2
θ_2
Level 2:
$\theta_1^*, \theta_2^* = \arg\min_{\theta_1, \theta_2} f_0^2(\theta_1, \theta_2)$
s.t. $f_j^2(\theta_1, \theta_2) = b_j, \;\; \theta_1 = \Gamma(\theta_2)$

Figure 0.2: *Schematische voorstelling van een hiërarchisch programmeringsprobleem. Laat f_0^1, f_i^1 en f_0^2, f_j^2 de twee objectieffuncties met bijbehorende beperkingen zijn. Beiden werken op een parameterruimte in $\mathbb{R}^2$ met parameters $\theta_1 \in \mathbb{R}$ en $\theta_2 \in \mathbb{R}$. Op het eerste niveau wordt θ_2 vast gehouden en geoptimaliseerd over θ_2 d.m.v. de functies f_0^1 en f_i^1. Voor elke waarde θ_2 bestaat er dan een unieke oplossing indien het probleem convex is, voorgesteld door $\Gamma(\theta_2) = \theta_1^*$. Op een tweede niveau wordt er dan geoptimaliseerd over deze parameter-ruimte $\{(\theta_1, \theta_2) \mid \Gamma(\theta_2) = \theta_1\}$ met behulp van de kostenfunctie f_0^2 en eventuele beperkingen f_j^2.*

$x \in \mathbb{R}^D$ en $y_i \in \mathbb{R}$ continu, en stel dat het model kan geschreven worden als $f(x) = w^T x$ met onbekende parameter vector $w \in \mathbb{R}^D$. Laat de matrix $X \in \mathbb{R}^{N \times D}$ en de vector $Y \in \mathbb{R}^N$ gedefinieerd zijn als $X = (x_1, \ldots, x_N)^T$ en $Y = (y_1, \ldots, y_N)^T$. De klassieke methode van kleinste kwadraten om dan de onbekende parameters te zoeken gegeven de observaties $\mathscr{D}$ is dan om de volgende kostenfunctie te minimaliseren:

$$\hat{w} = \arg\min_{w} \mathscr{J}(w) = \frac{\gamma}{2} \sum_{i=1}^{N} \left(w^T x_i - y_i\right)^2 .$$

De oplossing kan analytisch berekend worden door oplossing van het stelsel lineaire vergelijkingen

$$\left(X^T X\right) \, w = X^T Y.$$

Deze tekst beschouwt complexere vormen van zulke formuleringen die de model formulering uitbreidt naar niet-lineaire impliciete voorstellingen door het gebruik van het primair-duale argument zoals gebruikt in Sectie A.

Een reeks primair-duale kernfunctie machines wordt afgeleid, elk met een verschillende kostenfunctie. De volgende afleidingen worden gegeven voor het geval van regressie:

- [SVM] De standaard SVM voor regressie wordt bekomen door het aannemen van een kostenfunktie van de volgende vorm

  $$\ell_\varepsilon(e) = \max\left(0, |e| - \varepsilon\right).$$

- [LS-SVM] Door het beschouwen van een kleinste kwadraten kostenfunctie bekomt men een variant van de SVM die efficiënt kan berekend worden door het oplossen van een verzameling lineaire vergelijkingen. Een ander voordeel van deze formulering is zijn sterke relatie met de theorie van splines en Gaussiaanse processen en de interpretatie van de oplossing als een convolutie van de ruis met de gegeven kernfunctie.

- [hSVM] Integratie van de Huber-kostenfunctie resulteert in een formulering die het midden houdt tussen de twee voorgaande formuleringen. De klassieke motivatie van de Huber-kostenfunctie als een methode voor het bekomen van schattingen ongevoelig (“robust”) voor a-tyische observaties vormt een surplus.

- [SVT] De Support Vector Tube (SVT) is geformuleerd vanuit een andere doelstelling. Deze associeert met elke gegeven invoerobservatie een interval van de reële getallen waarin het gross van de mogelijke overeenkomstige uivoerobservaties mag verwacht worden. De SVT construeert een minimaal complexe begrenzing (“tube”) waar alle observaties in passen.

- [ν-SVT] Deze kernfunctie machine is een uitbreiding van de SVT waarin uitzonderingen worden toegelaten: in uitzonderlijke gevallen kunnen gegeven observaties buiten de tube toegelaten worden. De parameter ν geeft dan een indicatie hoeveel uitzonderingen toegelaten worden.

In het geval van classifictie worden de standaard SVM en LS-SVM classificator besproken.

In vele gevallen is het mogelijk voorkennis in de vorm van gekende structuur uit te buiten in het leeralgoritme. De volgende gevallen zijn uitgewerkt:

- [Semi-parametrische structuur] Het geschatte model kan mogelijk een vermenging zijn van een lineair deel met overeenkomstige parameters en een niet-parametrisch deel gesteund op kernfuncties. Laat elke observatie x bestaan uit een deel $x^P \in \mathbb{R}^d$ gebruikt voor het parameterisch model (met parameters $\beta \in \mathbb{R}^d$) en een deel $x^K \in \mathbb{R}^D$ voor het niet-parametrisch stuk f_K als volgt

  $$f(x) = f_K\left(x^K\right) + \beta^T x^P.$$

  De schatting van dit soort modellen kan efficiënt gebeuren gebruik makende van het primair-duale argument.

- [Additive Models] Het gebruik van additieve modellen levert vaak een praktisch evenwicht tussen een interpreteerbaar resultaat en een flexibele modelstructuur. Laat elke observatie x bestaan uit verschillende componenten $x^{(p)}$ met $p =$

$1,\ldots,P$. In vele gevallen geven modellen van de volgende vorm een accurate benadering van het bestudeerde fenomeen:

$$f(x) = \sum_{p=1}^{P} f_p\left(x^{(p)}\right) + b,$$

met f_p een serie van deelfuncties telkens gebaseerd op de overeenkomende componenten. Een additioneel voordeel van deze model structuur is dat theoretische resultaten aantonen dat schatting van deze modellen nauwkeuriger (in welbepaalde zin, zie later) kan gebeuren.

- [Puntsgewijze ongelijkheden] Vaak zijn er kwalitatieve regels in de vorm van ongelijkheden voorhanden waaraan de geschatte modellen moeten voldoen. Indien deze ongelijkheden geformuleerd kunnen worden in termen van een aantal concrete punten, kan het primair-duale argument gebruikt worden om een overeenkomstig leeralgoritme te bouwen.

- [Gecensureerde observaties] In bepaalde gevallen zijn de observaties gecensureerd. Bijvoorbeeld een meter kan maar tot een bepaalde waarde uitgelezen worden door technische beperkingen. De kostenfunctie kan overeenkomstig hiermee aangepast worden wat leidt tot een nieuwe kernfunctie methode.

Het laatste hoofdstuk van dit deel beschrijft dan het verband van de beschreven methodologie met de klassieke resultaten splines in de context van ruizige observaties, Gaussiaanse processen en Bayesiaanse technieken, wavelets, inverse problemen, vealgemeende kleinste kwadraten methoden en andere methoden.

Deel γ

Het tweede deel focust zich op de computationele aspecten van de gebruikte vorm van complexiteitscontrole of regularisatie. In eerste instantie worden verschillende vormen van complexiteitscontrole beschreven. We maken een onderscheid tussen parametrische modellen waar complexiteit uitgedrukt kan worden in termen van de norm van de parameters, en niet-parameterische kernfunctie methoden waar een maat van complexiteit bijvoorbeeld kan uitgedrukt worden in de maximale variatie die een functie vertoont op de gegeven dataset. In het eerste geval gebruikt men meestal de 2-norm van de parameter vector (“ridge regression”). Het volgende voorbeeld is klassiek. Beschouw opnieuw de lineaire model structuur $f(x) = w^T x$. We bestuderen de kostenfunctie

$$\hat{w} = \arg\min_{w} \mathscr{J}_\gamma(w) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} \left(w^T x_i - y_i\right)^2,$$

waar de ontwerpparameter $\gamma \geq 0$ de afweging bepaalt tussen de complexiteitsterm $w^T w$ en de empirische kost $\sum_{i=1}^{N} \left(w^T x_i - y_i\right)^2$. De optimale schatting kan analytisch berekend worden door oplossing van het stelsel lineaire vergelijkingen

$$\left(X^T X + \frac{1}{\gamma} I_D\right) w = X^T Y,$$

waar $I_D \in \mathbb{R}^{D\times D}$ de eenheidsmatrix voorstelt. Een analyse in de vorm van de evolutie van de bias (verwachtte afwijking van de echte functie) en variantie (onzekerheid op de geschatte functie) in functie van de ontwerpparameter γ is gegeven in de literatuur voor deze lineaire schatter. Deze tekst geeft een gelijkaardige afleiding voor de LS-SVM schatter in de vorm van bias en variantie. Verder is de relatie van deze ontwerpparameter met de signaal-ruis verhouding uitgewerkt door het bestuderen van gerelateerde regularisatieschemas genaamd Ivanov en Morozov regularisatie.

Huidige aandacht gaat meer en meer naar het gebruik van de 1-norm daar deze resulteert in oplossingen waar vele waarden nul zijn (spaarsheid van de parameters). Dit voorkomen van nullen in de oplossingsvector in het lineaire geval wordt geïnterpreteerd als een vorm van selectie van invoervariabelen. In het geval van niet-parametrische kernfunctie methoden voor additieve modellen stellen we het gebruik van de maat van maximale variatie voor. De componenten met een bijbehorende maximale variatie van nul duiden aan dat deze componenten niet wezenlijk bijdragen tot het geleerde model. Zodoende is er een niet-parametrische vorm van structuurdetectie bekomen. Verdere toepassingen van het principe van maximale variatie is bekomen in de context van het behandelen van missende waarden in de observaties.

Hoofdstukken 7 en 8 beschouwen het probleem van selectie van een optimale ontwerpparameter die een afweging maakt tussen complexiteit en empirische performantie (typisch genoteerd door een Griekse γ). Hiervoor worden modelselectiecriteria beschouwd als validatie, kruis-validatie en anderen. Beschouw bijvoorbeeld opnieuw het lineaire probleem zoals in vorige paragraaf, optimaliseren van de ontwerpparameter γ met betrekking tot de performantie op een validatiedataset $\mathscr{D}^v = \left\{\left(x_j^v, y_j^v\right)\right\}_{j=1}^n$ (met $x_j^v \in \mathbb{R}^D$ en $y_j^v \in \mathbb{R}$) resulteert in het volgende probleem

$$\min_{w,\gamma} \mathscr{J}^v(w) = \frac{1}{2}\sum_{j=1}^n \left(w^T x_j^v - y_j^v\right)^2 \quad \text{s.t.} \quad \left(X^T X + \frac{1}{\gamma} I_D\right) w = X^T Y.$$

Om complexere vormen van dit soort problemen formeel neer te schrijven, wordt het mechanisme van hiërarchisch programmeren gebruikt waarbij over w en γ wordt geoptimaliseerd met betrekking tot meerdere niveaus (zie vorig deel). Hiervoor worden de Karush-Kuhn-Tucker condities voor optimaliteit afgeëist aan het optimalisatie probleem. Hoewel dit soort problemen vaak niet meer convex is (zoals in dit geval), kunnen er efficiënte benaderingen van dit probleem gezocht worden zoals aangetoond in het proefschrift.

Een andere aanpak van dit probleem is gevonden door de invoering van een herparametrisering van de afweging tussen het belang van complexiteit en empirische kost. Laat de vector $c = (c_1,\ldots,c_N)^T \in \mathbb{R}^N$ de rol spelen van de ontwerpparameter

γ in de ridge-regressie formulering gegeven als

$$\hat{w} = \arg\min_{w} \mathscr{J}_c(w) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{N}\left(w^T x_i - y_i - c\right)^2.$$

De optimale schatting $\hat{w}$ is analytische gegeven voor elke vaste c als volgt

$$\left(X^T X + I_D\right)\ w = X^T(Y - c),$$

zodat voor elke mogelijke c er exact één globaal optimale oplossing bestaat. De voorgestelde herparametrisering leidt in het algemeen tot convexe modelselectie problemen. Dit pad is gevolgd voor het bouwen van nieuwe kernfunctie gebaseerde leeralgoritmen waar het primair-duale argument niet direct kan worden toegepast. Een belangrijke toepassing van het beschreven mechanisme is bekomen als een algoritme dat constructief in een maximaal stabiele oplossing resulteert.

Deel σ

Het laatste deel behandelt de vraag wat een goede kernfunctie kan zijn voor een welbepaalde taak. Vooreerst worden de relaties tussen gewogen regularisatieschema's, gewogen kleinste kwadraten en opgelegde lineaire structuur enerzijds, en het ontwerp van kernfuncties anderzijds beschreven. Daarna wordt uitgewijd hoe het mechanisme van structuurdetectie gebruik makende van de maat van maximale variatie zich leent tot het selecteren van een relevante kernfunctie gegeven een verzameling alternatieven.

Als laatste wordt het verband bestudeerd tussen het gebruik van isotropische kernfuncties (op basis van de wederzijdse afstand) en oorzakelijke filters. Dit resulteert in een convexe aanpak voor het leren van de kernfuncties uit gegevens op basis van het realizeren van de geschatte tweede orde karakteristieken van de observaties.

Conclusies

Dit proefschrift verdedigt hoofdzakelijk twee standpunten in het onderzoek naar het ontwerp van goede leeralgoritmen. Ten eerste is er geargumenteerd dat de taken van het ontwerp van een leermachine, de gebruikte maat van complexiteit en het bepalen van de ontwerpparameters in het algemeen, op vele manieren gerelateerd zijn. Het blijkt dat de studie van de interactie tussen genoemde onderwerpen efficiënt en consistent kan uitgevoerd worden door een invalshoek van optimalisatie te nemen. Concreet werd de taak van automatische modelselectie van ontwerpparameters bekeken als een hiërarchisch programmeringsprobleem.

Ten tweede tonen we aan dat het primair-duale argument zoals oorspronkelijk gebruikt in de formulering van SVMs een sterk formalisme verschaft voor het bouwen van nieuwe leeralgoritmen. Dit is aangetoond door het uitwerken en bestuderen van verschillende formuleringen voor het leren van nieuwe complexe taken, en het relateren en contrasteren van de methode met bestaande methodologiën. Een belangrijk resultaat is dat er aangetoond is dat structuur en voorkennis gemakkelijk kan ingebracht worden in het leeralgoritme door het gebruik van het primair-duale argument.

Appendices

Appendix A bespreekt de taak van het schatten van het ruisniveau in de data zonder dat er expliciet gesteund wordt op een geschat model. Hiervoor werd er een voorstelling van de data uitgewerkt op basis van de paarsgewijze verschillen tussen in- en uitvoerobservaties respectievelijk. Daar deze voorstelling van een differogram nadruk legt op de lokale eigenschappen van de data kunnen er eenvoudig eigenschappen zoal het ruisniveau worden afgeleid.

Appendix B geeft een korte bespreking van het software project LS-SVMlab dat de bestaande methodologie met betrekking tot LS-SVMs implementeert. In het kort worden de belangrijke bouwblokken van deze software voor MATLAB/C besproken.

List of Symbols

The following notation is used througout the text

Operators

$\triangleq$	By definition
$\succeq, \preceq$	Generalized Inequalities
$\arg\min_x \mathscr{J}$	Argument x minimizing the cost-function $\mathscr{J}$
$\arg\max_x \mathscr{J}$	Argument x maximizing the cost-function $\mathscr{J}$
$\text{Prob} : S \subset \mathbb{R}^D \to [0,\ \ 1]$	Probability
$P : \mathbb{R}^D \to [0,\ \ 1]$	Cumulative Distribution Function (cdf)
$p : \mathbb{R}^D \to \mathbb{R}^+$	Probability Density Function (pdf)
$\mathsf{Alg} : \mathscr{D} \to \mathscr{F}$	Algorithm mapping a dataset to an estimated function
$\mathsf{Modsel} : \mathscr{F} \to \mathbb{R}$	Model selection criterion
$\mathscr{R} : P \to \mathbb{R}$	Risk of an estimate given a distribution
$\mathscr{F} : \mathscr{F} \to \mathscr{F}$	Fourier transform of a function

Variables

$\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{e}$	Random variables
U, S, Ω	Matrices
Y, X	Vectors of observations
x	Vector of a single input observation
y	Single input observation
$\gamma, \lambda, \pi, \mu$	Hyper-parameters
D	Dimension of input vector
P	Number of parameters
N	Number of observations in training set
n	Number of observations in validation set
D_{eff}	Effective number of freedom
$\mathscr{M}$	Maximal variation

Sets

$\mathbb{R}$	Real numbers
$\mathbb{R}^d$	Vector of real numbers
$\mathbb{R}^{d\times n}$	Matrix of real numbers
$\mathbb{N}$	Set of positive integers
$\mathbb{T}$	Set of time-instances
$\mathscr{S}_a$	Affine set
$\mathscr{S}_c$	Convex set
$\mathscr{C}$	Cone
$\mathscr{D}$	Dataset $\{(x_i, y_i)\}_{i=1}^N$
$\mathscr{T}$	Dataset used for training purposes
$\mathscr{V}$	Dataset used for validation purposes
$\mathscr{F}$	Set of functions f
$\mathscr{H}$	Hilbert space of functions
$\mathbb{S}$	A set of indices
$\mathbb{P}_i$	Set of missing values of the ith datapoint
$\mathscr{F}_{\varphi,(P)}$	Class of Componentwise SVM models
$\mathscr{F}_{\varphi}$	Class of SVM models
$\mathscr{F}_{\varphi,T}$	Class of SVT models
$\mathscr{F}_{\varphi,P}$	Class of SVM models including parametric terms
$\mathscr{F}_{\omega}$	Class of linear parametric models
$\mathscr{E}$	Set of error terms
$\mathscr{A}$	Set of assumptions

Distributions

$\mathscr{N}$	Standard distribution
$\mathscr{U}$	Uniform distribution
χ^2	Chi-squared distribution
$\mathscr{L}$	Laplace distribution
$\mathscr{W}$	Wishart distribution

Abbrevitions

ν-SVT	Nu (ν) Support Vector Tube
ALS	Least Squares estimator based on Alternatives
Areg	Additive Regularization trade-off Scheme
cSVM	Componentwise Support Vector Machine
cLS-SVM	Componentwise Least Squares Support Vector Machine
CDF	Cumulative Distribution Function
hSVM	Huber-loss based Support Vector Machine
KKT	Karush-Kuhn-Tucker conditions for optimality
LASSO	Least Absolute Shrinkage Selection Operator
LS-SVM	Least Squares Support Vector Machine
OLS	Ordinary Least Squares estimator
PDF	Probability Density Function
pLS	Plausible Least Squares estimator
pSVM	Support Vector Machine with a parametric component
RR	Ridge Regression
SVM	Support Vector Machine
SVT	Support Vector Tube
TMSE	Total Mean Square Error

Contents

Chapter 1

Problems and Purposes

A broad overview is presented of a number of principles lying at the core of the process of induction of mathematical models from a finite set of observational data. Together with this general elaboration, recent advances in the area of kernel machines relevant to the presented research are sketched. Section 1.1 discusses the general setting of learning from data by induction, while Section 1.2 surveys the various approaches which give a sound foundation for doing so. Section 1.3 synthesizes a brief overview of various directions of the current research in machine learning using kernel methods. Section 1.4 then discusses the main contributions of the conducted research.

1.1 Learning

The science of learning plays a key role in the fields of statistics, data mining and artificial intelligence, intersecting with areas of engineering and other disciplines. The functional approach as e.g. used in (Bousquet and Elisseeff, 2002; Bousquet *et al.*, 2004) is employed to sketch a cross-section of this intertwined fields. Though this point of view is not exclusive, its strength may be found in its inherent relationship with convex optimization as showed next, its use in the problem of model analysis and model selection and its formal language.

Learning algorithms

A learning algorithm can be described as a mapping Alg from a set of given observations $\mathscr{D}$ and a collection of prior knowledge and assumptions represented as $\mathscr{A}$, to an optimal estimate belonging to the class $\mathscr{F}$:

$$\mathsf{Alg} : \mathscr{D} \times \mathscr{A} \rightarrow \mathscr{F}. \tag{1.1}$$

Let this mapping act as a definition of the process of *inference* (in this text). In statistical literature, this mapping is also known as an *estimation function* or an *estimator*. This formalization of a learning algorithm is denoted alternatively as a *learning machine*. The details of doing inference are explained in some detail in the case of supervised learning where the given set of training samples contains inputs as well as observed responses. The other cases (unsupervised, transductive learning and experimental or interactive data) are only marginally considered in the text.

Mapping Alg**:** As the learning algorithm is considered to be a uniquely defined mapping, some important assumptions (or restrictions) are imposed inherently. The most important is that there is exactly one estimate corresponding with a given dataset and a set of assumptions. Although quite restrictive with respect to methods employing global optimization techniques (as e.g. multi-layer perceptrons), this limitation will enable proper definition of a number of concepts as (global) sensitivity and stability. In this setup, the question can be formulated whether the mapping can be defined uniquely for any set of observations and assumptions. This general question is approached in this work by the extension of the primal-dual methodology to define learning algorithms for a variety of assumptions, as e.g. in terms of the noise conditions or the structure to be imposed on the algorithm.

Optimality**:** Somewhat central in the description of the learning algorithm as a mapping is the issue of optimality: the training dataset and the set of assumptions is mapped onto one and only one estimate which is the best among alternatives. The major concern is the purpose of the algorithm. One currently distinguishes between the often overlapping and sometimes conflicting objectives of (i) *Prediction* (what is the expected response of new observations), (ii) *Explanation* (what can be said about the generating mechanism underlying the observations), (iii) *Denoising* or smoothing (which part of the observations is due to external and unknown influences). Apart from these aims, an adequate definition of optimality is founded in a theory of inference (induction). The following section will elaborate on this issue. Inherently connected to the principle at hand is a set of rules to conduct calculations. Consider for example the classical practice of inference where one employs the notion of (relative) frequencies to translate the notion of likelihood. A complete different set of mathematical operations is used in e.g. Bayesian inference methods where computations are performed on (families of) distribution functions. Often, the theoretical foundation of the inductive technique translates into a measure of likeliness. From a practical perspective, a mathematical norm is to be optimized to find the estimate which is most consistent with the data or which captures optimally the chance regularity in the observations. More on this matter of norms in Subsection 1.2.7.

Data $\mathscr{D}$**:** Consider a set of N given observations

$$\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}, \tag{1.2}$$

of the input values $x_i \in \mathbb{D}^D$ in the D dimensional domain $\mathbb{D}^D$ and the corresponding output values $y_i \in \mathbb{D}$. Alternative denominators are respectively explanatory or independent variables, covariates, regressors or features, and outcome, response or dependent variable. One typically differentiates between various types of domains of the observed values. Consider the univariate case. An observation (say x) may be a *continuous* variable (e.g. $x \in \mathbb{R}$), *binary* variable (e.g. $x \in \{-1,1\}$), *categorical* variable which may either be a *nominal* (e.g. $x \in \{$Jazz, Pop, Classical, other$\}$), or an *ordered* variable (e.g. $x \in \{$Bad, Good, Superb, Exquisite$\}$), or a *sequence*. As a prototype of the latter, consider the series $\{x_t\}_{t\in\mathbb{T}}$ where $\mathbb{T}$ denotes a set of time instances.

Furthermore, an observation may be missing (we will only consider here the case that x is missing completely at random and no (external or conditional) knowledge can be exploited for predicting the unknown value, see (Rubin, 1976). Alternatively, the data observation may be known only partly due to a censoring mechanism. Consider the example of a clinical test on the reliability of a transplantation. An observation may be censored due to an unexpected car accident of the patient under study.

Assumptions $\mathscr{A}$**:** Assumptions (inexact) and prior knowledge (exact) come in different flavors:

- prior knowledge may be *qualitatively* (e.g. “the underlying function is strictly monotonically increasing”)
- some *quantitative* properties may be known (e.g. “the noise has a standard deviation of 3.1415”)
- prior *distributions* may be employed to express knowledge on the problem at hand (e.g. “the parameters are distributed as a χ^2 distribution with a certain degrees of freedom”)
- what is called *latent knowledge* embodies the set of results, theorems and (future) advances which may be of relevance to the problem at hand (e.g. “the arithmic mean is in the limit Gaussian distributed under mild regularity conditions and has bounded deviation for finite samples due to Hoeffding’s concentration inequality”).

Estimation Class $\mathscr{F}$**:** A particularly important case of prior knowledge is the representation of the members of the estimation class (denoted as models, estimated mappings or estimates). One distinguishes between parametric and non-parametric estimators as explained in the following subsection. Apart from this issue, the representation of the final estimate may be used to embed the known structure of the problem at hand. One can for example postulate a causal auto-regressive model representation in the case of sequential data. Another example is encountered when working with a (discrete) decision tree or with a real valued decision rule.

The distinction in output type has led to a naming convention for the learning task and the estimation class. Major classes in this respect include the class of regressors ($f_a : \mathbb{D}^D \rightarrow \mathbb{R}$), of classifiers ($f_c : \mathbb{D}^D \rightarrow \{-1,1\}$), of multi-class classifiers (e.g. $f_m : \mathbb{D}^D \rightarrow \{$Jazz, Pop, Classical, other$\}$) and the class of ordinal regressors (e.g. $f_o : \mathbb{D}^D \rightarrow \{\text{Bad, Good, Superb, Exquisite}\}$). This text will mainly focus on the first two choices, but later chapters will repeatedly touch upon the other cases. Apart from mentioned characterizations, one also distinguishes between linear versus nonlinear and parametric versus nonparametric models.

***Analysis*:** The analysis of the result of the learning algorithm and the mapping (1.1) itself is a major source of active research. A large set of notions have been defined over time in order to quantify different aspects. Important topics include the notions of *consistency* (does the estimate converge to the true quantity when $N \rightarrow +\infty$), *bias/variance* (what can be expected of the distribution of the estimates based on finite and noisy samples (mean/variance)) or *sensitivity/stability* (how is the estimate perturbed when modifying the dataset). These notions are formalized lateron.

This manuscript is organized around a set of principal guidelines which are re-occurring in the text at various places and under different disguises

Tools from convex optimization theory and linear algebra. This research mainly differs from the classical methodology of multi-layer perceptrons and artificial neural networks by putting the first property of convexity of the resulting optimization problems. Together with tools from linear algebra, a language is provided which enables the proper formulation and analysis of various nonlinear algorithms.

Model representations and residuals. Once the parameters of the problem, or the predictor in the non-parametric case are known, the characteristics of the (stochastic model of the) residuals are known. Although sounding rather obvious at first sight, this issue has some profound implications as motivated throughout the text.

Prior knowledge as constraints. This issue stresses the importance of prior knowledge (either qualitative or quantitative) to achieve better performance of the models. The primal-dual characterization is seen to be highly apropriate for supporting this guideline.

1.1.1 Probability, dependencies and correlations

Dependencies and correlations make up the heart of classical probability theory and statistical practice (Spanos, 1999). A brief overview of the basic machinery is given. Probability theory is often considered in a purely mathematical setting of measure theory as proposed in the seminal work (Kolmogorov, 1933). Let S be a the sample space. Let $\mathscr{B}$ be a collection of subsets of S representing the events of interest, (let $\mathscr{B}$ be a σ-field). Consider a function $\text{Prob} : \mathscr{B} \rightarrow [0,1]$ which satisfies the fundamental axioms

- $\text{Prob}(S) = 1$,
- $\text{Prob}(A) \geq 0$ for all sets $A \subset S$,
- $\text{Prob}(\bigcup A_i) = \sum_i \text{Prob}(A_i)$ if the sequence of subsets $\{A_i\}$ is a finite or countable set containing pairwise disjoint elements of $\mathscr{B}$.

This interpretation, abbreviated as the statistical space $(S, \mathscr{B}, P)$, reduces mathematical probability theory to the study of sets and measure theory (Kolmogorov, 1933). As a prototype, consider the space $(\mathbb{R}, \mathscr{B}_\mathbb{R}, P)$ where the events of interest are described as $\mathscr{B}_\mathbb{R} = \{B_x = [-\infty, x] \subset \mathbb{R} \mid x \in \mathbb{R}\}$. An intuitive explanation of the function P becomes then $P(x) = \text{Prob}(x \in B_{x'}) = \text{Prob}(x' \leq x)$. In general, any space $(S, \mathscr{B}, P)$ can be mapped onto $(\mathbb{R}, \mathscr{B}_\mathbb{R}, P_X)$ using a function $X : S \to \mathbb{R}$. This function (or its image) is referred to as a random variable. Let the cumulative distribution function (cdf) of the random variable be defined as $P_X : \mathbb{R} \to [0,1]$ such that $P_X(x) = \text{Prob}(\{s : X \leq x\})$. The subscript $_X$ of the function P_X is omitted with some abuse of notation in the cases in which the context makes it clear which random variable is involved. The derivative $p(x) = \partial P(x)/\partial x$, if it exist, is referred to as the probability density function (pdf). The expected value operator $E : X \to \mathbb{R}$ is defined as

$$E[X] = \int x dP(x) = \int x p(x) dx. \tag{1.3}$$

Example 1.1 gives a simple example of one family of distribution functions and two empirical estimators used to recover respectively the cdf and the pdf.

One proceeds by defining the notions of dependency and its weak variant correlation. Let X, X_1 and X_2 be univariate random variables with (cumulative) distributions functions $P(X)$, $P_1(X_1)$ and $P_2(X_2)$ respectively. Let the joint distribution denoted as $P_{12}(X_1, X_2)$ be defined analogously. The random variables X_1, X_2 are independent if the following relation holds

$$P(X_1, X_2) = P(X_1)P(X_2). \tag{1.4}$$

This motivates the definition of N independently and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_N$

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i). \tag{1.5}$$

An equivalent definition of independency is given as follows, for any well-defined functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$

$$E\left[f(X_1), g(X_2)\right] = E\left[f(X_1)\right] E\left[h(X_2)\right]. \tag{1.6}$$

Consider the special case where g and h are the functions $f(x) = x - E[X_1]$ and $g(x) = x - E[X_2]$ one obtains the covariation coefficient (or covariance) $c(X_1, X_2) \triangleq E[(X_1 - E[X_1])(X_2 - E[X_2])]$. The correlation coefficient corresponds to the normalized covariation as follows

$$\rho(X_1, X_2) \triangleq \frac{c(X_1, X_2)}{\sqrt{c(X_1, X_1)c(X_2, X_2)}}. \tag{1.7}$$

It follows that a zero covariance or zero correlation coefficient is a necessary (but not a sufficient) condition for independence. If a ± 1 correlation coefficient is obtained, the relationship between X_1 and X_2 is strictly linear. Finally, let the conditional probability $P(X_1 \mid X_2)$ be defined as

$$P(X_1 \mid X_2) \triangleq \frac{P(X_1, X_2)}{P(X_2)}. \tag{1.8}$$

This elaboration provides sufficient information to most theoretical concepts which are used throughout the text.

1.1.2 Parametric vs. non-parametric

Classical statistical inference starts with the model designer postulating explicitly and a priori a statistical model purporting to describe the stochastic mechanism underlying the observed data. Parametric model inference is concerned with the inference of the (limited) set of unknown parameters in the postulated statistical model. The class of parametric linear models is then defined as

$$\mathscr{F}_\omega = \left\{ f : \mathbb{R}^D \rightarrow \mathbb{R} \;\middle|\; f(x) = \omega^T x, y_i = f(x_i) + e_i \right\}, \qquad e_i \sim F(\theta), \tag{1.9}$$

where $F(\theta)$ denotes a distribution function determined up to a few parameters θ. This paradigm was the main subject of interest of the statistical literature and has had a profound impact on related domains as system identification.

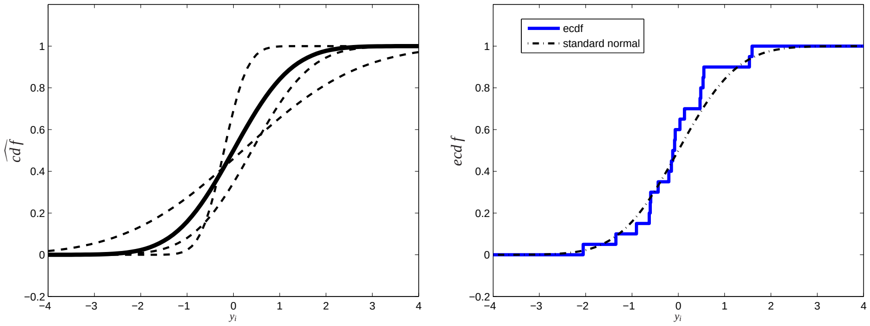
In contrast non-parametric (also called distribution-free) techniques do not postulate a parameterized family of statistical models underlying the observed data, but do instead define the class of estimators implicitly by imposing proper restrictions. Consider for example (and in contrast to $\mathscr{F}_\omega$) the non-parametric class of continuous functions with bounded higher order Lipschitz derivatives defined as

$$\mathscr{F}_L = \left\{ f : \mathbb{R}^D \rightarrow \mathbb{R} \;\mid\; \frac{\partial^d f(x)}{\partial x^d} \leq L_d, \forall x \in \mathbb{R}^D \right\}. \tag{1.10}$$

This definition commonly acts as a mathematical translation of the denominator *sufficiently smooth*. The non-parametric approach often has a specific goal (as prediction) but avoids to characterize the underlying generating mechanisms explicitly.

This terminology originates from statistical inference of density functions (Silverman, 1986) (see Example 1.1), but is used deliberately throughout many fields as e.g. in function approximation (e.g. to differentiate between parametric linear models versus non-parametric smoothing splines). The use of an implicitly defined broad class as in non-parametric estimators is often regarded as a safeguard against misspecification. However, the question which approach will obtain the highest statistical adequacy cannot be answered straightforwardly.

It is well-known that the early literature on robustness towards gross-errors, see Subsection 1.3.2, was motivated by the undue reliance of classical parametric inference on



(a) (b)

Figure 1.1: *Descriptions of the the cumulative distribution function (cdf) of a sample based on the parametric and non-parametric paradigm respectively.* **(a)** *The normal cdf model for different values of the mean μ and the variance σ. Typically, one uses the maximum likelihood method to estimate the mean and the variance from the sample.* **(b)** *The empirical cdf function is a theoretical sound method to summarize all information regarding the distribution from the finite sample. The disadvantage of this method are discontinuities which prohibit the proper derivation of an empirical probability density counterpart.*

the assumption of normality. Although a vague difference exist (robustness considers deviations from parametric models, non-parametric methods consider implicit model definitions), modern literature on robustness is in great pains to distinguish itself from non-parametric methods (Hampel *et al.*, 1986; Spanos, 1999). To side-step these issues, this text will take the convention to distinguish between (non-) parametric model (representations) and (non-) parametric noise models where the latter corresponds to the robustness approach. This convention makes it possible to speak of non-parametric models with contaminated parametric models that require robust methods.

Example 1.1 **[Representations of distributions]** The difference between the parametric and the non-parametric paradigm is illustrated readily by the following example in the field of density estimation. Let **Y** be a univariate random variable with samples $\{y_i\}_{i=1}^N$. Consider on the one hand the parametric approach where a family of densities (say the Normal distribution) is postulated.

$$\hat{F}(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right). \tag{1.11}$$

The task of inference amounts to finding the optimal parameters (the mean μ and the variance σ^2) from the observations. Employing the technique of maximum likelihood, one arrives at the arithmic mean and the sample variance as the preferable estimate, see also Example 1.2.

One has at least two non-parametric approaches: the empirical cumulative distribution (ecdf) estimator and the histogram, see e.g. (Rao, 1983; Silverman, 1986; Scott, 1992) for a broad account of the issue. For a given realization of the sample the empirical cdf (ecdf) is defined as (Billingsley, 1986)

$$\hat{F}(y) = \frac{1}{N}\sum_{k=1}^{N} I(y \leq y_k), \quad \text{for} \quad -\infty < y < \infty, \tag{1.12}$$

where the indicator function $I(y \leq y_k)$ equals 1 if $y \leq y_k$ and 0 otherwise. This estimator has the following properties: (i) it is uniquely defined; (ii) its range is $[0,1]$; (iii) it is non-decreasing and continuous on the right; (iv) it is piecewise constant with jumps at the observed points, i.e. it enjoys all properties of its theoretical counterpart, the cdf. Furthermore, $\sup_y \left|F(y) - \hat{F}(y)\right| \to 0$ with probability one as stated in the Glivenko-Cantelli Theorem (see e.g. (Billingsley, 1986)). While the ecdf is a theoretical sound tool, its practical applicability is obstructed as the corresponding estimated pdf cannot be computed straightforwardly (the ecdf is not differentiable) and its extension to the multivariate case is more involved.

The Parzen kernel approach represents any unknown but sufficiently smooth density function as the sum of density kernels (Parzen, 1970).

$$\hat{F}(y;h) = \frac{1}{Nh}\sum_{i=1}^{N} K_h\left(\frac{y_i - y}{h}\right), \tag{1.13}$$

where $h \in \mathbb{R}_0^+$ denotes the bandwidth and $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is the so-called Parzen kernel function. An univariate example of the latter is

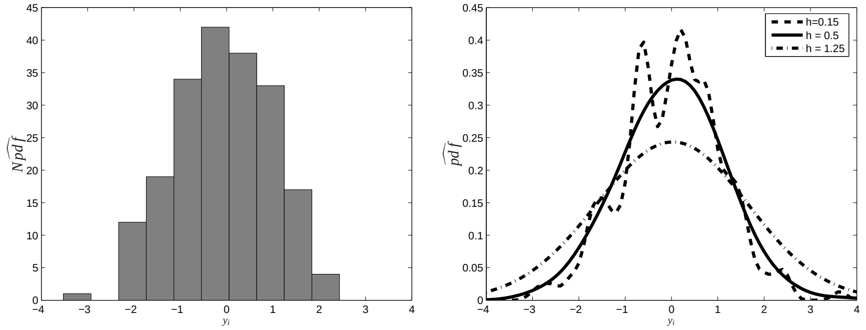
$$K_h(y, y_i) = \frac{1}{h\sqrt{2\pi}} \exp\left(\frac{-(y - y_i)^2}{2h^2}\right). \tag{1.14}$$

Figure 1.1 and 1.2 illustrate the different approaches of the parametric, the empirical cdf, the histogram and the Parzen window.

1.2 Generalization and Inference

Somewhat central in the discussion of induction from observational data lies at the concept of generalization. A model which is generalizing well will provide good predicted responses corresponding with new data-samples. Generalization acts as a bridge between properties of the estimate based on the observations and the expected global optimality principle. The intention of this text is not to advocate one principle over any other but rather to place the discourse in its historical and scientific context. Inference was motivated from different points of view throughout history. As summarized by (Vapnik, 1998)

> “Although the arms consisted mostly of mathematical symbols, the discussion is essentially philosophical in nature”.



(a) (b)

Figure 1.2: *Illustration of the difference between the histogram and the Parzen window estimator for the assessment of the probability density estimation (pdf) of a sample of size 100 i.i.d. sampled from the standard normal distribution.* **(a)** *The histogram method using 10 equidistant bins,* **(b)** *The Parzen window with three different bandwidths h. When h is too small, the estimate exhibits too much variability (under-smoothing). In the case h is too big, too little detail of the distribution is recovered (over-smoothing).*

1.2.1 Summary and descriptive statistics

The early history of statistics mainly focused on the description of data-samples by the use of so-called summary statistics (Pearson, 1902). Modern statistics criticized this approach (Fisher, 1922) for its lack of mathematical rigor and its ill-defined foundations. As was put by J. Williams, see also (Rice, 1988; Spanos, 1999)

> “We must be careful not to confuse data with the abstractions we use to analyze them”, J. Williams, 1842-1910.

This type of reasoning on the raw data gained renewed interest and a better justification with the advent of exploratory data analysis (EDA) (Tukey, 1977). The research on EDA deals with methods of describing and summarizing data that are in the form of a set of samples or batches. These procedures are useful in revealing the structure of the observed data. In the absence of a stochastic model, the methods are useful for purely descriptive purposes. Important tools here are the empirical cdf, the histogram and related methods (see example 1.1), the arithmic mean, median and quantiles readily summarized in a boxplot and the QQ-plot (Tukey, 1977). The latter is a very useful tool for the comparison and advice of distribution functions underlying the data. Common goals of EDA are to inspect the data on atypical observations and to get an initial idea on the class of stochastic models governing the relationships in the observed dataset.

The difference between descriptive statistics and non-parametric or even parametric statistics is in many cases very subtle and even artificial. Consider e.g. the case of the mean statistic as in example 1.2 which cannot be assigned uniquely to the class of descriptive or model based approaches. Moreover, visualization techniques and summary statistics do often exploit (hidden) assumptions which impose an implicit model on the data. For example the simple t-plot of the data over the indices do suggest a certain ordering or explanation on the observations. Those issues convert the distinction between descriptive and (non-)parametric models into a purely philosophical discussion.

1.2.2 Function approximation

Many complex functions that occur in mathematics cannot be used directly in computer simulations. This starting point motivated the elaboration of a subfield of mathematics concerned with the approximation of functions using simple schemes as polynomials. The study of the theory and the application of this type of problems is embodied in the literature on function approximation, see e.g. (Powell, 1981). The cornerstones of this research were set out by the work of Chebychev two centuries ago, see e.g. (Chebyshev, 1859). Typical for this approach is the lack of any reference to a probabilistic setting and the use of worst-case analysis often translated in the use of an L_∞ norm. Although approximation algorithms are used throughout the sciences and in many industrial and commercial fields, the theory has become highly specialized and abstract.

Important results where described in various directions, including the study and construction of (orthogonal) basis functions and their representational power. This lead to the study of fractional functions which have had a severe impact on the literature on system identification due to (Wiener, 1949), the construction of the non-parametric splines models as described e.g. in (Schumaker, 1981) which are discussed in the context of observational data including error terms in (Craven and Wahba, 1979; Wahba, 1990) and revised in Section 5.1. The construction of localized basis functions gained renewed interest through the theoretical and practical application of wavelets, see e.g. (Daubechies, 1988) for a complete account.

1.2.3 Maximum likelihood

A more stochastic setting was proposed under the framework of Maximum Likelihood (ML) for the purpose of fitting probability laws to the data as elaborated mainly due to sir R.A. Fisher (Fisher, 1922). The main intuition goes as follows. One starts by postulating a class of statistical generating models governing the chance regularities underlying the data. The different elements of this family are enumerated using a finite set of parameters which ought to be recovered by the observed samples.

The maximum of the likelihood $p(\mathbf{X}|\theta)$ of a parameter θ characterizing an element from a finite dimensional class of probabilistic laws, given a set of observations generically denoted as **X** is denoted as

$$\theta_{ml} = \arg\max_{\theta} p(\mathbf{X}|\theta) = \arg\min_{\theta} \sum_{i=1}^{N} \log p(\mathbf{X}_i|\theta). \tag{1.15}$$

The application of the ML in the context of fitting a Gaussian distribution with unknown mean to the observed data is discussed in the following example.

Example 1.2 **[Estimating location parameters, I]** The estimation of location parameters of a density from a set of i.i.d. samples is central in the field of statistics. The following derivation shows the similarity between the mean location estimator and the least squares method.

Let $\{y_i\}_{i=1}^N$ be sampled i.i.d. from a random variable **Y** with pdf $p_{\mathbf{Y}} = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(y_i-\mu)^2/\sigma^2\right)$. The maximum likelihood estimator of the location parameter μ becomes

$$\begin{aligned} \hat{\mu} &= \arg\max_{\mu} \log \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y_i-\mu)^2}{\sigma^2}\right) \\ &= \arg\min_{\mu} \sum_{i=1}^{N} (y_i-\mu)^2 \\ &\Leftrightarrow 1_N^T 1_N \hat{\mu} = 1_N^T Y, \end{aligned} \tag{1.16}$$

where $Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$. The last equation follows from the normal equations of the least squares estimate. From this it follows that the arithmic mean possesses the properties of the maximum likelihood estimator in the case a Normal distribution may be assumed. (Fisher, 1922), see e.g. (Rice, 1988; Spanos, 1999). See also Example 3.3 for a similar argument in the case of the Median.

An important issue in the theory of statistical inference becomes how the estimator behaves on average. This is often approached by the development of approximations to the sampling distribution of estimates by using limiting arguments as the sample size increases. Then there are a number of important concepts to qualify the properties of the estimator, including

Consistency An estimate $\hat{\theta}$ is called consistent in probability if for any $\varepsilon > 0$ arbitrarily small

$$\lim_{N\to\infty} P\left(\left|\hat{\theta} - \theta_0\right| > \varepsilon\right) \to 0, \tag{1.17}$$

where θ_0 is the true parameter of the underlying parametric probabilistic rule. Under reasonable conditions, the ML estimate θ_{ml} is consistent (Cramer, 1946).

Fisher Information Matrix The (Fisher) information matrix of an estimate $\hat{\theta}$ is defined as

$$I(\theta) = E\left[\frac{\partial \log p(X|\theta)}{\partial \theta}\right]^2 = -E\left[\frac{\partial^2 \log p(X|\theta)}{\partial \theta^2}\right], \tag{1.18}$$

under appropriate smoothness conditions. The large sample distribution of a maximum likelihood estimate is approximatively normal $\theta_{ml} \sim \mathcal{N}\left(\theta_0, \frac{1}{N} I(\theta_0)\right)$.

Bias A concept which will play an important role in the sequel is the decomposition of the expected Mean Squared Error (MSE) in bias and variance. The reach of this definitions were extended to the case of finite data samples. The bias-variance decomposition follows from the following equality

$$\mathrm{MSE}(\hat{\theta} - \theta_0) = E[\hat{\theta} - \theta_0]^2 = E\left[\hat{\theta} - E[\hat{\theta}]\right]^2 + (E[\hat{\theta}] - \theta_0)^2, \tag{1.19}$$

where the terms of the right hand side are referred to as the *variance* and the *bias* of the estimate respectively. In the case of ML, the estimator θ_{ml} is asymptotically unbiased following the previous item whenever the true probabilistic law is contained in the parametric class of distributions. Bias and variance of the estimator constitute a principal tool for the analysis of estimators in the case of a finite number of observations.

Efficiency The efficiency of an estimate $\hat{\theta}$ with respect to an alternative $\tilde{\theta}$ is defined as

$$\mathrm{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\mathrm{MSE}(\hat{\theta} - \theta_0)}{\mathrm{MSE}(\tilde{\theta} - \theta_0)} = \frac{E\left[\hat{\theta} - E[\hat{\theta}]\right]^2 + (E[\hat{\theta}] - \theta_0)^2}{E\left[\tilde{\theta} - E[\tilde{\theta}]\right]^2 + (E[\tilde{\theta}] - \theta_0)^2}, \tag{1.20}$$

which reduces to the fraction of the variances when both $\hat{\theta}$ and $\tilde{\theta}$ are unbiased estimates. A classical result is that in the case of i.i.d. data-samples a lower-bound holds. Let $\{\mathbf{X}_i\}_{i=1}^N$ be an i.i.d sample and let $\hat{\theta}$ be any unbiased estimate

$$E[\hat{\theta} - \theta_0]^2 \geq \frac{1}{N\, I(\theta_0)}, \tag{1.21}$$

which is known as the Cramer-Rao inequality (Cramer, 1946). The inequality holds asymptotically exactly in the case of ML estimates θ_{ml} under appropriate regularity conditions. An important caveat arises in the case of a finite number of samples where biased estimators exists which do improve on the bound even in the prototypical case of estimating location parameters (Stein, 1956).

Sufficiency An estimate $\hat{\theta}$ is called sufficient if it contains all information in the sample about θ_0. Formally

$$P(\theta_0 \mid \mathscr{D}) = P(\hat{\theta} \mid \mathscr{D}) \Leftrightarrow \exists P_\theta, P_{\mathscr{D}} \quad \text{s.t.} \quad P(\mathscr{D} \mid \theta_0) = P_\theta(\hat{\theta}, \theta) P_{\mathscr{D}}(\mathscr{D}), \tag{1.22}$$

where the righthandside provides a convenient way for identifying sufficient estimators. The Rao-Blackwell theorem states the following inequality: let θ_s be a sufficient estimate and let $\hat{\theta}$ be any estimate, then $E[\theta_s - \theta_0]^2 \leq E[\hat{\theta} - \theta_0]^2$ under regularity conditions, see e.g. (Rao, 1965).

1.2.4 Bayesian inference

Bayesian inference is concerned with the calculus of distribution functions representing degrees of belief in the phenomena under study. This is opposed to the classical view of probability and distributions as the limit of relative frequencies. One can think of the former methodology as a formalization of a purely rational judge, while the latter originates more from the analysis of rules of chance. The Bayesian method is constructed around the following equality referred to as Bayes' rule:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}, \tag{1.23}$$

where the terms are respectively called the *posterior* ($p(A|B)$), the *likelihood* ($p(B|A)$), the *prior* ($p(A)$) and the *evidence* ($p(B)$) which normalizes the right hand side. Mathematical, philosphical as well as practical issues of the Bayesian methodology were covered in detail in (Jaynes, 2003).

This general law may be applied readily to the parametric estimation problem of a model with parameters $\theta \in \Theta$. Let A be replaced by the parameter space Θ and substitute B by the observations $\mathscr{D}$ and the assumptions $\mathscr{A}$. Then one can readily express the posterior of the parameters given the data and an appropriate prior distribution on the possible parameters Θ. Maximizing this posterior results in the MAP (maximum a posterior) estimate

$$\hat{\theta} = \arg\max_{\theta \in \Theta} p(\theta|\mathscr{D},\mathscr{A}) = \frac{p(\mathscr{D}|\theta,\mathscr{A})p(\theta|\mathscr{A})}{p(\mathscr{D},\mathscr{A})}. \tag{1.24}$$

Although a decade or more older than the first glimpses of maximum likelihood (see Laplace), Bayesian inference has not overruled the classic statistical methodology sofar, mainly due to practical problems as slow sampling schemes (Gibbs and Markov Chain Monte Carlo), see e.g. (O'Hagen, 1988), oversimplifications or the enduring question of the optimal prior. Current research on those topics however narrows swiftly the gaps, see e.g. (Rasmussen, 1996; MacKay, 1998).

1.2.5 Statistical learning theory

The goal of statistical learning theory is to study and to formalize, in a statistical framework, the property of learning algorithms Alg (Bousquet *et al.*, 2004). In particular, most results take the form of so-called error bounds which amount to a worst case analysis. Although existing for over 40 years, the theory of statistical learning only gained the status of a major player in the field of inference from observational data since a decade or so. This is mainly due to the introduction, analysis and practical significance of the Support Vector Machine the kernel methods (Vapnik, 1998).

In statistical learning theory, one investigates under which conditions empirical risk minimization results into consistent estimates minimizing the theoretical risk. The key idea for creating effective methods of inference from small sample-sizes is formulated in the following main principle due to (Vapnik, 1998):

> "If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available

> information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem."

Although intuitive at first sight, it is somewhat in contrast with the paradigm of classical statistics where one tries to recover the probabilistic rules governing the data generation. The classical results from (Vapnik, 1998) may also be considered as a generalization to the Glivenko-Cantelli theorem towards finite numbers of data-samples stating that relative frequencies will converge to the underlying probability.

A crucial principle then is to consider a class of hypotheses with a restricted capacity. As was put by (Bousquet *et al.*, 2004),

> "Surprisingly as it may seem, there is no universal way of measuring simplicity (or complexity) and the choice of a specific measure inherently depends on the problem at hand. It is actually in this choice that the designer of the learning algorithm introduces knowledge about the specific phenomenon under study. This lack of universal best choice can actually be formalized in what is called the *No free lunch* theorem. [...] If there is a priori no restriction on the possible phenomena that are expected, generalization would be impossible and any algorithm would be beaten by another on some phenomenon. [...] The core assumption enabling generalization in this framework is that both given training dataset and future sample points are independently distributed using identical distributions (i.i.d.)."

The main theory describes the case of binary functions (classifications). Let $\mathbf{X} \in \mathbb{R}^D$ be a random variable with fixed but unknown cdf $P_\mathbf{X}$ and let $\mathbf{Y} \in \{-1,1\}$ be a binary random variable with fixed but and unknown cdf $P_\mathbf{Y}$. and let the theoretical risk $\mathscr{R}$ of any mapping $f : \mathbb{R}^D \to [0,1]$ be defined as follows

$$\mathscr{R}(f, P_{\mathbf{XY}}) = \int I\left(f(x)y \leq 0\right) dP_{\mathbf{XY}}, \tag{1.25}$$

where $I(x \leq 0)$ equals one if $(x \leq 0)$ and zero otherwise. The Bayes classifier $f^* : \mathbf{X} \to \mathbf{Y}$ becomes

$$f^*(x) = \operatorname{sign}\left(E\left[\mathbf{Y}|\mathbf{X}=x\right]\right). \tag{1.26}$$

The classifier f^* is proven to achieve the minimal risk over all mappings f. In this setting, one typically possesses a finite number of data-samples of the random variable denoted as $\mathscr{D} = \{(x_i, y_1)\}_{i=1}^N \subset \mathbb{R}^D \times [-1,1]$. The empirical risk based on this data sample becomes

$$\hat{\mathscr{R}}(f, \mathscr{D}) = \sum_{i=1}^N I\left(f(x_i)y_i \leq 0\right). \tag{1.27}$$

Now, statistical learning theory considers the question under which conditions the empirical risk $\hat{\mathscr{R}}$ will converge to the true risk $\mathscr{R}$ in general, formally

$$\sup_f \left|\mathscr{R}(f) - \hat{\mathscr{R}}(f)\right| \xrightarrow{\text{Prob}} 0. \tag{1.28}$$

More specifically, the convergence of the estimate minimizing the empirical risk to the Bayes classifier is discussed. Extensions to various related induction tasks in the occurence of a finite number of data-samples are discussed e.g. in (Vapnik, 1998; Bousquet *et al.*, 2004).

Necessary and sufficient conditions for convergence were expressed relying on various measures of capacity, including

Growth Function The growth function $S_{\mathscr{F}}(N)$ is the maximum number of different ways into which N points can be divided into two classes with an $f \in \mathscr{F}$.

VC-dimension The VC-dimension is the size of the largest number of samples which can be divided arbitrarily (shattered) in different classes using functions of the class $\mathscr{F}$. Formally, the VC-dimension of a class $\mathscr{F}$ is the largest N such that $S_{\mathscr{F}}(N) = 2^N$.

Covering Number A measure which is computable more easily is the covering number. This number corresponds to the size (capacity) of the function class $\mathscr{F}$ as measured by the Hamming distance based on the training dataset.

Rademacher Complexity The Rademacher complexity denotes the expected worst-case risk over the class of $f \in \mathscr{F}$ when assigning random labels to the dataset, or formally $\mathscr{R}_c(\mathscr{F}) = E \sup_{f\in\mathscr{F}} \frac{1}{2}\sum_{i=1}^{N} I(f(x_i)\sigma_i < 0)$, where $\{\sigma_i\}_{i=1}^{N}$ sampled at random from $\{-1,1\}^N$ with $p_{-1} = p_1 = 0. > 5$. The advantage of this measure over the others is that an empirical approximation can be computed straightforwardly.

This measures are used to construct bounds on the deviation of the empirical and theoretical risk minimizer, see e.g. (Vapnik, 1998; Shawe-Taylor and Cristianini, 2004). See also Theorem 3.2 and 3.4.

1.2.6 Hypothesis testing

To complete this overview, a brief description is given of one of the most important but also one of the most confusing parts of statistical inference. The difficultness of the theory and practice of hypothesis testing is mainly due to the phenomena that (a) numerous new concepts are needed before one is able to define the problem adequately, and (b) there is no single method available for constructing good tests under different circumstances which is comparable to the maximum likelihood estimator in estimation. While an historical account (as e.g. given in (Spanos, 1999)) has at least the advantage of a strict ordering, the subject is here only touched from the viewpoint of model testing.

Somewhat central in the theory and practice of hypothesis testing is a problem dependent definition of a null-hypothesis H_0. The procedure of testing proceeds with the derivation of the corresponding distribution of the estimate based on the finite number of (noisy) data samples in case the null-hypothesis were valid. If expressed without explicit reference to the unknown parameters except the null-hypothesis (pivotal function) by a proper normalization, a test statistic $T : \mathscr{D} \times H_0 \to \mathbb{R}$ is obtained. This test statistic expresses how much a sample realization of the null-hypothesis can deviate from the expected outcome. The final test decides whether the estimate from the observations is unlikely to be sampled from the test statistic. Applying the test statistic on the observed data results in the so-called p-value, defined as

$$p \triangleq P\left(c_0 \geq T(\mathscr{D}) \;\middle|\; H_0\right) \tag{1.29}$$

where c_0 denotes the distribution of the test statistic for any sample realization of the null-hypothesis. If p is small enough, the test would advocate rejection of the null-hypothesis. Opposed to this original formulation due to sir R.A. Fisher was the relative procedure of hypothesis testing as proposed by Neyman and Pearson (Neyman and Pearson, 1928). The key to their approach was the introduction of the notion of an alternative hypothesis H_1 to supplement the notion of the null-hypothesis and thus transform testing into a choice amongst different hypotheses. The design of a test amounts then to the derivation of a proper normalized indicator function $T : \mathscr{D} \Rightarrow \mathbb{R}$ which separates the null and the alternative hypothesis properly. Let $\mathscr{R}_0 \subset \mathbb{R}$ be defined such that for a pre-specified significance level $\alpha \in \mathbb{R}_0^+$ the following relation holds:

$$\begin{cases} P\left(T(\mathscr{D}) \notin \mathscr{R}_0 \;\middle|\; H_0\right) = \alpha \\ P\left(T(\mathscr{D}) \in \mathscr{R}_0 \;\middle|\; H_1\right) = \varepsilon, \end{cases} \tag{1.30}$$

where $\varepsilon \in \mathbb{R}_0^+$ is as small as possible.

Example 1.3 **[Hypothesis tests for input selection]** The following classical result is widely known as the z-test, see e.g. (Rice, 1988). Given an i.i.d. sample of a univariate Gaussian distribution $\{x_i\}_{i=1}^N$. Consider the problem to decide whether a location parameter is zero ($\mu = 0$). Assume the second moment (variance) σ^2 is given. Consider the following test statistic

$$z = \frac{\sqrt{N}(\hat{\mu} - \mu)}{\sigma} \sim \mathscr{N}(0,1) \tag{1.31}$$

where $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N$. Then its p-value is defined as

$$z = P\left(c_0 \geq T(\mathscr{D}) \;\middle|\; c_0 \sim \mathscr{N}(0,1)\right) \tag{1.32}$$

expressing an absolute likelihood of the null-hypothesis. Alternatively, a relative likelihood based test can be constructed. Consider the alternative hypothesis H_1 that $\mu \neq 0$. Again t ($T = t$) is derived as a good indicator function separating the two hypotheses. The threshold c_α of the test statistic T for a specified significance level α does not depend on any unknown parameter and is e.g. tabulated in various textbooks. Given this specifications, the final test is summarized as follows

$$T(\mathscr{D}) \geq c_\alpha \Rightarrow P(H_1) = 1 - \alpha,\; P(H_0) = \alpha. \tag{1.33}$$

1.2.7 Towards an optimization perspective

While the formulation of appropriate optimality principles giving sound foundations to the conducted inference often differ from a theoretical as well as practical perspective, the construction of the corresponding learning algorithm often coincides in large extents. We stress the fact that those apparent correspondences do not streamline the interpretation of the results. This issue motivates the further coexistence of the various approaches. A similar point of view was adopted in the book (Boyd and Vandenberghe, 2004).

The discrepancy between two objects can be expressed using different norms, each with its own characteristics and properties. The following enumeration is restricted to the norms of vectors.

L_1: The one-norm or L_1 started history due to Laplace some decades before the classical work by Gauss. Although obscured in scientific history in favour of the L_2 norm and L_1, it regained recently interest due to efficient ways to calculate the corresponding minimizer. This norm played a crucial role due to its relation to the median location estimator (Andrews *et al.*, 1972), in the recent formulation of SVMs (Vapnik, 1998) and kernel machines (Schölkopf and Smola, 2002), its theoretical properties for density estimation (Devroye and Györfi, 1985), and the property that its minimizer typically presents zeros in the solution parameter ("sparseness") as exploited in e.g. LASSO (Tibshirani, 1996).

L_2: This measure gained a central role in all different approaches towards the task of inference from data since the semimal work of Gauss two centuries ago. Its importance was confirmed by the works of (Fisher, 1922) and the central place of the corresponding central distribution, see e.g. (Jaynes, 2003) for a complete account. Its central role triggered the formulation of LS-SVMs (Suykens *et al.*, 2002*b*) as a general methodology based on SVMs extending its reach from classification to regression and unsupervised learning.

L_∞: The L_∞ norm came forth of the worst-case analysis in function-approximation problems as formulated in the classical works of Chebychev (Chebyshev, 1859). In theoretical and practical statistics its importance is given in results as the central Glivenko-Cantelli therorem, see e.g. (Vapnik, 1998), and in the test-statistics as Kolmogorov-Smirnoff (Conover, 1999). In the context of primal-dual kernel machines this measure lies at the basis of Support Vector Tubes (SVT) in Section 3.5 and the measure of maximal variation, see Section 6.4.

L_p: The previous norms were generalized in the formulation of the so-called Minkowski norms. This was exploited towards the modeling in the context of high dimensional and functional data, see e.g. (Verleysen, 2003).

L_0: It is argued that the use of the L_0 is most appropriate for obtaining sparseness and doing input selection (Weston *et al.*, 2003). However, it results in non-convex and even NP hard combinatorial optimization problems in most cases.

L_H: An optimal trade-off between robustness and efficiency while preserving the convexity property was found in the formulation od the Huber loss-function (Huber, 1964; Andrews *et al.*, 1972).

ON: The issue that the use of L_1 norms and L_0 norms leads to sparseness in the solution vector triggered a research to how the resulting sparseness is related to the structure of the true solution. Following (Donoho and Johnstone, 1994), an oracle estimator which is defined as the minimizer of the *Oracle Norm* (ON) equals the estimator containing the true sparseness while minimizing the theorethical L_2 risk. A number of different norms were proposed (Donoho and Johnstone, 1994; Fan, 1997; Antoniadis and Fan, 2001) with corresponding inequalities bounding the deviation from the oracle estimator. Norms as the Smoothly Clipped Absolute Deviation (SCAD) were incorporated in kernel machines in (Pelckmans *et al.*, 2004, *In press*).

KL: There exist a whole range of criteria measuring the discrepancy between objects of theoretical nature as well as originating from a practical need, In general, those need not to be norms in the strict sense (not satisfying the triangularity constraint). An important example of such a measure in a theoretical probabilistic context is the Kullback-Leibler divergence (Conover, 1999) measuring the discrepancy between distributions. Recent advances in system identification result in a norm between different dynamical systems based on the cepstrum (De Cock *et al.*, 2003). Other examples include dedicated measures used in text processing, see e.g. (Joachims, 2002).

Minimax: Somewhat related to this discussion is the frequent occurence of minimax methods. Those quantify the relationship between objects in terms of a discrepancy measure and a similarity measure similarly. Those typically occur in a setting of unsupervised learning as in PCA (Jollife, 1986), a worst case analysis (El Ghaoui and Lebret, 1997; Goldfarb and IYengar, 2003) and in a transductive setting, see e.g. (Lanckriet *et al.*, 2004)

1.3 Research in Machine Learning

Apart from the central issue of inference and generalization, literature in the machine learning domain focuses on many different issues. While often motivated from practical concerns, those directions make up the field mature and lead to a globally complete set of tools for handling a wide spectrum of problems. This section is by no means exhaustive and only a selection of representative publications are cited.

1.3.1 Modeling and estimation

While the generic theory and research on learning, inference or estimation has become fairly standard, an increasing demand for algorithms building models in highly specific settings is noted. Differences in applications of the modeling paradigm can be attributed to the presence of different assortments of prior knowledge typically studied from a Bayesian perspective, see e.g. (Jaynes, 2003) for a complete account. However, prior knowledge often comes under the disguise of known noise models or known model structures which can also be incorporated using other approaches as shown in this text. Those forms often originate from the assumption of a specific generating model, see e.g. (Shawe-Taylor and Cristianini, 2004) translating these issues in the methodology of kernel machines. Consider e.g. the cases of the analysis of survival rates in observed data, see e.g. (Klein *et al.*, 1997), and the handling of longitudinal data, see e.g. (Molenberghs *et al.*, 1997).

1.3.2 Robust inference

Somewhat at the outset of theory of inference is a body of research involved with estimation problems in the context of contaminated observations. This motivated the research of a methodology which is highly robust towards the occurrence of such outliers in the observations as instantiated by (Huber, 1964), see e.g. (Andrews *et al.*, 1972). Important tools include different measures of influence and their empirical counterparts (Tukey, 1977). New contributions in this field towards the description of robust model selection criteria were described in (De Brabanter *et al.*, 2002*a*). Section 3.6 discusses some extensions of kernel machines towards this context.

1.3.3 Model selection and analysis

Analysis of the result of one individual estimator is a crucial task in the process of building a good model from observations. Given a battery of results from different estimators, the issue of model selection deals with the question which estimate is to be favorized.

Somewhat similar to the case of the mapping (1.1), one can formalize the model selection criterion as a mapping from the assumptions, the algorithm and the given observations to an estimate of the generalization performance. Note that the assumptions $\mathscr{A}$ and the algorithm Alg are frequently parameterized by a vector $\Theta = (\Theta_1, \Theta_2)$. Model selection is typically used to decide which value for Θ leads to the best performing models. Consider for example the assumption that the noise level equals σ_e^2 which correspond with a fixed regularization parameter. One typically optimizes the model selection criterion over this value σ_e^2 to let the corresponding model obtain the best possible performance:

$$\mathscr{J}_{\mathsf{Modsel}} : \mathscr{A}(\Theta_1) \times \mathsf{Alg}(\Theta_2) \times \mathscr{D} \to \mathbb{R}. \tag{1.34}$$

The task of model selection typically amounts to the following optimization problem

$$\hat{\Theta} = \arg\min_{(\Theta_1,\Theta_2)} \mathscr{J}_{\mathsf{Modsel}}(\Theta_1, \Theta_2) \tag{1.35}$$

The determination of regularization constants and other hyper-parameters as the kernel parameters is important in order to achieve good generalization performance with the trained model and is an important problem in statistics (Hastie *et al.*, 2001) and learning theory (Vapnik, 1998; Suykens *et al.*, 2003*a*). Several methods have been proposed including validation (Val) and cross-validation (CV) (Stone, 1974; Burman, 1989), generalized cross validation (Golub *et al.*, 1979), Akaike information criteria (Akaike, 1973), Mallows C_p (Mallows, 1973), minimum description length (Rissanen, 1978), bias-variance trade-off (Hoerl and Kennard, 1970), L-curve methods (Hansen, 1992) and many others. For classification problems in pattern recognition, the Receiver Operating Characteristic (ROC) curve has been proposed for model selection (Hanley and McNeil, 1982). In the context of non-Gaussian noise models and outliers, robust counterparts have been presented in (De Brabanter *et al.*, 2002*b*; De Brabanter *et al.*, 2002*a*; De Brabanter *et al.*, 2003). Translation of a priori knowledge (e.g. norm of the solution, norm of the residuals or the noise variance) into an appropriate regularization constant has been described respectively as the secular equation (Golub and van Loan, 1989), in Morozov's discrepancy principle (Morozov, 1984) and (Pelckmans *et al.*, 2004*d*). In the specific context of kernel machines amongst others (Chapelle *et al.*, 2002) proposed criteria with bounds on the generalization error based on geometrical concepts (VC bounds, optimal margin and support vector span (Schölkopf and Smola, 2002)) to determine the regularization constant. A bound based on the leave-one-out cross-validation error was introduced in (Kearns, 1997). Bounds on the generalization error with analysis of the approximation and sample error were investigated in (Cucker and Smale, 2002). Efficient methods for calculating the leave-one-out cross-validation criterion for some kernel algorithms based on the matrix inversion lemma were described e.g. by (Van Gestel *et al.*, 2002; Cawley and Talbot, 2003). In general, the optimization of criteria for determination of unknown regularization constants often leads to non-convex optimization (or even non-smooth) and computationally intensive schemes (depending on the model selection scheme). In (Chapelle *et al.*, 2002) the determination of the tuning parameter is determined via solving alternating convex problems. Related research can be found in the literature about learning the kernel, see e.g. (Herrmann and Bousquet, 2003; Lanckriet *et al.*, 2004).

One of the most tempting and active research tracks in the statistical science and in machine learning is concerned with the question which inputs may/should or can be used in order to explain or predict optimally the observed dependent variable. Let $I \in \mathbb{R}^{D\times D}$ be a diagonal indicator matrix $I = \operatorname{diag}(\iota_1, \ldots, \iota_D)$ with $\iota_d \in \{0,1\}$ for all $d = 1, \ldots, D$. Let $\ell(f, \mathscr{D})$ denote generically a suitable measure for the performance of a function f on a dataset $\mathscr{D}$ with N observations (x_i, y_i). Then the input selection problem may be formalized as the problem of selecting an appropriate matrix I such that the corresponding estimate

$$\hat{f}_I = \arg\min_f \sum_{i=1}^N \ell\left(f(Ix_i) - y_i\right) \quad \text{s.t.} \quad f \in \mathscr{F}, \tag{1.36}$$

optimizes a suitable model selection problem. The method of Analysis Of Variance (ANOVA) constitutes a body of research on this topic in the dedicated case of linear parametric models satisfying the Gauss-Markov equations. Hypothesis tests make up the primary tools of the ANOVA practitioner, see e.g. (Neter *et al.*, 1974). The research on input selection for non-parametric models more shifted towards the regularization paradigm (Girosi *et al.*, 1995), especially since the advent of sparse regularization criteria in the form of LASSO (Tibshirani, 1996), SURE (Donoho and Johnstone, 1994) and basis pursuit (Friedman and Tukey, 1974; Friedmann and Stuetzle, 1981; Chen *et al.*, 2001), see Subsection 6.1.2.

1.3.4 Structured data and applications

Although the initial theory was restricted to one of the most simple problems of binary classification of numerical vectors, extension of the methodology and the analysis towards other data structures constitute now a full body of literature. These investigations were largely driven by specific case studies.

OCR Initial research on SVMs was driven by the problem of Optical Character Recognition (OCR) which triggered the research on fast (approximative) implementations and on the incorporation of invariances (as rotations are translations of the image) in the learning machine (Decoste and Schölkopf, 2002).

Text This type of application driven research was somewhat pioneered by the literature on text mining using SVMs and kernel methods. Results and different applications are surveyed in (Herbrich, 2001; Joachims, 2002). This body of literature relies heavily on the formulation of appropriate distance measures defined on strings, graphs and trees. Typical tasks include the automatic classification of web adresses (URLs) and the identification of unsolicited e-mail (spam).

Generative Models It is often the case that one has some kind of prior knowledge of the process generating the observations. For example DNA sequences have been generated through evolution in a series of modifications from ancestor sequences. This information in the form of invariances, features or distances that we expect it to contain may be used to tune the learning algorithm to the specific task. The discussion on this topic mainly concentrates on the design of an appropriate kernel, amongst which the probabilistic models leading to the so-called *p*-kernel and the Fisher kernel, see e.g. (Shawe-Taylor and Cristianini, 2004) for an overview. A noteworthy contribution in this context is (Bach and Jordan, 2004), applying this mechanism towards the characterization of time-series.

While previous methods rely on the derivation and construction of appropriate distance measures and equivalent kernels, many applications require a more elaborate modification to the learning machine itself.

Identification of Nonlinear Systems The case where the observations are a sequence sampled over time is generally coined as system identification. Initial examples of the application of kernel methods to system identification tasks and nonlinear time series analysis were given by (Mukherjee *et al.*, 1997; Mattera and Haykin, 2001; Müller *et al.*, 1999). A first approach towards the problem of non-linear control using kernel methods was coined in (Suykens *et al.*, 2001).

New results on the fitting of nonlinear time time-series were discussed in (Fan and Yao, 2003; Dodd and Harris, 2002). Further investigations on the topic concentrated more via the closely related Gaussian Processes, see e.g. (Kocijan *et al.*, 2003). The identification task of black-box models from input and output data was investigated by the author and others in (Goethals *et al.*, 2005*a*; Goethals *et al.*, 2004*b*; Goethals *et al.*, 2004*c*), combining linear subspace identification techniques (Vanoverschee and De Moor, 1996) with kernel based LS-SVMs, see also (Suykens *et al.*, 2002*b*).

Bio-informatics The field of kernel methods found a successful application area in the field of bio-informatics. This research is concerned with the integration of mathematical, statistical, and computer methods to analyze biological, biochemical, and biophysical data. The field of Bio-informatics, which is the merging of molecular biology with computer science, is essential to the use of genomic information in understanding human diseases and in the identification of new molecular targets for drug discovery. Investigations typically concern the processing of data from micro-array experiments representing the gene expression coefficients corresponding to the abundance of mRNA in a sample. A collection of results sampling the ongoing research on the topic using kernel machines can be found in (Schölkopf *et al.*, 2004). Recent advances using LS-SVM based approaches are published in (De Smet, 2004; Pochet *et al.*, 2004).

Other applications where described in various survey works including (Schölkopf *et al.*, 2001; Suykens *et al.*, 2002*b*; Shawe-Taylor and Cristianini, 2004) and others.

1.3.5 Large datasets and online estimation

With the advent of fast computers and cheap measurement devices, an ever growing collection of data is available. Mining for knowledge in this flood is not only a theoretical quest but also requires adapted numerical methods to get informative results in a reasonable time interval. Let N be the size of the training set. Large scale algorithms may be categorized in one of the following classes, where the size constraints are only indicative. This small overview follows the survey (Hamers, 2004).

Numerical $(2,000 < N < 20,000)$ In case the size of the dataset to be analyzed is not overwhelming, one often can formulate computationally tractable algorithms to compute the exact estimate. Consider e.g. the case where dependencies have a strictly local character. In case one does not need an explicit global model description but only a number of predictions on given data-points, fast counterparts may be formulated. This idea was applied in the framework of localized wavelets (Daubechies, 1988) and later exploited in the context of kernels (Genton, 2001; Hamers, 2004). For an overview of efficient numerical algorithms for large scale applications, see e.g. (Golub and van Loan, 1989) and (Van Dooren, 2004). Iterative approaches as the Krylov subspaces often lead to a less memory intensive approach and applied in (Suykens *et al.*, 2002*b*;

Hamers, 2004). Methods for the trading the accuracy of the solution for speed are generally based on low-rank approximations. A classical result there is the Sherman-Morisson-Woodbury formula described in the field of Fredholm equations, see e.g. (Press *et al.*, 1988), and the Nÿstrom low rank approximation, see e.g (Suykens *et al.*, 2002*b*) for its application on LS-SVMs.

Decomposition techniques $(10,000 < N < 50,000)$ In case the dataset is even too large to process in batch, a recursive approach may be advocated. Here the assumption is that the model provides an effective representation of the optimal solution thus far and a relatively simple updating rule is available to update the optimal model with respect to a new chunk of data. This approach is quite popular in the case of SVMs, denoted as chunking (Vapnik, 1998) and in the case of one-sample chunks as sequential minimal optimization (SMO) (Platt, 1999). Another noteworthy approach goes under the name of Successive Over-relaxation (SOR) (Mangasarian and Musicant, 1999).

Sampling $(N > 20,000)$ When an overwhelming amount of data is available which would saturates the memory of the computer as well as the monopolizes the cpu far too long, one may still obtain sensitive results by using an appropriate sampling mechanism. While statistical literature has a long tradition in sampling schemes (Rubinstein, 1981), the application towards kernel methods is still premature. A notable effort was described using a Renyi entropy based sampling mechanism (Girolami, 2002) and combined with Nÿstrom low rank approximation to highly workable and efficient algorithm under the name of fixed size LS-SVM, see (Suykens *et al.*, 2002*b*; Espinoza *et al.*, 2004).

Ensembles $(N > 20,000)$ Another class of practical algorithms in the case of large scale estimation constitute of committees of submodels each based on a subsample of the data. These go under the name of fancy names as bagging (Breiman, 1996), boosting (Rätsch, 2001) and others, see e.g. (Bishop, 1995).

Recursive Estimation Recursive extensions to the LS-SVM formulation and the closely related kernel PCA based on tracking the dominant eigenspace of a kernel matrix growing simultaneously in the number of rows and the number of columns are proposed and benchmarked in (Hoegaerts, 2005).

Hardware $(N > 20,000)$ The last decade witnessed an emergence of the research on analog implementations of data processing techniques as neural networks and associative memories, see e.g. the special issue of IEEE Transactions on Neural Networks, vol 4, number 3, may 1993. In line with this field, efforts were made to port the formulation of SVMs (Anguita *et al.*, 2003) and LS-SVMs (Anguita *et al.*, 2004) to hardware implementations often enabling the fast processing of huge datasets.

Database When the size of the collection of observations grows unboundedly, the problem how to organize and memorize the samples becomes increasingly important. This problem forms a major concern in the computer science part of

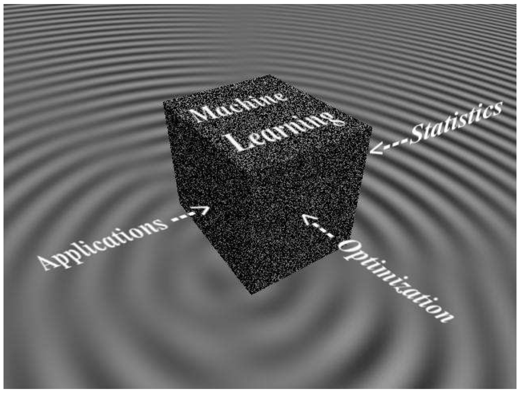


Figure 1.3: *This research on machine learning and kernel machines is driven by stimuli from convex optimization theory, various application areas and the issues raised during development of LS-SVMlab and results in the area of classical statistics.*

the research in machine learning and artificial intelligence. For a general starting point, see e.g. (Bertino *et al.*, 2001).

1.4 Contributions

The Ph.D. research of the author can be summarized from various perspectives. In order to overview the main advances, we divide into four different categories (1) published contributions which are surveyed in the present dissertation, (2) new research results which complete the dissertation and enhance the streamline of the text, (3) published research results which are not described explicitly in the present text as they do not fit into the main pressented story, (4) other forms of contributions of the research of the author as the development and support of the toolbox LS-SVMlab.

The synthesis of the Ph.D. research assimilated in the dissertation is twofold:

α-γ-σ The main structure of the text reflects the hypothesis that the questions concerning the optimal learning algorithm ("α"), the best regularization trade-off ("γ") and the characteristics of the smoothing kernel ("σ") are interrelated in many possible ways (see Figure 1.4) and should be addressed together.

Primal-Dual Argument The second hypothesis which is motivated throughout the thesis argues that the primal-dual argument based on convex optimization theory is not an ad hoc methodology, but can be centralized as a most powerful tool for the design of new kernel machines. Moreover the method is presented as a valuable alternative to the parametric modeling strategy (Figure 1.5 illustrates both methodologies).

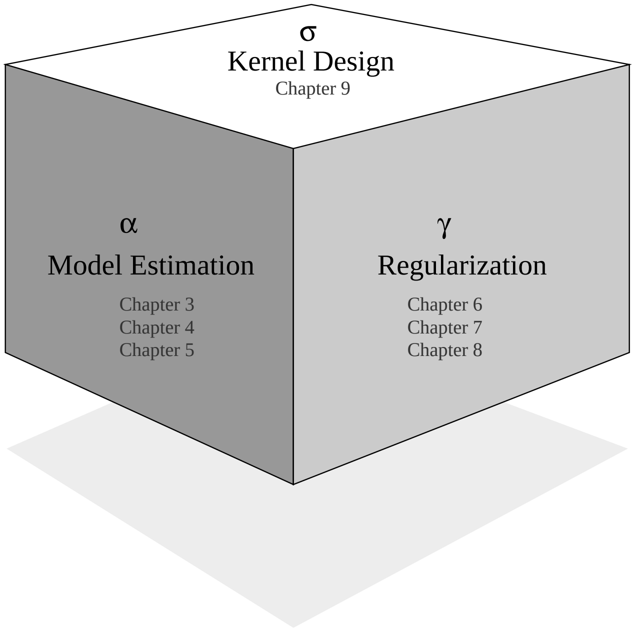


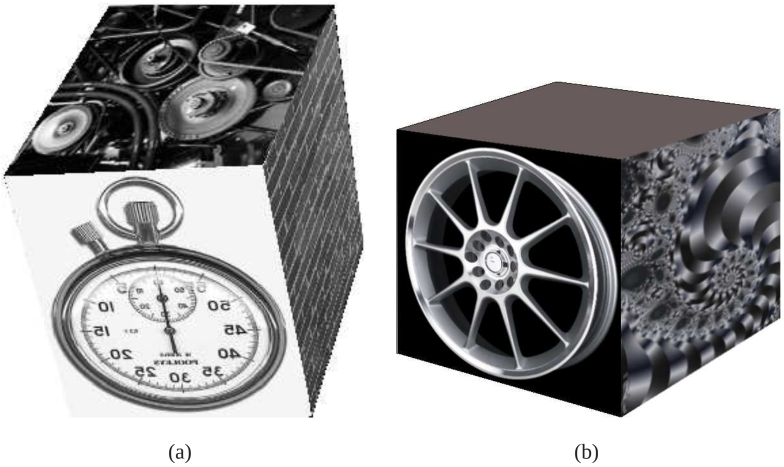
Figure 1.4: *The main theme of the text manifests itself in three interrelated ways. Part α studies the design of primal-dual kernel machines and extends results towards the incorporation of extra structure in the modeling process itself. Part γ then discusses the issue of regularization and its relation to imposing structure. An important advance in that context is made in the formulation of a methodology to automate model selection and tuning the regularization trade-off. Part σ finally discusses the relationship between regularization and the design of kernels and proposes an approach assisting the user in the choice of an appropriate kernel.*

This work mainly builds on tools and results in (Suykens *et al.*, 2002*b*; Boyd and Vandenberghe, 2004; Vapnik, 1998; Wahba, 1990) and takes essentially an optimization perspective towards the construction of new learning algorithms.

1.4.1 Contributions: published and in the dissertation

The text is built around a set of original results obtained by the author during the Ph.D. work. Only a subset of the published results are discussed in some detail to preserve a consistent story.

Hierarchical programming problems Multi-objective optimization problems are typically approached using a Pareto or scalarization approach. The hierarchical programming approach takes a different approach by not solving for the joint



(a)

(b)

Figure 1.5: *The research of primal-dual kernel machines inherits ingredients of the* **(a)** *parametric modeling paradigm (represented by the cube including the clock watch) and the non-parametric paradigm constituting of a series of individual tools.* **(b)** *The primal-dual framework (represented as the cube on the right) is a coherent approach towards many modeling tasks. While the inner mechanism is rather complex, the use of the method is rather intuitive (as e.g. the wheel). More specifically, a primal-dual model has simultaneously a primal (parametric) and a dual (non-parametric) representation.*

multiple objectives, but they do consider instead the different cost-functions at a different level. A typical occurrence of such a problem is found in the task of automatic model selection. This view was introduced in (Pelckmans *et al.*, 2003*b*) and further elaborated in (Pelckmans *et al.*, 2004*e*; Pelckmans *et al.*, 2004*c*; Pelckmans *et al.*, 2005*c*; Pelckmans *et al.*, 2004*b*).

Primal-dual Kernel Machines Many new learning machines based on kernels make use of results in convex optimization theory. This motivates the definition of a very broad class of machines where the primal-dual argument is put central. Important instances are then found as the SVMs and the LS-SVMs. This view follows directly from the work (Suykens *et al.*, 2002*b*). This perspective was taken as the main tool for designing new kernel machines in most publications of the author. Figure 1.6 gives a schematic overview of the presented research on primal-dual kernel machines.

Structured Primal-Dual Kernel Machines The primal-dual argument is elaborated as a strong tool for incorporating prior knowledge in the learning task. We studied prior knowledge in the form of modelstructure as estimating additive

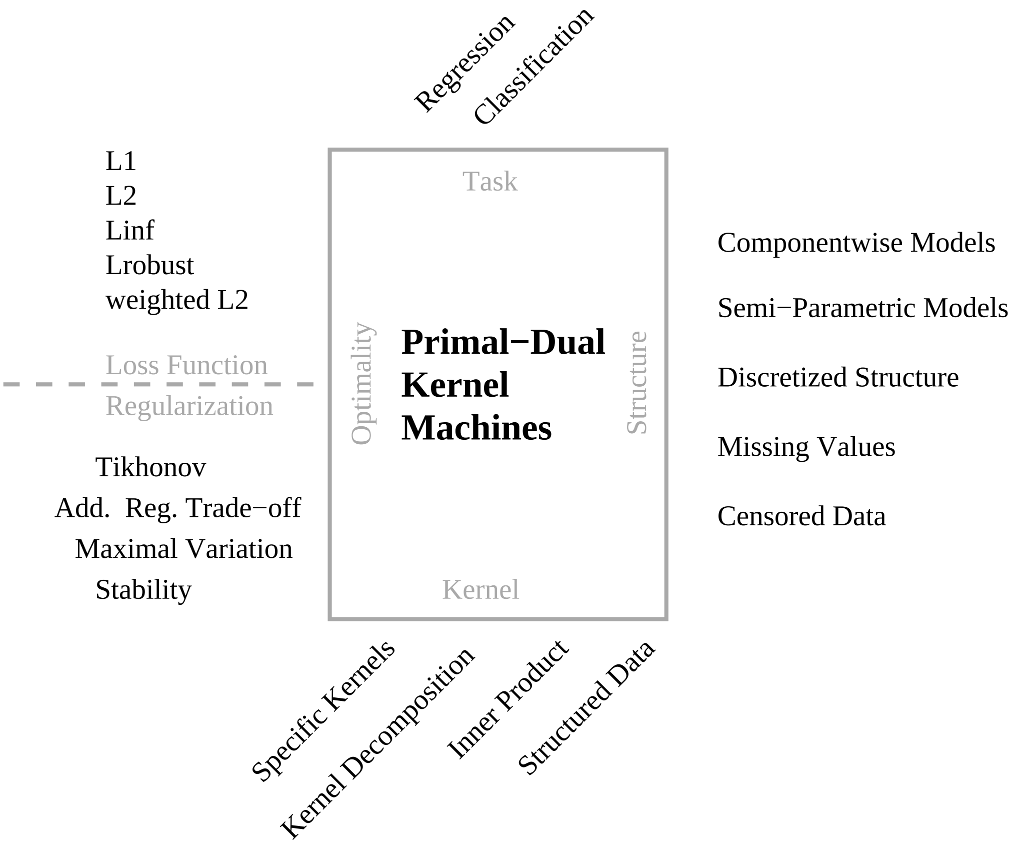


Figure 1.6: *Contributions on primal-dual kernel machines as presented in this text can be organized as illustrated. The different issues of the study of optimality in different tasks, the exploitation of structure in the learning process and the study of the role of the kernel correspond roughly with the different parts and chapters.*

models (Goethals *et al.*, 2005*a*; Goethals *et al.*, 2004*b*; Pelckmans *et al.*, 2004, *In press*; Pelckmans *et al.*, 2005*c*; Pelckmans *et al.*, 2005*b*; Pelckmans *et al.*, 2005*e*), semi-parametric models, learning in the context of given inequalities (Pelckmans *et al.*, 2004*g*) and others.

Advances in regularization or complexity control Somewhat central into the theory and practice of primal-dual kernel machines as well as SVMs is the issue of complexity control or regularization. Two new regularization schemes and their relation with the classical Tikhonov regularization were studied (Pelckmans *et al.*, 2004*d*). A main result is the formulation of the one-to-one relation between noise level and the regularization constant in LS-SVMs.

Differogram and estimators for the noise level A different approach towards the task of model selection and determining the regularization trade-off was initiated in (Pelckmans *et al.*, 2003*a*). Here, the noise level was put forward as a single parameter controlling the necessary amount of smoothing to be applied on the data. In order to estimate this parameter from observations, a data representation constituting of all mutual differences between observations was proposed. This so-called differogram cloud contains information on the second-order moments and the variance present in the data. The differogram method and various applications towards the task of model selection were further studied in (Pelckmans *et al.*, 2004*a*), together with extensions to robust estimators and spatio-temporal data.

Maximal Variation and structure detection New advances for structure detection for componentwise kernel machines were based on similar principles as the LASSO estimator in the linear parametric case. Here an appropriate regularization scheme is designed to detect components in the final predictor which do not contribute actively. The main difference is that structure detection does not follow from the sparseness of the parameters itself, but from the total amount a specific component variates over the training set, i.e. contributes to the model on the given dataset. Hereto, a measure of total variation (Pelckmans *et al.*, 2004, *In press*) and maximal variation (Pelckmans *et al.*, 2005*c*) was used (Pelckmans *et al.*, 2005*e*).

Kernel machines for handling missing data A recent result was achieved for handling missing values amongst the data observations. The handling of partially missing observations is approached by using additive models. A worst-case approach was taken in (Pelckmans *et al.*2005*c*) based on the measure of maximal variation. This research was elaborated in (Pelckmans *et al.*2005*b*) where the worst-case approach was contrasted to a method based on a modified empirical risk functional.

Fusion and automatic model selection The problem of model selection gained a crucial status into the theory and especially in the practice of applicability of linear and nonlinear learning algorithms. Past research of the author focussed especially on the optimization aspect: given a model selection criterion, how to optimize this criterion on the dataset. Though such a problem are in many cases computationally hard, appropriate relaxations can be devised (Pelckmans *et al.*, 2003*b*; Pelckmans *et al.*, 2004*b*).

Additive Regularization Trade-off and LS-SVM substrates An efficient approach to the problem of automatic model selection was studied in (Pelckmans *et al.*, 2003*b*) by using an appropriate re-parameterization of the hyper-parameter under study. This paper considered regularization trade-off tuning with respect to validation and cross-validation.

Hierarchical kernel machines and stable learning machines It was argued in (Pelckmans *et al.*, 2004*e*; Pelckmans *et al.*, 2005*c*) that the formulation of additive regularization trade-off could be used to emulate the use of slightly different optimality criteria while inheriting the main advantages of LS-SVM

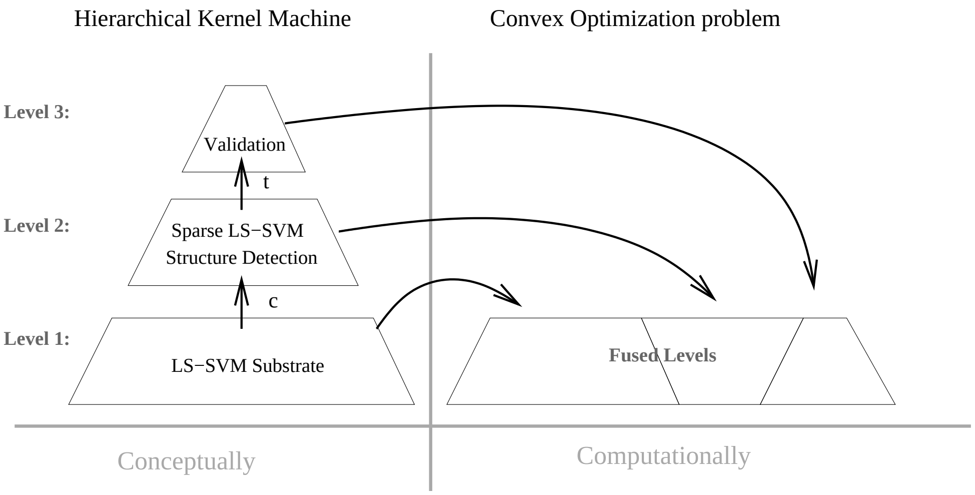


Figure 1.7: *Illustration of the idea behind hierarchical kernel machines. On the conceptual level, different hierarchical levels are formulated, each with their own optimality principles and free variables. Computationally, all corresponding conditions for optimality are fused into one constrained optimization problem.*

formulations. This led to the concept of hierarchical kernel machines. A special instance was described where algorithmic stability was maximized during learning itself (Pelckmans *et al.*, 2004*c*). Figure 1.4.1 gives a schematical representation of such a hierarchical kernel machine. In (Pelckmans *et al.*, 2004*c*), the use of a representation similar as the *L*-curve was elaborated, displaying information on the trade-off between empirical performance and stability.

1.4.2 Contributions: new results in the dissertation

A variety of new results were added to bridge the gaps and to glue the main results together. We emphasize the following results.

Positive OR constraints A first new contribution is the formulation of a specific kind of quadratic constraints, denoted as positive OR constraint stating that at most one of two positive variables may be non-zero. This type of constraints often occur in hierarchical programming problems. It is shown that this kind of constraints may often be embedded in a quadratical programming problem without losing the global property of convexity.

Sensitivity interpretation The perspective of convex optimization theory towards the construction of learning machines reveals a strong relation between the dual model representation and the sensitivity of the estimate to given observations.

Support Vector Tubes and ν-Support Vector Tubes In addition to the standard kernel machines, we studied a new formulation built for the task of predicting intervals for given covariates. This leads to a non-parametric generalization of quantile interval estimators. A robust version turns out to correspond largely to a ν-SVM and is called the ν-SVT.

Efficient iterative algorithm for semi-parametric LS-SVMs and robust SVMs In addition to the sound formulation of the structured and robust kernel machine given in Section 4.1 and Subsection 3.6.1, an efficient algorithm is elaborated for calculating the estimate in the case of large datasets.

Kernel machines for handling censored data The mentioned results were employed to design a primal-dual kernel machine capable of handling observations which are censored. Censoring can occur due to sensor limitations or other physical phenomena as an unexpected failure of the data sample.

Relation semi-parametric LS-SVMs and generalized Least Squares regression In addition to the relations of the LS-SVM with other well known techniques as regularization networks, smoothing splines and others, the relationship with the standard generalized least squares estimator is noted.

Alternative Least Squares A new result is stated in the context of linear parametric models advancing the popular practice of LASSO estimators. The alternative least squares method results in an estimator making use of only one single input variable among the proposed alternatives.

Bias-variance trade-off for LS-SVMs The classical study of the impact of regularization in bias and variance in the context of linear ridge regression is migrated to a context of nonlinear kernel models. The main difference is that bias and variance are not expressed in terms of the parameters but in the prediction itself.

Fusion of ridge-regression and stepwise regression with validation The task of automatic model selection using the hierarchical programming approach is applied to the task of learning the regularization trade-off and input selection in ridge-regression and least squares respectively. Appropriate convex approximations to the problem are described resulting in a practical and efficient approach of model selection in those cases.

Plausible Least Squares The formulation of plausible least squares illustrates how one can use the fusion argument beyond the context of classical model selection. Instead the use of a significance test is embedded into an estimation problem. Given the sample distribution of the parameter estimation using a resampling procedure, plausible least squares estimates the least complex parameter vector (in L_1 sense) which cannot be rejected given the samples.

Fusion of LS-SVMs and SVMs with validation Similar formulations are derived for selection of the regularization trade-off in SVMs and LS-SVMs. A relaxation to the former is elaborated resulting in fast and reliable estimates of the regularization trade-off solving a convex problem.

A modified loss function approach to additive regularization The additive regularization trade-off is seen to provide an efficient and convex approach towards the task of model selection in ridge regression and LS-SVMs. A different perspective towards this scheme is given where the trade-off expresses local modifications to the loss function.

Relation weighting schemes and model structure with kernel design This dissertation reports new advances in the study of good kernel designs. We state results relating specific weighting schemes of errors and regularization, and model structures with the form of the kernel. Those results are proven using tools from optimization theory.

Kernel decompositions and structure detection A practical method for detecting appropriate kernel designs given a finite set of alternatives is formulated related to the method of structure detection using the measure of maximal variation.

Realization approach to kernel design The relation of smoothing kernels with smoothing filters is used to design a technique to derive the form of the kernel from the data observations itself. The implicitly used criterion for selecting the kernel is based on the sample covariance in the data. In correspondence to classical stochastical realization theory, the technique is build on a matrix decomposition of the sample covariance matrix.

Various new examples give a theoretical or practical support of the concerning elaboration. We especially spent some effort to illustrate the usability of the studied results.

A χ^2 density estimator Given the formulation of second order cone programming problems, a probability density estimator is formulated which builds on the classical result of histosplines but uses a more appropriate χ^2-measure instead.

Learning machine based on Fourier feature space map In order to make the concept of the feature space map less mysterious, a concrete mapping is studied where data samples are mapped onto the corresponding Fourier coefficients. Furthermore, it is shown that the application of a low-pass filter on the estimate corresponds with the use of the classical RBF kernel. Though relying heavily on published results, the context of this example in primal-dual kernel machines and the employed techniques are original.

Learning machine based on Wavelet feature space map Equivalently, an explicit feature space mapping is based on the wavelet decomposition, showing that results on wavelets can easily be migrated to a context of kernel machines and SVMs.

A robust location estimator based on the modified loss function approach The modified loss function interpretation to additive regularization trade-off is used to design a robust location estimator. The modifications to the classical empirical mean based on a least squares estimator are determined using the technique of the quantile-quantile plot. We exploit the classical result that a linear relation of the theoretical and empirical quantiles implicates a Gaussian distribution.

Kernel machine for handling colored noise schemes Most results rely (at least in theory) on the property of i.i.d. of the data-samples. This example shows however that one can design kernel machines with the noise following a known coloring scheme by using the primal-dual argument.

Modeling discontinuities It is illustrated how one can incorporate a finite set of known discontinuities in the estimates using semi-parametric primal-dual kernel machines. This example is extended to the task of learning where an infinite set of discontinuities can be modeled by building a partially explicit feature space mapping.

Relation RBF-kernel and AR(1) representation A classical result concerning autoregressive models of first order and the convolution with an exponential function is interpreted into a kernel context. This example illustrates the equivalence between prediction with smoothing filters and modeling with smoothing kernels.

1.4.3 Contributions: Ph.D. research

During the doctoral research active contributions were made to various related fields. The following contributions are only marginally touched in the dissertation as they do not fit straightforwardly into the presented story.

Robust Model Selection criteria Robust inference is concerned with the task of estimation and prediction in the context of atypical observations or outliers. Contributions to the literature in this field were made by formulating robust model selection criteria together with a theoretical as well as practical assessment of their performance. Robust cross-validation measures were described in (De Brabanter *et al.*, 2002*b*) and extensions of different information criteria as Akaike's were described in (De Brabanter *et al.*, 2003). The report (De Brabanter *et al.*, 2002*a*) discusses the robust model selection criteria in more detail. The extension of the robust kernel based methodology towards the estimation of nonlinear ARX models in the context of outliers was discussed in (De Brabanter *et al.*, 2004). Here, various new tools as nonlinear influence functions and empirical assessment of the robustness of nonlinear methods were proposed. More details may be found in the dissertation (De Brabanter, 2004).

Identification of nonlinear systems A fruitful field for research on learning in the context of known structure was found in the literature on non-linear system identification. The high potential of this cross-fertilization was shown in (Espinoza *et al.*, 2004) where a generic primal-dual kernel method was shown to perform very well on a benchmark dataset denoted as the *Silverbox Data* consisting of a real-life nonlinear system (Schoukens *et al.*, 2003). Further advances for the identification of general problems where reported in (De Brabanter *et al.*, 2003; De Brabanter *et al.*, 2004) where robustness issues are studied with respect to model selection of nonlinear ARX problems and of the identification task itself using LS-SVMs respectively.

Identification of Hammerstein and Hammerstein-Wiener systems A further contribution was made in this direction by the construction and study of learning algorithms for the identification of Hammerstein models consisting of a sequence of a non-linear static model and a linear dynamical system. The publications (Goethals *et al.*, 2005*a*) and (Goethals *et al.*, 2004*a*) study this task by combining a primal-dual formulation succeeded by a linear Auto-Regressive model with eXogenous variables (ARX). While the method ressembles the classical over-parameterization technique, new elements were introduced in the form of model complexity control or regularization (Pelckmans *et al.*, 2005*a*) and a primal-dual argument enabling a very broad and flexible representation of the nonlinear model. In (Goethals *et al.*, 2004*b*), extension are studied to the classical N4SID subspace identification method towards the identification of Hammerstein models where the nonlinearity is again represented as a kernel machine. The subspace intersection method was employed towards the identification of Hammerstein-Wiener systems consisting of a sequence of a static nonlinearity, a linear dynamic model and again a nonlinear static function, see (Goethals *et al.*, 2004*c*) and (Goethals *et al.*, 2005*b*). A thorough discussion of the subject may be found in the Ph.D. dissertation (Goethals, 2005).

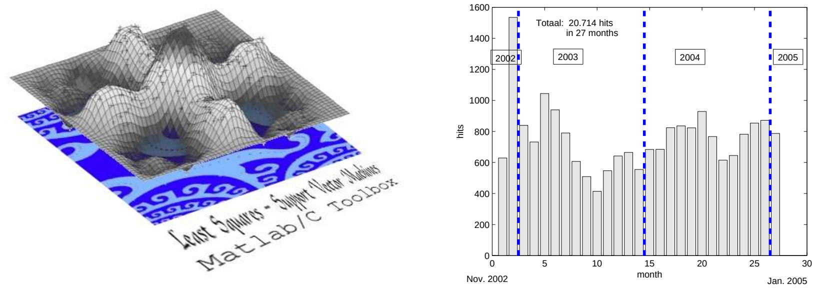
1.4.4 Contributions: other output

LS-SVMlab

During the start of the research, we concentrated on a Matlab/C implementation of the algorithms related to LS-SVMs. The methodology was embodied into a toolbox called LS-SVMlab which can be found at

http://www.esat.kuleuven.ac.be/sista/lssvmlab/

including a full tutorial (Pelckmans *et al.*, 2002*a*). A demonstration was presented at NIPS 2002 (Pelckmans *et al.*, 2002*b*). The toolbox includes extensions to multi-class classification tasks, Bayesian interpretation, adequate preprocessing, model selection and model tuning, handling of large scale algorithms, unsupervised learning tasks and other. More details on the update are given in Section B.1. Figure 1.8.b reports some measures of the impact of this toolbox. The goal of this toolbox was the practical support of the (Suykens *et al.*, 2002*b*). The toolbox was used e.g. in the project SOFT4s regarding software simulators for replacing expensive sensors (De Moor *et al.*, 2002) and in various publications as (Espinoza *et al.*, 2004; Pochet *et al.*, 2004) and others.



(a) (b)

Figure 1.8: **(a)** *Main theme of the LS-SVMlab website.* **(b)** *Number of visits of the website. The number of downloads of the toolbox in the 27 months of existence equals* 11.581. *This may be compared with the approximate* 500.000 *hits of the classical website* http://www.kernel-machines.org *and the approximate* 27.000 *visits of the LS-SVMlab site.*

Industrial Projects

During the Ph.D, the author collaborated in two industrial projects:

Soft4S In the context of the chemical process industry, the monitoring of the details of a process can be expensive due to very expensive sensors or the need for time-consuming manual investigation of chemical samples. The aim of the Soft4s project is to develop a simulator of such a sensors based on a series of less-expensive measuring sensors. The main contribution of the author in this project was the application of the software LS-SVMlab for this goal. Other advances were reported including the application of Bayesian input selection, handling of huge datasets and modeling of dynamic behaviour of the process under study, see (De Moor *et al.*, 2002) for more details.

ELIA The other project concerns the forecast of expected electricity consumption on various locations. An important application of LS-SVMs was found in the modeling on the dependence of load on the daily temperature. Further concerns were the occurence of periodical variations, nonstationarities and clustering of different stations.

1.4.5 Chapter-by-chapter overview

The main theme of the text manifests in many interrelated ways each discussed in the four chapters. Figure 1.4 highlights the global setup of the dissertation.

Introduction Part I discusses the general setting of the research and introduces a set of definitions useful in the remainder of the text.

$\boldsymbol{\alpha}$ Part II studies the formulation and properties of primal-dual kernel machines in some detail. The character α refers to the common symbol of the dual representation of the modeling technique.

$\boldsymbol{\gamma}$ Part III examines the impact of the concept of complexity control or regularization in the construction of algorithms. The Greek symbol γ refers to the typical trade-off between complexity and empirical performance by the regularization constant in the studied modeling strategies.

$\boldsymbol{\sigma}$ Part IV discusses the impact of the shape and the properties of the employed kernel and proposes various methods to assist the user in the choice of an appropriate kernel. The symbol σ refers to the typical parameter also called the bandwidth determining the amount of smoothness of the final estimate via the kernel.

Finally, a number of conclusive remarks and directions towards future work are described.

Part I, chapter 1: Problems and Purposes

The first chapter presents an overview of a number of principles lying at the core of the process of induction of mathematical models from a finite set of observational data. Section 1.1 discusses the general setting of learning from data or induction, while Section 1.2 survey the various approaches which give a sound foundation for doing so. Section 1.3 synthesizes a brief overview of the various directions of the current research in machine learning using kernel methods.

Part I, chapter 2: Techniques from Convex Optimization Theory

As motivated in the previous chapter, the following text will essentially take an optimization point of view. Moreover, convex optimization theory gives rise to the primal-dual argument explored in this work. The following chapter reviews some important results from the theory and discusses the renewed interest for convex optimization.

The first Section surveys a number of definitions which are necessary for a clear exposition of the subject. More specifically, the reach of the theory of convex optimization problems is properly defined. Section 2.2 then reviews the machinery of dual problems in the sense of Lagrange. Section 2.3 discusses the problem from a more practical point of view, while in Section 2.4 a number of useful extensions are reported. Subsection 2.4.4 specifically introduces then the problem of hierarchical programming.

Part II, chapter 3: Primal-Dual Kernel Machines

This chapter presents an overview of the application of the primal-dual optimization framework to the inference of regression functions and classification rules from a finite set of observed data-samples. The aim of the chapter is to provide a sound and general basis towards the design of algorithms relying on the theory of constrained optimization. While historical breakthroughs mainly focussed on the case of classification, this chapter mainly considers the regression case.

Section 3.2 discusses general parametric and classical kernel based methods, while Section 3.3 studies one of the most straightforward formulations leading to the standard Least Squares Support Vector Machine (LS-SVM). This formulation is studied in some detail as it will play a prototypical role in the remainder. Section 3.4 then proceeds with the derivation of the Support Vector Machine (SVM) for regression. Section 3.5 gives a variation on the theme by proposing a primal-dual kernel machine for interval estimation, coined as the Support Vector Tube (SVT). Section 3.6 considers a number of extensions of the previous methods to the context of outliers, and Section 3.7 reports a number of results in the context of classification.

Part II, chapter 4: Structured Primal-Dual Kernel Machines

It is common intuition that the incorporation of prior knowledge into the problem's formulation will lead to improvements of the final estimate with respect to naive applications of an off-the-shelf method. The following chapter shows the flexibility of the primal-dual optimization framework for incorporating this knowledge into the estimation problem.

While extensive discussions and analysis are far beyond the scope of this text, the relevance of this chapter is found in the fact that the remainder of the treatise and some commonly formulated commentaries on the method frequently touch on these subjects.

Various types of structural information are considered, including semi-parametric model structures (Section 4.1), additive models (Section 4.1), pointwise structure (Section 4.1) in the form of inequalities and its extension towards handling censored observations (Section 4.1).

Part II, chapter 5: Relations with other Modeling Methods

This chapter takes the opportunity to frame the preceding discussion into a broader context and to review various related approaches. While differences were mainly conceived in the conjectured assumptions and the way of deriving the results, the final formulations frequently present many correspondences. However, different interpretations of the results seem to support the coexistence of the individual approaches.

Methods close to the formulation of LS-SVMs include different variational approaches as smoothing splines (Section 5.1), the approach of Gaussian processes (Section 5.2) and Kriging methods in the context of spatial analysis (Section 5.3). Relationships with other methods methods as system-identification, wavelets, the theory of inverse problems and the weighted least squares approach are described in Section 5.4.

Part III, chapter 6: Regularization Schemes

Capacity control or regularization amounts to the artificial shrinkage of the solution-space in order to obtain increased generalization. This topic re-occurs under many disguises and in many domains. The purpose of this chapter is both to motivate, to analyze and to include regularization schemes in the process of model estimation.

Section 6.1 surveys results in the context of linear parametric models. Section 6.2 extends the results on the bias-variance result for LS-SVMs for regression. Section 6.3 extends the classical regularization scheme in primal-dual kernel machines to various other classical schemes. The measure of maximal variation for componentwise models was introduced in Section 6.4 and various applications of this idea are presented.

Part III, chapter 7: Fusion of Training with Strong Measures

The amount of regularization is often determined by a set of constants which should be set by the user a priori. The (meta-) problem of setting those is often classified as a problem of model selection and considered as being solved. However, a procedure for the automatic optimization of these hyper-parameters given model selection criterion and model training procedure is highly desirable, at least in practice. This chapter unfolds a framework for this purpose based on optimization theory.

Section 7.1 introduces the problem and the proposed solution towards it. Various applications of this issue towards model selection problems in linear parametric models are given. Section 7.2 studies the problem of model selection in the case of LS-SVMs and SVMs.

Part III, chapter 8: Additive Regularization Trade-off Scheme

This chapter elaborates on the results of the previous chapter, but rather takes a different approach towards the problem of fusion. Instead of considering existing training procedures, a flexible formulation employing an additive regularization trade-off scheme is taken as the basis for fusion. The resulting substrate is found much easier to proceed with whenever more complex model selection criteria are involved.

The basic ingredients are introduced in Section 8.1 and various relations are discussed. Section 8.2 then proceeds with the study of the fusion argument in the context of an LS-SVM regressor with additive regularization trade-off. Furthermore, the concept of an hierarchical kernel machine is introduced, leading to the construction of kernel machines maximizing their own stability (Section 8.3).

Part IV, chapter 9: Kernel Parameterizations and Decompositions

The generalization performance of kernel machines in general often depends crucially on the choice of the (shape of the) kernel and its parameters. The following chapter shows the relationship between the issue of regularization and the choice of the kernel. Furthermore, the idea of kernel decompositions is proposed to approach the problem of the choice of the kernel. Finally, relations with techniques from the field of system identification are elaborated. Given observed moments, the task of stochastic realization amounts to finding those internal (kernel) structures effectively realizing this empirical characterization. This results in a tool which can assist the user in the decision for a good (shape of the) kernel.

Section 9.1 and Section 9.1.3 introduce a formal argument relating the regularization scheme and a weighting term in the loss function respectively with the form of the kernel using a primal-dual argument. Then Section 9.2 proceeds with the elaboration of a method for searching compact kernel decompositions based on the method of maximal variation. Section 9.4 then discusses a method for recovering the shape of the kernel from the observed second order moments in the univariate case and is also extended to the multivariate case.

Appendix A: Differogram

This appendix reviews the result of the differogram for estimating the noise level without relying exlicitly on an estimated model. The differogram cloud constitutes of a representation of the data in terms of the mutual distances amongst input- and output samples respectively. The behaviour of this representation towards the origin is then proven to be closely related with the noise level. The use of a parametric differogram model is used to estimate the noise level accurately. The main difference with existing methods is that there is no need for an extra hyperparameter whatever.

Appendix B: LS-SVMlab

While the presented research is rather methodological in nature, much effort was spent on the practical abilities of the methods and on increasing the userfrinedliness of the tools by elaborating a MATLAB/C toolbox called LS-SVMlab. The content and implementation details of the Matlab/C toolbox are discussed qualitatively and some details are given about the interface.

Chapter 2

Convex Optimization Theory: A Survey

As motivated in the previous chapter, the thesis will essentially take an optimization point of view as primal-dual optimization aspects lie somewhat at the core of the approach. This chapter reviews some important results from optimization theory and discusses the renewed interest for convex optimization. The first section surveys a number of definitions which are necessary for a clear exposition of the subject. More specifically, the scope of the theory of convex optimization problems is properly defined. Section 2.2 then reviews the machinery of dual problems in the sense of Lagrange. Section 2.3 discusses the problem from a more practical point of view, while in Section 2.4 a number of useful extensions are reported. Subsection 2.4.4 then introduces then the problem of hierarchical programming.

2.1 Convex Optimization

While the mathematics of convex optimization has been studied for about a century, several recent developments have stimulated new interest in the topic (Boyd and Vandenberghe, 2004). The first is the recognition that interior-point methods - developed in the 1980s to solve linear programming problems - can be used to solve general convex optimization problems as well (Nesterov and Nemirovski, 1994). The second development is the discovery that convex optimization problems beyond least squares and linear programming are more prevalent in practice than was previously thought. Furthermore there are great practical as well as theoretical advantages to recognizing or formulating a problem as a convex optimization problem. Moreover practical reliable and highly automated implementations exist for solving those problems efficiently. This motivation is readily summarized in the following quote due to (Rockafellar, 1993)

> "In fact the great watershed in optimization isn't between linearity and non-linearity, but convexity and non-convexity."

The remainder of the text primarily focuses on convex problems. A crash course is synthesized based on (Boyd and Vandenberghe, 2004) and (Rockafellar, 1970).

2.1.1 Convex sets and functions

Convex analysis, the mathematics of convex sets, functions and optimization problems is a well-developed subfield of mathematics, see e.g. (Rockafellar, 1970). Let $d \in \mathbb{N}$ be a positive integer denoting the dimensionality of the variables of a problem. Consider the following definitions of subsets of $\mathbb{R}^d$:

$$\begin{cases} \mathscr{S}_a = \{x \mid x = \beta x_1 + (1-\beta)x_2, x_1, x_2 \in \mathscr{S}_a, \beta \in \mathbb{R}\} \\ \mathscr{S}_c = \{x \mid x = \theta x_1 + (1-\theta)x_2, x_1, x_2 \in \mathscr{S}_c, \theta \in [0,\ \ 1] \subset \mathbb{R}\} \\ \mathscr{C} = \{x \mid x = \theta x_1,\ \ x_1 \in \mathscr{C}_k,\ 0 \leq \theta \in \mathbb{R}\}, \end{cases} \tag{2.1}$$

respectively denoted as an affine set, a convex set and a cone. The last is used to define the generalized inequality as follows (Luenberger, 1969),

$$x \succeq_k z \Leftrightarrow x - z \in \mathscr{C}_k. \tag{2.2}$$

Consider the cone $\mathscr{S}_c^+ = \mathbb{R}^{d,+}$, then the generalized inequality '$\succeq_k$' corresponds with the inequality '$\geq$'. Another well-known example is the semi-positive cone denoted as $\mathscr{C}_{pd}$, herefor let $A, B \in \mathbb{R}^{d \times d}$ be any symmetric matrices ($A^T = A, B^T = B$) and the following ordering is defined

$$A \succeq_{\mathscr{C}_{pd}} B \Leftrightarrow A - B \succeq_{\mathscr{C}_{pd}} 0 \Leftrightarrow A - B \text{ positive semi-definite}. \tag{2.3}$$

see e.g. (Alizadeh and Goldfarb, 2003; Boyd and Vandenberghe, 2004).

A function $f : \mathbb{R}^d \to \mathbb{R}$ is called convex if it satisfies the following property

$$\forall x_1, x_2 \in \mathbb{R}^d, \forall 0 \leq \theta \leq 1,\ f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2), \tag{2.4}$$

also referred to as Jensen's inequality. Let $f' : \mathbb{R}^d \to \mathbb{R}$ denote the first derivative of f over x. From the previous inequality, it follows that $f(x) \geq f(x_0) + f'(x_0)(x - x_0)$ for all $x, x_0 \in \mathbb{R}^D$ and that a global minimum is attained in $x^* \in \mathbb{R}^d$ if $f'(x^*) = 0$. This result shows that from local information on a convex function, one can derive global properties of it.

2.1.2 Convex optimization problems

Definition 2.1. [Convex Optimization Problem] *Let $m,p \in \mathbb{N}$ be positive integers and $b_i \in \mathbb{R}$ for all $i = 1,\dots,m,\dots,m+p$. Consider a well-defined generalized ordering associated with a cone $\mathscr{C}_k$, represented as '$\preceq_k$'. A mathematical optimization problem has the form*

$$p^* = \min_{x\in\mathbb{R}^D} f_0(x) \quad s.t. \quad \begin{cases} f_i(x) \preceq_k b_i & \forall i = 1,\dots,m \\ f_j(x) = b_j & \forall j = m+1,\dots,m+p. \end{cases} \tag{2.5}$$

where $f_k : R^D \to \mathbb{R}$ for all $k = 0,\dots,m+p$. The function f_0 is referred to as the objective function, the functions f_i for all $i = 1,\dots,m$ and f_j for all $j = m+1,\dots,m+p$ denote the inequality and the equality functions respectively. The vector $(b_1,\dots,b_m,\dots,b_{m+p})^T \in \mathbb{R}^{m+p}$ represent the bounds. An optimization problem is convex if it can be written in the form (2.5) with f_i convex functions for all $i = 0,1,\dots,m,\dots,m+p$ as the domain satisfying the constraints then is convex.

The convention is adopted to omit the domain $\mathbb{R}^D$ from the formulation as any restriction on x is explicified in the proper set of constraints. A conjugate function can be associated to a convex problem as follows:

Definition 2.2. [Conjugate Function] *Let $f : \mathbb{R}^D \to \mathbb{R}$ be a function. The conjugate function $f^* : \mathbb{R}^D \to \mathbb{R}$ then is defined as*

$$f^\star(y) = \sup_{x\in\mathbb{R}^D} \left(y^T x - f(x)\right). \tag{2.6}$$

Consider e.g. the function $f_Q(x) = \frac{1}{2}x^T Qx$ with $Q = Q^T \succeq 0$ symmetric and strictly positive definite. The maximum of $y^T x - \frac{1}{2}x^T Qx$ follows from taking the derivative towards y, resulting in the dual function $f_Q^* = f_{Q^{-1}} : \mathbb{R}^d \to \mathbb{R}$ defined as $f_Q^* = \frac{1}{2}y^T Q^{-1} y$.

2.1.3 Standard convex programming problems

A number of classes of convex programming problems occur frequently and received the following naming convention. Let $N_a, N_b, N_c \in \mathbb{N}$ be positive integers, let $A \in \mathbb{R}^{N_a\times d}$, $B \in \mathbb{R}^{N_b\times d}$ and $C \in \mathbb{R}^{N_c\times d}$ be matrices, let $a \in \mathbb{R}^{N_a}$, $b \in \mathbb{R}^{N_b}$ and $c \in \mathbb{R}^{N_c}$ denote vectors, let $Q \in \mathbb{R}^{N_a\times N_a}$ be a symmetric positive definite matrix and let $q \in \mathbb{R}^N$ be a given vector.

LS An unconstrained *Least Squares* (LS) problem can be written in the form

$$\min_x \|Ax - a\|_Q^2 = (Ax-a)^T Q(Ax-a). \tag{2.7}$$

If Q were the identity matrix $I_N \in \mathbb{R}^{N_a\times N_a}$, the ordinary least squares problem is obtained. Taking the first order conditions for optimality result in the equations

$$(A^T QA)x = A^T Qa, \tag{2.8}$$

which result in the unique global optimum $x^* \in \mathbb{R}^d$ of (2.7) if $A^T QA$ is of full rank. This set of equations can be solved with highly standard and reliable numerical methods, see e.g. (Golub and van Loan, 1989).

LP A *Linear Programming* (LP) problem then can be written as

$$\min_x a^T x \quad \text{s.t.} \quad \begin{cases} B_i x \leq b_i & \forall i = 1, \ldots, N_b \\ C_j x = c_j, & \forall j = 1, \ldots, N_c. \end{cases} \tag{2.9}$$

This class of problems was studied intensively in the literature on operations research (Dantzig, 1963; Bellman and Kalaba, 1965). See e.g. (Todd, 2002) for an historic account.

QP A *Quadratic Programming* (QP) problem can be written in the following standard form

$$\min_x \frac{1}{2} x^T Qx + q^T x \quad \text{s.t.} \quad \begin{cases} B_i x \leq b_i & \forall i = 1, \ldots, N_b \\ C_j x = c_j, & \forall j = 1, \ldots, N_c, \end{cases} \tag{2.10}$$

which is convex if and only if Q is positive definite and there exist a feasible solution x satisfying the constraints. Research on this type of problems was stimulated by e.g. the Markovitz portfolio problem (Markowitz, 1956).

SDP A *Semi-definite Programming problem* (SDP) takes the following form. Let $X \in \mathbb{R}^{d \times d}$ be a matrix of unknowns.

$$\min_X \ \text{tr}(AX) \quad \text{s.t.} \quad \begin{cases} \text{tr}(B_i X) = b_i & \forall i = 1, \ldots, N_b \\ CX \succeq 0, \end{cases} \tag{2.11}$$

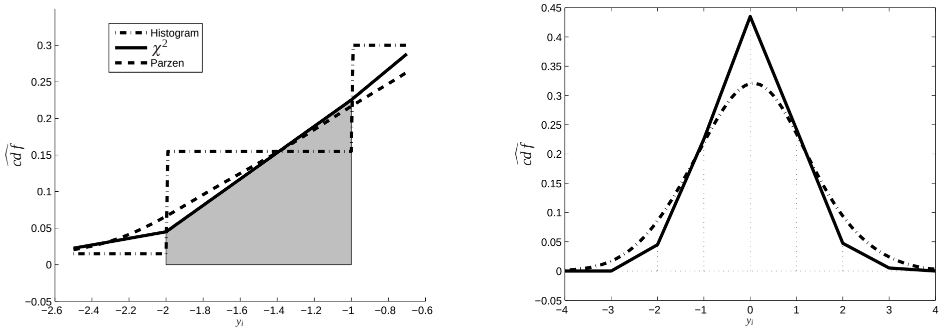
where the last constraint is referred to as a Linear Matrix Inequality (LMI). This formulation has found a rich variety of applications in e.g. problems of Model Predictive Control (MPC), see e.g. (Boyd *et al.*, 1994) and as illustrated by the popularity of the LMI lab toolbox in this community.

SOCP A problem takes the form of a *Second Order Cone Programming* (SOCP) problem if it can be written as follows

$$\min_x q^T x \quad \text{s.t.} \quad \begin{cases} \|Ax - a\|_2^2 \leq B_i x - b_i & \forall i = 1, \ldots, N_b \\ C_j x = c_j, & \forall j = 1, \ldots, N_c \end{cases} \tag{2.12}$$

The constraint $\|Ax - a\|_2^2 \leq Bx - b$ is called a second order cone constraint since it is the same as requiring that $(Ax - a, Bx - b)$ lies in the second order cone $\mathscr{S}_k = \{(x,t) \mid x^2 \leq t, x \in \mathbb{R}^d, t \in \mathbb{R}\}$. See e.g. (Lobo *et al.*, 1998).

Various other classes exist as the Quadratical constrained Quadratical Programming (QCQP) problems (Lobo *et al.*, 1998) or geometric programming problems (Boyd and Vandenberghe, 2004).



(a) (b)

Figure 2.1: *Illustrative example relating three different methods for univariate density estimation qualitatively. The classical histogram method is prone to non-continuous artifacts by construction. The Parzen window estimator results in smooth estimates but is based on an ad hoc L_2 optimality criterion. The proposed χ^2 approach makes a trade-off between both approaches as it is based on a clear optimality principle and enforces continuity on the knots.* **(a)** *shows a detail of the estimates of the three methods, while* **(b)** *illustrates the global difference of the χ^2 approach with respect to the Parzen window. From the figure and the optimality principle (2.13) it is immediately clear that the χ^2 estimator is more flexible towards modeling data concentrations (peaks).*

Example 2.1 [a χ^2 density estimator] An example of the application of this class of optimization methods towards the task of density estimation is given following the setup of Example 1.1. Let $\{y_i\}_{i=1}^N$ be i.i.d. sampled from a random variable $\mathbf{Y} \in \mathbb{R}^D$ with smooth density function $p_{\mathbf{Y}} : \mathbb{R}^D \to [0,1]$. Assume a disjoint but complete partitioning of the support of the random variable with contiguous sets $\mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_s$ such that $\bigcup_{i=1}^s \mathscr{S}_i = \text{support}(\mathbf{Y})$. Let f_i denote the number of samples in the set $\mathscr{S}_i$ such that $N = \sum_{i=1}^r f_i$. A common method in the case of grouped data is the minimum χ^2-estimator (Rao, 1983; Press *et al.*, 1988). Under the assumption that $p_{\mathbf{Y}}$ can be described by the element of a parameteric family $\{p_\theta | \theta \in \Theta\}$ with a set Θ of finite dimension, then the chi-squared estimator takes the following form

$$\hat{\theta} = \min_{\theta} \sum_{i=1}^{r} \frac{\| f_i - N f_A(\mathscr{S}_i, p_\theta) \|_2^2}{N f_A(\mathscr{S}_i, p_\theta)} \quad \text{s.t.} \quad f_A(\mathscr{Y}, p_\theta) = 1, \tag{2.13}$$

where the function f_A is defined as $f_A(\mathscr{S}, p) = \int_{y \in \mathscr{S}} p(y) dy$.

Consider the univariate case where $\mathbf{Y} \in \mathbb{R}$. Let the sets be described as $\{S_{(i)} | S_{(i)} = [b_{(i)}, b_{(i+1)}], b_{(i)} < b_{(i+1)}\}$. The minimum $b_{(1)}$ and maximum $b_{(r+1)}$ describe the extrema of the support of the distribution. Instead of a parametric family of density functions, consider the (non-parametric) piecewise linear models

$$p_c(y) = c_{(i)} + \frac{\left(y - b_{(i)}\right)\left(c_{(i+1)} - c_{(i)}\right)}{b_{(i+1)} - b_{(i)}} \quad \text{where} \quad b_{(i)} \leq y \leq b_{(i+1)}, \;\; c_{(i)} \geq 0. \tag{2.14}$$

Let $c = \left(c_{(1)}, \ldots, c_{(r)}, c_{(r+1)}\right)^T \in \mathbb{R}^{r+1}$ and $b = \left(b_{(1)}, \ldots, b_{(r)}, b_{(r+1)}\right)^T \in \mathbb{R}^{r+1}$ be vectors. In this case, the function f_A can then be written as follows

$$\begin{aligned} f_A\left(\mathscr{S}_{(i)}, c_{(i)}, c_{(i+1)}\right) &= \frac{1}{2}\left(b_{(i+1)} - b_{(i)}\right)\left(c_{(i+1)} + c_{(i)}\right) = A_i c \\ \text{s.t.} \quad A_i &= \frac{1}{2}\left[0_{i-1} \quad \left(b_{(i+1)} - b_{(i)}\right) \quad \left(b_{(i+1)} - b_{(i)}\right) \quad 0_{r-i}\right]. \end{aligned} \tag{2.15}$$

Let b be given and c be unknowns to the problem and $f = (f_1, \ldots, f_r)^T \in \mathbb{R}^r$. Then the chi-squared estimator with respect to the non-parametric model class of the piecewise linear models may be formulated as

$$\hat{c} = \min_c \sum_{i=1}^{r} \frac{\| f_i - NAc\|_2^2}{NAc} \quad \text{s.t.} \quad f \geq 0_r, \;\; c \geq 0_{r+1}, \;\; 1_N^T Ac = 1, \tag{2.16}$$

where $A = (A_1, \ldots, A_r) \in \mathbb{R}^{r \times r+1}$. This problem can be written as a convex SOCP problem as follows. Let $t_i \geq \frac{\| f_i - NA_i c\|_2^2}{NA_i c}$ which can be rewritten (see e.g. (Lobo *et al.*, 1998)) as $t_i + NA_i c \geq \left\| \begin{bmatrix} 2(f_i - NA_i c) \\ t_i - A_i c \end{bmatrix} \right\|_2^2$. The optimization problem becomes

$$\hat{c} = \min_c \sum_{i=1}^{r} t_i \quad \text{s.t.} \quad \left\| \begin{bmatrix} 2(f_i - NA_i c) \\ t_i - NA_i c \end{bmatrix} \right\|_2^2 \leq t_i + NA_i c, \;\; f \geq 0_r, \;\; c \geq 0_{r+1}, 1_N^T Ac = 1. \tag{2.17}$$

This problem can be solved efficiently as a SOCP problem using e.g. the Matlab toolbox SeDuMi as described in (Sturm, 1999).

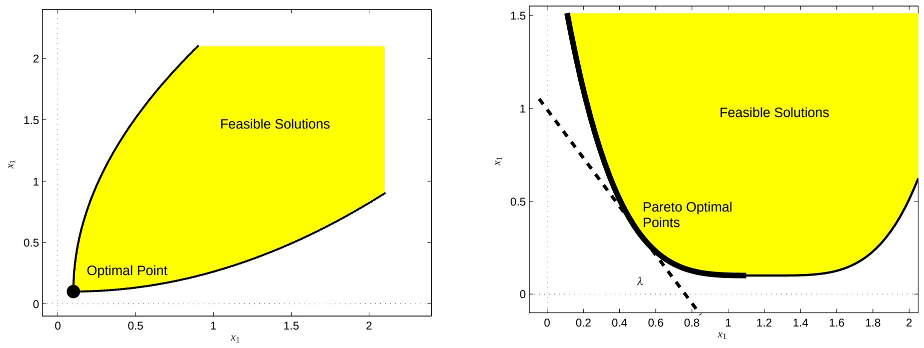
This approach differs from more classical methods as a (finite sample) optimality principle is postulated. Figure 2.1 illustrates the qualitative difference of this estimator and the classical histogram technique and the Parzen window estimator. The described method is closely related to the method of histosplines, but uses a χ^2 measure instead of the constraint that the bin-area should equal the empirical frequency exactly (Rao, 1983).

2.1.4 Multi-criterion optimization

The following discussion of optimization with more than one objective is surveyed as in (Luenberger, 1969) and (Boyd and Vandenberghe, 2004).

Definition 2.3 (Multi-criterion optimization problems). *A multi-criterion or vector optimization problem is defined as a programming problem*

$$p^* = \min_{x \in \mathbb{R}^D} f_0(x) \quad s.t. \quad \begin{cases} f_i(x) \preceq_k b_i & \forall i = 1, \ldots, m \\ f_j(x) = b_j & \forall j = m+1, \ldots, m+p. \end{cases} \tag{2.18}$$



(a) (b)

Figure 2.2: *Schematical illustration of the problem of multicriterion optimization for* $D = 2$. **(a)** *Feasible solutions (fill) with optimal solution with the inequality* $(x_1^*, x_2^*) \preceq (x_1, x_2)$ *if* $x_1^* \leq x_1$ *and* $x_2^* \leq x_2$. **(b)** *Feasible solutions without an optimal point, but with a collection of Pareto optimal points (thick line) which are all solutions to a scalarized problem with scalarization terms* λ.

where $f_0 : R^D \rightarrow \mathbb{R}^Q$ *where* $Q > 1$. *and* $f_k : R^D \rightarrow \mathbb{R}$ *for all* $k = 1, \ldots, m + p$. *The functions* f_i *for all* $i = 1, \ldots, m$ *and* f_j *for all* $j = m+1, \ldots, m+p$ *denote the inequality and the equality functions respectively. The vector* $(b_1, \ldots, b_m, \ldots, b_{m+p})^T \in \mathbb{R}^{m+p}$ *represent the bounds.*

The optima to multi-criterion problems are defined as follows:

Definition 2.4. [Optimal and Pareto Optimal] *The meaning of an optimal point* $x^* \in \mathbb{R}^D$ *satisfying the constraints can be translated as follows. For all* $x \in \mathbb{R}^D$ *which satisfy the constraints, the inequality* $f_0^q(x^*) \leq f_0^q(x)$ *holds for all* $q = 1, \ldots, Q$. *For a Pareto optimal point* $x^\star \in \mathbb{R}^D$ *satisfying the constraints, one has for all* $x \in \mathbb{R}^D$ *which is feasible that if* $f_0^q(x^*) \leq f_0^q(x)$ *for all* $q = 1, \ldots, Q$, *then* $f_0^q(x^*) = f_0^q(x)$ *for all* $q = 1, \ldots, Q$.

Note that not every multi-criterion problem has an optimal element, but if it exists, it is unique. Pareto optimal points always exist, but are often not unique, see also Figure 2.2. In the case the problem (2.18) consists of convex functions f_k for all $k = 0, \ldots, m+p$, for every Pareto-optimal point $x^\star \in \mathbb{R}^D$, there consist a parameter $\lambda \in \mathbb{R}^q$ with $\lambda \succeq_{k^*} 0$ such that it is the unique minimizer to

$$p^* = \min_{x \in \mathbb{R}^D} \lambda^T f_0(x) \quad \text{s.t.} \quad \begin{cases} f_i(x) \preceq_k b_i & \forall i = 1, \ldots, m \\ f_j(x) = b_j & \forall j = m+1, \ldots, m+p, \end{cases} \tag{2.19}$$

which is a one-dimensional (scalar) optimization problem which can be solved using standard techniques. The set of Pareto optima x^* may be found by exploring all such scalarization vectors λ. This scalarization technique is hevily used in the remainder e.g. in the discussion of regularization schemes (see e.g. Chapter 6).

2.2 The Lagrange Dual

Th following definition follows the exposition in (Boyd and Vandenberghe, 2004). Let $\alpha = (\alpha_1, \ldots, \alpha_m, \ldots, \alpha_{m+p})^T \in \mathbb{R}^{m+p}$ be a vector of Lagrange multipliers associated with the m inequalities and the p equalities where $\alpha_i \geq 0$ for all $i = 1, \ldots, m$. Then the Lagrangian $\mathscr{L} : \mathbb{R}^D \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ of the optimization problem (2.5) is defined as follows.

$$\mathscr{L}(x; \alpha) = f_0(x) + \sum_{i=1}^{m} \alpha_i (f_i(x) - b_i) + \sum_{j=m+1}^{m+p} \alpha_j (f_j(x) - b_j). \tag{2.20}$$

The Lagrange dual function is defined as the infimum over x,

$$g(\alpha) = \inf_x \left(f_0(x) + \sum_{i=1}^{m} \alpha_i (f_i(x) - b_i) + \sum_{j=m+1}^{m+p} \alpha_j (f_j(x) - b_j) \right). \tag{2.21}$$

which can be proven to be concave even if the problem (2.5) is not convex. Furthermore, the inequality $g(\alpha) \leq p^* \leq f_0(x)$ holds for any $\alpha \geq 0, \beta$ and feasible x (satisfying the constraints). In the case the (in)equalities can be written in matrix form ($Bx \leq b, Cx = c$) as previously (consider e.g. the QP), then the dual can be written in function of the conjugate function $f_0^* : \mathbb{R}^{m+p} \to \mathbb{R}$ of f_0 as defined in (2.6). Let the vector α be subdivided in two disjunct parts as follows $\alpha^b = (\alpha_1, \ldots, \alpha_{N_b})^T \in \mathbb{R}^{N_b,+}$ and $\alpha^c = (\alpha_{N_b+1}, \ldots, \alpha_{N_b+N_c})^T \in \mathbb{R}^{N_c}$.

$$\begin{aligned} g(\alpha) &= \inf_x \left(f_0(x) + \alpha^{b^T}(Bx - b) + \alpha^{c^T}(Cx - c) \right) \\ &= -\alpha^{b^T} b - \alpha^{c^T} c - f_0^*(-\alpha^{b^T} B - \alpha^{c^T} C). \end{aligned} \tag{2.22}$$

In the case of an LP as in (2.9), this simplifies to

$$\begin{aligned} g(\alpha) &= \inf_x \left(a^T x + \alpha^{b^T}(Bx - b) + \alpha^{c^T}(Cx - c) \right) \\ &= -\alpha^{b^T} b - \alpha^{c^T} c + \inf_x (a^T - \alpha^{b^T} B - \alpha^{c^T} C)\, x \qquad (2.23) \\ &= \begin{cases} -\alpha^{b^T} b - \alpha^{c^T} c & \text{if} \quad a = B^T \alpha^b + C^T \alpha^c \\ -\infty & \text{elsewhere.} \end{cases} \qquad (2.24) \end{aligned}$$

The best lower-bound using the Lagrangian on the cost given by $f_0(x)$ for x a feasible function is then obtained as

$$d^* = \max_\alpha g(\alpha) \quad \text{s.t.} \quad \alpha_i \geq 0, \ \forall i = 1, \ldots, N_b, \tag{2.25}$$

referred to as the Lagrange dual problem. Strong duality is said to hold when the duality gap $p^* - d^*$ is zero. Convex problems have the property of strong duality under mild regularity conditions (Slater's condition). Also the following result holds (see e.g. von Neumann, (Rockafellar, 1970)).

Lemma 2.1. [Saddlepoint Interpretation, e.g. (Rockafellar, 1970)] *If a vector $(x^*;\alpha^*) \in \mathbb{R}^d \times \mathbb{R}^{N_b} \times \mathbb{R}^{N_c}$ forms a saddlepoint of the Lagrangian such that*

$$(x^*;\alpha^*) = \arg\max_{\alpha}\min_{x} \mathscr{L}(x;\alpha) = \arg\min_{x}\max_{\alpha} \mathscr{L}(x;\alpha)$$
$$s.t. \quad \alpha_i \geq 0 \ \ \forall i = 1,\ldots,N_b, \qquad (2.26)$$

then x^* is the optimum of the primal problem (2.5), α^* gives the optimum to (2.25) and strong duality holds.

This Lemma will form the basis to the framework of primal-dual kernel machines.

2.2.1 Conditions for optimality

In the case of a convex problem (2.5) with differential objective function and constraint function satisfying Slaters condition, the so-called Karush-Kuhn-Tucker (KKT) conditions are both necessary and sufficient conditions for a vector $(x^*;\alpha^*)$ to be a global optimum to the primal problem (2.5) and to the dual problem (2.25):

$$\text{KKT} = \begin{cases} \left.\dfrac{\partial \mathscr{L}(x;\alpha^*)}{\partial x_i}\right|_{x_i = x_i^*} = 0 & \forall i = 1,\ldots,d & (a) \\ f_i(x^*) \leq b_i & \forall i = 1,\ldots,m & (b) \\ f_j(x^*) = b_j & \forall j = m+1,\ldots,m+p & (c) \\ \alpha_i^* \geq 0 & \forall i = 1,\ldots,m & (d) \\ \alpha_i^*\left(f_i(x^*) - b_i\right) = 0. & \forall i = 1,\ldots,m & (e) \end{cases} \qquad (2.27)$$

In case the optimization problem is not convex, these conditions are only necessary.

Remark 2.1. Note that in the case no inequalities occur in the convex programming problem, the first order conditions are both necessary and sufficient (Luenberger, 1969; Nocedal and Wright, 1999; Boyd and Vandenberghe, 2004).

2.2.2 Sensitivity interpretation

When strong duality holds, the optimal dual variables contain useful information about the sensitivity of the optimum with respect to perturbations of the constraints. Let $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_m,\ldots,\varepsilon_{m+p})^T \in \mathbb{R}^{m+p}$ be a vector containing small perturbation terms and let the function $\overline{p} : \mathbb{R}^{m+p} \to \mathbb{R}^d$ be defined as follows

$$\overline{p}^*(\varepsilon) = \min_x f_0(x) \quad \text{s.t.} \quad \begin{cases} f_i(x) \leq b_i + \varepsilon_i & \forall i = 1,\ldots,m \\ f_j(x) = b_j + \varepsilon_j. & \forall j = m+1,\ldots,m+p. \end{cases} \qquad (2.28)$$

This perturbed problem preserves convexity of the original problem (2.5). Let (α^*, β^*) be the optimal to the dual unperturbed problem. Then the following inequality holds

$$\overline{p}^*(\varepsilon) \geq d^* - \sum_{i=1}^{m} \alpha_i^* \varepsilon_i - \sum_{i=m+1}^{m+p} \alpha_j^* \varepsilon_j. \tag{2.29}$$

By strong duality, it follows that the derivative

$$\frac{\partial \overline{p}^*(\varepsilon)}{\partial \varepsilon_i} = -\alpha_i. \quad \forall i = 1, \ldots, m, \ldots, m+p \tag{2.30}$$

see e.g. (Rockafellar, 1970).

2.2.3 Dual standard programming problems

The dual of the standard programming problems itemized in Subsection 2.1.3 are reviewed. Let $0_D = (0, \ldots, 0)^T \in \mathbb{R}^D$ be a vector of zeros of length $D \in \mathbb{N}$. Let $Q \in \mathbb{R}^{d \times d}, A \in \mathbb{R}^{N_a \times d}, B \in \mathbb{R}^{N_b \times d}, C \in \mathbb{R}^{N_c \times d}$ be matrices and $q \in \mathbb{R}^d, a \in \mathbb{R}^{N_a}, b \in \mathbb{R}^{N_b}, c \in \mathbb{R}^{N_c}$ be vectors as in Subsection 2.1.3.

LP* Following equation (2.22), the dual function to the problem (2.9) is given as

$$\begin{aligned} g(\alpha) &= -\alpha^{b^T} b - \alpha^{c^T} c + \inf_x (a + B^T \alpha^b + C^T \alpha^c)^T x \\ &= \begin{cases} -\alpha^{b^T} b - \alpha^{c^T} c & (a + B^T \alpha^b + C^T \alpha^c) = 0_d \\ -\infty & \text{otherwise,} \end{cases} \end{aligned} \tag{2.31}$$

such that the dual problem can be written as

$$\max_\alpha - \left(\alpha^{b^T} b + \alpha^{c^T} c \right) \quad \text{s.t.} \quad \begin{cases} a^T = -B^T \alpha^b - C^T \alpha^c \\ \alpha^b \geq 0_{N_b}. \end{cases} \tag{2.32}$$

Moreover, strong duality holds.

QP* The dual function to the problem (2.10) is given as

$$d^* = \max_\alpha g(\alpha) = (B^T \alpha^b + C^T \alpha^c + q)^T Q^{-1} (B^T \alpha^b + C^T \alpha^c + q) - b^T \alpha^b - c^T \alpha^c$$
$$\text{s.t.} \quad \alpha_i^b \geq 0, \forall i = 1, \ldots, N_b \tag{2.33}$$

More detailed derivation of this problem will re-occur in chapter 4.

SOCP* Consider the primal SOCP problems that can be rewritten in the following form

$$p^* = \min_x a^T x \quad \text{s.t.} \quad \begin{cases} x \succeq_k 0 \\ Cx = c, \end{cases} \tag{2.34}$$

with $\succeq_k$ associated with the proper (pointed) second order cone (Boyd and Vandenberghe, 2004). The dual problem to the problem (2.12) is given as

$$d^* = \max_\alpha -c^T\alpha^c \quad \text{s.t.} \quad \begin{cases} C^T\alpha^c + a = \alpha^b \\ \alpha^b \succeq_k^* 0. \end{cases} \tag{2.35}$$

where $\succeq_k^*$ is the generalized inequality corresponding with the dual cone k^* which equals the original cone $\mathscr{C}_k = \mathscr{C}_{k^*}$ in the case of the quadratic cone.

SDP* Let $G, F_1, \ldots, F_d$ be a set of matrices such that $G, F_1, \ldots, F_d \in \mathbb{R}^{D\times D}$ for $D \in \mathbb{N}$. Consider the primal SDP problem without equality constraints

$$p^* = \min_x a^T x \quad \text{s.t.} \quad x_1F_1 + \cdots + x_dF_d + G \preceq 0. \tag{2.36}$$

The dual problem can then be written as

$$d^* = \max_\Gamma -\operatorname{tr}(G\Gamma) \quad \text{s.t.} \quad \begin{cases} \operatorname{tr}(F_i\Gamma) = a_i \quad \forall i = 1, \ldots, d \\ \Gamma \succeq 0, \end{cases} \tag{2.37}$$

where $\Gamma \in \mathbb{R}^{D\times D}$ is a matrix containing the Lagrange multipliers.

Duality has a profound basis (Luenberger, 1969; Rockafellar, 1970) and has lead to a number of interesting results both theoretically (feasibility study) as practically (e.g. in learning theory, see later chapters), (Boyd and Vandenberghe, 2004).

2.3 Algorithms and Applications

2.3.1 Algorithms

A short summary of the main numerical algorithms for solving convex optimization problems is given. While initial research following in the streamline of the seminal work of (Dantzig, 1963) mainly focussed on simplex methods in the area of operations research (Bellman and Kalaba, 1965), later investigations concentrate more on efficient barrier methods as the interior point methods.

Since the seminal work of (Karmarkar, 1984) there has been a concentrated effort to develop efficient interior-point methods for linear programming (LP). More recently, researchers have begun to appreciate important properties of these interior-point methods beyond their efficiency for LP (Nesterov and Nemirovski, 1994). Major advantages include the fact that they extend gracefully to nonlinear convex optimization problems. New interior-point algorithms for problem classes such as SDPs or second-order cone programming (SOCPs) (Nesterov and Todd, 1997) are now approaching the extreme efficiency of modern linear programming codes proving the notable efforts SDPack, see e.g. (Alizadeh and Goldfarb, 2003) for pointers, and SeDuMi (Sturm, 1999). Another class of methods relies on the exploitation of the primal and the dual problem formulation. In general primal-dual optimization algorithms try to find the global optimum by minimizing the gap between the optimum of the primal and the dual. Most state-of-the art implementations use ingredients of both interior point as well as from primal-dual methods (Sturm, 1999). Recent advances describe methods to highly increase the efficiency of the methods by exploiting structure in the matrices at hand.

2.3.2 Applications and the design of algorithms

Renewed interest for the theory of convex optimization was stimulated amongst others by the reformulation of a number of estimation problems as a convex optimization problem. While the initial literature mainly focussed on control problems as surveyed in (Boyd *et al.*, 1993; Boyd *et al.*, 1998), a fruitful field of application is found into the practice of estimation and identification and more specifically in the design of kernel machines wich are explicitly based on an optimality principle as initiated by (Boser *et al.*, 1992; Cortes and Vapnik, 1995; Vapnik, 1998), see the remainder of the text. More theoretical and mathematical applications were formulated in the form of convex relaxations to hard combinatorial constraints, see e.g. (Grötschel *et al.*, 1988; Boyd and Vandenberghe, 2004).

A significant obstacle to the widespread use of the methodology remains: the high level of experience in both convex optimization and numerical algebra required to use it. Recent advances in the theory aim at lowering the barrier of using the methods for the unexperienced. Disciplined Convex Programming (DCP) approaches this problem by proposing a formal ruleset and conventions in order to derive proper convex programs from the problem at hand (Grant, 2004).

The present text may be seen from a similar perspective as it illustrates the use of the primal-dual optimization framework for the construction of various non-trivial estimation tasks.

2.4 Extensions

This section describes a number of examples of optimization problems which can be cast as convex problems. As those results will re-occur in the remainder of the text under various disguises, they are treated here somewhat generically.

2.4.1 Robust and stochastic programming

Let $0.5 < \eta < 1$ be a fixed confidence level. Let $A \in \mathbb{R}^D$ and $B \in \mathbb{R}^{N_b \times D}$ be a vector and a matrix respectively . Let B_i be samples of a random variable with Gaussian distribution with mean $\overline{B}_i$ and variance Σ_i such that $B_i \sim \mathcal{N}(\overline{A}_i, \Sigma_i)$. Consider the following stochastic programming problem.

$$\min_x A^T x + a \quad \text{s.t.} \quad \text{Prob}(B_i x \leq b_i) \geq \eta \;\; \forall i = 1, \ldots, N_b. \tag{2.38}$$

Consider for a moment the ith constraint and let $u = B_i x$, $\overline{u} = \overline{B}_i x$ and σ^2 denote $\mathrm{var}(u) = \mathrm{var}(B_i x)$. Let $\phi(x)$ denote the cdf of the standard normal $\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-t^2/2) dt$. The ith constraint (2.38) can be normalized to the standard distribution as follows.

$$\mathrm{Prob}\left(\frac{u-\overline{u}}{\sigma} \leq \frac{b_i - \overline{u}}{\sigma}\right) \geq \eta \quad \Leftrightarrow \quad \frac{b_i - \overline{u}}{\sigma} \geq \phi^{-1}(\eta) \tag{2.39}$$

$$\Leftrightarrow \quad \overline{B}_i x + \phi^{-1}(\eta) \|\Sigma^{1/2} x\|_2 \leq b_i, \tag{2.40}$$

and as $\varphi^{-1}(\eta) > 0$ as $\eta > 0.5$, this inequality has the form of a second order cone constraint:

$$\min_x Ax + a \quad \text{s.t.} \quad \phi^{-1}(\eta)\|\Sigma^{1/2}x\|_2 \leq b_i - \overline{B}_i, \;\; \forall i = 1, \ldots, N. \tag{2.41}$$

Application of this kind of formulations is found e.g. in stochastic Markovitz portfolio problem (Goldfarb and IYengar, 2003). Recent advances in machine learning cast robust counterparts of SVMs as SOCPs using similar results (Trafalis and Alwazzi, 2003).

2.4.2 Quadratical constraints

Consider the following quadratical form

$$x^T H x + f^T x \tag{2.42}$$

with $H \in \mathbb{R}^{D\times D}$ and $f \in \mathbb{R}^D$. This kind of constraints is hard to cast as constraints into an efficient optimization algorithm. A classical relaxation method for such quadratic forms is based on semidefinite programming (Grötschel *et al.*, 1988). let H_f denote the matrix

$$H_f = \begin{bmatrix} H & 0.5f \\ 0.5f^T & 0 \end{bmatrix} \in \mathbb{R}^{(D+1)\times(D+1)}. \tag{2.43}$$

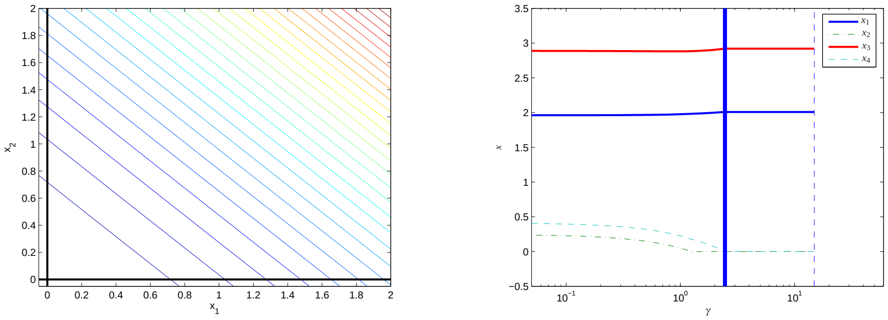
One can rewrite the cost function of (2.42) as follows

$$x^T H x + f^T x = \begin{bmatrix} x \\ 1 \end{bmatrix}^T H_f \begin{bmatrix} x \\ 1 \end{bmatrix}. \tag{2.44}$$

Consider the reparameterization of the problem (2.42) based on the new set of variables $Z \in \mathbb{R}^{(D+1)\times(D+1)}$ related with x as follows (Nesterov, 1998)

$$\begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^T = Z \quad \Leftrightarrow \quad \begin{bmatrix} x \\ 1 \end{bmatrix}^T H_f \begin{bmatrix} x \\ 1 \end{bmatrix} =< H_f, Z >= \sum_{i,j=1}^{D+1} H_{f,ij} Z_{ij}. \tag{2.45}$$

From this overparameterization, it is clear that the matrix Z should be symmetric positive definite and rank one. The common relaxation then consists of omitting the rank one constraint which is hard to impose. The former positive semi-definite constraint is denoted as $Z \succeq 0$. Such a relaxations can be cast as a convex semi-definite programming problem, see e.g. (Zhang, 2000).



(a)

(b)

Figure 2.3: *A four-dimensional example* $x = (x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4$ *is studied where* $H \in \mathbb{R}^{4\times4}$ *is strictly positive definite and and the positive OR constraints* $x_1x_3 = 0$ *and* $x_2x_4 = 0$ *are to be satisfied.* **(a)** *displaying the Hessian* H^TH *and its augmented counterpart* $(H^TH + \gamma N)$ **(b)** *the evolution of the estimates when ranging* γ *from* 0.5 *to* 50*. From the figure it becomes apparent that the positive OR constraints are satisfied when* $\gamma \geq 30$ *(solid vertical line). The dashed vertical line indicates the value of* γ *where the problem becomes non-convex.*

2.4.3 Positive OR-constraints

A special class of quadratical constraints is considered.

Definition 2.5. [Positive OR Constraint] *A positive OR-constraint between scalars* $x_1, x_2 \in \mathbb{R}$ *is defined as follows*

$$x_1\, x_2 = 0, \quad x_1, x_2 \geq 0. \tag{2.46}$$

Let x denote the vector $[x_1, x_2]^T \in \mathbb{R}^2$. The quadratic constraint (2.46) is equivalent to $x^TNx = 0$ where $N = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Although this class of constraints does clearly not describe a convex set, one can approach such constraints efficiently if they are embedded in a quadratical programming problem. Consider the following prototypical problem:

$$\mathscr{J}_N(x) = x^THx + f^Tx \quad \text{s.t.} \quad x^TNx = 0, \ \ x \geq 0 \tag{2.47}$$

where $H = \begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix} \in \mathbb{R}^{2\times2}$ is positive definite and $f = \begin{bmatrix} f_1 & f_2 \end{bmatrix}^T \in \mathbb{R}^2$.

Example 2.2 **[Augmented Hessian Relaxation]** A technique based on augmenting the Hessian is considered. Let $\gamma \geq 0$ be a positive constant, the following modified problem to (2.47) is studied

$$\mathscr{J}_{N,\gamma}(x) = \left(x^T H x + f^T x\right) + \gamma \left(x^T N x\right) \quad \text{s.t.} \quad x^T N x \leq 0, \;\; x \geq 0, \tag{2.48}$$

which may be seen as a bi-criterion optimization problem with trade-off constant γ. This problem is convex whenever the following condition is satisfied

$$H^T H + \gamma N \succeq 0, \tag{2.49}$$

see e.g. (Boyd and Vandenberghe, 2004). The term $x^T N x$ is bounded below by 0 by construction, such that the problem (2.48) reduces to the problem (2.47) when the optimum $x^T N x = 0$ is attained. This ensures the property that the modified cost-function acts as an upper-bound to the cost of the original problem.

Formally, the modified problem (2.48) shares its first order conditions for optimality as given by the KKT conditions with the necessary conditions for optimality of the non-convex problem (2.47). This can be seen by relating the problem (2.48) with the Lagrangian of the QP problem (2.47) given as

$$\mathscr{L}_H(x) = x^T H x + f^T x + \lambda\left(x^T N x\right) - \alpha^T x \tag{2.50}$$

with multipliers $\alpha \in \mathbb{R}_0^{+,D}$ and $\lambda \in \mathbb{R}_0^+$. Figure 2.3 illustrates this issue.

A four-dimensional example $x = (x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4$ is studied where $H \in \mathbb{R}^{4\times 4}$ is strictly positive definite and and the positive OR constraints $x_1 x_3 = 0$ and $x_2 x_4 = 0$ are to be satisfied. **(a)** displaying the Hessian $H^T H$ and its augmented counterpart $H^T H + \gamma N$ **(b)** the evolution of the estimates when ranging γ from 0.5 to 50. From the figure it becomes apparent that the positive OR constraints are satisfied when $\gamma \geq 30$ (solid vertical line). The dashed vertical line indicates the value of γ where the problem becomes non-convex.

2.4.4 Hierarchical programming problems

An hierarchical programming (HP) problem amounts to the simultaneous optimization of different objectives defined on a common set of variables. Here every level considers the optimization of all variables constrained to the intersection of the solution spaces corresponding to the previous levels, with respect to its own cost-function (Pelckmans *et al.*, 2003*b*; Pelckmans *et al.*, 2005*c*). This approach is to be opposed to the more standard approach as the scalarization technique to infer Pareto optima (see Subsection 2.1.4).

Consider for instance a two-level HP problem. Let both objectives $\mathscr{J}^{(1)}$ and $\mathscr{J}^{(2)}$ act on the variables x and θ. Let the level one cost-function $\mathscr{J}_\theta^{(1)}(x)$ describe an optimum x^* corresponding to a certain θ which is provided by the user. Let the level two cost-function $\mathscr{J}^{(2)}(\theta, x)$ act on x and θ where x^* is to be a solution to $J_\theta^{(1)}$ for the optimal θ^*. Formally,

$$\begin{cases} \textbf{Level 1}: & (x^* \mid \theta) = \arg\min_x \mathscr{J}_\theta^{(1)}(x) \\ \textbf{Level 2}: & (x_\theta, \theta^*) = \arg\min_{x,\theta} \mathscr{J}^{(2)}(x, \theta) \quad \text{s.t.} \quad x_\theta = (x \mid \theta). \end{cases} \tag{2.51}$$

The following example illustrates how one can formulate and solve hierarchical programming problems using results from convex optimization.

Consider on a first level an LP of dimension $D \in \mathbb{N}_0$ with $N \in \mathbb{N}_0$ inequality constraints and no equality constraints. Let $B \in \mathbb{R}^{N \times D}$ be a given matrix and $u = (u_1, \dots, u_N)$ be a fixed but unknown vector.

$$\textbf{First Level: } x^* = \arg\min_{x \in \mathbb{R}^D} \mathscr{J}_D = a^T x \quad \text{s.t.} \quad B_i x \leq u_i \;\; \forall i = 1, \dots, N. \tag{2.52}$$

The Karush-Kuhn-Tucker conditions provide necessary and sufficient conditions for x to be a solution to (2.52). Let $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^{+,N}$ be a vector of positive Lagrange multipliers:

$$\text{KKT}(x,u;\alpha) = \begin{cases} a = -B^T \alpha & & (a) \\ u_i - B_i x \geq 0 & \forall i = 1, \dots, N & (b) \\ \alpha_i \geq 0 & \forall i = 1, \dots, N & (c) \\ \alpha_i (u_i - B_i x) = 0. & \forall i = 1, \dots, N & (d) \end{cases} \tag{2.53}$$

Let $F \in \mathbb{R}^{n \times D}$ be a given matrix with $n \in \mathbb{N}_0$ rows and $f = (f_1, \dots, f_n)^T \in \mathbb{R}^n$ be a given vector. On a second level, consider the problem of choosing u such that x satisfies $Fx - f$ optimally in an L_2 sense. Let $\upsilon = (\upsilon_1, \dots, \upsilon_N)^T \in \mathbb{R}^N$ be a variable vector, then the problem on the second level can be written as follows:

$$\textbf{Second Level: } (\hat{\upsilon}, \hat{x}) = \arg\min_{x, \upsilon \in \mathbb{R}^D} \mathscr{J}_F^{(2)} = \|Fx - f\|_2^2$$
$$\text{s.t.} \quad x \text{ solves (2.52) with } u = \upsilon. \tag{2.54}$$

Using the KKT conditions, the problem equals

$$\textbf{Second Level: } (\hat{u}, \hat{x}, \hat{\alpha}) = \arg\min_{x, u, \alpha \in \mathbb{R}^D} \mathscr{J}_F^{(2)} = \|Fx - f\|_2^2 \quad \text{s.t.} \quad \text{KKT}(x,u;\alpha). \tag{2.55}$$

One refers to this approach as *fusion* of a first level problem with a second level. In general this amounts to multi-criterion optimization which builds a construction based on the explicit description of the solution-space of previous levels, hence the name *Hierarchical programming problem*. This method can be contrasted with the Pareto (Pareto, 1971) multi-criterion approach.

The hierarchical programming problem (2.55) is convex up to the complementary slackness conditions (2.53.d) which belong to the class of positive OR constraints as discussed in the previous subsection. Hierarchical optimization problems have a natural application in the task of model selection as discussed in Chapters 7 and 8.

Remark 2.2. Note that this programming paradigm is already employed in various derivations. As a first example, consider the saddlepoint approach for constructing the dual problem as surveyed in Section (2.2) and (2.26) for constructing the dual problem. The saddlepoint is computed as the solution to the problem $\max_\theta \min_x$ where the $\max_\theta$ is taken over the solution-space of the optimum to the minization. Another manifestation of the hierarchical programming approach is found in the analysis of the least squares estimator (see Subsection 3.2 and 6.1) as employed in the derivation of the hat matrix and smoother matrix (Lemma 3.2 and 3.4) where the solution-space of the least squares estimator is made explicit for the purpose of statistical analysis (see e.g. (Rao, 1965)) as well as from a numerical point of view (see e.g. (Golub and van Loan, 1989)).

Example 2.3 **[Hierarchical programming with a QP]** The following example is prototypical. Let $Q, q \in \mathbb{R}^N$ be given vectors, $x \in \mathbb{R}$ the unknown parameter and let $c \in \mathbb{R}$ act as a fixed but unknown hyper-parameter. Consider the following QP optimization problem $\mathscr{J}_c^{(1)}$ on the first level

$$\textbf{Level 1:}\quad \min_x \mathscr{J}_c^{(1)}(x) = \frac{1}{2}\|Qx - q\|_2^2 \quad \text{s.t.} \quad x \leq c. \tag{2.56}$$

The Lagrangian then becomes

$$\mathscr{L}_c(x;\alpha) = \frac{1}{2}(Qx-q)^T(Qx-q) + \alpha(x-c), \tag{2.57}$$

where $\alpha \in \mathbb{R}^+$ is a single positive Lagrange multiplier. Necessary and sufficient conditions for the optimal solution x^* to (2.56) are given as follows

$$\mathrm{KKT}_{(2.56)}(x;\alpha,c)\begin{cases} Q^TQx - Q^Tq + \alpha = 0 & (a)\\ x - c \leq 0 & (b)\\ \alpha \geq 0 & (c)\\ \alpha(c-x) = 0 & (d) \end{cases} \tag{2.58}$$

Let $F, f \in \mathbb{R}^n$ be vectors. On the second level, one can e.g. consider the following hierarchical programming problem:

$$\textbf{Level 2:}\quad \min_{x;\alpha,c} \mathscr{J}^{(2)}(x;c) = \frac{1}{2}\|Fx - f\|_2^2 \quad \text{s.t.} \quad \mathrm{KKT}_{(2.56)}(x;\alpha,c). \tag{2.59}$$

The necessary conditions for optimality become

$$\mathrm{KKT}(x,\alpha,c;r,s,t,l)\begin{cases} \frac{\partial \mathscr{L}}{\partial x} = 0 \rightarrow & F^TFx - f^TF = l\alpha + Q^TQr - s\\ \frac{\partial \mathscr{L}}{\partial \alpha} = 0 \rightarrow & l(c-x) = r + t\\ \frac{\partial \mathscr{L}}{\partial c} = 0 \rightarrow & \varepsilon\alpha = s\\ & Q^TQx - Q^Tq + \alpha = 0\\ & x - c \leq 0\\ & \alpha \geq 0\\ \text{comp. slackn.} & l\alpha(x-c) \leq 0 \qquad\qquad (g)\\ \text{comp. slackn.} & s(c-x) = 0\\ \text{comp. slackn.} & t\alpha = 0, \end{cases} \tag{2.60}$$

where $r \in \mathbb{R}$ and $l, s, t \in \mathbb{R}^+$ are the associated multipliers of the Lagrangian
$\mathscr{L}(x,\alpha,c;r,s,t,l) = \mathscr{J}^2(x;c) - r(Q^TQx - Q^Tq + \alpha) + s(x-c) - t\alpha + l(\alpha(x-c))$.

To overcome the non-convex complementary slackness constraint (2.58.d), the following relaxation is proposed. Let $\varepsilon > 0$ be a constant such that

$$\begin{bmatrix} F^T F & \varepsilon \\ \varepsilon & 0 \end{bmatrix} \succeq 0, \tag{2.61}$$

such that the problem remanins convex, then the following relaxation $\mathscr{J}^{(2')}$ is convex and the solution $(\hat{x}, \hat{\alpha}, \hat{c})$ does satisfy the conditions (2.58).

$$\min_{x;\alpha,c} \mathscr{J}^{(2')}(w) = \frac{1}{2}\|Fx - f\|_2^2 + \varepsilon(c-x)\alpha \quad \text{s.t.} \quad \begin{cases} Q^T Qx - Q^T q + \alpha = 0 & (a) \\ x - c \leq 0 & (b) \\ \alpha \geq 0. & (c) \end{cases} \tag{2.62}$$

After constructing the Lagrangian $\mathscr{L}'(x,\alpha,c;r,s,t)$ of problem (2.62) with multipliers $r \in \mathbb{R}$ corresponding with (2.62.a) and $0 \leq s,t \in \mathbb{R}^+$ corresponding with the inequalities (2.62.bs), the following conditions for optimality holds:

$$\text{KKT}_{(2.62)}(x,\alpha,c;r,s,t) \begin{cases} \dfrac{\partial \mathscr{L}'}{\partial x} = 0 \to & F^T F x - f^T F = \varepsilon\alpha + Q^T Q r - s \\ \dfrac{\partial \mathscr{L}'}{\partial \alpha} = 0 \to & \varepsilon(c-x) = r + t \\ \dfrac{\partial \mathscr{L}'}{\partial c} = 0 \to & \varepsilon\alpha = s \qquad (c) \\ & Q^T Q x - Q^T q + \alpha = 0 \\ & x - c \leq 0 \\ & \alpha \geq 0 \\ \text{comp. slackn.} & s(c-x) = 0 \\ \text{comp. slackn.} & t\alpha = 0. \end{cases} \tag{2.63}$$

By comparing conditions (2.60) and (2.63), the only difference between the original problem and the relaxation is the role of the unknown l (Lagrange multiplier) in the former and ε (hyper-parameter) in the latter, together with the occurence of the equality $l(\alpha(x-c))$ in (2.63.g). However, from condition (2.63.c) it follows that condition (2.60.g) is always satisfied for $\varepsilon \neq 0$, and thus the optimum to (2.62) satisfies the KKT conditions (2.58). As the solution to the KKT conditions of (2.63) is identical for any value of ε, the relaxation provides necessary and sufficient conditions for the problem (2.59).

This example may be seen as an application of the augmented Hessian approach discussed in the previous subsection.

Part I

α

Chapter 3

Primal-Dual Kernel Machines

This chapter presents an overview of the application of the primal-dual optimization framework to the inference of regression functions and classification rules from a finite set of observed data-samples. The aim of the chapter is to provide a sound and general basis towards the design of algorithms relying on the theory of convex optimization. While historical breakthroughs mainly focussed on the case of classification, this chapter mainly considers the regression case. Section 3.2 discusses general parametric and classical kernel-based methods, while Section 3.3 studies one of the most straightforward formulations leading to the standard Least Squares Support Vector Machine (LS-SVM). This formulation is studied in some detail as it will play a prototypical role in the remainder. Section 3.4 then proceeds with the derivation of the Support Vector Machine (SVM) for regression. Section 3.5 gives a variation on the theme by proposing a primal-dual kernel machine for interval estimation, coined as the Support Vector Tube (SVT). Section 3.6 considers a number of extensions of the previous methods to the context of outliers and Section 3.7 reports a number of results in the context of classification.

3.1 Some Notation

Before going into the subject, some notation is introduced. Let $\mathbf{X} \in \mathbb{R}^D$ and $\mathbf{Y} \in \mathbb{R}$ be random variables as described in Subsection 1.1.1. Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}$ be a collection of observed i.i.d. data-samples as in Subsection 1.1. Let there be a mapping $f : \mathbb{R}^D \rightarrow \mathbb{R}$ such that $E[\mathbf{Y}|\mathbf{X} = x] = f(x)$ and $\mathrm{var}[\mathbf{Y}|\mathbf{X} = x] < \infty$ for all $i = 1, \ldots, N$. In most cases, the vector (x_i, y_i) is sampled from the random vector $(\mathbf{X}, \mathbf{Y})$, but one often makes the assumption that $\mathrm{var}(\mathbf{X}) \ll \mathrm{var}(\mathbf{Y})$ such that the samples x can be considered to be deterministic.

Let x_i^d denote the dth variable of the ith sample with $1 \leq d \leq D$. One can organize these values into a matrix as $X = (x_1, \ldots, x_N)^T \in \mathbb{R}^{N \times D}$. Let a superscript denote the column or the variable and let a subscript denote a sample index. Then $X_i = x_i$ and X^d contains the samples of the dth variable. Let $Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$ and $e = (e_1, \ldots, e_N)^T \in \mathbb{R}^N$ be vectors.

3.2 Parametric and Non-parametric Regression

3.2.1 Regression as conditional mean

The regression estimate which is optimal in the expected integrated square error sense corresponds with the conditional mean, see e.g. (Hastie *et al.*, 2001) and references

$$f(x) = E\left[\mathbf{Y}|\mathbf{X} = x\right] = \int y p_{\mathbf{Y}|\mathbf{X}}(y|x) dy = \int y \frac{p_{\mathbf{XY}}(x,y)}{p_{\mathbf{X}}(x)} dy. \tag{3.1}$$

This result is somewhat similar to the optimal Bayes classifier (1.26), see e.g. (Hastie *et al.*, 2001) for a survey.

3.2.2 Parametric regression estimates

It is instructive to relate the general formulation (3.1) to the linear least squares problem. The following stochastic model underlying the chance regularities is postulated classically.

Lemma 3.1. [Gauss-Markov Conditions] *Let $\{x_i\}_{i=1}^N$ be samples from the random variable $\mathbf{X}$ such that $E[\mathbf{X}^2] \ll E[\mathbf{Y}^2]$. Let $\omega \in \mathbb{R}^D$ be fixed (deterministic) but unknown. A linear model is postulated $f(x) = \omega^T x$ to underly the observations $\mathscr{D}$ such that the relation*

$$y_i = \omega^T x_i + e_i \tag{3.2}$$

holds where the noise sequence $\{e_i\}_{i=1}^N$ sampled from the random variable $\mathbf{e}$ satisfies the Gauss-Markov conditions, see e.g. (Rao, 1965; Neter et al.*, 1974):*

(i.i.d.) *Let the sequence $\{e_i\}_{i=1}^N$ be a sequence of i.i.d. samples from the random variable $\mathbf{e}$*

(zero mean) *$E[\mathbf{e}|X = x] = E[\mathbf{e}] = 0$ for all $x \in \mathbb{R}^D$*

(uncorrelated) *Let $0 < \sigma_e^2 < \infty$, then $E[e_i e_j] = \delta_{ij}\sigma_e^2$ where $\delta_{ij} = 1$ if $i = j$ and zero elsewhere.*

The parameter vector $\omega \in \mathbb{R}^D$ can be estimated in least squares sense

$$\hat{w} = \arg\min_w \sum_{i=1}^N (w^T x_i - y_i)^2, \tag{3.3}$$

which also equals the maximum (log) likelihood (ML) estimate following from the assumption that **e** possesses a Gaussian distribution and thus $y_i \sim \mathscr{N}(\omega^T x_i, \sigma_e^2)$ (Fisher, 1922), see e.g. (Rice, 1988). The global solution is characterized by its first order conditions of optimality

$$(X^T X)\, w = X^T Y, \tag{3.4}$$

which are referred to as the normal equations. Due to the Gauss-Markov theorem, the estimator $\hat{w}$ solving (3.4) possesses the BLUE property (Best Linear Unbiased Estimator) under the given assumptions (Neter *et al.*, 1974; Rice, 1988). The least squares estimator has the following interpretation via the hat matrix.

Lemma 3.2. [Hat matrix] *Assume the function underlying the observations $\mathscr{D}$ takes the form of (3.2) and the errors are satisfy the Gauss-Markov conditions. The least squares smoother can be written as a linear operator H as follows*

$$\hat{Y} = HY \;\; with \;\; H = X(X^T X)^{-1} X^T. \tag{3.5}$$

where $H \in \mathbb{R}^{N\times N}$ is referred to as the hat matrix. The following properties hold

1. *H is symmetric positive semi-definite (denoted as $H \succeq 0$)*
2. *The rank of H provides a measure of the (effective) dimensions of the fitted model*
3. *H is idempotent i.e. $H = H^2$.*

The proofs can be found in any statistical work concerning linear regression, see e.g. (Rao, 1965; Neter *et al.*, 1974).

Example 3.1 [Loss functions and noise distributions] An illustrative example is given of the parameter estimation task in the context of different noise models and using estimators employing different norms. Four different estimators are considered using the convex cost-functions defined as follows

$$\begin{cases} \mathscr{J}_1(w) & = \sum_{i=1}^N |w^T x_i - y_i| & (a) \\ \mathscr{J}_H(w) & = \sum_{i=1}^N \ell_H(w^T x_i - y_i) & (b) \\ \mathscr{J}_2(w) & = \sum_{i=1}^N (w^T x_i - y_i)^2 & (c) \\ \mathscr{J}_\infty(w) & = \max_{i=1}^N |w^T x_i - y_i| & (d) \end{cases} \tag{3.6}$$

where the Huber loss function ℓ_H is defined later-on in (3.59) and the constant c in the Huber loss function is as commonly fixed as $c = 1.345\sigma_e^2$ (Huber, 1964). Consider the linear model (3.2) with $D = 5$, $N = 100$, X_d taken random and independently from the interval $[-1,1]^N$ for all $d = 1,\dots,D$ and ω chosen uniformly from $[-5,5]^D$. Four different noise models were added (i) a Laplacian noise model $e_i \sim \mathscr{L}(0,1.5)$, (ii) a Gaussian noise model $e_i \sim \mathscr{N}(0,1)$ and (iii) a contaminated noise model with a Gaussian nominal model and 5% outliers with variance 10, (iv) a uniform noise model $e_i \sim \mathscr{U}([-1.5,1.5])$. The performance is expressed in the mean squared error of the estimate $\hat{w} = \arg\min_w \mathscr{J}(w)$ to the true parameter ω. The boxplots [1] of Figure 3.1 show the results of a Monte Carlo study with 1000 iterations.

[1] A boxplot is a compact representation of a distribution, based on a number of order statistics as the median. From top to bottom, a boxplot displays respectively the upper outliers, upper-quartile plus 1.5 inter-quartile range, the upper-quartile , the median, lower-quartile, lower-quartile minus 1.5 times inter-quartile range, and all lower outliers.

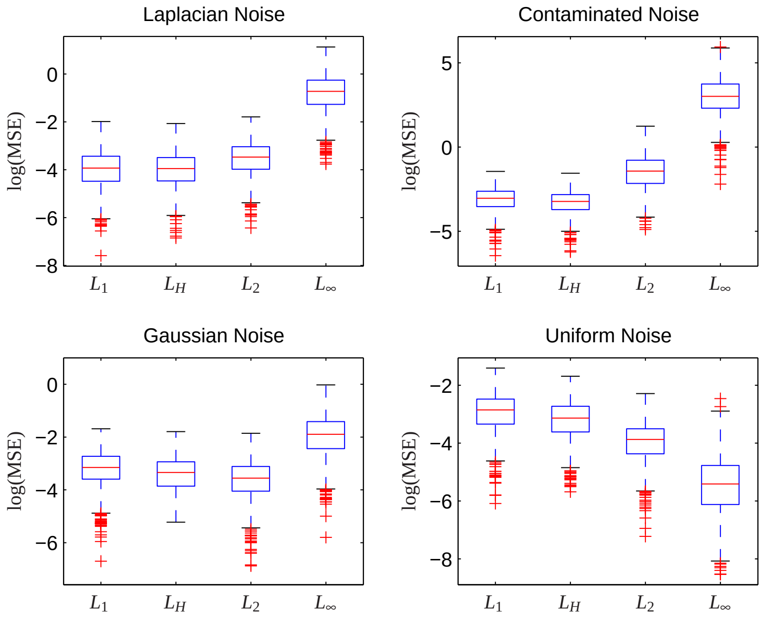


Figure 3.1: *Numerical results of a Monte Carlo study relating the average MSE of the estimate and the true parameter corresponding with a specific noise model and chosen norm. This simulation results emphasize the importance of choosing an appropriate norm corresponding to the underlying noise model.*

This numerical example illustrates the fact that the choice of the most efficient lossfunction depends on the underlying distribution of the perturbations. More specificly, this figure supports the theoretical results of maximum likelihood relating the optimal costfunction $(L_2, L_1, L_\infty, L_H)$ to the corresponding noise model (respectively (i), (ii), (iii) and (iv)).

3.2.3 Non-Parametric regression estimates

Consider the Parzen window estimator (Parzen, 1962) for non-parametric density estimation (see example 1-1, 1.13 and 1.14). The Nadarya-Watson non-parametric kernel regression estimator then follows immediately from (3.1):

$$\hat{f}(x) = \int y \frac{p_{\mathbf{XY}}(x,y)}{p_{\mathbf{X}}(x)} dy = \frac{\sum_{i=1}^N K(x,x_i) y_i}{\sum_{j=1}^N K(x,x_j)}, \tag{3.7}$$

where $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ denotes a nonnegative weight function centered around zero with bandwidth h as defined in example 1.1, see e.g. (Watson, 1964). This estimator has various optimality properties as described e.g. in (Rao, 1983) and often acts as a tool for exploratory data analysis and for testing procedures.

3.3 L_2 Kernel Machines: LS-SVMs

Consider the following class of models linear in the parameters

$$\mathscr{F}_\varphi = \left\{ f(x) = \omega^T \varphi(x) \;\middle|\; \omega \in \mathbb{R}^{D_\varphi} \right\}, \tag{3.8}$$

where the mapping $\varphi : \mathbb{R}^D \to \mathbb{R}^{D_\varphi}$ is fixed but unknown and can be infinite dimensional. Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$ satisfy the relation $y_i = f(x_i) + e_i$ where $f : \mathbb{R}^D \to \mathbb{R}$ is fixed and e_i is i.i.d. sampled from a random variable $\mathbf{e}$ with a fixed but unknown distribution satisfying $E[\mathbf{e}|\mathbf{X} = x] = 0$ and $E[\mathbf{e}^2] = \sigma_e^2 < +\infty$. Extensions of this model towards additional parametric terms (as the so-called intercept term) are discussed extensively in the following chapter. This description of the model is referred to as the *primal model* being related to the to the following primal optimization problem. Consider the regularized least squares loss function with hyper-parameter $\gamma > 0$:

$$(\hat{w}, \hat{e}) = \arg\min_{w,e} \mathscr{J}_\gamma(w, e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^N e_i^2$$
$$\text{s.t.} \quad w^T \varphi(x_i) + e_i = y_i, \;\; \forall i = 1, \ldots, N, \tag{3.9}$$

which is also referred to as ridge regression in feature space, see also (Saunders *et al.*, 1998). The Lagrangian of this constrained optimization problem becomes

$$\mathscr{L}_\gamma(w, e; \alpha) = \mathscr{J}_\gamma(w, e_i) - \sum_{i=1}^N \alpha_i \left(w^T \varphi(x_i) + e_i - y_i \right). \tag{3.10}$$

The first order (necessary and sufficient) conditions for optimality are given as

$$\text{KKT}(w, e; \alpha) \begin{cases} \dfrac{\partial \mathscr{L}_\gamma}{\partial w} = 0 \to \quad w = \sum_{i=1}^N \alpha_i \varphi(x_i) & & (a) \\ \dfrac{\partial \mathscr{L}_\gamma}{\partial e_i} = 0 \to \quad \gamma e_i = \alpha_i & \forall i = 1, \ldots, N & (b) \\ \dfrac{\partial \mathscr{L}_\gamma}{\partial \alpha_i} = 0 \to \quad w^T \varphi(x_i) + e_i = y_i & \forall i = 1, \ldots, N. & (c) \end{cases} \tag{3.11}$$

Eliminating the possibly infinite dimensional parameter w and the residuals e, one obtains an equivalent dual system expressed in the Lagrange multipliers using matrix formulations as

$$\left(\Omega + \frac{1}{\gamma} I_N \right) \alpha = Y, \tag{3.12}$$

where $\alpha = (\alpha_1, \ldots, \alpha_N)^T \in \mathbb{R}^N$ is a vector, $I_N \in \mathbb{R}^{N \times N}$ denotes the identity matrix and $\Omega \in \mathbb{R}^{N \times N}$ represents the kernel matrix defined as follows. Let Φ_N denote the mapped training data points $\Phi_N = (\varphi(x_1), \ldots, \varphi(x_N))^T \in \mathbb{R}^{N \times D_\varphi}$, then one defines the kernel matrix as $\Omega = \Phi_N^T \Phi_N$. Let a Mercer kernel function $K : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ be defined as an inner product

$$\varphi(x_i)^T \varphi(x_j) \triangleq K(x_i, x_j) \quad \forall x_i, x_j \in \mathscr{D}. \tag{3.13}$$

The following subsection elaborates on the duality between the kernel function and the mapping φ.

The final estimate $(\hat{w}, \hat{e}) = \arg\min_{w,e} \mathscr{J}_\gamma(w, e)$ can be evaluated in a new point $x_* \in \mathbb{R}^D$ in terms of the multipliers and the inner product $K(x_i, x_*) = \varphi(x_i)^T \varphi(x_*)$ as follows

$$\hat{f}(x_*) = \sum_{i=1}^N \hat{\alpha}_i K(x_i, x_*) = \Omega_{\mathscr{D}}(x_*)^T \hat{\alpha}, \tag{3.14}$$

where $\hat{\alpha}_i$ solve (3.12) for all $i = 1 \ldots, N$. Here, the mapping $\Omega_{\mathscr{D}} : \mathbb{R}^D \times \mathscr{D} \rightarrow \mathbb{R}^N$ is defined as $\Omega_{\mathscr{D}}(x_*) = (K(x_i, x_*), \ldots, K(x_N, x_*))^T \in \mathbb{R}^N$.

Lemma 3.3. *The dual problem to (3.9) becomes*

$$\max_\alpha \mathscr{J}_\gamma^D(\alpha) = \frac{1}{2} \alpha^T \left(\Omega + \frac{1}{\gamma} I_N \right) \alpha - Y^T \alpha, \tag{3.15}$$

from which not only the training solutions (3.12) follow, but also the Hessian $\mathscr{H}_e = \left(\Omega + \frac{1}{\gamma} I_N \right)$ *can be derived readily.*

A detailed derivation on the variance of the estimator can be found in subsection 6.2. Similar to the Hat matrix described in Lemma 3.2, one can reformulate the LS-SVM as a linear operator as follows

Lemma 3.4. [Smoother Matrix] *The estimated values $\hat{Y}$ of the given training datapoints Y using the model class (3.8) and regularized least squares cost-function (3.9) follow from the linear operator $S_\gamma \in \mathbb{R}^{N \times N}$ which is defined as follows*

$$\hat{Y} = S_\gamma Y \;\; \textit{where} \;\; S_\gamma = \Omega \left(\Omega + \frac{1}{\gamma} I_N \right)^{-1}. \tag{3.16}$$

The following properties hold:

1. *S_γ is symmetric positive semi-definite $S_\gamma = S_\gamma^T \succeq 0$ (Boyd and Vandenberghe, 2004).*
2. *The smoother matrix has a shrinking nature, meaning that $S_\gamma^2 \preceq S_\gamma$ or $S_\gamma^2 - S_\gamma$ is negative definite. Note the difference with the Hat matrix (see Lemma 3.2) which is idempotent.*
3. *The rank of the smoother matrix $\Gamma(S_\gamma) \leq N$ is an indication of the* number of degrees of freedom *or the* effective number of parameters *as argued in (Mallows, 1973). This motivated the following definition*

$$D_{\text{eff}} = \text{tr}(S_\gamma) = \sum_{i=1}^N \frac{\lambda_i}{\lambda_i + \gamma^{-1}}, \tag{3.17}$$

where $\Lambda = (\lambda_1, \ldots, \lambda_N)^T \in \mathbb{R}^N$ *denotes the eigenvalues of the kernel matrix* $\Omega \in \mathbb{R}^{N\times N}$.

Note that the smoother matrix is also positive definite, and consists as such of the elements of a positive definite function which is sometimes referred to as the dual kernel (Hardle, 1990; Girosi *et al.*, 1995). The smoother matrix has an important role into various model selection criteria as the PRESS statistic (Allen, 1974) and the generalized cross-validation measure (Golub *et al.*, 1979).

3.3.1 Mercer theorem and kernel trick

The Mercer theorem (Mercer, 1909; Aronszajn, 1950) was formulated as follows

Theorem 3.1. [Mercer Theorem] *Let* $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ *be in* $L^2(C)$ *where* C *denotes a compact subset of* $\mathbb{R}^D$. *To guarantee that the function* $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ *has an expansion of the form*

$$K(x,y) = \sum_{j=1}^{\infty} a_j \phi_j(x)^T \phi_j(y) \;\; \forall x,y \in \mathbb{R}^D, \tag{3.18}$$

with positive coefficients $a_j \geq 0$, *a set of mappings* $\{\phi_j : \mathbb{R}^D \to \mathbb{R}^{D_\varphi}\}_{j=1}^{\infty}$ *and* $D_\varphi \in \{\mathbb{N}_0, +\infty\}$, *it is necessary and sufficient that*

$$\int_C \int_C K(x,y) g(x) g(y) dx dy \geq 0, \tag{3.19}$$

be valid for any function $g : \mathbb{R}^D \to \mathbb{R}$ *in* $L^2(C)$.

This means that any kernel function K corresponds with an inner product in a corresponding feature space

$$\exists \;\; \varphi : \mathbb{R}^D \to \mathbb{R}^{D_\varphi} \quad \text{s.t.} \quad K(x,y) = \varphi(x)^T \varphi(y) \;\; \forall x,y \in C, \tag{3.20}$$

as long as the function K is positive semi-definite. This classical result was introduced in the literature by (Aizerman *et al.*, 1964). The consequence is that if one fixes a kernel function K, one works explicitly with a feature space which is induced by this kernel. As such, there is no need for the mapping φ to be defined explicitly as long as the model can be expressed completely in terms of inner-products between (mapped) data-points. This principle is often referred to as the *kernel trick* (Vapnik, 1998), see e.g. (Schölkopf and Smola, 2002).

3.3.2 Primal-dual interpretation

One can now properly define the concept of a primal-dual kernel machine.

Definition 3.1. [Primal-dual Kernel Machines] *A primal-dual kernel machine consists of a model formulation which possesses a primal and a dual representation in the sense of (Lagrangian) optimization theory. The primal representation is used to formulate the optimality principle underlying the model as a constrained optimization problem based on the training-set and all available prior knowledge, while the dual representation refers to the characterization of the problem in the Lagrange multipliers enabling the application of the kernel trick.*

Note the difference with the primal-dual optimization methods in the context of algorithms for (generic) convex optimization problem as described in Section 2.3. It is instructive to discuss the conditions for optimality (3.11) in detail as those will re-occur in most derivations of primal-dual kernel machines.

1. Condition (3.11.a) relates the parameters w of the fitted model to the finite set of Lagrange multipliers. This condition goes along the same lines as the Representer theorem (Craven and Wahba, 1979), see Section 5.1. Note that this relation holds as long as the L_2 norm of the parameters ($w^T w$) is considered. It will be crucial in all primal-dual kernel machine formulations.
2. Condition (3.11.b) states that the ith Lagrange multiplier is proportional to the ith residual e_i with a factor γ. This property is specific to the use of the L_2 loss function. It will be important in the realization approach for learning the kernel as elaborated in Chapter 9.2.2.
3. Condition (3.11.c) repeats the original constraints.

Advantages of the use of primal-dual derivations of kernel machines include the properties following from the derived KKT conditions for optimality (as the box constraints in the case of SVM) and the sensitivity interpretation related to the Lagrange multipliers (as elaborated next) following from the theory of convex optimization At this stage, one can state the duality between the estimated parameter w and the residuals e_i more clearly. Eliminating the Lagrange multipliers α_i from condition (3.11.a) using condition (3.11.b) results into the equation

$$\hat{w} = \gamma \sum_{i=1}^{N} \hat{e}_i \varphi(x_i), \tag{3.21}$$

stating that the model (parameters) and the noise terms are not only related via the model definition, but also in a more direct way.

Example 3.2 **[Learning Machine based on Fourier Decompositions]** Consider the case of a finite mapping of the observed data into a feature space using the Fourier decomposition. Let $\{x_i\}_{i=1}^N$ be equidistantly sampled on the interval $[0,2\pi]$ such that $x_i = 2\pi \frac{i-1}{N}$ for all $i = 1,\ldots,N$. Define the mapping to feature space $\varphi : [0,2\pi] \rightarrow \mathbb{R}^{2N+1}$ as follows (Vapnik, 1998)

$$\textbf{mapping:} \;\; \varphi(x) = \left(\frac{1}{\sqrt{2}}, \sin(x), \ldots, \sin(Nx), \cos(x), \ldots, \cos(Nx)\right), \tag{3.22}$$

such that the feature space has a dimensionality of $D_\varphi = 2N+1$. The corresponding inner product becomes

$$\textbf{kernel:}\ \ K(x_i,x_j) = \frac{1}{2} + \sum_{k=1}^{N} \big(\sin(kx_i)\sin(kx_j) + \cos(kx_i)\cos(kx_j)\big)\,. \tag{3.23}$$

Let $w = (w_0, w_1, \ldots, w_N, w_{N+1}, \ldots, w_{2N}) \in \mathbb{R}^{2N+1}$ be the parameter vector. The primal linear model then becomes

$$\textbf{function:}\ \ f(x) = w^T\varphi(x) = \frac{w_0}{\sqrt{2}} + \sum_{k=1}^{N} w_k \sin(kx) + \sum_{k=1}^{N} w_{N+k}\cos(kx). \tag{3.24}$$

Consider the ridge regression loss function

$$\textbf{cost:}\ \ \mathscr{J}_{\mathscr{F}}(w) = \frac{1}{2}w^Tw + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad w^T\varphi(x_i) + e_i = y_i. \tag{3.25}$$

The dual solution follows from solving (3.15) and the optimum takes the form

$$\textbf{Dual estimate:}\ \ \hat{f}(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x). \tag{3.26}$$

where α_i for all $i = 1, \ldots, N$ are the Lagrange multipliers characterizing the dual solution. The estimated model $\hat{f}$ has Fourier coefficients

$$\textbf{Primal estimate:}\ \ \big(\mathscr{F}\hat{f}\big)(k) = \sum_{i=1}^{N} \alpha_i\ \left(\sin(kx_i) + \cos(kx_i)\right). \tag{3.27}$$

Example 9.1 further studies this setting towards the context of more elaborate regularization schemes and infinite feature space mappings.

A similar primal-dual derivation formed the basis towards new interpretations of unsupervised learning problems for kernel PCA following (Schölkopf and Smola, 2002) in (Suykens *et al.*, 2003*b*), see also (Suykens *et al.*, 2002*b*) for extra results on Kernel Canonical Correlation Analysis (KCCA) and Kernel Partial Least Squares (KPLS).

3.3.3 Sensitivity interpretation

This subsection studies the relationship of the dual representation and the sensitivity of the solution to small perturbations in the observations. The following definition is taken from Hampel (Hampel, 1974; Hampel *et al.*, 1986).

Definition 3.2. [Influence Function] *Let A denote a statistical functional mapping a random vector* $(\mathbf{X},\mathbf{Y})$*, and a distribution P on* $\mathbb{R}$*. The influence function of A with the (theoretical) nominal model* $P(\mathbf{X},\mathbf{Y})$ *underlying a dataset* $\mathscr{D}$ *and a pointmass distribution* Δ *is then defined as*

$$\text{IF}(A,P,\Delta) = \lim_{\varepsilon\downarrow 0} \frac{A\big((\mathbf{X},\mathbf{Y}),(1-\varepsilon)P(\mathbf{X},\mathbf{Y}) + \varepsilon\Delta, \mathscr{A}\big) - A\big((\mathbf{X},\mathbf{Y}),P\big)}{\varepsilon}. \tag{3.28}$$

The most important empirical versions are the sensitivity curve (Tukey, 1977), and the Jackknife (Tukey, 1958), based on addition and replacement respectively. The latter is considered. Let $\mathscr{D}^{-i}$ denote the dataset without the ith sample.

$$\hat{\mathrm{IF}}(\mathsf{Alg}, \mathscr{D}, \delta_i) = \lim_{\delta \to 0} \frac{\mathsf{Alg}\left(\mathscr{D}, \mathscr{A}\right) - \mathsf{Alg}\left(\{\mathscr{D}^{-i}, (x_i, y_i + \delta_i)\}, \mathscr{A}\right)}{\delta}. \tag{3.29}$$

This statistical concept is closely related to the perturbation and sensitivity interpretation of the Lagrange multipliers as reviewed in subsection 2.2.2. Let $\mathsf{Alg}^* : \mathscr{D} \times \mathscr{A} \times \mathbb{R} \to \mathbb{R}^D$ be defined as follows

$$\mathsf{Alg}^*(\mathscr{D}, \mathscr{A}, \delta_i) = \arg\min_{w,e} J_\gamma(w, e, \delta_i) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{k=1}^{N} e_k^2$$
$$\text{s.t.} \quad \begin{cases} w^T \varphi(x_j) + e_j = y_j & \forall j \neq i \\ w^T \varphi(x_i) + e_i + \delta_i = y_i, \end{cases} \tag{3.30}$$

returning the optimum when varying the ith constraint by adding a perturbation δ_i on the ith constraint.

Lemma 3.5. [Sensitivity of LS-SVMs] *The sensitivity of the estimate $\hat{\alpha}_i$ on the ith data-sample is given as follows*

$$\left. \frac{\partial \mathsf{Alg}^*(\mathscr{D}, \mathscr{A}, \delta)}{\partial e_i} \right|_{\delta=0} = \lim_{\delta \to 0} \frac{\mathsf{Alg}^*(\mathscr{D}, \mathscr{A}, 0) - \mathsf{Alg}^*(\mathscr{D}, \mathscr{A}, \delta)}{\delta} = -\hat{\alpha}_i. \tag{3.31}$$

The sensitivity of the estimate $\hat{w}$ and the prediction $\hat{f}(x_*)$ with $x_* \in \mathbb{R}^D$ is thus given as

$$\begin{cases} \frac{\partial \hat{w}}{\partial e_i} = \hat{\alpha}_i \varphi(x_i) \\ \frac{\partial \hat{f}(x_*)}{\partial e_i} = \hat{\alpha}_i K(x_i, x_*). \end{cases} \tag{3.32}$$

From this, the estimated model (3.14) can be interpreted as the sum of the empirical influences of the given data-samples.

3.3.4 Bounding the L_2 risk

This formulation was also coined also as kernel ridge regression (Saunders *et al.*, 1998), under which name it received considerable attention from a statistical learning point of view (Shawe-Taylor and Cristianini, 2004). Hence the following theorem

Theorem 3.2. [Bounding the L_2 Risk] *Let $0 < \varepsilon \ll N$ be a constant. Let $f : \mathbb{R}^D \to \mathbb{R}$ be contained in the class $\mathscr{F}_\varphi$ (3.8) with bounded norm $B \in \mathbb{R}^+$ such that $\|w\|_2^2 \leq B$. Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$ be sampled i.i.d. from a fixed but unknown distribution $P_{\mathbf{XY}}$. Let the L_2 risk of a function f be defined as*

$$\mathscr{R}_2(f, P_{\mathbf{XY}}) = \int (f(x) - y)^2 \, dP_{\mathbf{XY}}(xy). \tag{3.33}$$

Its empirical counterpart may be defined as follows

$$\hat{\mathscr{R}}_2(w,\mathscr{D}) = \frac{1}{N}\sum_{i=1}^{N}\left(w^T\varphi(x_i) - y_i\right)^2. \tag{3.34}$$

If the mapped data-points $\{\varphi(x_i)\}_{i=1}$ are contained in a ball with radius R and origin zero, one can bound the risk as follows

$$\text{Prob}\Bigg(\left|\mathscr{R}_2(w,P_{\mathbf{XY}}) - \hat{\mathscr{R}}_2(w,\mathscr{D})\right| \\ \leq \frac{16RB}{N}\left(B\sqrt{\text{tr}(\Omega)} + \|Y\|_2\right) + 12(RB)^2\sqrt{\frac{\ln(2/\varepsilon)}{2N}}\Bigg) \geq (1-\varepsilon), \tag{3.35}$$

where $\Omega = \Phi_N\Phi_N^T$ denotes the kernel matrix as before and Y denote the vector containing the N observed outputs.

From this result, it follows that the estimator (3.9) also minimizes the theoretical risk if $N \to \infty$ and $B < \infty$. Traditional statistics often prefers the analysis of this estimator from the point of view of bias-variance trade-off as elaborated in Chapter 6.

3.4 L_1 and ε-loss Kernel Machines: SVMs

Instead of the common L_2-based approach, an L_1 norm based norm is sometimes preferred, although it is both practically as theoretically less covenient. Use of the L_1 norm can be motivated as an appropriate noise scheme (e.g. Laplacian distribution, see e.g. example 3.1) can be assumed or the method should be more robust to outliers than a least squares estimator. The derivations are summarized in the following Lemma.

Lemma 3.6. [SVMs for regression] *Consider the model class $\mathscr{F}_\varphi$ of (3.8). Let the ε-loss function be defined as $|e|_\varepsilon = \max(0,|e|-\varepsilon)$ (Vapnik, 1998). The regularized ε-loss estimate follows from solving the optimization problem*

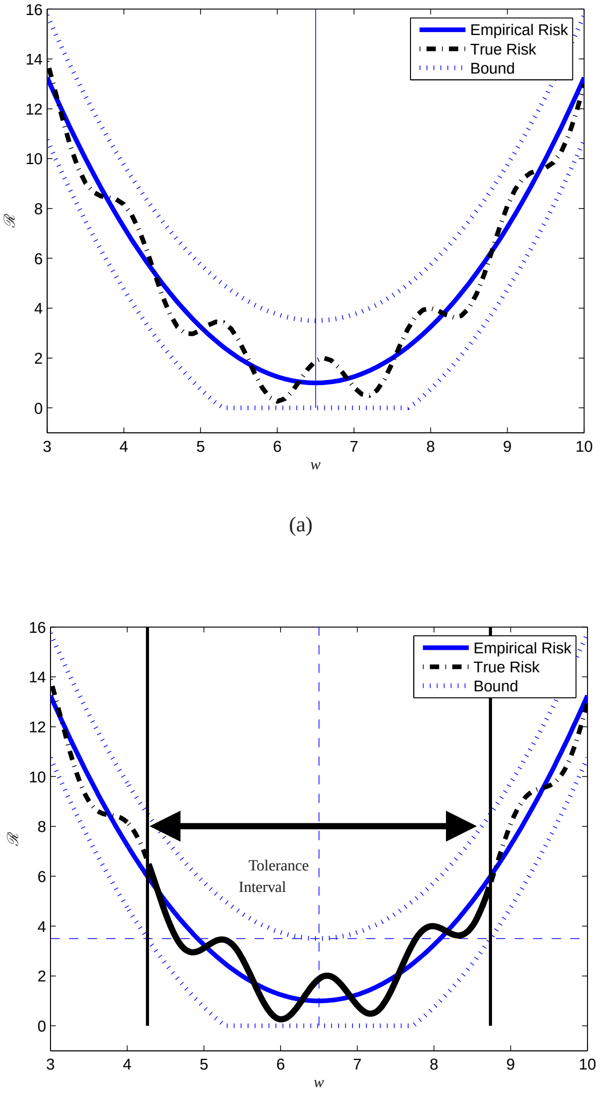
$$(\hat{w},\hat{e}) = \arg\min_{w,e} \mathscr{J}_{C,\varepsilon}(w,e_i) = \frac{1}{2}w^Tw + C\sum_{i=1}^{N}|e_i|_\varepsilon \\ \textit{s.t.}\quad w^T\varphi(x)_i + e_i = y_i \;\; \forall i = 1,\ldots,N. \tag{3.36}$$

This is equivalent to the dual optimization problem

$$\max_{\alpha^+,\alpha^-} -\frac{1}{2}(\alpha^- - \alpha^+)^T\Omega(\alpha^- - \alpha^+) + Y^T(\alpha^- - \alpha^+) - \varepsilon 1_N^T(\alpha^- + \alpha^+) \\ \textit{s.t.}\quad (\alpha_i^- + \alpha_i^+) \leq C,\ \forall\alpha_i^+,\alpha_i^- \geq 0,\ \forall i = 1,\ldots,N \tag{3.37}$$

where $\alpha^- = (\alpha_1^-,\ldots,\alpha_N^-)^T \in \mathbb{R}^N$ $\alpha^+ = (\alpha_1^+,\ldots,\alpha_N^+)^T \in \mathbb{R}^N$ are the positive Lagrange multipliers. The resulting function $\hat{f}$ can be evaluated at a new point $x_* \in \mathbb{R}^D$ as

$$\hat{f}(x_*) = \Omega_{\mathscr{D}}(x_*)^T(\hat{\alpha}^- - \hat{\alpha}^+). \tag{3.38}$$



(a)

(b)

Figure 3.2: *Illustration of the principle behind bounding the empirical risk.* **(a)** *Statistical learning theory provides bounds on the worst case deviation of the risk of a function in terms of the empirical risk and the capacity of the functions.* **(b)** *Using the upper bound (3.35), the empirical risk minimizer will converge to the theoretical risk minimizers when $N \longrightarrow \infty$ and $B < \infty$. If minimal empirical risk is attained (dashed vertical line in), the minimizer of the true risk must satisfies the interval indicated by the black arrows with high probability.*

Proof. One can reformulate the ε-loss $\max(0,|e_i-\varepsilon|)$ by using the slack variables $\xi_i = \max(0,|e_i-\varepsilon|) \in \mathbb{R}^{+,N}$ as follows

$$\xi_i \quad \text{s.t.} \quad -(\xi_i+\varepsilon) \le w^T\varphi(x_i) - y_i \le (\varepsilon+\xi_i), \;\; \xi_i \ge 0. \tag{3.39}$$

Employing this change of variables in the cost function (3.36), the Lagrangian becomes

$$\begin{aligned}\mathscr{L}_{C,\varepsilon}(w,\xi_i;\alpha^+,\alpha^-,\beta) = \frac{1}{2}w^Tw + C\sum_{i=1}^N \xi_i + \sum_{i=1}^N \alpha_i^+\left[(w^T\varphi(x_i)-y_i)-(\xi_i+\varepsilon)\right]\\ + \sum_{i=1}^N \alpha_i^-\left[-(w^T\varphi(x_i)-y_i)-(\xi_i+\varepsilon)\right] - \sum_{i=1}^N \beta_i\xi_i,\end{aligned} \tag{3.40}$$

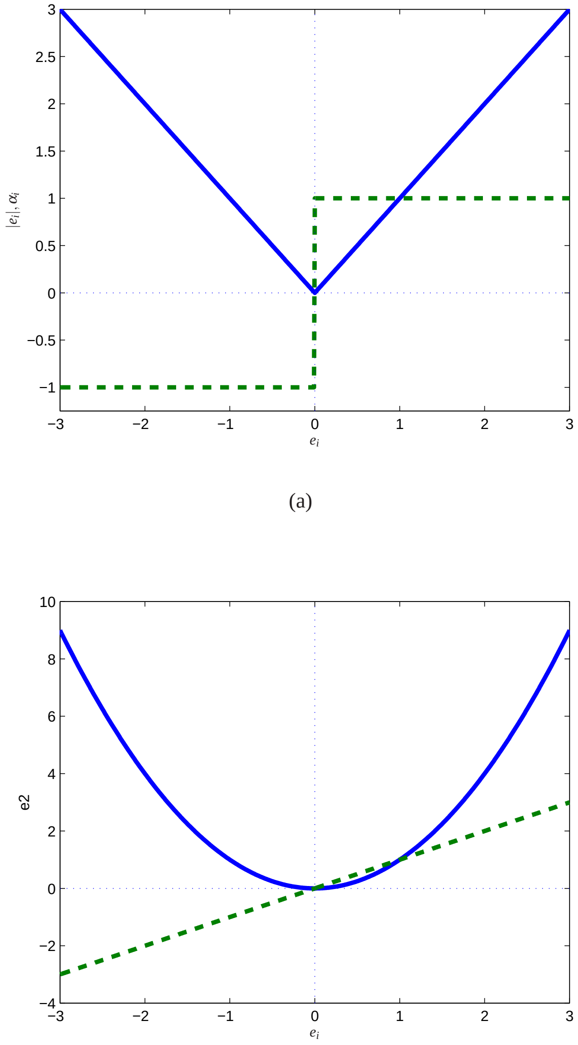
with positive multipliers $\alpha^+ = (\alpha_1^+,\ldots,\alpha_N^+)^T \in \mathbb{R}^{+N}$, $\alpha^- = (\alpha_1^-,\ldots,\alpha_N^-)^T \in \mathbb{R}^{+N}$ and $\beta = (\beta_1,\ldots,\beta_N)^T \in \mathbb{R}^{+N}$. The necessary and sufficient conditions for optimality are given as

$$\text{KKT}\left\{\begin{array}{llll} \dfrac{\partial \mathscr{L}_{C,\varepsilon}}{\partial w} = 0 \rightarrow & w = \sum_{i=1}^N(\alpha_i^- - \alpha_i^+)\varphi(x_i) & & (a)\\ \dfrac{\partial \mathscr{L}_{C,\varepsilon}}{\partial \xi_i} = 0 \rightarrow & C = \alpha_i^- + \alpha_i^+ + \beta_i & \forall i = 1,\ldots,N & (b)\\ & \alpha_i^+,\alpha_i^-,\beta_i \ge 0 & \forall i = 1,\ldots,N & (c)\\ & -(\xi_i+\varepsilon) \le w^T\varphi(x_i) - y_i \le (\xi_i+\varepsilon) & \forall i = 1,\ldots,N & (d)\\ & \xi_i \ge 0 & \forall i = 1,\ldots,N & (e)\\ & \alpha_i^+\left[(w^T\varphi(x_i)-y_i)-(\xi_i+\varepsilon)\right] = 0 & \forall i = 1,\ldots,N & (f)\\ & \alpha_i^-\left[-(w^T\varphi(x_i)-y_i)-(\xi_i+\varepsilon)\right] = 0 & \forall i = 1,\ldots,N & (g)\\ & \beta_i\xi_i = 0. & \forall i = 1,\ldots,N & (h)\end{array}\right. \tag{3.41}$$

Alternatively, one can reformulate the optimization problem (3.36) as a saddle-point problem $\min_{w,\xi}\max_{\alpha^+,\alpha^-,\beta}$ or in its dual form as in (3.37) after elimination of the primal unknowns w,ξ and the dual multipliers β. The obtained estimator $\hat{w}^T\varphi(\cdot)$ can be evaluated in a new point using only the dual variables as in (3.38). □

This formulation was coined as the Support Vector regressor (SVM regressor) (Vapnik, 1998). Note the correspondence between the dual representation of the solution to the L_2 (3.14) and the L_1 kernel machine (3.38). The representer theorem states that this correspondence is not a coincidence. In the language of SVM, the non-sparse Lagrange multipliers α_i are denoted as support values and the corresponding vectors x_i are called support vectors. Note that sparseness here results from the use of the 1-norm and the inequalities. Following the complementary slackness conditions (3.41.fg), the support vectors are located outside or on the maximal margin boundary $\hat{f}(x) \pm \varepsilon$.

Example 3.3 [Estimating location parameters, II] As an application of this result, reconsider the setting of Example 1.2 of a sample $\{y_i\}_{i=1}^N$ sampled from an univariate random variable $\mathbf{Y}$. Let the pdf of $\mathbf{Y}$ be a Laplacian such that $p_{\mathbf{Y}}(y) = \mathscr{L}(\mu,\sigma) = \frac{1}{2\sigma}\exp(-|y-\mu|/\sigma)$. This distribution occurs e.g. as the distribution of the mutual



(a)

(b)

Figure 3.3: *The solid lines indicate the L_1* **(a)** *and the L_2* **(b)** *loss function used for the estimation of location. The dashed line in* **(a)** *represents the values of the term $(\alpha_i^+ - \alpha_i^-)$ in the L_1 estimator corresponding with the residual term e_i. The dashed line in* **(b)** *represents the Lagrange multiplier α_i of the dual of the L_2 estimator corresponding with the residual term e_i, see Example 1.2. Note the correspondence of the dashed line with the theoretical influence function of the mean and the median.*

differences between two independent variates with identical exponential distributions (Abramowitz and Stegun, 1972).

The maximum likelihood estimator of the location parameter μ then becomes

$$\begin{aligned} \hat{\mu} &= \arg\max_{\mu} \log \prod_{i=1}^{N} \frac{1}{2\sigma} \exp\left(\frac{-|y_i - \mu|}{\sigma} \right) \\ &= \arg\min_{\mu} \sum_{i=1}^{N} |y_i - \mu| \\ &= \arg\min_{\mu,e} \sum_{i=1}^{N} e_i \quad \text{s.t.} \quad -e_i \le y_i - \mu \le e_i, \end{aligned} \tag{3.42}$$

which can be cast as an LP problem, see also Chapter 2. The Lagrangian becomes $\mathscr{L}_1(\mu, e; \alpha^+, \alpha^-) = \sum_{i=1}^N e_i + \sum_{i=1}^N \alpha_i^+(\mu - y_i - e_i) + \sum_{i=1}^N \alpha_i^-(-\mu + y_i - e_i)$ with positive multipliers $\alpha^+ = (\alpha_1^+, \ldots, \alpha_N^+)^T \in \mathbb{R}^{+,N}$ and $\alpha^- = (\alpha_1^-, \ldots, \alpha_N^-)^T \in \mathbb{R}^{+,N}$. Necessary and sufficient conditions are given by the Karush-Kuhn-Tucker conditions:

$$\mathrm{KKT}(\mu, e; \alpha^+, \alpha^-) = \begin{cases} \dfrac{\partial \mathscr{L}_1}{\partial e_i} = 0 \rightarrow & 1 = \alpha_i^+ + \alpha_i^- & \forall i = 1, \ldots, N & (a) \\ \dfrac{\partial \mathscr{L}_1}{\partial \mu} = 0 \rightarrow & \sum_{i=1}^N \alpha_i^+ = \sum_{i=1}^N \alpha_i^- & \forall i = 1, \ldots, N & (b) \\ & -e_i \le y_i - \mu \le e_i & \forall i = 1, \ldots, N & (c) \\ & \alpha_i^+, \alpha_i^- \ge 0 & \forall i = 1, \ldots, N & (d) \\ & \alpha_i^+(\mu - y_i - e_i) = 0 & \forall i = 1, \ldots, N & (e) \\ & \alpha_i^-(\mu - y_i + e_i) = 0. & \forall i = 1, \ldots, N & (f) \end{cases} \tag{3.43}$$

From the complementary slackness constraints (3.43.ef), it follows that α_i^+ and α_i^- can only be non-zero simultaneously when $\mu = y_i$. Furthermore, the relation $\alpha_i^+(1 - \alpha_i^-) = 0$ holds elsewhere. In case all samples y_i were different, the equality $y_i = \mu$ can only be attained for a single y_i, say y_μ. In summary,

$$\begin{cases} \alpha_i^+ = I(\mu - y_i > 0), \;\; \alpha_i^- = I(\mu - y_i < 0) & \text{if} \;\; y_i \ne \mu \\ \alpha_i^+ = \alpha_i^- = 0.5 & \text{if} \;\; y_i = \mu \\ \sum_{i=1}^N I(\mu - y_i > 0) = \sum_{i=1}^N I(\mu - y_i < 0), & \end{cases} \tag{3.44}$$

where the indicator function $I(x > 0)$ equals one if $x > 0$ and zero else. If N were odd, condition (3.43.b) ensures that $(N-1)/2$ number of data-points are strictly lower than μ and $(N-1)/2$ are strictly larger such that $\hat{\mu} = y_{((N+1)/2)}$ If N were even, $N/2$ data-points are strictly lower than μ and $N/2$ are strictly larger, and $\hat{\mu} = \left(y_{((N-1)/2)} + y_{((N+1)/2)} \right)/2$. As such the median would correspond with maximum likelihood estimate whenever a Laplacian distribution may be postulated.

Figure 3.3.a illustrates the connection between the loss function $|e_i|$ and the value of the corresponding Lagrange multipliers $(\alpha_i^+ - \alpha^-)$. Following Subsection 3.3.3, one sees the connection between $(\alpha_i^+ - \alpha^-)$ and the sensitivity of the values e_i in the median estimator. Figure 3.3.b shows the case of the L_2 location estimator and the Lagrange multipliers α_i corresponding with e_i, again suggesting the sensitivity interpretation. For a complete account of robust location estimators and influence functions, see e.g. (Andrews *et al.*, 1972) and the survey in (De Brabanter, 2004).

3.5 L_∞ Kernel Machines: Support Vector Tubes

A slightly different setting is considered. Example 3.4 considers the most basic case (without covariates) in some detail.

Example 3.4 **[Tolerance bounds]** Let $\mathscr{D}_{\mathbf{Z}}$ denote a set $\{z_i\}_{i=1}^N \subset \mathbb{R}$ sampled i.i.d from a random variable $\mathbf{Z}$ with cdf $P_{\mathbf{Z}}$. Given an interval $[-t,t] \subset \mathbb{R}$, one can give a bound on the probability that future samples $z_* \in \mathbf{Z}$ sampled from the same distribution will lie inside the interval. Let $T : \mathbb{R}^+ \rightarrow \mathscr{S} \subset \mathbb{R}$ be elements of the following class

$$\mathscr{F}_T = \left\{ T : \mathbb{R}^+ \rightarrow \mathscr{S} \;\middle|\; T(t) = [-t,\ t], 0 \leq t \in \mathbb{R} \right\}. \tag{3.45}$$

Let the true risk and its empirical counterpart be defined respectively as

$$\begin{cases} \mathscr{R}_T^1(I_t, P_{\mathbf{Z}}) = \int I_t(|z| > t) dP_{\mathbf{Z}}(z) \\ \hat{\mathscr{R}}_T^1(t, \mathscr{D}_{\mathbf{Z}}) = \frac{1}{N} \sum_{i=1}^N I_t(|z_i| > t), \end{cases} \tag{3.46}$$

where $I_t(|z| > t)$ equals one if $z \notin [-t,\ \ t]$ and zero otherwise. If t is chosen with zero empirical risk ($\hat{\mathscr{R}}_T^1(t, \mathscr{D}_{\mathbf{Z}}) = 0$), and after constructing the cdf and the empirical cdf (ecdf) of the dataset $\mathscr{D}_{|\mathbf{Z}|} = \{|z_i|\}_{i=1}^N$. Then the application of classical results (Vapnik, 1998) gives the following results

- Due to the Glivenko-Cantelli theorem, the ecdf of $|z_i|$ will converge to the true cdf when $N \rightarrow \infty$ such that

$$\lim_{N\rightarrow\infty} \sup_{z \geq 0} \left| P_{\mathbf{Z}}(z) - \hat{P}_{\mathbf{Z}}(z) \right| \xrightarrow{P} 0. \tag{3.47}$$

- Application of the law of the iterated logarithm gives

$$\text{Prob}\left(\lim_{l\rightarrow\infty} \sup_{N>l} \mathscr{R}_T^1(I_t, P_{\mathbf{Z}}) < \sqrt{\frac{\ln\ln N}{2N}} \right) = 1. \tag{3.48}$$

- From the Kolmogorov-Smirnoff bound, the following inequality can be derived

$$\text{Prob}\left(\mathscr{R}_T^1(I_t, P_{\mathbf{Z}}) > \varepsilon \right) < 2\exp(-2\varepsilon^2 N), \tag{3.49}$$

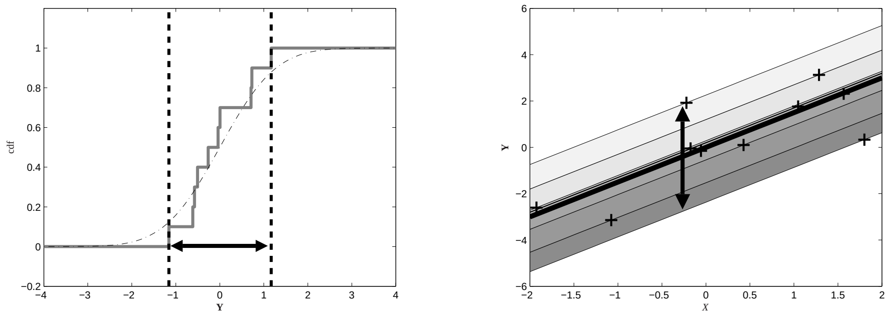
  which hold for finite sample sets with size N.

- A related result originates from the theory of random variables and order statistics known as the formulation of tolerance intervals:

$$\text{Prob}\left(\mathscr{R}_T^1(I_t, P_{\mathbf{Z}}) > \varepsilon \right) \leq N\varepsilon^{N-1} - (N-1)\varepsilon^N, \tag{3.50}$$

  where $0 < \alpha < 1$ is the confidence level, see e.g. (Rice, 1988), Chapter 3, Example E.

Given a set of samples $\mathscr{D}$ from a random vector $(\mathbf{X},\mathbf{Y})$ with joint distribution $P_{\mathbf{XY}}$. let $\mathbf{Z}$ be a random variable defined as $\mathbf{Z} = \mathbf{Y} - f(\mathbf{X})$ with $f : \mathbb{R}^D \rightarrow \mathbb{R}$ a fixed function. The transformed dataset then becomes $\mathscr{D}_{\mathbf{Z}} = \{(x_i, z_i)\}_{i=1}^N$ where $z_i = y_i - f(x_i)$. The



(a) (b)

Figure 3.4: **(a)** *Illustration of the intuition behind the interpretation of interval estimation in a univariate sample as explained in Example 3.4.* **(b)** *In the setting of regression, the conditional distribution $P(\mathbf{Y}|\mathbf{X}=x)$ may be estimated by the empirical cdf estimator based on the residuals $e_i = y_i - f(x_i)$ of the data observations (black crosses), resulting in an uncertainty region as indicated by the gray zones. The solid black line indicates the expected conditional density $\hat{f}(x) = E[\mathbf{Y}|\mathbf{X}=x]$. The black arrow indicates the height of the region with zero empirical risk.*

marginal probability of the random vector $(\mathbf{X},\mathbf{Z})$ over $\mathbf{X}$ becomes $\text{Prob}(\mathbf{Z} \leq z, \mathbf{X} \in \mathbb{R}^d) = \text{Prob}(\mathbf{Z} \leq z) = P_{\mathbf{Z}}(z)$. The results of Example 3.4 may be used to derive bounds on the marginal risk and the marginal empirical risk of the tube defined as follows

$$\begin{cases} \mathscr{R}_T(I_t, P_{\mathbf{XY}}) = \int I\left(y \notin T(x)\right) dP_{\mathbf{YX}}(yx) = \int I_t\left(|z| > t\right) dP_{\mathbf{Z}}(z) \\ \hat{\mathscr{R}}_T(w,t,\mathscr{D}) = \frac{1}{N}\sum_{i=1}^N I(y_i \notin T(x_i)), \end{cases} \tag{3.51}$$

where $I_t(y \notin T(x))$ equals one if $y \notin [w^T \varphi(x) - t, \;\; w^T \varphi(x) + t]$ and zero otherwise. Subsection 3.6.2 gives a more detailed derivation which incorporates the complexity of the tube. Consider the task of approximating the unknown support of $P_{\mathbf{XY}}$. As in practice one typically distinguishes between the unknown response variable Y and the inputs $\mathbf{X}$ which happen to be given, a support may be expressed as a function of the given dependent variable $\mathbf{X} = x$. To simplify matters further, the following family of support functions is considered

$$\mathscr{F}_{\varphi,T} = \{T(w,t) = w^T \varphi(x) \pm t \mid w \in \mathbb{R}^{D_\varphi}, t \in \mathbb{R}^+\}. \tag{3.52}$$

In a practical setting, those result may be used as follows. Let $T(w,t)$ be an element of $\mathscr{F}_{\varphi,T}$ with empirical risk zero. *Let $\{(x_j,y_j)\}_{j=1}^N \subset \mathbb{R}^D \times \mathbb{R}$ be drawn i.i.d. according to the same distribution $P_{\mathbf{XY}}$ underlying $\mathscr{D}$. In this case the output samples y_j will on average lie inside the interval $T(x_j)$ with high probability.* This result shifts the focus of the point estimator $\hat{f}$ to the interval estimator $[\hat{f}-\hat{t},\hat{f}+\hat{t}]$ denoted as the support tube. As such, the proposed support vector tube is closely related to results in novelty detection algorithms (Tax and Duin, 1999). Figure 3.5 illustrates the principle behind the Support Vector Tube on a one-dimensional example. The primal-dual derivation is summarized in the following Lemma.

Lemma 3.7. [Support Vector Tubes] *Consider the class of support tubes $\mathscr{F}_{\varphi,T}$ defined in (3.52). Let $\mu > 0$ be a hyper-parameter. The smallest tube of minimal complexity is found as the solution to the following optimization problem*

$$(\hat{w},\hat{t}) = \arg\min_{w,t} \mathscr{J}_\mu(w,t) = \frac{1}{2}w^T w + \mu t$$
$$\text{s.t.} \quad -t \le w^T\varphi(x_i) - y_i \le t, \ \ \forall i = 1,\dots,N. \qquad (3.53)$$

The dual problem becomes

$$(\hat{\alpha^+},\hat{\alpha^+}) = \arg\max_{\alpha^+,\alpha^-} -\frac{1}{2}(\alpha^- - \alpha^+)^T\Omega(\alpha^- - \alpha^+) + (\alpha^- - \alpha^+)^T Y$$
$$\text{s.t.} \quad (\alpha^- + \alpha^+)^T 1_N = \mu, \ \alpha^+, \ \alpha^- \ge 0_N. \qquad (3.54)$$

The resulting tube can be evaluated in a new point $x_* \in \mathbb{R}^D$ as follows

$$\hat{T}(x_*) = \Omega_{\mathscr{D}}(x_*)^T(\hat{\alpha}^- - \hat{\alpha}^+) \pm \hat{t}, \qquad (3.55)$$

where $\hat{\alpha}^+$ and $\hat{\alpha}^-$ solve (3.54) and $\hat{t}$ can be recovered from the KKT conditions (3.57.fg).

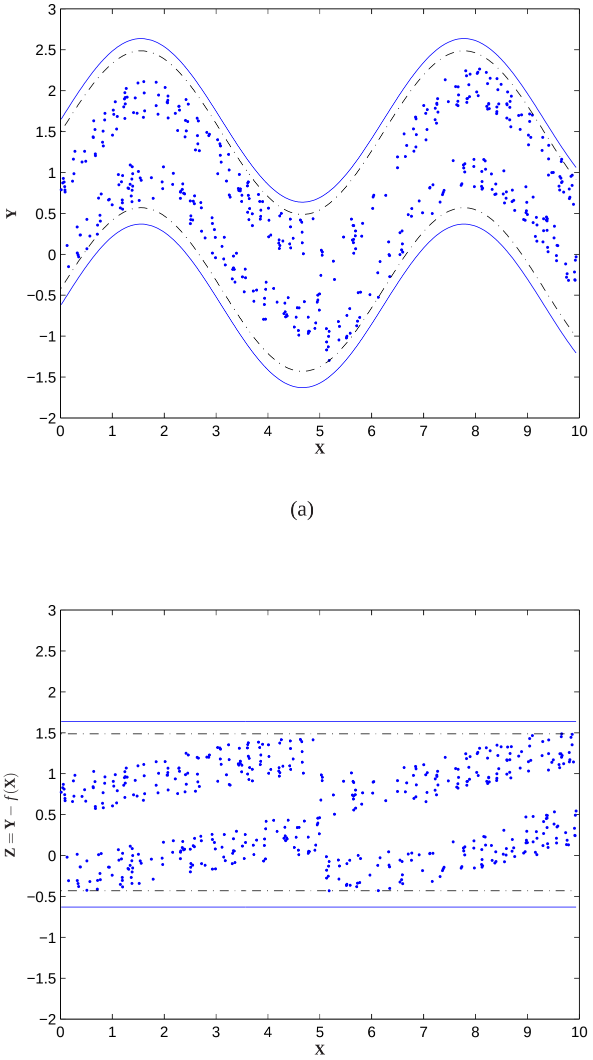
Proof. The Lagrangian becomes

$$\mathscr{L}_\mu(w,t;\alpha^+,\alpha^-) = \frac{1}{2}w^T w + \mu t + \sum_{i=1}^N \alpha_i^+\left[(w^T\varphi(x_i) - y_i) - t\right]$$
$$+ \sum_{i=1}^N \alpha_i^-\left[-(w^T\varphi(x_i) - y_i) - t\right], \qquad (3.56)$$

with positive multipliers $\alpha^+ = (\alpha_1^+,\dots,\alpha_N^+)^T \in \mathbb{R}^{+N}$, $\alpha^- = (\alpha_1^-,\dots,\alpha_N^-)^T \in \mathbb{R}^{+N}$. The necessary and sufficient conditions for optimality are given as

$$\text{KKT}\begin{cases} \frac{\partial \mathscr{L}_\mu}{\partial w} = 0 \to & w = \sum_{i=1}^N(\alpha_i^- - \alpha_i^+)\varphi(x_i) & & (a) \\ \frac{\partial \mathscr{L}_\mu}{\partial t} = 0 \to & \mu = \sum_{i=1}^N\left(\alpha_i^- + \alpha_i^+\right) & \forall i = 1,\dots,N & (b) \\ & \alpha_i^+,\alpha_i^- \ge 0 & & (c) \\ & -t \le w^T\varphi(x)_i - y_i \le t & \forall i = 1,\dots,N & (d) \\ & \alpha_i^+\left[(w^T\varphi(x_i) - y_i) - t\right] = 0 & \forall i = 1,\dots,N & (f) \\ & \alpha_i^-\left[-(w^T\varphi(x_i) - y_i) - t\right] = 0. & \forall i = 1,\dots,N & (g) \end{cases}$$
$$(3.57)$$

The saddle-point interpretation leads to the dual problem (3.54). The parameter t can be recovered from the complementary slackness conditions (3.57.fg) by the equality $w^T\varphi(x_i) - y_i = t$ which hold when $\alpha_i^+ > 0$. □



(a)

(b)

Figure 3.5: *Illustration of the Support Vector Tube.* **(a)** *Let $\mathscr{D}$ be a sample of a joint distribution $P_{\mathbf{XY}}$ with bounded support.* **(b)** *Consider the transformed data $\mathbf{Z} = \mathbf{Y} - f(\mathbf{X})$. The solid line shows an absolute upper- $U_{\mathbf{Z}}$ respectively lower-bound $u_{\mathbf{Z}}$ of the support of $\mathbf{Z}$ such that $P(u_{\mathbf{Z}} < \mathbf{Z} < U_{\mathbf{Z}}) = 1$. The dashed line shows the empirical counterpart.*

3.6 Robust Inference of Primal-Dual Kernel Machines

3.6.1 Huber's loss function

Definition 3.3. [Contaminated Noise Model, (Huber, 1964)] *The general gross-error model or ρ-contamination noise model is defined as the union of the nominal noise model F_0 and an arbitrary continuous distribution G. Let $0 \leq \rho \ll 1$ be the first parameter of contamination:*

$$\mathscr{F}(F_0,\rho) = \{F|F(x) = (1-\rho)F_0(x)+\rho G(x)\}. \tag{3.58}$$

This contamination scheme describes the case where the data occurs with large probability $(1-\rho)$ according to the (ideal) nominal model. Outliers occur with probability ρ according to the distribution G.

A robust way to handle this family of noise models in parametric models is the use of the so-called Huber loss function which is a combination of an L_2 norm for obtaining efficiency and L_1 for the sake of robustness. The loss function is defined as follows

$$\ell_H(e) = \begin{cases} \frac{e^2}{2} & |e| \leq c \\ c|e| - \frac{c^2}{2} & |e| > c, \end{cases} \tag{3.59}$$

where c is a constant depending on the noise level σ_e. A good initial estimate for c was proposed as $\hat{c} = 1.483\ \ \mathrm{MAD}(\mathscr{D})$ where $\mathrm{MAD}(\mathscr{D})$ is the Median Absolute Deviation of the estimated residuals $\mathrm{MAD}(\mathscr{D}) = \mathrm{median}\left(\{e_i = y_i - \hat{f}(x_i)\}_{i=1}^N\right)$. Robust statistics for non-parametric techniques were studied in (Hettmansperger and McKean, 1994). Analogously, one can consider this family of noise models for non-parametric primal-dual kernel machines as proposed in (Vapnik, 1998). The primal-dual derivations are summarized in the following lemma.

Lemma 3.8. [Primal-Dual Kernel Machine with Huber-loss (Vapnik, 1998)] *Consider the class of models $\mathscr{F}_\varphi$. Let $c, \nu \in \mathbb{R}_0$ be positive constants and $r = (r_1,\ldots,r_N)^T \in \mathbb{R}^N$ be slack-variables modeling the outliers. Then the kernel machine based on the Huber loss function is equivalent to the following optimization problem*

$$(\hat{w},\hat{e},\hat{r}_i) = \underset{w,e,r}{\arg\min}\, \mathscr{J}_{c,\gamma}(w,e,r) = \frac{1}{2}w^Tw + \gamma\left(c\sum_{i=1}^N r_i + \frac{1}{2c}\sum_{i=1}^N e_i^2\right)$$
$$s.t. \quad -r_i \leq w^T\varphi(x) + e_i - y_i \leq r_i. \tag{3.60}$$

The dual problem becomes

$$(\hat{\alpha}^+,\hat{\alpha}^-) = \underset{\alpha^+,\alpha^-}{\arg\max} -\frac{1}{2}(\alpha^- - \alpha^+)^T\left(\Omega + \frac{c}{\gamma}I_N\right)(\alpha^- - \alpha^+) + Y^T(\alpha^- - \alpha^+)$$
$$(\alpha_i^+ + \alpha_i^-) = \gamma c,\ \alpha_i^+, \alpha_i^- \geq 0,\ \ \forall i = 1,\ldots,N. \tag{3.61}$$

and the estimate of a new data point can be written as $\hat{f}(x_*) = \Omega(x_*)(\hat{\alpha}^+ - \hat{\alpha}^-)$ where $\hat{\alpha}^+, \hat{\alpha}^-$ solves (3.61).

Proof. The Lagrangian of the cost-function becomes

$$\begin{aligned}\mathscr{L}_{c,\gamma}(w,e,r;\alpha^+,\alpha^-) &= \frac{1}{2}w^Tw+\gamma\left(c\sum_{i=1}^N r_i+\frac{1}{2c}\sum_{i=1}^N e_i^2\right)\\ &+\sum_{i=1}^N \alpha_i^+\left[(w^T\varphi(x_i)+e_i-y_i)-r_i\right]+\sum_{i=1}^N \alpha_i^-\left[-(w^T\varphi(x_i)+e_i-y_i)-r_i\right],\end{aligned} \tag{3.62}$$

with positive Lagrange multipliers $\alpha^+,\alpha^-\in\mathbb{R}^{+,N}$. The Karush-Kuhn-Tucker conditions for optimality become

$$\text{KKT}\left\{\begin{array}{llll} \frac{\partial\mathscr{L}_{c,\gamma}}{\partial w}=0\to & w=\sum_{i=1}^N(\alpha_i^- - \alpha_i^+)\varphi(x_i) & & (a)\\ \frac{\partial\mathscr{L}_{c,\gamma}}{\partial e_i}=0\to & \gamma e_i=c(\alpha_i^- - \alpha_i^+) & \forall i=1,\ldots,N & (b)\\ \frac{\partial\mathscr{L}_{c,\gamma}}{\partial r_i}=0\to & \gamma c=\alpha_i^+ + \alpha_i^- & \forall i=1,\ldots,N & (c)\\ & \alpha_i^+,\alpha_i^-\geq 0 & & (d)\\ & -r_i\leq w^T\varphi(x_i)+e_i-y_i\leq r_i & \forall i=1,\ldots,N & (e)\\ & \alpha_i^+\left[(w^T\varphi(x_i)+e_i-y_i)-r_i\right]=0 & \forall i=1,\ldots,N & (f)\\ & \alpha_i^-\left[-(w^T\varphi(x_i)+e_i-y_i)-r_i\right]=0. & \forall i=1,\ldots,N & (g)\end{array}\right. \tag{3.63}$$

Substituting the conditions (3.63.abc) and maximizing over the Lagrange multipliers α^+,α^- results in the dual problem (3.61). □

The following algorithm can be used in practice to speedup the computations.

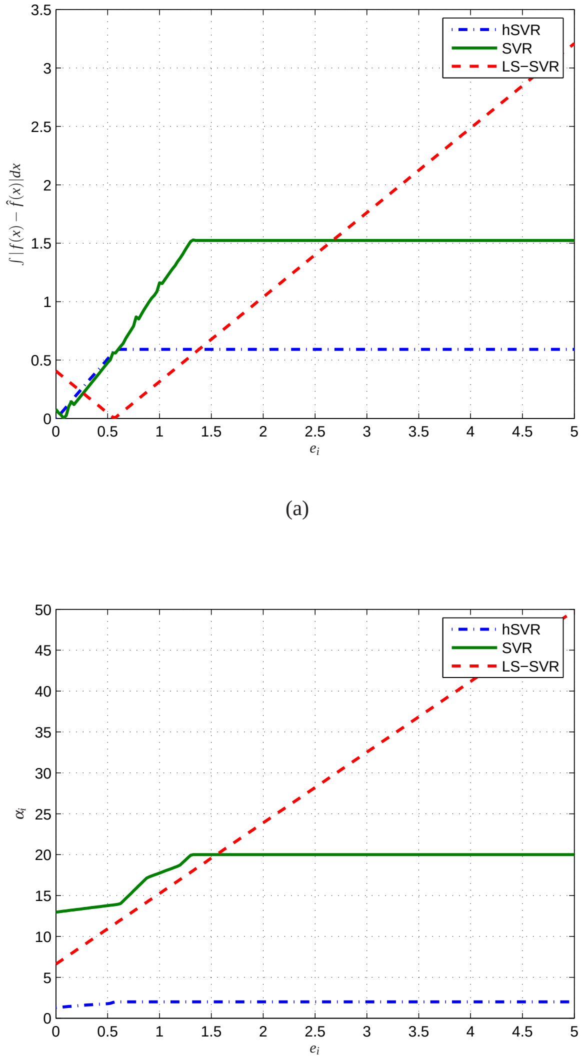
Algorithm 3.1. [Iteratively Re-weighted Robust LS-SVM] *An iteratively re-weighted algorithm based on the weighted LS-SVM regressor is proposed to solve the optimization problem (3.60) efficiently. The algorithm was first proposed as a standalone formulation of a robust LS-SVM for regression (Suykens* et al.*, 2002*a*). It is based on following reformulation of the regularized least squares cost-function (3.9) using the adaptive weighting terms $\Gamma=(\Gamma_1,\ldots,\Gamma_N)^T\in\mathbb{R}^N$:*

$$(\hat{w},\hat{e})=\arg\min_{w,e}\mathscr{J}'_{c,\Gamma}(w,e)=\frac{1}{2}w^Tw+\frac{1}{2}\sum_{i=1}^N\Gamma_i e_i^2$$

$$\text{s.t.}\quad\begin{cases} w^T\varphi(x_i)+e_i=y_i & (a)\\ \Gamma_i e_i^2=\ell_H(e_i)=\frac{e^2}{2} \quad |e|\leq c & (b)\\ \Gamma_i e_i^2=\ell_H(e_i)=c|e|-\frac{c^2}{2} \quad |e|>c. & (c)\end{cases} \tag{3.64}$$

By alternating over the constraints (3.64.a) and (3.64.bc), one obtains an iterative algorithm for solving the problem as follows:

- *If the weighting Γ were known, one can obtain the solution to (3.64) by solving a linear system following the primal-dual derivation of the LS-SVR as described in*



(a)

(b)

Figure 3.6: *Empirical assessment of the influence of outliers on the Huber based SVM regressor, the standard SVM regressor and the LS-SVM regressor.* **(a)** *Effect of the global performance of the estimators when ranging the error e_i on the ith output from* 0 *to* 5. **(b)** *Influence on the ith Lagrange multiplier α_i of the estimators when ranging the error e_i on the ith output from* 0 *to* 5.

Subsection 3.3 (see (Suykens et al., *2002*a)*). Let $D_\Gamma = \operatorname{diag}(\Gamma_1, \ldots, \Gamma_N) \in \mathbb{R}^{N \times N}$ be a diagonal matrix, then the weighted LS-SVR results from*

$$(\Omega + D_\Gamma)\,\alpha = Y. \tag{3.65}$$

From the necessary and sufficient conditions for optimality it follows that $\Gamma_i \hat{e}_i = \hat{\alpha}_i$ where $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)^T \in \mathbb{R}^N$ solve (3.65). The estimated function $\hat{f}$ can be evaluated in any point $x_* \in \mathbb{R}^D$ as $\hat{f}(x_*) = \Omega_{\mathscr{D}}(x_*)\hat{\alpha}$ (Suykens et al., *2002*a*).*

- *Given the estimates $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_N)^T \in \mathbb{R}^N$, the weightings Γ can be recomputed by solving the equations*

$$\Gamma_i e_i^2 = \ell_H(e_i) = \begin{cases} \frac{e_i^2}{2} & |e_i| \le c \\ c|e_i| - \frac{c^2}{2} & |e_i| > c. \end{cases} \quad \forall i = 1, \ldots, N, \tag{3.66}$$

  for Γ_i. From this equality, it follows that $|\hat{\alpha}_i| \le \gamma c$ for all $i = 1, \ldots, N$.

- *The algorithm then goes as follows:*

  1. *Initiate $\Gamma^{(t)} = \gamma 1_N$ for $t = 0$*
  2. *Compute $\alpha^{(t)}$ from (3.65) and $\Gamma^{(t)}$*
  3. *Recompute the parameters Γ^* by using equation (3.66)*
  4. *Let $0 \le \rho \ll 1$ be a factor to decrease the speed of convergence and to avoid instabilities. Then $\Gamma^{(t+1)} = \rho\Gamma^{(t)} + (1-\rho)\Gamma^*$*
  5. *Let $t = t+1$ and iterate steps 2-5 until the algorithm converges.*

A further convergence analysis of this algorithm is extended to future work.

It turns out that only a very few iterations are needed in practice (Suykens *et al.*, 2002*a*) and the solution follows much faster than from the QP formulation implemented by a general purpose solver.

Example 3.5 **[Comparison of Robust inference Machines]** A simple example is given to illustrate the effective robustness of the different approaches. A dataset is generated as follows $y_i = \operatorname{sinc}(x_i) + e_i$ where x_i is taken from the interval $[-3,\ 3]$, $N = 100$ and e_i is taken from a contaminated Gaussian distribution. Consider the standard LS-SVM regressor (Section 3.3), SVM for regression (Section 3.4), Huber based SVM regressor (subsection 3.6.1) and respectively. In the first example, the ith error term e_i is grown from zero to 10 and the corresponding prediction error $\int |\operatorname{sinc}(x) - \hat{f}(x)| dx$ is computed for the four estimators. Figure 3.6.a reports the evolution of the global performance of the different estimators while the error e_i becomes more outlying. Figure 3.6.b gives the corresponding evolution of the ith Lagrange multiplier.

Let e_i be distributed as follows $e_i \sim (1-\rho)\,\mathscr{N}(0, 0.1) + \rho\,\mathscr{U}([-10, 10])$ with $0 \le \rho \ll 1$ the factor of contamination. Figure 3.7.a gives the empirical influence function when the factor of contamination ρ grows. Figure 3.7.b reports the performance in the case $\rho = 0$, showing that the robustness of the hSVR and the SVR comes at the price of efficiency and performance in the uncontaminated case with respect to the LS-SVR. While the qualitative behavior is typical for the estimators, the quantitative properties (slope, breakdown point etc.) depend on the chosen hyper-parameters. The hyper-parameters where tuned using 10-fold cross-validation on the uncontaminated case and were fixed throughout the experiment for clarity of illustration.

3.6.2 ν- Support Vector Tubes

A relaxation of the finite support assumption is considered based on the contaminated noise model (3.58). The primal-dual derivations are summarized in the following lemma.

Lemma 3.9. [ν-Support Vector Tubes] *Consider the tube $T(x) = w^T\varphi(x) \pm t$ where w and t are to be estimated. Let $\nu, \mu \in \mathbb{R}_0^+$ be constants.*

$$(\hat{w}, \hat{t}, \hat{r}_i) = \arg\min_{w,t,r} \mathscr{J}_{\nu,\mu}(w,t,r) = \frac{1}{2}w^Tw + \nu\left(\sum_{i=1}^N r_i + \mu t\right)$$
$$s.t. \quad \begin{cases} -t - r_i \leq w^T\varphi(x) - y_i \leq t + r_i & \forall i = 1,\ldots,N, \\ r_i \geq 0 & \forall i = 1,\ldots,N. \end{cases} \tag{3.67}$$

The dual problem becomes

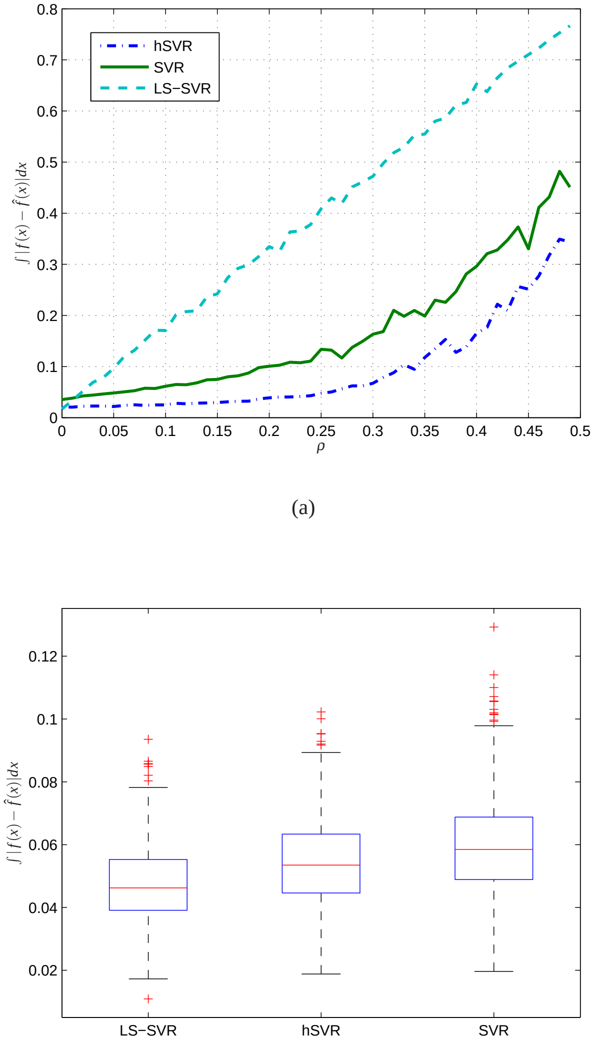
$$(\hat{\alpha}^+, \hat{\alpha}^-) = \arg\max_{\alpha^+,\alpha^-} -\frac{1}{2}(\alpha^+ - \alpha^-)^T\Omega(\alpha^+ - \alpha^-) + (\alpha^+ - \alpha^-)^TY$$
$$s.t. \quad \begin{cases} 0_N \leq \alpha^+, \alpha^- \\ (\alpha_i^+ + \alpha_i^-) \leq \nu & \forall i = 1,\ldots,N \\ (\alpha_i^+ + \alpha_i^-)^T 1_N = \nu\mu, \end{cases} \tag{3.68}$$

and the estimate of a new datapoint can be written as $\hat{f}(x_*) = \Omega(x_*)(\hat{\alpha}^+ - \hat{\alpha}^-)$ where $\hat{\alpha}^-$ and $\hat{\alpha}^+$ solve (3.68).

Proof. The Lagrangian of the cost-function becomes

$$\mathscr{L}_{\nu,\mu}(w,r,t;\alpha^+,\alpha^-) = \frac{1}{2}w^Tw + \nu\left(\sum_{i=1}^N r_i + \mu t\right) - \sum_{i=1}^N \beta_i r_i$$
$$-\sum_{i=1}^N \alpha_i^+\left[(w^T\varphi(x_i) - y_i) - t - r_i\right] - \sum_{i=1}^N \alpha_i^-\left[-(w^T\varphi(x_i) - y_i) - t - r_i\right], \tag{3.69}$$

with Lagrange multipliers $\alpha^+, \alpha^-, \beta \in \mathbb{R}^N$. The Karush-Kuhn-Tucker conditions for



(a)

(b)

Figure 3.7: *Empirical assessment of the performance of the Huber based SVR, the standard SVR and the LS-SVR.* **(a)** *Empirical influence function of the global performance of the estimators when increasing the factor of contamination ρ from* 0 *to* 50%. **(b)** *Global Performance of the estimators in the case of uncontaminated data e_i is approximatively normally distributed. The LS-SVR obtains the best performance with the lowest variance.*

optimality become

$$\text{KKT}\begin{cases}
\dfrac{\partial \mathscr{L}_{\nu,\mu}}{\partial w} = 0 \rightarrow & w = \sum_{i=1}^N (\alpha_i^+ - \alpha_i^-)\varphi(x_i) & & (a) \\
\dfrac{\partial \mathscr{L}_{\nu,\mu}}{\partial t} = 0 \rightarrow & \nu\mu = \sum_{i=1}^N (\alpha_i^+ + \alpha_i^-) & \forall i = 1,\ldots,N & (b) \\
\dfrac{\partial \mathscr{L}_{\nu,\mu}}{\partial r_i} = 0 \rightarrow & \nu = \beta_i + \alpha_i^+ + \alpha_i^- & \forall i = 1,\ldots,N & (c) \\
& \alpha_i^+, \alpha_i^-, \beta_i \geq 0 & & (d) \\
& -r_i - t \leq w^T\varphi(x_i) - y_i \leq t + r_i & \forall i = 1,\ldots,N & (e) \\
& r_i \geq 0 & \forall i = 1,\ldots,N & (f) \\
& \alpha_i^+ \left[(w^T\varphi(x_i) - y_i) - t - r_i\right] = 0 & \forall i = 1,\ldots,N & (g) \\
& \alpha_i^- \left[-(w^T\varphi(x_i) - y_i) - t - r_i\right] = 0 & \forall i = 1,\ldots,N & (h) \\
& \beta_i r_i = 0. & \forall i = 1,\ldots,N & (i)
\end{cases} \tag{3.70}$$

Substituting the conditions (3.70.abc) and maximizing over the Lagrange multipliers α^+, α^- results in the dual problem (3.68). □

The naming convention ν-SVT follows from the fact that the primal problem and the dual derivation goes along the same lines as the ν-SVM (Schölkopf and Smola, 2002), although the setting is different. This observation triggers the following result, which follows from the Karush-Kuhn-Tucker conditions.

Lemma 3.10. [Sparseness in ν-SVTs] *The hyper-parameter μ is a lower-bound to the number of nonzero Lagrange multipliers and serves as an upper-bound to the number of outliers o_i outside the tube.*

Proof. This follows from the observation that for all $i = 1,\ldots,N$, the values of α_i^+ and α_i^- cannot be nonzero simultaneously when $t > 0$. Furthermore, conditions (3.70.cd) guarantee that α_i^+ and α_i^- lie in the interval $[0,\nu]$ (also referred to as box constraints). The second statement follows from the complementary slackness condition (3.70.i). □

An analysis of the finite sample behavior of the robust SVT follows along the same lines as that of the Support Vector Machine for regression (Shawe-Taylor and Cristianini, 2004).

Theorem 3.3. [Risk of ν-SVTs, (Shawe-Taylor and Cristianini, 2004)] *Let $B \in \mathbb{R}_+$ and $0 < \varepsilon \ll 1$ be fixed. Consider the class $\mathscr{F}_{\varphi,T}$ with bounded norm $\|w\|_2^2 \leq B$. Let $\mathscr{D} = \{(x_i,y_i)\}_{i=1}^N$ drawn i.i.d. from a fixed but unknown distribution $P_{\mathbf{XY}}$. Let the risk $\mathscr{R}_T(w,t,P_{\mathbf{XY}})$ and its empirical counterpart $\hat{\mathscr{R}}_T(w,t,\mathscr{D})$ be defined as in (3.51). Then the following inequality holds for every element of the class $\mathscr{F}_{\varphi,T}$ with bounded norm $\|w\|_2 \leq B$ simultaneously:*

$$P\left(\left|\mathscr{R}_T(w,\tau,P_{\mathbf{XY}}) - \frac{\hat{\mathscr{R}}_T(w,t,\mathscr{D})}{\varepsilon-\tau}\right| \leq \frac{4B\sqrt{\mathrm{tr}(\Omega)}}{N(\varepsilon-\tau)} + 3\sqrt{\frac{\ln(2/\varepsilon)}{2N}}\right) \geq (1-\varepsilon), \tag{3.71}$$

where $\tau \in \mathbb{R}^+$ is such that $t < \tau$.

This result corresponds entirely with Theorem 7.49 in (Shawe-Taylor and Cristianini, 2004). This provides an upper-bound to the theoretical risk that a new point drawn according to $P_{\mathbf{XY}}$ will lie outside the Support Vector Tube with empirical risk as obtained (3.60).

3.7 Primal-Dual Kernel Machines for Classification

While the previous elaboration mainly focused on the case of regression, the past literature on kernel machines mainly considered the case of classification for a number of reasons which are properly summarized in the following quotation

> "(...) However, it was extremely lucky that at the first and the most important stage of developing the theory - when the main concepts of the entire theory had to be defined - simple sets of functions were considered. Generalizing these results obtained for estimation indicator functions (pattern recognition) to the problem of estimating real-valued functions (regressions, densities, etc.) was a purely technical achievement." (Vapnik, 1998).

Though a multitude of formulations and derivations exist, only two cases are elaborated in some detail.

3.7.1 Standard Support Vector Machines

Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$ be samples from the random vector $(\mathbf{X}, \mathbf{Y})$ such that $x_i \in \mathbb{R}^D$ and $y_i \in \{-1, 1\}$. Let us consider the hyperplane described as

$$\mathrm{Hp}(w) = \left\{ x \in \mathbb{R}^{d_\varphi} \;\middle|\; f(x) = w^T \varphi(x) = 0 \right\}, \tag{3.72}$$

where again $\varphi : \mathbb{R}^D \to \mathbb{R}^{D_\varphi}$ is a fixed but unknown mapping. Let $\mathscr{F}_{\mathrm{Hp}} = \{\mathrm{Hp}(w),\ w \in \mathbb{R}^{D_\varphi}\}$ be the class of hyperplanes which is considered in this case. The placement of a new point x_* with respect to the hyperplane $\mathrm{Hp}(w)$ can be determined as follows

$$\hat{y}_* = \mathrm{sign}\left[f(x_*)\right] = \mathrm{sign}\left[w^T \varphi(x_*)\right]. \tag{3.73}$$

The distance of any point $\varphi(x_*)$ to the hyperplane $\mathrm{Hp}(w) \in \mathscr{F}_{\mathrm{Hp}}$ is given as

$$d\left(\varphi(x_*), \mathrm{Hp}(w)\right) = \frac{|f(x_*)|}{\|f'(x_*)\|_2} \geq \frac{y_i(w^T \varphi(x_*))}{w^T w}. \tag{3.74}$$

Now consider the problem of finding the hyperplane with maximal margin:

$$(\hat{w}, \hat{m}) = \arg\max_{w,m} m \quad \text{s.t.} \quad d\left(\varphi(x_i), \mathrm{Hp}(w)\right) \geq m. \tag{3.75}$$

Without loss of generality one can change variables such that $w^T w = 1/m$. As such, one rewrite equation (3.75)

$$\hat{w} = \arg\min_{w} \frac{1}{2} w^T w \quad \text{s.t.} \quad y_i \left(w^T \varphi(x_i) \right) \geq 1 \;\; \forall i = 1, \ldots, N. \tag{3.76}$$

Moreover, it follows that the resulting margin equals $m = 1/w^T w$. A proper relaxation was formulated to the case where the data of the different classes is not strictly separable by an hyperplane from the class $\mathscr{F}_{\text{Hp}}$. After introducing the slack variables $\xi = (\xi_1, \xi_2, \ldots, \xi_N)^T$, one can write

$$(\hat{w}, \hat{\xi}) = \arg\min_{w,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \qquad \text{s.t.} \quad y_i \left(w^T \varphi(x_i) \right) \geq 1 - \xi_i \;\; \forall i = 1, \ldots, N. \tag{3.77}$$

The first notions of this strategy appeared in (Vapnik, 1982). This formulation of SVMs appeared first in literature in (Boser *et al.*, 1992) and was elaborated in (Vapnik, 1995).

Statistical learning theory provides lower-bounds on the generalization performance of such a maximal margin classifier. A central result is summarized in the following result due to (Vapnik, 1998).

Theorem 3.4. [Bounding the risk] *Let $0 < \varepsilon \ll 1$ be fixed. Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^D \times \{-1, 1\}^N$ be sampled i.i.d. from the fixed but unknown distribution $P_{\mathbf{XY}}$. Let the theoretical risk of a classifier be defined as*

$$\mathscr{R}(f, P_{\mathbf{XY}}) = \int I(y f(x) < 0) dP_{\mathbf{XY}}, \tag{3.78}$$

where $I(f(x)y < 0)$ is one if $f(x)y < 0$ and zero otherwise. Its empirical counterpart is defined as $\hat{\mathscr{R}}(w, \mathscr{D}) = \frac{1}{N} \sum_{i=1}^N \|y_i w^T \varphi(x_i)\|$. The following bound holds simultaneously for all hyperplanes with given VC-dimension c

$$P\left(\mathscr{R}(f, P_{\mathbf{XY}}) \leq \hat{\mathscr{R}}(w, \mathscr{D}) + \sqrt{\frac{c \ln(2N/c + 1) - \ln(\varepsilon/4)}{N}} \right) \geq (1 - \varepsilon). \tag{3.79}$$

Extensions to so-called ramp-functions (squared classification loss) were studied e.g. in (Cristianini and Shawe-Taylor, 2000). Alternative bounds were constructed using complexity measures as the (empirical) Rademacher complexity (Shawe-Taylor and Cristianini, 2004). A modified version of the primal-dual derivation (as can be found e.g. in (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000)) is given in Subsection 6.4.3.

3.7.2 LS-SVMs for classification

Consider the parametric assumption that both classes $C_{+1} = \{(x_i, y_i)\}_{y_i=+1}$ and $C_{-1} = \{(x_i, y_i)\}_{y_i=-1}$ are drawn from two different multivariate Gaussian distributions with equal variances, say $C_{+1} \sim \mathcal{N}(w_{+1}, I_d\sigma^2)$ and $C_{-1} \sim \mathcal{N}(w_{-1}, I_d\sigma^2)$. Some algebra shows (see e.g. (Friedman, 1989; Hastie *et al.*, 2001)) that $\mathrm{Hp}(w) = \frac{1}{2}(w_{+1} + w_{-1})$ describes the unique line such that $x \in \mathrm{Hp}(w) \Leftrightarrow P(y = +1|X = x) = P(Y = -1|X = x)$. Given a finite sample, the penalized maximum likelihood estimate results from the following optimization problem

$$(\hat{w}, \hat{e}) = \arg\min_{w,e} \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad y_i \left(w^T x_i \right) = 1 - e_i. \tag{3.80}$$

Employing the primal-dual optimization framework, it is readily seen (Suykens and Vandewalle, 1999; Suykens *et al.*, 2002*b*; Van Gestel *et al.*, 2002) that the solution is characterized by the following linear system

$$\left(\Omega^y + \frac{1}{\gamma} I_N \right) \alpha = 1_N, \tag{3.81}$$

where $\Omega^y \in \mathbb{R}^{N \times N}$ is the modified kernel matrix defined as $\Omega^y_{ij} = K(x_i, x_j) y_i y_j$ for all $i, j = 1, \ldots, N$ denotes the pointwise matrix product. The decision of a new point x_* is then made as follows

$$\hat{y} = \mathrm{sign} \left[\sum_{i=1}^{N} \hat{\alpha}_i y_i K(x_i, x_*) \right], \tag{3.82}$$

where $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)^T \in \mathbb{R}^N$ solve (3.81). This primal-dual derivation including bias term was coined as Least Squares SVM classifier (Suykens and Vandewalle, 1999). The dual solution is strongly related as kernel Fisher discriminant analysis (Baudat and Anouar, 2000), proximal SVM (Fung and Mangasarian, 2001) and Regularized Least Squares Classification (Rifkin, 2002).

Other kernel based approaches towards the task of classification include amongst others Parzen based classifiers as the naive Bayes classifier, see e.g. (Hastie *et al.*, 2001) and kernel logistic regression (Jaakkola and Haussler, 1999). Robust minimax extensions were studied in (Lanckriet *et al.*, 2002; Trafalis and Alwazzi, 2003).

Chapter 4

Structured Primal-Dual Kernel Machines

It is a common intuition that the incorporation of prior knowledge into the problem's formulation will lead to improvements of the final estimate with respect to naive applications of an off-the-shelf method. The following chapter shows the flexibility of the primal-dual optimization framework for decoding this knowledge into the estimation problem. Various types of structural information are considered, including semi-parametric model structures (Section 4.1), additive models (Section 4.1), imposing pointwise structure (Section 4.3) in the form of inequalities and its extension towards handling censored observations (Section 4.4).

4.1 Semi-Parametric Regression and Classification

4.1.1 Semi-parametric LS-SVMs for regression

Suppose the underlying function generating the data can be arbitrarily well approximated by a model contained in the following class

$$\mathscr{F}_{\varphi,P} = \Big\{ f : \mathbb{R}^D \times \mathbb{R}^{D_p} \to \mathbb{R} \ \Big| \ f(x,x^p) = w^T\varphi(x) + \beta^T x^p, \quad w \in \mathbb{R}^{d_\varphi}, \beta \in \mathbb{R}^{D_p} \Big\}, \tag{4.1}$$

where x represents the non-parametric dependent variable $x \in \mathbb{R}^D$ and $x^p \in \mathbb{R}^{D_p}$ denote the parametric dependent variable of dimension D_p. This setting reduces to the commonly considered case of the intercept (bias) term whenever one chooses $D_p = 1$ and $x^p = (1,\dots,1)^T \in \mathbb{R}^N$. Let $X_p \in \mathbb{R}^{N\times D_p}$ denote the matrix with D_p columns where each ith column contains the N samples of the ith parametric component for all

$i = 1,\dots,D_p$. Applications can be found in e.g. (Engle *et al.*, 1986) for the modeling of electricity demand.

As an example, consider again the regularized least squares cost function

$$(\hat{w},\hat{\beta},\hat{e}) = \arg\min_{w,\beta,e} \mathscr{J}(w,\beta,e) = \frac{1}{2}w^T w + \frac{\gamma}{2}e^T e$$
$$\text{s.t.} \quad w^T \varphi(x_i) + x_i^p \beta + e_i = y_i, \forall i = 1,\dots,N, \tag{4.2}$$

where $e = (e_1,\dots,e_N)^T \in \mathbb{R}^N$ is a vector. Let 0_{D_p} denote the vector of zeros $(0,\dots,0)^T \in \mathbb{R}^{D_p}$. The dual problem to this problem becomes similar as in (3.15)

$$\left[\begin{array}{c|c} 0_{D_p\times D_p} & {X_p}^T \\ \hline X_p & \Omega + \frac{1}{\gamma}I_N \end{array}\right] \left[\begin{array}{c} \beta \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0_{D_p} \\ \hline Y \end{array}\right], \tag{4.3}$$

where $\alpha \in \mathbb{R}^N$ are the Lagrange multipliers and $\Omega \in \mathbb{R}^{N\times N}$ denotes the kernel matrix as previously. Eliminating the Lagrange multipliers from the linear system (4.3), results in the following set of linear equations

$$\left[X_p^T \left(\Omega + \frac{1}{\gamma}I_N\right)^{-1} X_p\right] \beta = X_p^T \left(\Omega + \frac{1}{\gamma}I_N\right)^{-1} Y. \tag{4.4}$$

Note the correspondence with the generalized and weighted least squares regression where the errors obey a pre-specified correlation function (Mardia *et al.*, 1979; Wetherill, 1986).

From the conditions of optimality, it follows that the optimal model can be evaluated in a new point $(x_*, x_*^p) \in \mathbb{R}^D \times \mathbb{R}^{D_p}$ as follows

$$\hat{y} = \Omega_{\mathscr{D}}(x_*)^T \hat{\alpha} + \hat{\beta}^T x_*^p, \tag{4.5}$$

where $\Omega_{\mathscr{D}}(x_*) = (K(x_1,x_*),\dots,K(x_N,x_*))^T \in \mathbb{R}^N$ and $\hat{\alpha}$ and $\hat{\beta}$ solve (4.3). Furthermore, the conditions for optimality result into the property $\gamma e_i = \alpha_i$ and the orthogonality constraints $X^T\alpha = 0_{D_p}$ in case the parametric components are not regularized $\gamma_\beta = 0$. The following modification to the conjugate gradient algorithm provides an efficient implementation for the solution of the set of linear equations (4.3):

Algorithm 4.1. [Semi-parametric Models] *Given the set of linear equations (4.3), the conjugate gradient algorithm (CG) can be modified for solving this positive semi-definite linear system. First consider the positive definite matrix $A \in \mathbb{R}^{N\times N}$ and the vector $b \in \mathbb{R}^N$, Then the set of linear equations $Ax = b$ can be solved for x using CG as described in e.g. (Golub and van Loan, 1989; Nocedal and Wright, 1999). Having fixed this algorithm, one can cast the positive semi-definite problem (4.3) as two different, less complex and strictly definite sets of equations as follows. The convergence speed and the use of possible preconditioners (Nocedal and Wright, 1999) was investigated in the context of LS-SVMs (Hamers, 2004).*

1. *solve for $A \in \mathbb{R}^N$ in the linear system*

$$\left(\Omega + \frac{1}{\gamma} I_N\right) A = Y. \tag{4.6}$$

2. *solve for $B \in \mathbb{R}^{N \times D_p}$ in the linear system*

$$\left(\Omega + \frac{1}{\gamma} I_N\right) B = X_p. \tag{4.7}$$

3. *Let $S \in \mathbb{R}^{D_p \times D_p}$ be defined as follows*

$$S = X_p^T B. \tag{4.8}$$

4. *The parameters β then result from*

$$S\beta = B^T Y. \tag{4.9}$$

   Note that this problem may be ill-conditioned as the condition number of S is large.

5. *The Lagrange multipliers solving (4.3) can be recovered as*

$$\alpha = A - B^T \beta. \tag{4.10}$$

This algorithm corresponds with the derivation as in (Suykens et al.*, 1999; Suykens* et al.*, 2002*b*).*

This algorithm can be verified easily by eliminating the variable B and A and comparing the result with (4.4) and (4.3).

4.1.2 Semi-parameteric classification with SVMs

All machines described in the previous section can be extended with parameteric components which are not to be regularized explicitly. Consider the case of classification with SVMs as described in Subsection 3.7.1. Let an observation consist of a parametric term $x^P = \left(x^{(1)}, \ldots, x^{(P)}\right)^T \in \mathbb{R}^{D_P}$ and a term used for non-parametric modeling $x \in \mathbb{R}^D$. Consider the semi-parametric description of the hyperplane

$$\mathrm{Hp}(w, \beta) = \left\{ x \in \mathbb{R}^{d_\varphi} \,\middle|\, f(x) = w^T \varphi(x) + \beta^T x_p = 0 \right\}, \tag{4.11}$$

with parameters $\beta \in \mathbb{R}^{D_p}$. Then the modified distance measure of a point consisting of a parametric term $x_*^p \in \mathbb{R}^{D_P}$ and $x_* \in \mathbb{R}^D$ is adopted

$$d\left(\varphi(x_*), \mathrm{Hp}(w, \beta)\right) = \frac{|f(x_*)|}{\|f'(x_*)\|_2} = \frac{y_i(w^T \varphi(x_*) + \beta^T x_*^p)}{w^T w}. \tag{4.12}$$

which is invariant to the parameteric terms x^P. The resulting semiparameteric SVM is summarized in the following Lemma.

Lemma 4.1 (Semi-parameteric SVMs). *Consider the maximal margin classifier using a hyperpplane described in (4.11) and the modified distance function (4.12):*

$$(\hat{w}, \hat{\beta}, \hat{\xi}) = \arg\min_{w,\xi,\beta} \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i$$
$$s.t. \quad \begin{cases} y_i \left(w^T \varphi(x_i) + \beta^T x_i^p \right) \geq 1 - \xi_i & \forall i = 1, \ldots, N \\ \xi_i \geq 0 & \forall i = 1, \ldots, N. \end{cases} \tag{4.13}$$

The dual problem becomes

$$\hat{\alpha} \quad = \quad \arg\max_{\alpha} -\frac{1}{2} \alpha^T \Omega_Y \alpha \quad + \quad \alpha \quad s.t. \quad \begin{cases} \sum_{i=1}^N \alpha_i x_i^p = 0 & \forall p = 1, \ldots, P \\ 0 \leq \alpha_i \leq C & \forall i = 1, \ldots, N, \end{cases} \tag{4.14}$$

where $\alpha = (\alpha_1, \ldots, \alpha_N)^T \in \mathbb{R}^N$ are the Lagrange multipliers corresponding to the constraints in (4.13) and $\Omega_Y \in \mathbb{R}^{N \times N}$ is defined as $\Omega_{y,ij} = K(x_i, x_j) y_i y_j$ for all $i, j = 1, \ldots, N$.

The proof is omitted as it goes along the same lines as described in the previous chapter.

Remark 4.1. This result triggers the following observation. Let the parameteric terms consist of two variables which are (close to) collinear. It is clear that the solution to the primal problem (4.13) is numerical ill-conditioned as no form of regularization on the parameters is present. The dual problem (4.14) is not suffering this problem as the influence of the parameteric terms does only occur in the occurence of the equality constraints. The ill-conditioning however will reoccur if one is interested in the value of the estimated parameters by exploiting the complementary slackness conditions.

4.2 Estimating Additive Models with Componentwise Kernel Machines

Direct estimation of high dimensional nonlinear functions using a non-parametric technique without imposing restrictions faces the problem of the curse of dimensionality (Bellman and Kalaba, 1965). One way to quantify the curse of dimensionality is the optimal minimax rate of convergence $N^{-2l/(2l+D)}$ for the estimation of an l times differentiable regression function which converges to zero slowly if D is large compared to l (Stone, 1982). Several attempts were made to overcome this obstacle, including projection pursuit regression (Friedman and Tukey, 1974; Friedmann and Stuetzle, 1981) and kernel methods for dimensionality reduction (KDR) (Fukumizu *et al.*, 2004).

Another possibility to overcome the curse of dimensionality is to impose additional structure on the regression function. Additive models are very useful for approximating high dimensional nonlinear functions (Stone, 1985; Hastie and Tibshirani, 1990).

These methods and their extensions have become one of the widely used non-parametric techniques as they offer a compromise between the somewhat conflicting requirements of flexibility, dimensionality and interpretability. Traditionally, splines (Wahba, 1990) are commonly used in the context of additive models as e.g. in MARS (see e.g. (Hastie *et al.*, 2001)) or in combination with ANOVA (Neter *et al.*, 1974). Additive models were brought further to the attention of the machine learning community by e.g. (Vapnik, 1998; Gunn and Kandola, 2002).

The following approach was described in (Pelckmans *et al.*, 2004, *In press*). Some extra notation is introduced. Let x consist of P different components $x = \left(x^{(1)}, \ldots, x^{(P)}\right)$ where each component is defined as $x^{(p)} \in \mathbb{R}^{D^{(p)}}$ and $D^{(p)} \in \mathbb{N}$ for $p = 1, \ldots, P$. In the simplest case, let $P = D$, $D^{(p)} = 1$ and $x^{(p)} = x^p$ for all $p = 1, \ldots, D$.

Definition 4.1. [Additive Model] *An additive model consists of a sum of (possibly nonlinear) functions each based on one (or a set of) independent variable(s). Let $x \in \mathbb{R}^D$ represent a set of d components $\left(x^{(1)},...,x^{(P)}\right)$*

$$f(x) = \sum_{p=1}^{P} f^p\left(x^{(p)}\right), \tag{4.15}$$

where $f^p : \mathbb{R}^{D^{(p)}} \to \mathbb{R}$ are smooth functions.

The optimal rate of convergence for estimators based on this model is $N^{-2l/(2l+d)}$ where $d = \max_p \left(D^{(p)}\right)$ which is independent of D (Stone, 1985), and $l \in \mathbb{R}^+$ is a measure of the smoothness of the underlying function. Most state-of-the-art estimation techniques for additive models can be divided into two approaches (Hastie *et al.*, 2001):

- *Iterative approaches* use an iteration where in each step part of the unknown components are fixed while optimizing the remaining components. This is motivated as:
$$\hat{f}^{p_1}\left(x_k^{(p_1)}\right) = y_k - e_k - \sum_{p_2 \neq p_1} \hat{f}^{p_2}\left(x_k^{(p_2)}\right), \tag{4.16}$$
for all $k = 1, \ldots, N$ and $d_1 = 1, \ldots, D$. Once the $N-1$ components of the second term are known, it becomes easy to estimate the lefthandside. For a large class of linear smoothers, such so-called backfitting algorithms are equivalent to a Gauss-Seidel algorithm for solving a large ($ND \times ND$) set of linear equations (Hastie *et al.*, 2001). The backfitting algorithm (Hastie and Tibshirani, 1990) is theoretically and practically well motivated.

- *Two-stages marginalization approaches* construct in the first stage a general black-box pilot estimator (as e.g. a Nadaraya-Watson kernel estimator) and finally estimate the additive components by marginalizing (integrating out) for each component the variation of the remaining components (see e.g. (Linton and Nielsen, 1995)).

Although consistency of both is shown under certain conditions, important practical problems (number of iteration steps in the former) and more theoretical problems (the pilot estimator needed for the latter procedure is a too generally posed problem) are still left.

The framework of the primal-dual kernel machines does provide a one-stage alternative. For completeness, consider the case of the LS-SVM or L_2 kernel machine. The derivation however is extendable to any chosen loss function for fitting an additive model which includes a (parametric) bias term. The considered model class becomes

$$\mathscr{F}_{\varphi,(P)} = \left\{ f(x) = \sum_{p=1}^{P} {w_p}^T \varphi_p \left(x^{(p)} \right) + b \;\middle|\; w_p \in \mathbb{R}^{D_\varphi^{(p)}}, b \in \mathbb{R} \right\}, \tag{4.17}$$

where $\varphi_p : \mathbb{R}^{D^{(p)}} \to \mathbb{R}^{D_\varphi^{(p)}}$ is a fixed but unknown mapping to a space of dimension $D_\varphi^{(p)}$ (possibly infinite dimensional). Consider the modified regularization term

$$(\hat{w}_p, \hat{b}, \hat{e}) = \arg\min_{w_p,b,e} \mathscr{J}_\gamma^c(w_p,b,e) = \frac{1}{2}\sum_{p=1}^{P} {w_p}^T w_p + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2$$
$$\text{s.t.} \quad \sum_{p=1}^{P} w_p^T \varphi_p \left(x_i^{(p)} \right) + b + e_i = y_i, \;\; \forall i = 1,\ldots,N. \tag{4.18}$$

Constructing the Lagrangian gives

$$\mathscr{L}_\gamma^c(w_p,b,e;\alpha) = \mathscr{J}_\gamma^c - \sum_{i=1}^{N} \alpha_i \left(\sum_{p=1}^{P} w_p^T \varphi_p \left(x_i^{(p)} \right) + b + e_i - y_i \right), \tag{4.19}$$

with multipliers $\alpha = (\alpha_i,\ldots,\alpha_N)^T \in \mathbb{R}^N$. Taking the first order conditions for optimality gives

$$\text{KKT} \begin{cases} \dfrac{\partial \mathscr{L}_\gamma^c}{\partial w_p} = 0 \to & w_p = \sum_{i=1}^N \alpha_i \varphi_p \left(x_i^{(p)} \right) & \forall p = 1,\ldots,P \\ \dfrac{\partial \mathscr{L}_\gamma^c}{\partial e_i} = 0 \to & \gamma e_i = \alpha_i & \forall i = 1,\ldots,N \\ \dfrac{\partial \mathscr{L}_\gamma^c}{\partial b} = 0 \to & \sum_{i=1}^N \alpha_i = 0 & \\ \dfrac{\partial \mathscr{L}_\gamma^c}{\partial \alpha_i} = 0 \to & \sum_{p=1}^P w_p^T \varphi_p \left(x_i^{(p)} \right) + b + e_i = y_i, & \forall i = 1,\ldots,N. \end{cases} \tag{4.20}$$

By eliminating the primal variables w_p and e_i, one obtains the following dual linear system

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega_P + \frac{1}{\gamma} I_N \end{array} \right] \left[\begin{array}{c} b \\ \hline \alpha \end{array} \right] = \left[\begin{array}{c} 0 \\ \hline Y \end{array} \right], \tag{4.21}$$

where $\Omega_P = \sum_{p=1}^P \Omega^{(p)} \in \mathbb{R}^{N \times N}$ and $\Omega_{ij}^{(p)} = K_p\left(x_i^{(p)}, x_j^{(p)} \right) = \varphi_p \left(x_i^{(p)} \right)^T \varphi_p \left(x_j^{(p)} \right)$ is the inner product of the feature maps of the pth component evaluated on the points $x_i^{(p)}$

and $x_j^{(p)}$. Let $\Omega_{\mathscr{D}}^{(p)} = \left(K_p\left(x_1^{(p)}, x_*^{(p)}\right), \ldots, K_p\left(x_N^{(p)}, x_*^{(p)}\right)\right) \in \mathbb{R}^N$, then the pth estimated model f^p can be evaluated in a new point $x_* = \left(x_*^{(1)}, \ldots, x_*^{(P)}\right)$ as follows

$$\hat{f}_p(x_*) = \sum_{i=1}^{N} \hat{\alpha}_i K_p\left(x_i^{(p)}, x_*^{(p)}\right) = \Omega_{\mathscr{D}}^{(p)}\left(x_*^{(p)}\right)^T \hat{\alpha}. \tag{4.22}$$

The total function can be evaluated in a point x_* as follows

$$\hat{f}(x_*) = \sum_{p=1}^{P} \hat{f}_p\left(x_*^{(p)}\right) + \hat{b} = \sum_{p=1}^{P} \Omega_{\mathscr{D}}^{(p)}\left(x_*^{(p)}\right)^T \hat{\alpha} + \hat{b}. \tag{4.23}$$

Observe the fact that the unknowns $\hat{\alpha}$ are constant over the different components. This is unlike any parametric approach or a backfitting approach where each component is characterized by its own set of unknowns.

The set of linear equations (4.21) corresponds with a classical LS-SVM regressor where a modified kernel is used given as

$$K(x_k, x_j) = \sum_{p=1}^{P} K_p\left(x_k^{(p)}, x_j^{(p)}\right). \tag{4.24}$$

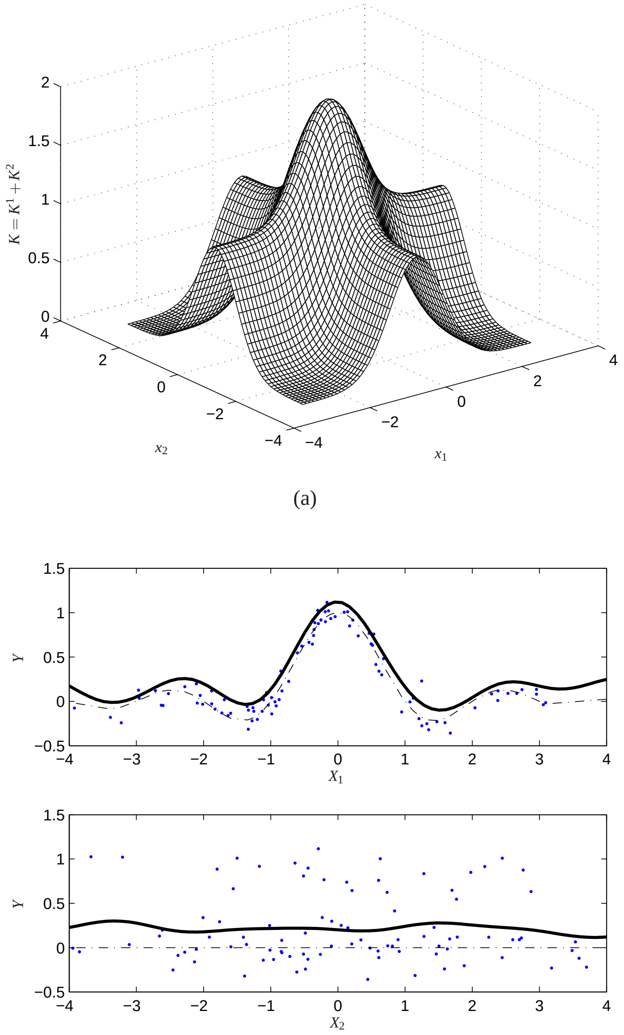
Figure 4.1 shows the modified kernel in case a one dimensional Radial Basis Function (RBF) kernel is used for all D (in the example, $D = 2$) components. This observation implies that componentwise LS-SVMs inherit results obtained for classical LS-SVMs and kernel methods in general. From a practical point of view, the previous kernels (and a fortiori componentwise kernel models) result in similar algorithms as considered in the ANOVA kernel decompositions as in (Vapnik, 1998; Gunn and Kandola, 2002).

$$K(x_k, x_j) = \sum_{d=1}^{D} K^d\left(x_k^{(d)}, x_j^{(d)}\right) + \sum_{d_1 \neq d_2} K^{d_1 d_2}\left((x_k^{(d_1)}, x_k^{(d_2)}), (x_j^{(d_1)}, x_j^{(d_2)})\right) + \ldots, \tag{4.25}$$

where the componentwise LS-SVMs only consider the first term in this expansion. The formal proof of the underlying theorem that the kernel of the union of two orthogonal subspaces equals the sum of the individual kernels corresponding with each subspace may be found in (Aronszajn, 1950). The derivation as such bridges the gap between the estimation of additive models and the use of ANOVA kernels.

4.3 Imposing Pointwise Inequalities

Consider the case where prior knowledge in the form of known (in)equalities are known to hold on a finite set of locations. This kind of discrete structure can be easily imposed during the learning process by adopting the primal-dual argument. This case was studied in some detail in (Pelckmans *et al.*, 2004*g*) and contrasted to various existing two-stages approaches as described in (Boor and Schwartz, 1977; Gaylord and Ramirez, 1991). The following example gives a further application of this research.



(a)

(b)

Figure 4.1: *Illustrations of the mechanism of componentwise LS-SVMs for fitting additive models.* **(a)** *Estimation of an additive model with a componentwise kernel machine and a RBF kernel corresponds with the use of a modification of the RBF kernel as displayed.* **(b)** *A simple example of the two components of a componentwise LS-SVM (solid lines) fitted 50 noisy data-samples with underlying additive model as illustrated by the dashed-dotted lines. The contributions of the two variables can be visualized explicitly due to the additive structure. It becomes that this example depends in a clear way on the first variable but not on the second one.*

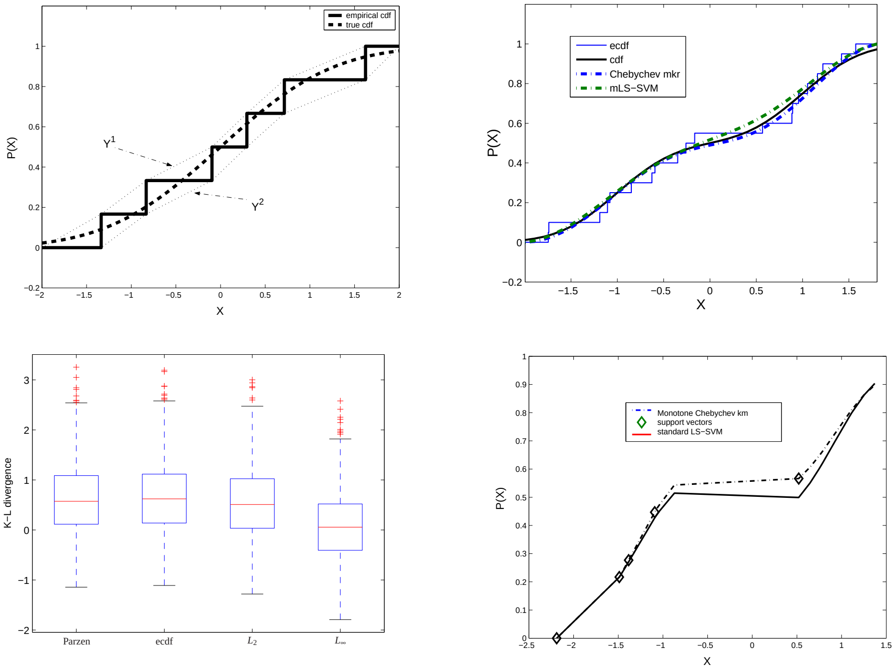


Figure 4.2: *Illustration of the use of monotone kernel machines in estimating the cumulative distribution function.* **(a)** *As the ecdf is discontinuous at the sample points, the estimated cdf should lie between the upper- (Y^1) and lower-curve (Y^2) where possible while being smooth.* **(b)** *Application of the smooth estimate of the ecdf on the artificial example of Subsection 4.1.* **(c)** *Boxplots of the results of a Monte Carlo simulation for estimating the cdf based on respectively the Parzen window, ecdf, the monotone LS-SVM smoother and the monotone Chebychev kernel regressor.* **(d)** *Comparison of the smooth monotone Chebychev kernel machine and its sparse representation (using only 5 support vectors) and a standard LS-SVM which is not guaranteed to be monotone in general.*

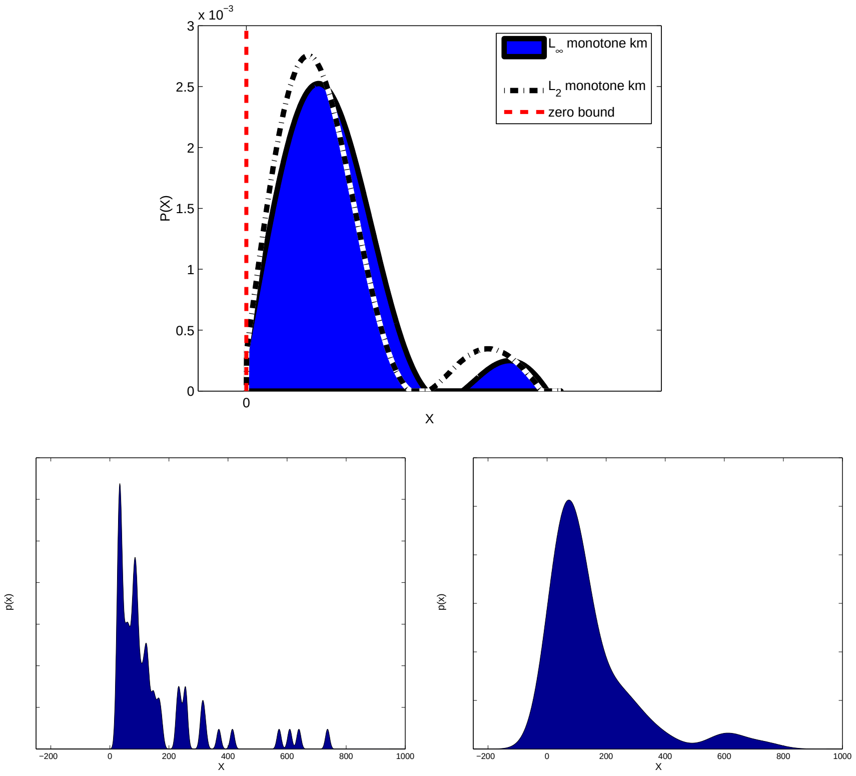


Figure 4.3: **(a)** *Density estimation of the suicide data using the derivative of the monotone Chebychev kernel regressor and the monotone LS-SVM technique. Both estimates reflect the trimodal structure as well as the positive support. A well-known drawback of the Parzen window estimator in this case is seen in that no single bandwidth parameter of the Parzen window results in both a strictly positive density (one has to under-smooth,* **(b)***) and a smooth trimodal structure (one has to over-smooth,* **(c)***).*

Example 4.1 **[Empirical distribution estimate]** In Example 1.1 and Example 2.1 different approaches were given to the task of univariate density estimation. Complementary to these examples, the techniques introduced in this section can be exploited for designing an estimator of a kernel based cdf estimator in the case of univariate data-samples. Then the empirical cdf estimator is defined as

$$\hat{P}(x) = \frac{1}{N}\sum_{i=1}^{N} I_{(x_i<x)}. \tag{4.26}$$

Now assume that the generating cdf is smooth. The best smoother in regularized L_1 sense which takes the structure of the cdf into account (at least at the sample points).

$$\mathscr{J}_C(w,e) = \frac{1}{2}w^T w + C\sum_{i=1}^N |e_i|$$
$$\text{s.t.} \quad \begin{cases} w^T\varphi(x_i) + e_i = \frac{1}{N}\sum_{j=1}^N I^2_{(x_j<x_i)} & \forall i = 1,\dots,N \\ w^T\varphi(x_i) \leq w^T\varphi(x_j) & \text{if} \quad x_i^1 \leq x_j^1 \text{ and } x_i^2 \leq x_j^2. \end{cases} \tag{4.27}$$

A primal-dual kernel machine can be derivation using the standard techniques, see (Pelckmans *et al.*, 2004*g*) where the equivalent univariate case is studied in some detail.

The technique based on the L_2 norm and the L_∞ norm was applied to generate a density estimate of the suicide data (see e.g. (Silverman, 1986)) by taking the numerical derivative of the smooth estimate. In this case the support of the data was known to have an exact lower bound at 0 which can be nicely incorporated in this framework as shown in Figure 4.3.b. A main advantage of this technique over the use of the Parzen kernel estimator becomes apparent in this study. As well known in literature, this strictly positive dataset manifests a tri-modal structure (Silverman, 1986). As shown in Figure 4.3.b and 4.3.c one cannot find a single bandwidth of the Parzen window estimator which result in a plausible density satisfying both constraints, while the monotone Chebychev kernel machine manages to do so in Figure 4.3. Remark that for convenience, the density function is displayed although no guarantees are given that the derivative of the estimated cdf is optimal.

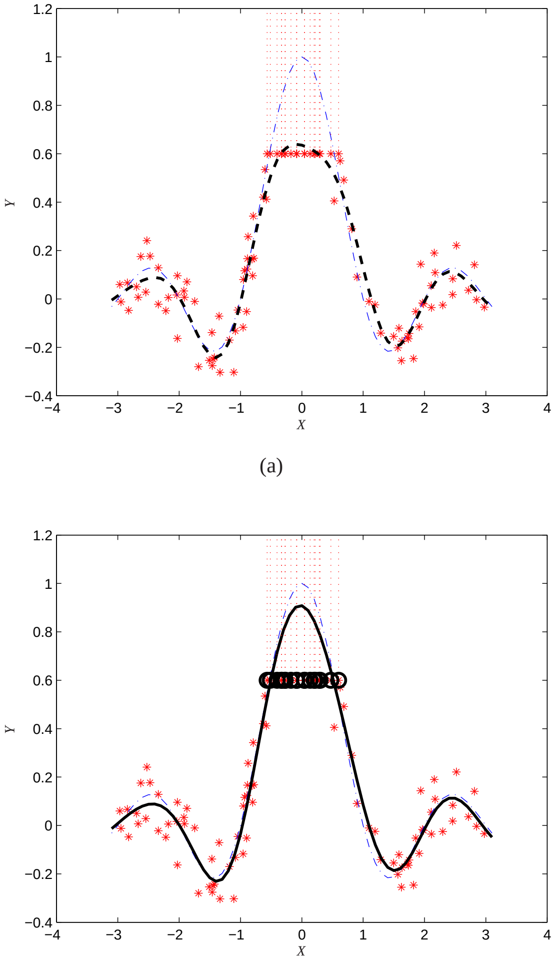
4.4 Censored Primal-Dual Kernel Regression

The case of incomplete or censored output observations is considered here. Let a data observation consist of a triple $(x_i, y_i^-, y_i^+) \in \mathbb{R}^D \times \mathbb{R} \times \mathbb{R}$ where the unknown (noisy) output observation is only known to be contained in the interval $y_i \in [y_i^-, y_i^+]$. For notational convenience, this notation differs somewhat from the one used in the literature on survival regression as employed e.g. in (Cox, 1972), where an extra indicator variable is used to indicate wether the observation is censored or not. This follows here from the fact is the range of the interval $[y_i^-, y_i^+]$ equals zero or not. Let then $\mathscr{D}_c = \{(x_i, y_i^-, y_i^+)\}_{i=1}^N \in \mathbb{R}^D \times \mathbb{R} \times \mathbb{R}$. Let the data be generated from $y_i = f(x_i) + e_i$ with $y_i \in [y_i^-, y_i^+]$ and e_i i.i.d. sampled from a fixed but unknown distribution. Let $Y^+ = (y_1^+, \dots, y_N^+)^T \in \mathbb{R}^N$ and $Y^- = (y_1^-, \dots, y_N^-)^T \in \mathbb{R}^N$. The primal-dual derivation of a modified least squares cost-function is summarized in the following Lemma.

Lemma 4.2. [Censored Primal-Dual Kernel Machines] *Let the class of estimators be contained in $\mathscr{F}_\varphi$ described in (3.8). Consider the modified regularized cost function*

$$(\hat{w}, \hat{b}, \hat{e}) = \arg\min_{w,b,e} \mathscr{J}_\gamma(w,b) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^N e_i^2$$
$$s.t. \quad y_i^- \leq w^T\varphi(x_i) + b + e_i \leq y_i^+, \quad \forall i = 1,\dots,N. \tag{4.28}$$

The dual problem becomes



(a)

(b)

Figure 4.4: *A simple example of a censored primal-dual kernel machine. Given a dataset based on the* sinc*-function including noise terms where observations of the output above* $y = 0.6$ *are censored and only known to be contained in the interval* $[0.6, \infty]$. *The lower bounds of the observations are denoted as asteriskses, while the intervals of the censored observations are given as dotted lines. The underlying* sinc *function is given as a dashed-dotted line.* **(a)** *The application of the standard LS-SVM discussed in Section 3.3. The solid line gives the estimate which follow the lower-bound of the intervals.* **(b)** *The application of the modified LS-SVM for censored observations. The circles indicate the sparse support vectors of the estimate. The main advantage of the latter is seen in the fact that one does not try to fit the censoring bound at* $y = 0.6$.

$$(\hat{\alpha}^+,\hat{\alpha}^-) = \underset{\alpha^+,\alpha^-}{\arg\max} -\frac{1}{2}(\alpha^- - \alpha^+)^T\left(\Omega+\frac{1}{\gamma}I_N\right)(\alpha^+ - \alpha^-)$$
$$+Y^{-T}\alpha^- - Y^{-T}\alpha^+ \quad s.t. \quad \begin{cases}\alpha^+,\alpha^- \geq 0_N\\ 1_N^T(\alpha^- - \alpha^+) = 0.\end{cases} \qquad (4.29)$$

The estimate can be evaluated in a new datapoint $x_* \in \mathbb{R}^D$ as $\hat{f}(x_*) = \Omega_{\mathscr{D}}(\hat{\alpha}^- - \hat{\alpha}^+) + \hat{b}$ where $\hat{\alpha}^+,\hat{\alpha}^-$ solve (4.29) and $\hat{b}$ may be recovered from the complementary slackness conditions.

Proof. The Lagrangian of the modified cost-function becomes

$$\mathscr{L}^c_\gamma(w,b,e;\alpha^+,\alpha^-) = \frac{1}{2}w^Tw + \frac{\gamma}{2}\sum_{i=1}^N e_i^2$$
$$-\sum_{i=1}^N \alpha_i^+\left(-(w^T\varphi(x_i)+b+e_i) - y_i^+\right) - \sum_{i=1}^N \alpha_i^-\left((w^T\varphi(x_i)+b+e_i) - y_i^-\right). \qquad (4.30)$$

The Karush-Kuhn-Tucker conditions become

$$\text{KKT} = \begin{cases}
\frac{\partial \mathscr{L}^c_\gamma}{\partial w} \to & w = \sum_{i=1}^N(\alpha_i^- - \alpha_i^+)\varphi(x_i) & & (a)\\
\frac{\partial \mathscr{L}^c_\gamma}{\partial b} \to & \sum_{i=1}^N(\alpha_i^- - \alpha_i^+) = 0 & & (b)\\
\frac{\partial \mathscr{L}^c_\gamma}{\partial e} \to & \gamma e_i = (\alpha_i^- - \alpha_i^+) & \forall i=1,\ldots,N & (c)\\
& \alpha_i^+,\alpha_i^- \geq 0 & \forall i=1,\ldots,N & (d)\\
& y_i^- \leq w^T\varphi(x_i)+b+e_i \leq y_i^+ & \forall i=1,\ldots,N & (e)\\
& \alpha_i^+\left(-(w^T\varphi(x_i)+b+e_i) - y_i^+\right) = 0 & \forall i=1,\ldots,N & (f)\\
& \alpha_i^-\left((w^T\varphi(x_i)+b+e_i) + y_i^-\right) = 0. & \forall i=1,\ldots,N & (g)
\end{cases}$$
$$(4.31)$$

Elimination of the primal variables w,b and e using the conditions (4.31.abc) leads to the dual formulation (4.29). If α_i^+ were nonzero, the equality $w^T\varphi(x_i)+b+e_i = -y_i^+$ (4.31.f) or the equivalent in (4.31.g) can be used to recover the term b implicit encoded in the dual formulation (4.29). □

The case of right censoring of the ith datapoint (denoted as $y_i^+ = \infty$) follows along the lines, but the multiplier α_i^+ equals zero and the optimization problem simplifies. The case of left censoring of the ith datapoint (denoted as $y_i^- = -\infty$) is analogous. Note that other loss functions (as e.g. the L_ε loss) can be treated along the same lines.

Lemma 4.3. [Sparseness in Censored Learning Machines] *Only censored observations where $y_i^- < y_i^+$ may lead to sparse support vectors.*

Proof. This follows readily from inspecting the complementary slackness conditions (4.31.fg). □

Figure 4.4 illustrates the difference on a simple example based on the sinc function.

Chapter 5

Relations with other Modeling Methods

This chapter takes the opportunity to situate the previous discussion within a broader context and to review various related approaches. While differences were mainly conceived in the conjectured assumptions and the way of deriving the results, the final formulations frequently present many correspondences. However, different interpretations of the results seem to support the coexistence of the individual approaches. Methods close to the formulation of LS-SVMs include different variational approaches as smoothing splines (Section 5.1), the approach of Gaussian processes (Section 5.2) and Kriging methods in the context of spatial analysis (Section 5.3). Relationships with other methods in system-identification, wavelets, the theory of inverse problems and the weighted least squares approach are described in Section 5.4.

5.1 Variational Approaches and Smoothing Splines

Spline methods have a long tradition concerning theoretical as well as practical aspects (Schoenberg, 1946), and extended their reach from a purely function approximation setting towards a nonlinear smoothing task. The latter is reviewed briefly in accordance with the exposition of (Wahba, 1990) in order to relate such smoothing splines to the proposed methodology. All of the splines discussed in the cited work may be obtained as solutions to variational problems, which makes the methodology at first sight different from the discrete optimization approach of the primal-dual kernel machines. The route followed by the work of G. Wahba differs from the main body of literature on spline methods as it adopts the Reproducing Kernel Hilbert Space framework as studied in (Aronszajn, 1950). For convenience, only the one dimensional case is considered, though extensions are made to two and three-dimensional smoothing problems, see e.g. (Dierckx, 1993).

Somewhat central is the following definition:

Definition 5.1. [Mercer theorem and Reproducing Kernel Hilbert Space (rkhs), (Mercer, 1909; Aronszajn, 1950)] *A real Reproducing Kernel Hilbert Space $\mathscr{H}$ (rkhs) is a Hilbert space (complete under an inner product $< \cdot, \cdot >$ and satisfying everywhere the triangular inequality) of real valued functions $f : \mathbb{R} \to \mathbb{R}$ with the property that for each $x \in \mathbb{R}$ there exist a functional $R_x : \mathbb{R} \to \mathbb{R}$ (by the Riesz representer theorem) such that $< R_x, f >= f(x)$ are bounded linear functionals. Furthermore, a unique reproducing kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ can be attached to a specific rkhs defined as*

$$k(x,y) =< R_x, R_y >, \tag{5.1}$$

which is a positive definite function (see also the Mercer Theorem 3.1). The converse also holds (a reproducing kernel constructs a unique rkhs).

At the core of the derivation of smoothing splines lies the description of an rkhs $\mathscr{H}^f$ endowed with an inner product (and hence a norm) involving derivatives as summarized as follows

Lemma 5.1. [Rkhs of Smooth Functions, (Wahba, 1990)] *The following Sobolev space is a rkhs*

$$\mathscr{H}^f = \Big\{ f : [0,1] \to \mathbb{R} \ \Big| \ f^{(r)} \ \textit{absolutely continuous} \\ \textit{for all } r = 0, \ldots, m-1, f^{(m)} \in L_2(\mathbb{R}) \Big\}. \tag{5.2}$$

Proof. The proof is sketched as follows (Wahba, 1990). Consider the *m*th order Taylor series approximation

$$f(x) = \sum_{r=0}^{m-1} \frac{x^r}{r!} f^{(r)}(0) + \int_0^1 \frac{(x-z)_+}{(m-1)!} f^{(m)}(z) dz \triangleq f_{m-1}(x) + f_m(x), \tag{5.3}$$

where $(z)_+ = z$ if $z > 0$ and zero otherwise. Let $\mathscr{H}^f$ be decomposed in two subspaces corresponding with the two terms in the right hand side of equation (5.3) such that $\mathscr{H}^f = \mathscr{H}_0^f + \mathscr{H}_m^f$. Consider the Sobolev function space

$$\mathscr{H}_m^f = \Big\{ f : [0,1] \to \mathbb{R} \ \Big| \ f^{(r)} \ \text{absolutely continuous,} \\ f^{(r)}(0) = 0 \ \text{ for all } r = 0, \ldots, m-1, f^{(m)} \in L_2(\mathbb{R}) \Big\}. \tag{5.4}$$

It follows that any function $f \in \mathscr{H}_m^f$ can be written as

$$f(x) = \int_0^1 \frac{(x-u)_+}{(m-1)!} f^{(m)}(u) du$$

$$\triangleq \int_0^1 G_m(x,u)\, f^{(m)}(u)du =< G_m(x,\cdot), f^{(m)} > \quad (5.5)$$

where $G_m(x,u)$ is the Green function for the problem $D^m f = g$ with D^m denoting the linear operator corresponding with the mth derivative, (Wahba, 1990). It then can be shown that the reproducing kernel corresponding with $\mathscr{H}_m^f$ becomes

$$K_m(x,y) = \int_0^1 G_m(x,u)G_m(u,y)du. \quad (5.6)$$

Let $\{\phi_r\}_{r=0}^{m-1}$ be a set of functions spanning the null-space of $\mathscr{H}_0^f$. The rkhs corresponding to the function space $\mathscr{H}^f$ and the corresponding kernel becomes

$$\begin{cases} \|f\|_{\mathscr{H}}^f = \sum_{r=0}^{m-1} f^{(r)}(0)^2 + \int_0^1 f^m(u)^2 du{+}3mm \\ K(x,y) = \sum_{r=0}^{m-1} \phi(x)_r \phi(y)_r + \int_0^1 G_m(x,u)G_m(u,y)du = G_m(x,y). \end{cases} \quad (5.7)$$

□

The representer theorem then states that the function $f \in \mathscr{H}^f$ minimizing the regularized cost-function can be represented as follows.

Theorem 5.1. [Representer Theorem, (Craven and Wahba, 1979)] *Suppose we are given a nonempty set $\mathscr{X} \subset \mathbb{R}^D$, a positive definite real-valued kernel function $K_m : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ being the reproducing kernel of a Hilbert space $\mathscr{H}_m^f$ of functionals $f : \mathscr{X} \to \mathbb{R}$. Let the null-space $\mathscr{H}_0^f$ of $\mathscr{H}_m^f$ spanned by a set of basis functions $\{\phi_d : \mathscr{X} \to \mathbb{R}\}_{d=1}^D$, let $\mathscr{H}^f$ denote the sum of the orthogonal spaces $\mathscr{H}^f = \mathscr{H}_0^f + \mathscr{H}_m^f$, let $\mathscr{D}$ be a training set $\{(x_i,y_i)\}_{i=1}^N$ i.i.d. sampled from $\mathscr{X} \times \mathbb{R}$, let $g : \mathbb{R}^+ \to \mathbb{R}$ be a strictly monotonically increasing real-valued function, $\ell : \mathbb{R} \to \mathbb{R}$ an arbitrary loss-function and a class of functions*

$$\mathscr{F} = \left\{ f \in \mathscr{H}^f \quad \middle| \quad f(x) = \sum_{d=1}^D w_d \phi_d(x_i) + \sum_{i=1}^\infty \beta_i K(x_i,x), x_i \in \mathscr{X}, w_d, \beta_i \in \mathbb{R}, \|f\|_{\mathscr{H}}^f < \infty \right\}, \quad (5.8)$$

where $\|\cdot\|_{\mathscr{H}}^f$ denotes the squared norm induced by the Hilbert space $\mathscr{H}_m^f$ of functionals f becoming $\|f\|_{\mathscr{F}}^{\mathscr{H}} = \sum_{ij=1}^\infty \beta_i \beta_j K_m(x_i,y_j)$. Consider a regularized loss function

$$\min_{f \in \mathscr{F}} \mathscr{J}(f) = g(\|f\|_{\mathscr{H}}^f) + \gamma \sum_{i=1}^N \ell\left(f(x_i) - y_i\right), \quad (5.9)$$

where g is a monotone function. Then any f minimizing the regularized loss function admits the representation of the form

$$\hat{f}(x^*) = \sum_{i=1}^N a_i K(x_i,x^*) + \sum_{d=1}^D w_d \phi_d(x^*), \quad (5.10)$$

where $a = (a_1,\ldots,a_N)^T \in \mathbb{R}^N$ and $w = (w_1,\ldots,w_D)^T \in \mathbb{R}^D$ be vectors of unknowns.

This theorem has a long tradition in functional analysis and variational methods and formed the basis of many methods as e.g. smoothing splines (Wahba, 1990) and was tuned towards kernel machines (Schölkopf *et al.*, 2001).

Let $\{\phi_d\}_{d=0}^{m-1}$ be an orthogonal set of basis functions spanning the subspace $\mathscr{H}_0^f$ such that $\phi_d(x) = \frac{x^d}{d!}$. Consider the cost-function

$$\min_{f\in\mathscr{H}^f} \mathscr{J}_{\text{splines}}(f) = \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int_0^1 f^{(m)}(u)^2 du. \tag{5.11}$$

Let $X \in \mathbb{R}^{N\times m}$ be a matrix containing the evaluations of these functionals in the data points such that $X_{id} = \frac{x_i^{d-1}}{(d-1)!}$ for all $d = 1,\ldots,m$ and $i = 1,\ldots,N$. In the case of the decomposition (5.3), the kernel K_m of $\mathscr{H}_m^f$ becomes

$$K_m(x,y) = \int_0^1 \frac{(x-u)_+^{m-1}(y-u)_+^{m-1}}{((m-1)!)^2} du, \tag{5.12}$$

and the solution of the optimization problem (5.11) follows from the solution to the set of linear equations

$$\begin{bmatrix} 0_{m\times m} & X' \\ X & \Omega_m + \frac{1}{\lambda} I_N \end{bmatrix} \begin{bmatrix} w \\ a \end{bmatrix} = \begin{bmatrix} 0_m \\ Y \end{bmatrix}, \tag{5.13}$$

where $\Omega_m \in \mathbb{R}^{N\times N}$ is the kernel matrix with elements $\Omega_{m,ij} = K_m(x_i,x_j)$. The estimated function $\hat{f}$ can then be evaluated in a new point $x_* \in [0,1]$ as follows

$$\hat{f}(x_*) = \sum_{i=1}^{N} \hat{a}_i K_m(x_i,x_*) + \sum_{r=0}^{m-1} \hat{w}_r \phi_r(x_*), \tag{5.14}$$

where $\hat{a} = (\hat{a}_1,\ldots,\hat{a}_N)^T \in \mathbb{R}^N$ and $\hat{w} = (\hat{w}_0,\ldots,\hat{w}_{m-1})^T \in \mathbb{R}^m$ solve (5.13). This rkhs derivation places the smoothing splines derivation into the context of kernel machines endowed with the specific kernel (5.12) which may be rewritten as (Vapnik, 1998)

$$K_m(x_i,x_j) = \sum_{d=0}^{m} \frac{C_m^d}{2m-d+1} \min(x_i,x_j)^{2m-d+1} |x_i - x_j|^d, \tag{5.15}$$

where C_m^d is the number of combinations of d elements taken m at a time.

The regularization term $\int_0^1 f^{(m)}(u)^2 du$ may be expressed alternatively using the Fourier expansion of f denoted as $\mathscr{F}f$ as follows

$$\int_0^1 f^{(m)}(x)^2 dx = \int_{\mathbb{R}} \frac{\mathscr{F}f(\lambda)^2}{\mathscr{F}g(\lambda)} d\lambda \tag{5.16}$$

where $\mathscr{F}f(\lambda) = \frac{1}{\sqrt{2\pi}} \int_0^1 f(x)\exp(-ix\lambda)dx$ and $\mathscr{F}g : \mathbb{R} \to \mathbb{R}^+$ is a positive symmetric function that tends to zero when $|\lambda| \to \infty$, see (Girosi *et al.*, 1995). Different choices for the low-pass filter $\tilde{g}$ may be considered. The case of thin-plate splines of order m is equivalent to the choice $\mathscr{F}g(\lambda) = 1/\lambda^{2m}$ (Duchon, 1977; Schumaker, 1981). In this case the null-space $\mathscr{H}_0$ is the vector space space of polynomials of degree at most $m-1$. It is interesting to contrast this derivation to Example 3.2, Example 9.1 and Lemma 9.1.

5.2 Gaussian Processes and Bayesian Inference

A stochastic process is defined as follows, see e.g. (Doob, 1953).

Definition 5.2. [Gaussian Process, (Doob, 1953)] *Consider a family of random variables $\mathbf{Z}_{\mathbb{T}} = \{\mathbf{Z}_t\}_{t\in\mathbb{T}}$ over an index set $\mathbb{T}$ with covariance function $E(\mathbf{Z}_t\mathbf{Z}_s) = \rho(t,s)$. If $\rho(t,t+u) = \rho(u)$, the process $\mathbf{Z}_{\mathbb{T}}$ is called stationary. The process $\mathbf{Z}_{\mathbb{T}}$ is a Gaussian process when any finite subset of variables is entirely described by its first two moments.*

Classically, the index set $\mathbb{T}$ represents a series of time instants (Wiener, 1949). A representation theory due to (Loeve, 1955) shows that there is an intimate connection between Gaussian processes (time series of second order) and reproducing kernel Hilbert spaces:

Theorem 5.2. [Covariance vs. Reproducing Kernel, (Loeve, 1955)] *A positive definite covariance function of a time series ρ generates a unique Hilbert space of which $K = \rho$ is the reproducing kernel.*

This is discussed in (Loeve, 1955; Parzen, 1961; Grenander and Rosenblatt, 1957). This result relates the Gaussian processes approach to the rkhs approach as summarized in the previous subsection, see also (Weinert, 1982) which makes extensive use of this result in the context of signal processing.

More recent work (O'Hagen, 1978; Neal, 1994) also approaches problems of static regression and classification using this machinery, but mainly differ by taking a Bayesian approach (Wahba, 1990; MacKay, 1998), see also subsection 1.2.4. Let the index set here be denoted as $\mathbb{X} \subset \mathbb{R}^D$ consisting of the deterministic inputs $\{x_i\}_{i=1}^N$ which are possibly higher dimensional and non-equidistantly sampled. One typically proceeds under the assumption of zero mean $E[\mathbf{Z}_{\mathbb{X}}|\mathbb{X} = x] = m(x) = 0$. Bayes' law then relates the posterior probability of the Gaussian process $P(\mathbf{Z}_{\mathbb{X}}|\mathscr{D},\mathscr{A})$ to the likelihood $P(\mathscr{D}|\mathbf{Z}_{\mathbb{X}},\mathscr{A})$, the prior $P(\mathbf{Z}_{\mathbb{X}}|\mathscr{A})$ and the evidence $P(\mathscr{D}|\mathscr{A})$ as follows

$$P(\mathbf{Z}_{\mathbb{X}}|\mathscr{D},\mathscr{A}) = \frac{P(\mathscr{D}|\mathbf{Z}_{\mathbb{X}},\mathscr{A})P(\mathbf{Z}_{\mathbb{X}}|\mathscr{A})}{P(\mathscr{D}|\mathscr{A})}, \tag{5.17}$$

see also Subsection 1.2.4. Let $x_* = x_{N+1}$ be the input data point to be evaluated, $y_* = y_{N+1}$ the response to be found and let $\mathscr{D}^*$ be defined as the extended dataset $\{\mathscr{D},(x_{N+1},y_{N+1})\}$. Let $Z \in \mathbb{R}^{N+1}$ be a realization of the Gaussian process $\mathbf{Z}_{\mathbb{X}}$ evaluated in the observed data points. Assume the $N+1$ observations y_i are versions of Z_i perturbed by i.i.d. noise such that $y_i = Z_i + e_i$ for all $i = 1,\ldots,N+1$. The problem of prediction using Gaussian processes then boils down to finding the realization $Z \in \mathbf{Z}_{\mathbb{X}}$ with maximal posterior probability.

To formalize the problem, the likelihood function and an appropriate prior of any realization Z is to be defined. The evidence is assumed to remain constant in the setup. Consider the prototypical case that $P(\mathscr{D}|Z,\mathscr{A}) \propto \Pi_{i=1}^{N+1}\exp(-\|Z_i - y_i\|/\gamma_1)$

and $P(Z|\mathscr{A}) \propto \exp(-Z^T \Sigma Z/\gamma_2)$ with $\Sigma \in \mathbb{R}^{N+1\times N+1}$ a positive definite matrix. The maximum a posteriori (MAP) Gaussian process realization Z follows then from

$$\max_{Z\in\mathbf{Z}_{\mathbb{X}}} \log P(Z|\mathscr{D},\mathscr{A}) = \arg\min_{Z} \frac{1}{\gamma_1} \sum_{i=1}^{N+1} \|Z_i - y_i\| + \frac{\lambda}{\gamma_2} Z^T \Sigma Z, \tag{5.18}$$

where γ_1, γ_2 and λ are appropriate hyper-parameters. After taking the first order optimality conditions and by application of the matrix inversion Lemma (Golub and van Loan, 1989), the solution of the predictor of x_* is seen to equal the results (3.12) and (3.14), see (O'Hagen, 1978). Note that the described paradigm resembles a parametric approach where the goal is to recover the generating model in contrast to e.g. the structural risk minimization based algorithms where one merely tries to predict with minimal risk (see also Subsection 1.1.2). Let $D^{(m)} \in \mathbb{R}^{N+1\times N+1}$ be the squared linear mth order differential operator. If $\Sigma = {D^{(m)}}^T D^{(m)}$, the derivation is equivalent to the (primal) cost-function at the basis of LS-SVMs for regression (see Section 3.3) and the cost-function (5.11) of smoothing splines.

A major advantage of the Gaussian process formulation is the ability of doing inference of uncertainties of the model (Wahba, 1990) and to optimize the model's hyper-parameters. The latter leads to the hierarchical evidence framework as introduced in (MacKay, 1992) and elaborated in the case of LS-SVMs in (Van Gestel *et al.*, 2002; Suykens *et al.*, 2002*b*). A thorough empirical assessment of the performance of Gaussian processes may be found in (Rasmussen, 1996) and of a Bayesian techniques applied on LS-SVMs in (Van Gestel *et al.*, 2002).

5.3 Kriging Methods

Spatial statistics is concerned with the analysis of observations scattered over the (geographical) space (Cressie, 1993). Recent advances cast the problem as a generalization to the Wiener-Kolmogorov theory of prediction in time-series (Wiener, 1949) and provide a flexible framework for smoothing and interpolation of spatial surfaces. Let again $\mathbb{X} \subset \mathbb{R}^D$ denote a spatial index set and $\mathbf{Z}_{\mathbb{X}}$ be a Gaussian process over this set. For notational convenience, let $\mathbf{Z}(x)$ denote the random variable $\mathbf{Z}_{\mathbb{X}}$ given the fact that $\mathbb{X} = x$. The random variable $Z(x)$ has a mean function $m : \mathbb{R}^D \rightarrow \mathbb{R}$ and covariance function $\rho : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ such that one can write

$$\begin{cases} E[\mathbf{Z}(x)] = m(x) \\ \operatorname{cov}(\mathbf{Z}(x_i)\mathbf{Z}(x_j)) = E\left[(\mathbf{Z}(x_i) - m(x))^T\,(\mathbf{Z}(x_j) - m(x))\right] = \rho(x_i - x_j). \end{cases} \tag{5.19}$$

Let the mean function $m(x)$ be parameterized linearly as $m(x) = \sum_{d=1}^{D} \beta_d \phi_d(x_i)$. Let $Z = (z_1, \ldots, z_N)^T \in \mathbb{R}^N$ contain the observed samples at the spatial points $\{x_i\}_{i=1}^N$. Let $X \in \mathbb{R}^{N\times D}$ be a matrix with idth entry $X_{id} = \phi_d(x_i)$ for all $i = 1, \ldots, N$ and $d = 1, \ldots, D$, and let $a \in \mathbb{R}^N$ be a vector of unknowns. Then the minimum mean square error unbiased predictor $\hat{Z}(x_*)$ is given as

$$\begin{cases} L^{-1}X\beta = L^{-1}Z \\ Ka = (Z - X\beta) \\ \hat{Z}(x_*) = k(x_*)^T a + x_*^T \beta, \end{cases} \tag{5.20}$$

where $K \in \mathbb{R}^{N \times N}$ is the covariance matrix with ijth entry $K_{ij} = \rho(x_i, x_j)$ and $k : \mathbb{R}^D \rightarrow \mathbb{R}^N$ a function such that $k(x_*) = (\rho(x_1, x_*), \ldots, \rho(x_N, x_*))^T$. Let L then be the Cholesky decomposition (Golub and van Loan, 1989) of the matrix K. This is a numerically reliable form (Ripley, 1988) of universal Kriging (Cressie, 1993). The variance of the estimate is given as follows

$$\begin{cases} \operatorname{var}\big(Z(x_*) - \hat{Z}(x_*)\big) = \rho(x_*, x_*) - \|e\|_2^2 + \|g\|_2^2 \\ Le = k(x_*) \\ L^{-1}Xg = X\beta - (L^{-1}X)^T e, \end{cases} \tag{5.21}$$

where $g, e \in \mathbb{R}^N$ are vectors (Ripley, 1988).

Remark 5.1. We emphasize the close relationship with the derivation of the semi-parametric LS-SVM formulation (see Section 4.1). The main difference is the interpretation where in the case of Kriging the kernel plays the role of the covariance of the stochastic terms while in the case of SVMs and LS-SVMs, the kernel are deterministic in nature. As such, Kriging methods are more related in nature to Gaussian processes (see Section 5.2).

5.4 And also

5.4.1 Wavelets

Wavelets are a family of orthogonal bases that can effectively compress signals with possible irregularities. Although wavelets constitute a large body of literature mainly situated in function approximation problems (Daubechies, 1988), the main ideas can also be recovered in a smoothing context as eg. (Donoho and Johnstone, 1994). An approach is sketched based on (Daubechies, 1992) and elaborated e.g. in (Yu *et al.*, 1998). What makes the wavelet expansion unlike the Fourier transform or RBF based expansion is that the wavelet functions (mother functions) are (i) localized in frequency *and* space (compactly supported), (ii) will allow for varying resolution parameters (iii) will favor sparse expansions and (iv) are orthonormal. Again the method is typically applied to functions with respect to the time-index, but do not impose a causal ordering and the extension to one-dimensional spatial indices is straightforward. For a thorough elaboration of the subject and its extensions to multivariate cases we refer the reader to (Daubechies, 1992).

The analysis starts from an appropriate definition of a so-called mother-function $\delta : \mathbb{R} \rightarrow \mathbb{R}$ which is localized in space as well as in frequency such that $\exists L$ such that

$\delta(x) = 0$ if $|x| > L$ and $\exists L_\xi$ such that $\mathscr{F}\delta(\xi) \downarrow 0$ if $|\xi| > L_\xi$. Different classical results as the Paley-Wiener theorem (Daubechies, 1992) state that functions cannot be both band- (finite support of $\mathscr{F}f$) and time-limited (finite support of f) at the same time. Much of the literature on wavelets is then concerned with the derivation and analysis of an appropriate basis making an optimal trade-off between band- and time-limiting. Consider then the dilated (by $a \in \mathbb{R}$) and translated (by a vector $b \in \mathbb{R}$) basis function.

$$\delta_{ab}(x) = \sqrt{a}\delta\left(\frac{ax-b}{a}\right). \tag{5.22}$$

A set of mathematical operations were proposed (Daubechies, 1992) to infer an orthonormal set of basis functions $\{\rho_{ab} : \mathbb{R} \to \mathbb{R}\}_{a,b}$ from the father δ_{ab}. In this case, one also refers to the method as multi-resolution analysis (Daubechies, 1992). Traditional choices for the mother functions ρ_{ab} with dilation a and translation b are (i) the Haar functions (Haar, 1910) (emphasizing localizations in space) and (ii) symmlets (Daubechies, 1992) emphasizing the band-limiting property. Let x be sampled equidistantly in the interval $[0,1]$, then the mother function and the scaled basis functions become respectively

$$\begin{cases} \rho^{\text{haar}}(x) = I_{[0,1]}(x)\left(-I(x < 0.5)\right) \\ \rho^{\text{haar}}_{ab}(x) = 2^{-0.5a}\rho^{\text{haar}}(2^{-a}x - b), \end{cases} \tag{5.23}$$

See also Figure 5.1.a. The relationship of this method with the discussed primal-dual kernel machines is illustrated in the following example.

Example 5.1 **[Learning Machine based on Wavelet Decomposition]** Consider the function space based on the orthonormal Haar wavelet bases:

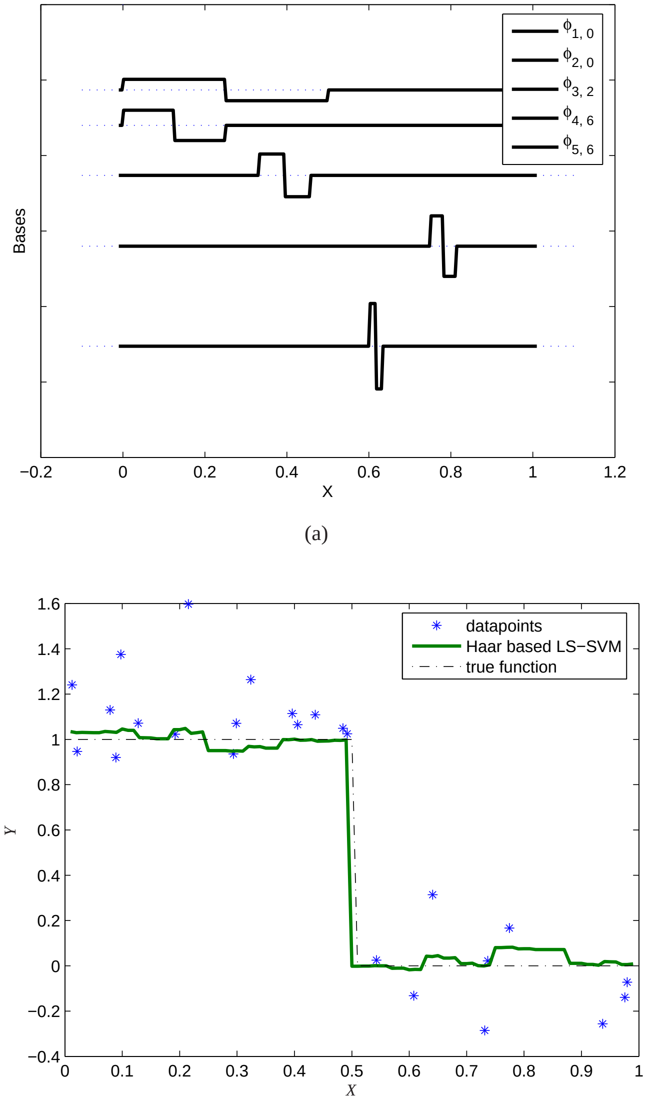
$$\mathscr{F}_S = \left\{ f : \mathbb{R} \to \mathbb{R} \;\middle|\; f(x) = \sum_{a=0}^{S}\sum_{k=0}^{S-1} w_{a,k}\rho^{\text{haar}}_{a,k2^{-a}}(x) \right\}, \tag{5.24}$$

where w contains the coefficients of the function for the different dilations s and translation $k2^{-s}$. A parametric approach as described in Lemma 6.1, is traditionally employed for the construction of the approximation.

The mechanism of primal-dual kernel machines comes into play e.g. when infinite bases expansions are considered or when one considers more complex regularization schemes which can be written as $w^T G^{-1} w$ as elaborated in Theorem 9.1. Consider the first case. The kernel corresponding with the infinite basis expansion becomes

$$K(x_i, x_j) = \sum_{a=0}^{\infty}\sum_{k=0}^{S-1} \rho^{\text{haar}}_{a,k2^{-a}}(x_i)^T \rho^{\text{haar}}_{a,k2^{-a}}(x_j) \tag{5.25}$$

which can be simplified considerably by exploiting the localized structure of the basis functions. An illustrative example was devised. Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$ contain $N = 25$ univariate input samples randomly chosen in the interval $[0,1]$. Let y_i then satisfy $y_i = I(x_i < 0.5) + e_i$ with e_i i.i.d. sampled from $\mathscr{N}(0, 0.2)$. The fit of the LS-SVM regressor on this dataset employing the kernel (5.25) is displayed in figure 5.1.b, clearly showing the



(a)

(b)

Figure 5.1: *Illustration of the Haar wavelet bases.* **(a)** *A sample of the set of Haar wavelet bases for the scales respectively* $0,\dots,4$ *and different translations.* **(b)** *An example of the fitted (solid line) indicator function (dashed-dotted line) sampled by* $N = 25$ *noisy observations (dots) using an LS-SVM regressor employing a kernel based on the infinite Haar basis expansion.*

ability to recover the discontinuity in the data. A disadvantage of the use of this specific wavelet kernel is that the solution is non-smooth in other locations.

The issue of wavelet kernels in smoothing tasks is discussed in more detail in (Amato *et al.*, 2004) and the abilities to recover discontinuities using wavelets expansions is reported in (Antoniadis and Gijbels, 2002). An alternative approach which do avoid the mentioned disadvantage is elaborated in Example 9.3. This example shows the potential path towards integration of wavelet based methods and the primal-dual kernel based methodology as described in the present work.

5.4.2 Inverse problems

Most linear inverse problems can be formulated as follows: let f and g be elements of a function (Hilbert) space(s) $\mathscr{F}$ and $\mathscr{G}$. Given a linear operator $L : \mathscr{F} \rightarrow \mathscr{G}$. Consider the equation $g = Lf$. The forward problem then amounts to solving for g given f. The inverse problem amounts to solving the equation for f given g. Consider as a typical example the integral operator which amounts to the problem

$$g(x) = \int_a^b K(x,y) f(y) dy, \tag{5.26}$$

referred to as the Fredholm equation of the first kind, see e.g. (Press *et al.*, 1988) for an introduction. Inverse and ill-posed problems are very important in several domains of applied science such as medical diagnosis, problems in vision, atmospheric remote sensing etc., see e.g. (Bertero *et al.*, 1988). The relevance of these problems has stimulated the development of theoretical and practical methods for determining approximative and numericallt reliable solutions (Hansen, 1998).

Fredholm equations of the first kind are often extremely ill-conditioned as may be understood as follows. Convolving the function f using the function K amounts in general to a smoothing operation which actually looses information. As such there is no direct way to recover all information by an inverse operation and one needs additional (external) knowledge on the solution in order to get a unique solution to the inverse problem (Press *et al.*, 1988). This concept is often referred to as regularization or capacity control and is treated extensively in the following Part, see e.g. (Backus and Gilbert, 1970; Tikhonov and Arsenin, 1977; Morozov, 1984; Neumaier, 1998).

5.4.3 Generalized least squares

As already noted in Section 4.1, a direct correspondence between the modeling of the parameters in a semi-parametric LS-SVM regressor and the classical Generalized Least Squares estimator (Mardia *et al.*, 1979) can be observed. The GLS estimator is well-described in statistical literature (e.g. see e.g. (Wetherill, 1986) and references). The estimator e.g. possesses the important BLUE (Best Linear Unbiased Estimator) property and appropriate efficient statistical tests were designed (Sen and Srivastava, 1990).

Part II

γ

Chapter 6

Regularization Schemes

Model complexity and regularization amounts to the artificial shrinkage of the solution-space in order to obtain increased generalization. The purpose of this chapter is both to motivate, to analyze and to discuss different regularization schemes in the process of model estimation. Section 6.1 surveys results in the context of linear parametric models. Section 6.2 gives results on the bias-variance trade-off for regression using LS-SVMs. Section 6.3 extends the well-known Tikhonov regularization scheme in primal-dual kernel machines to various other classical schemes. The measure of maximal variation for componentwise models is introduced in Section 6.4 and various applications of this idea are presented.

6.1 Regularized Parametric Linear Regression

Consider the class of linear models

$$\mathscr{F}_\omega = \left\{ f_\omega(x) = \omega^T x \;\middle|\; \omega \in \mathbb{R}^D, b \in \mathbb{R} \right\}. \tag{6.1}$$

Let the dataset $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$ satisfy $y_i = \omega^T x_i + b + e_i$ where $\{e_i\}_{i=1}^N$ is a sequence of uncorrelated i.i.d. samples with zero mean and bounded variance $E[e_i^2] = \sigma_e^2 < \infty$. For notational convenience, we do not include an intercept term in the derivations but assume a proper normalization of the data.

This section elaborates the discussion in Section 3.2.

6.1.1 Ridge regression

The use of an 2-norm based regularization scheme results in a mechanism which is convenient to analyze and to apply. Given the model class (6.1), ridge regression (Hoerl *et al.*, 1975) amounts to minimizing the following regularized cost function

$$\hat{w} = \arg\min_{w} \mathscr{J}_{\gamma}(w,b) = \frac{\gamma}{2}\|w\|_2^2 + \frac{1}{2}\sum_{i=1}^{N}\left(w^T x_i - y_i\right)^2 . \tag{6.2}$$

The modified normal equations become

$$\left(X^T X + \gamma I_D\right) w = X^T Y, \tag{6.3}$$

following from the first order conditions for optimality. This is seen as an application of the Tikhonov regularization scheme for function approximation (Tikhonov and Arsenin, 1977; Hansen, 1998).

6.1.2 LASSO

While Tikhonov regularization schemes based on $\|w\|_2^2$ are commonly used in order to improve estimates (statistically as well as numerically), interest in L_1-based regularization schemes has emerged recently as seen in the formulation and study of LASSO (Least Absolute Shrinkage and Selection Operator) estimators (Tibshirani, 1996), SURE (Stein Unbiased Risk Estimator) (Donoho and Johnstone, 1994) and basis pursuit (Friedmann and Stuetzle, 1981; Chen *et al.*, 2001) algorithms. Here one typically considers estimators of the form

$$\hat{w} = \arg\min_{w} \mathscr{J}_{\alpha}(w) = \sum_{i=1}^{N}\left(w^T x_i - y_i\right)^2 \quad \text{s.t.} \quad \|w\|_1 \leq \alpha, \tag{6.4}$$

where $\alpha \in \mathbb{R}^+$ is a hyper-parameter. The primal-dual optimization framework may be used to derive properties on the estimator regarding the obtained sparseness and the variance of the estimate (Osborne *et al.*, 2000).

The optimization problem (6.2) and (6.4) simplify considerably when the inputs are orthonormal:

Lemma 6.1. [Orthonormal Inputs, (Tibshirani, 1996)] *If the input matrix $X \in \mathbb{R}^{N\times D}$ is such that $X^T X = I_D$, the solutions to the ridge regression estimate (6.2) and the LASSO estimator (6.4) can be written as*

$$\begin{cases} \hat{w}_d^{\mathrm{rr}} = \frac{X_d^T Y}{1+\gamma} & \forall d = 1,\ldots,D \\ \hat{w}_d^{\mathrm{lasso}} = \operatorname{sign}(X_d^T Y)[X_d^T Y - \lambda]_+, & \forall d = 1,\ldots,D \end{cases} \tag{6.5}$$

respectively. Here λ is the Lagrange multiplier corresponding to the constraint $\|w\|_1 \leq \alpha$.

This result was extended towards more general regularization cost-functions as the hard- and soft- thresholding rule in (Donoho and Johnstone, 1994). A similar argument was used to compute efficiently the solution path of the LASSO estimator and the SVM classifier over all constants $\alpha > 0$ as e.g. in (Hastie *et al.*, 2004).

6.1.3 Least squares amongst alternatives

A stronger formulation regarding sparseness is considered. Given a set of observed input/output data-samples $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}$. Let one be interested in the linear model ($D > 1$) with minimal empirical risk only using one single input variable. This problem can be written as follows

$$\hat{w} = \arg\min_w \mathscr{J}_s(w) = \sum_{i=1}^N (w^T x_i - y_i)^2 \quad \text{s.t.} \quad w_g w_d = 0, \ \forall g \neq d, \tag{6.6}$$

from which it follows that at most one element of the parameter vector may be nonzero. The following result leads to a practical approach to this problem.

Lemma 6.2. [Embedding Least Squares amongst Alternatives] *The task of estimating the optimal predictor based on a single variable amongst given alternatives is considered. Formally, one searches the optimal model parameters w such that*

$$w_i \, w_j = 0, \ \forall i, j = 1, \ldots, D, \ i \neq j. \tag{6.7}$$

This quadratical constraints can be embedded in a least squares estimator as follows

$$(\hat{w}, \hat{t}) = \arg\min_{w,t} J(w) = \frac{1}{2} \left\| w^T x_i - y_i \right\|_2^2 \quad s.t. \quad \begin{cases} t^T \, 1_{D\times D} \, t \leq w^T w \\ -t_i \leq w_i \leq t_i \qquad \forall i = 1, \ldots, D, \end{cases} \tag{6.8}$$

where $1_{D\times D} \in \mathbb{R}^{D\times D}$ contains all ones.

Proof. Let $X = (x_1, \ldots, x_N)^T \in \mathbb{R}^{N\times D}$ and $Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$ be vectors. The Lagrangian of the constrained optimization problem (6.8) becomes

$$\begin{aligned} \mathscr{L}(w, t; \lambda, \alpha^+, \alpha^-) = {} & \frac{1}{2} \|Xw - Y\|_2^2 \\ & + \sum_{i=1}^D \alpha_i^- (-t_i - w_i) + \sum_{i=1}^D \alpha_i^+ (-t_i + w_i) + \frac{\lambda}{2} \left(t^T \, 1_{D\times D} \, t - w^T w \right), \end{aligned} \tag{6.9}$$

where $\alpha^+, \alpha^- \in \mathbb{R}^{+,D}$ and $\lambda \in \mathbb{R}^+$ are positive multipliers. Let $1_D \in \mathbb{R}^D$ denote the vector containing ones. The first order (necessary) conditions for optimality are given by the Karush-Kuhn-Tucker conditions (KKT), see e.g. (Boyd and Vandenberghe, 2004):

$$\begin{cases} \left(X^T X - \lambda I_D\right) w - X^T Y = \alpha^- - \alpha^+ & (a) \\ \alpha_i^- + \alpha_i^+ = \lambda 1_D^T \, t \qquad \forall i = 1, \ldots, D & (b) \\ -t_i \leq w_i \leq t_i \qquad \forall i = 1, \ldots, D & (c) \\ \alpha_i^+, \alpha_i^- \geq 0 \qquad \forall i = 1, \ldots, D & (d) \\ \alpha_i^- (t_i + w_i) = 0 \qquad \forall i = 1, \ldots, D & (e) \\ \alpha_i^+ (t_i - w_i) = 0 \qquad \forall i = 1, \ldots, D & (f) \\ \\ t^T \, 1_{D\times D} \, t \leq w^T w & (g) \\ \lambda \geq 0, \ \lambda (t^T \, 1_{D\times D} \, t - w^T w) = 0, & (h) \end{cases} \tag{6.10}$$

where the equalities (6.10.efh) are referred to as the complementary slackness constraints. By combining conditions (6.10.ef) and (6.10.b), it follows that $t_i = |w_i|$ for all $i = 1, \ldots, D$. From condition (6.10.g) it then follows that

$$t^T\ 1_{D\times D}\ t \leq w^T w = t^T t \ \ \Rightarrow \ \ t^T\left(1_{D\times D}\ - I_D\right) t \leq 0. \tag{6.11}$$

As the vector t and the matrix $\left(1_{D\times D}\ - I_D\right)$ contains all positive numbers, only $t^T\left(1_{D\times D}\ - I_D\right) t = 0$ is to be considered. As such, conditions (6.7) are satisfied in (any) optimum to (6.8). This concludes the proof. □

This task is elaborated in some detail here as it is closely related to the formulation and handling of positive OR-constraints (see Subsection 2.4.3) which play often an important role in hierarchical programming problems (see next Chapter).

The relationship with the least squares estimator when the relevant variable were known beforehand is given in the following lemma.

Lemma 6.3. [Relation to Univariate Least Squares] *Assume a γ^* exist such that $(X^T X - \lambda^* I_D) \succeq 0$ and that the constraint $\left(t^T\ 1_{D\times D}\ t\right) \leq w^T w$ is satisfied, then the prediction corresponds with the least squares predictor based on the variable with nonzero parameter only.*

Proof. Assume the single variate predictor uses finally one variable denoted as $X_{(1)} \in \mathbb{R}^N$ for prediction. Let then $X_{(0)} \in \mathbb{R}^{N\times(D-1)}$ be a vector denoting all other candidate variables. Condition (6.10.a) can then be rewritten as

$$\begin{bmatrix} (X_{(1)}^T X_{(1)} - \lambda) & X_{(1)}^T X_{(0)} \\ X_{(0)}^T X_{(1)} & (X_{(0)}^T X_{(0)} - \lambda I_{D'\times D'}) \end{bmatrix} \begin{bmatrix} w_{(1)} \\ w_{(0)} \end{bmatrix} = \begin{bmatrix} X_{(1)}^T Y \\ X_{(0)}^T Y \end{bmatrix} + \begin{bmatrix} \alpha_{(1)}^- - \alpha_{(1)}^+ \\ \alpha_{(0)}^- - \alpha_{(0)}^+ \end{bmatrix}, \tag{6.12}$$

where the parameters $w_{(1)} \in \mathbb{R}$ and $w_{(0)} \in \mathbb{R}^{D-1}$ correspond to $X_{(1)}$ and $X_{(0)}$ respectively. In the case the parameters $w_{(0)}$ are zero and $w_{(1)}$ is nonzero, the following property holds

$$\left(X_{(1)}^T X_{(1)}\right) w_{(1)} = X_{(1)}^T Y, \tag{6.13}$$

as $\alpha_{(1)}^+ - \alpha_{(1)}^- = \lambda w_{(1)}$ from application of (6.10.bef) and the property that $|w_{(1)}| = 1^T t$ in the solution to (6.10). Then note that (6.13) corresponds with the normal equations of the least squares problem $\min_w \left\|X_{(1)} w_{(1)} - Y\right\|_2^2$. If also $w_{(1)}$ were zero and thus $1^T t = 0$, the Lemma also holds as $\alpha^- + \alpha^- = 0_D$. □

This result is strongly related to the derivations of oracle inequalities as in (Donoho and Johnstone, 1994; Antoniadis and Fan, 2001).

Remark 6.1. Note that this result leads to an alternative practical approach to the problem (6.6). One can as well compute the least squares minimizer based on every individual individual variable and then pick the variable obtaining the best performance. This approach however becomes infeasible when more sets of alternatives are considered. Consider e.g. the task of estimating a model based on

10 variables where each individual variable belongs to a disjunct set of 2 candidates. Then the described combinatorial method should compute $2^{10} = 1024$ candidate least squares regressions, while the problem (6.8) would give the result by solving one QP.

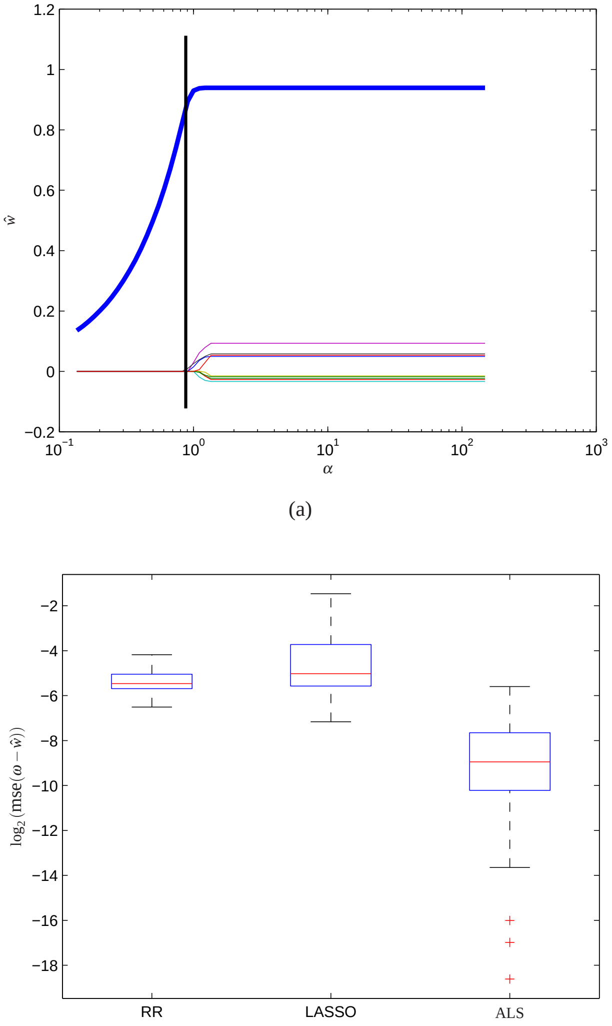
Sofar, we did not discuss the uniqueness of the solutions to (6.8) or (6.10) nor the choice of the Lagrange parameter λ satisfying (6.10.g). However, it turns out that the global optimum can be computed efficiently in many cases. In order to derive necessary conditions for uniqueness of local solutions to (6.8), consider the following modified formulation with fixed hyper-parameter $\gamma \in \mathbb{R}^+$

$$(\hat{w},\hat{t}) = \arg\min_{w,t} J_\gamma(w) = \frac{1}{2}\|Xw - Y\|_2^2 + \frac{\gamma}{2}\left(t^T\ 1_{D\times D}\, t - w^T w\right)$$
$$\text{s.t.} \quad -t_i \leq w_i \leq t_i \quad \forall i = 1,\ldots,D, \quad (6.14)$$

which is a convex problem as long as $(X^T X - \gamma I_D)$ is positive semi-definite (Boyd and Vandenberghe, 2004). The KKT conditions characterizing the global solution then corresponds to (6.10.a-f) with λ substituted by the given γ. Furthermore, if $\gamma \geq \lambda$ where λ solves (6.10) and $(X^T X - \lambda I_D)$ is positive semi-definite, it is easily seen that a solution to the original problem (6.8) follows uniquely as for increased values the cost of the term $t^T\ (1_{D\times D} - I_D)\ t \geq 0$ corresponding to γ is to be smaller than the cost corresponding to λ which is zero already. This results in a practical algorithmic approach to estimate the solution to the original problem (6.8) if it is unique.

Algorithm 6.1. [Least Squares amongst Alternatives] *Hereto, let σ^- denote the smallest eigenvalue of the sample covariance matrix $X^T X$. Then it is easily seen that $\gamma = \sigma^-$ is the largest value for which the problem (6.8) is convex. Furthermore, if the conditions $\hat{w}_i\hat{w}_j = 0$ of the solution vector $\hat{w}$ corresponding to $\gamma = \sigma^-$ are satisfied for all $i \neq j = 1,\ldots,N$ the problem is solved as if λ were found exactly. If not so, the problem (6.8) is not convex and one can use local optimization strategies to search the global solution.*

A Monte Carlo simulation study was conducted. In each iteration, a dataset $\mathscr{D} = \{(x_i,y_i)\}_{i=1}^N$ was generated with $N = 50$ and $D = 20$. The outputs were generated as $y_i = \omega_1 x_i^{(1)} + e_i$ with $e_i \sim \mathscr{N}(0,0.5)$ and ω_1 chosen in the interval $[-5,\ 5]$. The LASSO estimator was tuned using the validation performance on a disjunct part of the data, while the final performances of the estimate resulting from the tuned LASSO estimator and from the proposed method respectively were quantified as the mean squared error between the estimate and the true parameter vector $\omega = (\omega_1, 0\ldots,0)^T \in \mathbb{R}^{20}$. Figure 6.1.a shows the evolution diagram of the LASSO estimator in a single iteration step by ranging the hyper-parameter α from which the structure of the true parameter vector may be recovered. Panel 6.1.b then reports the results of the Monte Carlo study with 1000 iterations comparing the tuned Ridge Regression (RR) estimator, the tuned LASSO estimator and the proposed Alternative Least Squares method (ALS) of Algorithm 6.1. In addition to this results, the proposed relaxation succeeded in recovering the structure of the true parameter in 97.34% of the iterations, while the LASSO recovered on the average 35.23% of the underlying structure. This example shows the benefits of the proposed formulation in this specific case.



(a)

(b)

Figure 6.1: *Illustration of the Alternative Least Squares (ALS) as in Algorithm 6.1. Panel* **(a)** *shows the evolution diagram of the LASSO estimator in a single iteration step by ranging the hyper-parameter* α *from which the structure of the true parameter vector may be recovered. Panel* **(b)** *reports the results of the Monte Carlo study with* 1000 *iterations comparing the tuned Ridge Regression (RR) estimator, the tuned LASSO estimator and the proposed ALS method. In addition to this results, the proposed relaxation succeeded in recovering the structure of the true parameter in* 97.34% *of the iterations, while the LASSO recovered on the average* 35.23% *of the underlying structure.*

6.1.4 Bridge regression

The use of other norms for the regularization term from Ridge Regression to general Minkowski norms has been discussed under the name of bridge regression (Frank and Friedman, 1993; Fu, 1998; Antoniadis and Fan, 2001). The general Minkowski norm is defined as

$$\|w\|_p = \left(\sum_{d=1}^{N} w_d^p \right), \tag{6.15}$$

which is convex (satisfying the triangular inequality) whenever $p \geq 1$. The bridge regression estimator then becomes

$$(\hat{w}, \hat{b}) = \arg\min_{w,b} \mathscr{J}_{\psi}^{p}(w,b) = \|w\|_p + \frac{\psi}{2} \sum_{i=1}^{N} \left(w^T x_i + b - y_i \right)^2, \tag{6.16}$$

which is a convex problem whenever $p \geq 1$. It is mostly solved using an iteratively re-weighted algorithm where one uses the following reformulation

$$(\hat{w}, \hat{b}; g) = \arg\min_{w,b} \mathscr{J}_{\psi}^{g}(w,b) = \sum_{d=1}^{D} g_w^d \, w_d^2 + \frac{\psi}{2} \sum_{i=1}^{N} \left(w^T x_i + b - y_i \right)^2$$
$$\text{s.t.} \quad g_w^d \, w_d^2 = w_d^p, \; \forall d = 1, \ldots, D, \tag{6.17}$$

which is solved for w and b. The hyper-parameters $g_w = (g_w^1, \ldots, g_w^D)^T \in \mathbb{R}^N \in \mathbb{R}^D$ are consequently adjusted correspondingly, see e.g. (Fu, 1998). This procedure corresponds with a particular instance of the Gauss-Seidel algorithm, see e.g. (Hastie and Tibshirani, 1990). The use of p-norms different than L_2 or L_1 may be usefull in problems involving higher dimensional data, see e.g. (Frank and Friedman, 1993).

6.1.5 Shrinkage estimators for parametric large margin classifiers

Similar estimators were introduced recently in order to automatically select features in parametric large margin classifiers (Weston *et al.*, 2003; Bhattacharya, 2004). The following estimator was proposed.

$$(\hat{w}, \hat{b}) = \arg\min_{w,b} \mathscr{J}_C(w,b) = \|w\|_1 + C \sum_{i=1}^{N} \left[1 - y_i (w^T x_i + b) \right]_+, \tag{6.18}$$

where $C \in \mathbb{R}^+$ acts as a hyper-parameter.

6.2 The Bias-Variance Trade-off

A classical tool to analyze the generalization performance in the form of the total Mean Squared Error (MSE) of the estimate with respect to the true model was found in the bias-variance trade-off (Hoerl *et al.*, 1975; Hastie *et al.*, 2001). Recently, this analysis was introduced for the SVM classifier (Valentini and Dietterich, 2004). The discussion is extended to the LS-SVM regressor as follows.

Let the observed data $\mathscr{D}$ satisfy the relation $y_i = f^*(x_i) + e_i$ where $f^* : \mathbb{R}^d \to \mathbb{R}$ is a smooth function and the errors $\{e_i\}_{i=1}^N$ satisfy the Gauss-Markov conditions described in Definition 3.1. The vector $Y^* = (f^*(x_1), \ldots, f^*(x_N))^T \in \mathbb{R}^N$ denotes the true function $f^* : \mathbb{R}^D \to \mathbb{R}$ evaluated in the training points which is typically unknown in practice. Let $\hat{Y} = (\hat{f}(x_1), \ldots, \hat{f}(x_N))^T \in \mathbb{R}^N$ denote the estimator $\hat{f}$ resulting from the LS-SVM estimate $\hat{f}$ evaluated on the training data. The total MSE can be decomposed as

$$\mathrm{MSE}(\hat{Y}, Y^*) = E\left[\hat{Y} - Y^*\right]^2 = E\left[\hat{Y} - E[\hat{Y}]\right]^2 + \left[E[\hat{Y}] - Y^*\right]^2,$$

where the two last terms are denoted as the variance and the bias respectively. The bias, covariance and the total mean squared error are then derived for the LS-SVM smoother similar to the derivation in (Hoerl and Kennard, 1970; Hoerl *et al.*, 1975).

Let $E[\hat{Y}]$ denote the expected predicted smoothed data given the used model definition using any realization of the noise terms $\{e_i\}_{i=1}^N$ in the data. The bias can then be written as

$$\begin{aligned}
\mathrm{Bias}(\hat{Y}, Y^*) = Y^* - E[\hat{Y}] &= Y^* - \Omega[\Omega + I_N \gamma^{-1}]^{-1} E[Y] \\
&= Y^* - \Omega[\Omega + I_N \gamma^{-1}]^{-1} Y^* \\
&= Y^* - [\Omega + I_N \gamma^{-1} - I_N \gamma^{-1}][\Omega + I_N \gamma^{-1}]^{-1} Y^* \\
&= Y^* - Y^* + \gamma^{-1}[\Omega + I_N \gamma^{-1}]^{-1} Y^* \\
&= \gamma^{-1}[\Omega + I_N \gamma^{-1}]^{-1} Y^*. \qquad (6.19)
\end{aligned}$$

Let the singular value decomposition of $\Omega \in \mathbb{R}^{N \times N}$ be denoted as $\Omega = U^T S U$ where $U^T U = I_N$ and $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_N) \in \mathbb{R}^{N \times N}$ denote the eigenvalues of Ω.

The trace of the squared bias becomes

$$\begin{aligned}
\mathrm{tr}[\mathrm{Bias}(\hat{Y}, Y^*)\,\mathrm{Bias}(\hat{Y}, Y^*)^T] &= \gamma^{-2} \mathrm{tr}\left[(\Omega + I_N \gamma^{-1})^{-1} Y^* {Y^*}^T (\Omega + I_N \gamma^{-1})^{-1}\right] \\
&= \gamma^{-2} {Y^*}^T (\Omega + I_N \gamma^{-1})^{-2} Y^* \\
&= \gamma^{-2} \sum_{i=1}^N \frac{p_i^2}{(\sigma_i + \gamma^{-1})^2}, \qquad (6.20)
\end{aligned}$$

where $p_i = {Y^*}^T U_i$ and $U_i \in \mathbb{R}^N$ denotes the ith column of U. The covariance of the estimate can be written as follows

$$\mathrm{Cov}(\hat{Y}, \hat{Y}) = E[\hat{Y} \hat{Y}^T] = \Omega(\Omega + I_N \gamma^{-1})^{-1} E[Y Y^T](\Omega + I_N \gamma^{-1})^{-T} \Omega^T. \qquad (6.21)$$

The total variance can be written then as follows

$$\begin{aligned}
\mathrm{tr}\left(\mathrm{Cov}(\hat{Y}, \hat{Y})\right) &= \sigma_e^2 \mathrm{tr}[\Omega(\Omega + I_N \gamma^{-1})^{-2} \Omega] \\
&= \sigma_e^2 \sum_{i=1}^N \frac{\sigma_i^2}{(\sigma_i + \gamma^{-1})^2}. \qquad (6.22)
\end{aligned}$$

The total mean squared error can be computed as

$$\begin{aligned}\mathrm{TMSE}(\hat{Y},Y^*) &= \mathrm{tr}\left[\mathrm{Cov}(\hat{Y},\hat{Y})\right] + \mathrm{tr}\left[\mathrm{Bias}(\hat{Y},Y^*)\,\mathrm{Bias}(\hat{Y},Y^*)^T\right] \\ &= \sigma_e^2 \sum_{i=1}^N \frac{\sigma_i^2}{(\sigma_i+\gamma^{-1})^2} + \gamma^{-2}\sum_{i=1}^N \frac{p_i^2}{(\sigma_i+\gamma^{-1})^2} \\ &= \sum_{i=1}^N \frac{\sigma_e^2\sigma_i^2+\gamma^{-2}p_i^2}{(\sigma_i+\gamma^{-1})^2}. \end{aligned} \tag{6.23}$$

From this expressions, it is possible to make the bias-variance trade-off explicit when the true function f^* or Y^* were known. The bias-variance decomposition for the LS-SVM smoother is illustrated in figure 6.2.

Lemma 6.4. [Optimality of Regularization in LS-SVMs] *Let the bias and variance be formulated as in (6.20) and (6.22). There exists a $\gamma < \infty$ (or $\gamma^{-1} > 0$) resulting in a lower TMSE with respect to $\gamma = \infty$.*

Proof. The proof follows from the following inequality

$$\frac{\partial\, \mathrm{tr}(\mathrm{bias}(\hat{Y},Y^*)\ \mathrm{bias}(\hat{Y},Y^*)^T)}{\partial\gamma^{-1}}\Big|_{\gamma^{-1}=0} < -\frac{\partial\, \mathrm{tr}(\mathrm{Cov}(\hat{Y},\hat{Y}))}{\partial\gamma^{-1}}\Big|_{\gamma^{-1}=0} \tag{6.24}$$

when $\gamma^{-1} = 0$. This result shows that there exists a nonzero amount of regularization leading to a minimal TMSE. □

In practice, regularization is more important for this nonlinear setting as for the linear parametric ridge regression case.

6.3 Tikhonov, Morozov and Ivanov Regularization

6.3.1 Regularization schemes

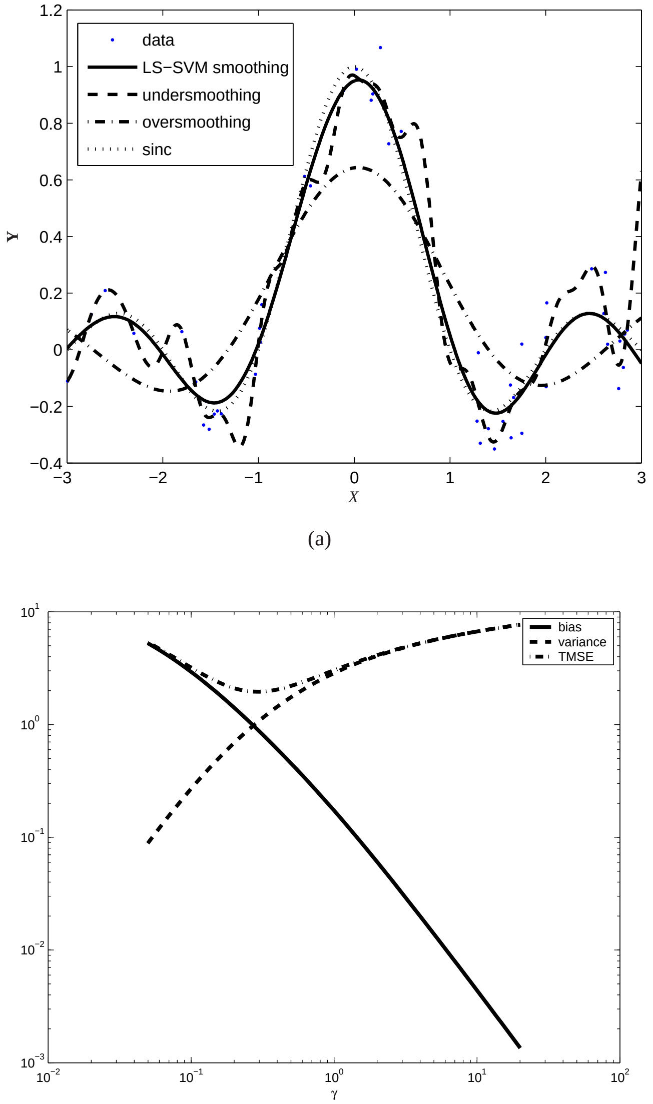
The Tikhonov scheme (Tikhonov and Arsenin, 1977), Morozov's discrepancy principle (Morozov, 1984) and Ivanov Regularization scheme (Ivanov, 1976) are discussed simultaneously to stress the correspondences and the differences. The cost functions are given respectively as

- Tikhonov, see Chapter 3 and Section 4.1:

$$\min_{w,e} \mathscr{J}_T(w,e) = \frac{1}{2}w^Tw + \frac{\gamma}{2}\sum_{i=1}^N e_i^2 \ \text{ s.t. } \ w^T\varphi(x_i)+e_i = y_i, \quad \forall i=1,\dots,N. \tag{6.25}$$

- Morozov's discrepancy principle (Morozov, 1984), where the minimal 2-norm of w realizing a fixed noise level σ_e^2 is to be found:

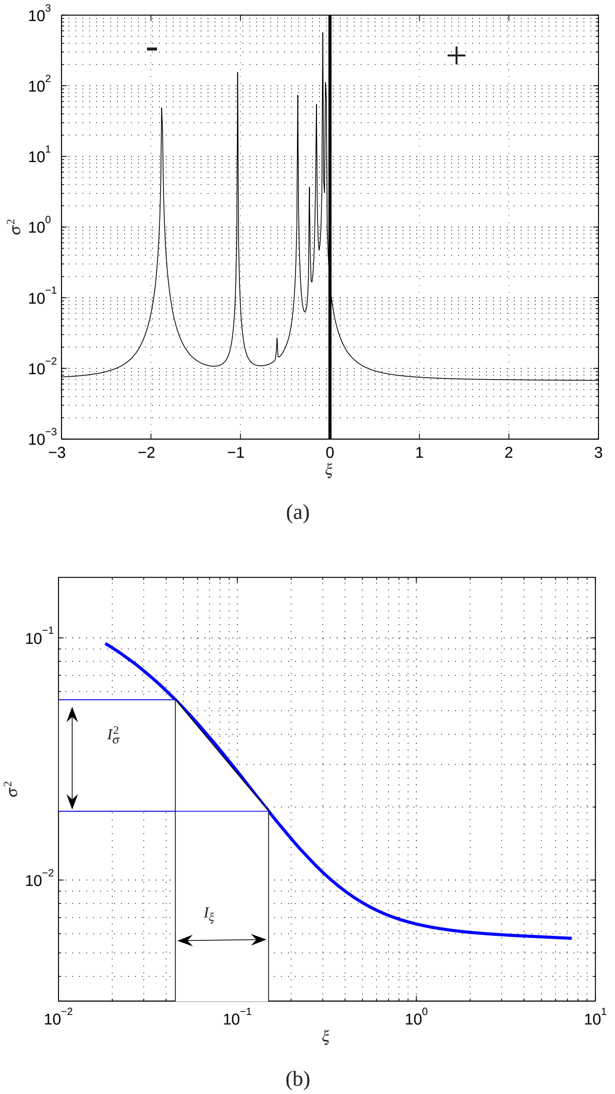
$$\min_{w,e} \mathscr{J}_M(w) = \frac{1}{2}w^Tw \ \text{ s.t. } \ \begin{cases} w^T\varphi(x_i)+e_i = y_i, & \forall i = 1,\dots N \\ \frac{1}{N}\sum_{i=1}^N e_i^2 = \sigma_e^2. \end{cases} \tag{6.26}$$



(a)

(b)

Figure 6.2: *Illustration of the bias-variance trade-off.* **(a)** *A dataset based on the relation $y_i = \mathrm{sinc}(x_i) + e_i$ with $e_i \sim \mathscr{N}(0, 0.1)$ was generated. Different values for γ in the applied LS-SVM smoother leads to over-smoothing (dashed-dotted line), undersmoothing (dashed line) and an optimal trade-off between bias and variance (solid line).* **(b)** *Theoretical values for the bias (solid line), the variance (dashed line) and the total MSE (dashed dotted line) of an LS-SVM smoother.*



(a)

(b)

Figure 6.3: *Illustration of a typical behavior of the Morozov secular equation (6.35.a).* **(a)** *If ξ is positive, the secular equation is monotonically decreasing. If ξ is negative, the function grows unbounded (poles) when $\xi = -1/(2\sigma_i)$.* **(b)** *As the secular equation is monotonically decreasing for $\xi > 0$, a positive interval I_{ξ} will be mapped uniquely to an interval I_{σ}^2.*

- Ivanov (Ivanov, 1976) regularization amounts at solving for the best fit with a 2-norm on w smaller than π^2:

$$\min_{w,e} \mathscr{J}_I(e) = \frac{1}{2}e^T e \text{ s.t. } \begin{cases} w^T\varphi(x_i) + e_i = y_i, & \forall i = 1,\ldots N \\ w^T w \leq \pi^2. \end{cases} \tag{6.27}$$

This formulation is also referred to as the trust-region subproblem (Rockafellar, 1993; Nocedal and Wright, 1999) employed in the context of optimization theory.

The Lagrangians become respectively

$$\begin{cases} \mathscr{L}_T(w,e;\alpha) & = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^N e_i^2 - \sum_{i=1}^N \alpha_i(w^T\varphi(x_i) + e_i - y_i) \\ \mathscr{L}_M(w,e;\alpha,\xi) & = \frac{1}{2}w^T w + \xi(\sum_{i=1}^N e_i^2 - N\sigma^2) - \sum_{i=1}^N \alpha_i(w^T\varphi(x_i) + e_i - y_i) \\ \mathscr{L}_I(w,e;\alpha,\xi) & = \frac{1}{2}e^T e + \xi(w^T w - \pi^2) - \sum_{i=1}^N \alpha_i(w^T\varphi(x_i) + e_i - y_i). \end{cases} \tag{6.28}$$

The conditions for optimality are

Condition	Tikhonov	Morozov	Ivanov
$\frac{\partial\mathscr{L}}{\partial w} = 0$	$w = \sum_{i=1}^N \alpha_i\varphi(x_i)$	$w = \sum_{i=1}^N \alpha_i\varphi(x_i)$	$w = \frac{1}{2\xi}\sum_{i=1}^N \alpha_i\varphi(x_i)$
$\frac{\partial\mathscr{L}}{\partial e_i} = 0$	$\gamma e_i = \alpha_i$	$2\xi e_i = \alpha_i$	$e_i = \alpha_i$
$\frac{\partial\mathscr{L}}{\partial \alpha_i} = 0$	$w^T\varphi(x_i) + e_i = y_i,$	$w^T\varphi(x_i) + e_i = y_i,$	$w^T\varphi(x_i) + e_i = y_i$
	–	$\sum_{i=1}^N e_i^2 = N\sigma^2$	$w^T w \leq \pi^2$
	–	$\xi \geq 0$	$\xi \geq 0$

(6.29)

for all $i = 1,\ldots,N$. After elimination of the parameter vector w, the Tikhonov conditions result in the following set of linear equations as classical, see Chapter 3,

$$\textbf{Tikhonov}: \quad \left(\Omega + \frac{1}{\gamma}I_N\right)\alpha = Y. \tag{6.30}$$

Re-organizing the sets of constraints of the Ivanov scheme results in the following sets of linear equations where an extra nonlinear constraint relates the Lagrange multiplier ξ with the hyper-parameter σ^2 as follows

$$\textbf{Morozov}: \quad \left(\Omega + \frac{1}{2\xi}I_N\right)\alpha = Y \text{ s.t. } \alpha^T\alpha \leq N\sigma^2, \ \xi \geq 0. \tag{6.31}$$

Similarly, the Morozov scheme has a dual problem which can be rewritten as follows. Let $\tilde{\alpha} = \frac{1}{2\xi}\alpha$, then

$$\textbf{Ivanov}: \quad \left(\Omega + \frac{1}{2\xi}I_N\right)\tilde{\alpha} = Y \text{ s.t. } \tilde{\alpha}^T\Omega\tilde{\alpha} \leq \pi^2, \ \xi \geq 0, \tag{6.32}$$

and the dual representation may be evaluated at a new point as $\hat{f}(x_*) = \Omega_N(x_*)^T \tilde{\alpha}$.

One now can rephrase the optimization problem (6.26) in terms of the Singular Value Decomposition (SVD) of Ω (Golub and van Loan, 1989). For notational convenience, the bias term b is omitted from the following derivations. The SVD of Ω is given as

$$\Omega = USU^T \quad \text{s.t.} \quad U^T U = I_N, \tag{6.33}$$

where $U \in \mathbb{R}^{N \times N}$ is orthonormal and $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$ with $\sigma_1 \geq \cdots \geq \sigma_N$. Using the orthonormality property of the SVD, the conditions (6.31) can be rewritten as

$$\begin{cases} \alpha = U\left(S + \frac{1}{2\xi} I_N\right)^{-1} p \quad \text{s.t.} \quad \frac{1}{4\xi^2}\alpha^T \alpha \leq N\sigma^2, \;\; \xi \geq 0 \\ \tilde{\alpha} = U\left(S + \frac{1}{2\xi} I_N\right)^{-1} p \quad \text{s.t.} \quad \tilde{\alpha}^T \Omega \tilde{\alpha} \leq \pi^2, \;\; \xi \geq 0 \end{cases} \tag{6.34}$$

where $p = U^T Y \in \mathbb{R}^N$. Eliminating of the dual variables $\alpha \in \mathbb{R}^N$ and $\tilde{\alpha} \in \mathbb{R}^N$ respectively leads to the equalities

$$\begin{cases} \frac{1}{4\xi^2}\alpha^T \alpha = \sum_{i=1}^N \left(\frac{p_i}{2\xi\sigma_i + 1}\right)^2 \leq N\sigma^2 & (a) \\ \sum_{i=1}^N \frac{\sigma_i p_i^2}{(\frac{1}{2\xi}\sigma_i + 1)^2} \leq \pi^2. & (b) \end{cases} \tag{6.35}$$

One refers to the equations in (6.35) as the secular equations (Golub and van Loan, 1989; Neumaier, 1998). Now the largest value of ξ (smallest fitting term) satisfying this relation can be searched using e.g. a bisection algorithm (Press *et al.*, 1988). As can be seen from the expressions (6.35) and Figure 6.3, the relation between $\sigma^2(\pi^2)$ and $\xi \geq 0$ is strictly monotone and there is exactly one ξ corresponding with a given noise level σ^2 (or π^2).

6.3.2 Differogram

In (Pelckmans *et al.*, 2003*a*; Pelckmans *et al.*, 2004*a*), a model free noise variance estimator denoted as a differogram method was elaborated. Appendix A gives details on this estimator and relates it to a series of other estimators. The following example shows a direct use of this method towards the estimation of the regularization trade-off.

Example 6.1 The Morozov regularization scheme (6.26) has various practical implications including the following. Given prior information or a reliable estimate of the noise level, one can transform this knowledge into an appropriate regularization parameter $\xi \geq 0$. Let $\sigma_e : \mathscr{D} \to \mathbb{R}$ be an estimator of the noise variance in the dataset $\mathscr{D}$ such that $\sigma_e(\mathscr{D}) = \hat{\sigma}_e^2$ with variance σ_v^2. Let $\alpha \in \mathbb{R}^+$ be a fixed constant determining the relative width of the interval. Given the interval $[\hat{\sigma}_e \pm \alpha\sigma_v^2]$, one may determine the corresponding interval of regularization terms as $I_\xi = [\hat{\xi}^-, \hat{\xi}^+]$ and one can marginalize over this region. See also Figure 6.3.b. Let $\hat{I}_\xi$ be a finite subset of I_ξ, then

$$\hat{f}(x_*) \quad = \quad \int_{\xi \in I_\xi} f_\xi(x_*) dP_\xi$$

$$\begin{aligned} &= \sum_{i=1}^{N} \int_{\xi \in I_\xi} \Omega_N^T(x_*)\hat{\alpha}_\xi \, dP\xi \\ &\approx \sum_{i=1}^{N} \sum_{\xi \in \hat{I}_\xi} \left(\Omega_N^T(x_*)\hat{\alpha}_\xi \right) p_\xi , \qquad (6.36) \end{aligned}$$

where f_ξ parameterized with $\hat{\alpha}_\xi$ solves the LS-SVM cost-function (3.9) corresponding with a regularization parameter $\gamma = 2\xi$ and $p_\xi > 0$ are weighting terms corresponding with the distribution on $\hat{I}_\xi$ such that $\int_{\xi \in \hat{I}_\xi} p_\xi d\xi = 1$. A similar result is also derived in Algorithm 8.1.

A distribution free approach towards the estimation of the noise variance without the explicit construction of a model was discussed in (Pelckmans *et al.*, 2004*a*) called the differogram. The key idea is to infer properties of the observed data on the cloud of mutual differences of the data-points defined as $\Delta_{x,ij} = \|x_i - x_j\|_2$ and $\Delta_{y,ij} = \|y_i - y_j\|_2$, instead of on the data itself. Figure 6.4.a illustrates the effect of the chosen noise level on the validation set of an artificial regression example. Figure 6.4.b shows the differogram cloud of the higher dimensional data of the Boston housing dataset and its resulting variance estimate. Section 9.4.3 discusses the differogram method in more detail in a slightly different context.

6.4 Regularization Based on Maximal Variation

6.4.1 Maximal variation

Consider again the setting as in Section 4.2 of componentwise models where a datapoint is reorganized as a set of P components such that $x = \left(x^{(1)}, \ldots, x^{(P)}\right)$. In (Pelckmans *et al.*, 2004, *In press*) the use of the following criterion is proposed:

Definition 6.1. [Maximal Variation] *Let $x_i^{(p)}$ be samples of the random variable $\mathbf{X}^{(p)} \in \mathbb{R}^{D_p}$ with a finite range such that $\exists L_x^p$ with $-L_x^p \leq \mathbf{X}^{(p)} \leq L_x^p$. The maximal variation of a function $f_p : \mathbb{R}^{D_p} \to \mathbb{R}$ is defined as*

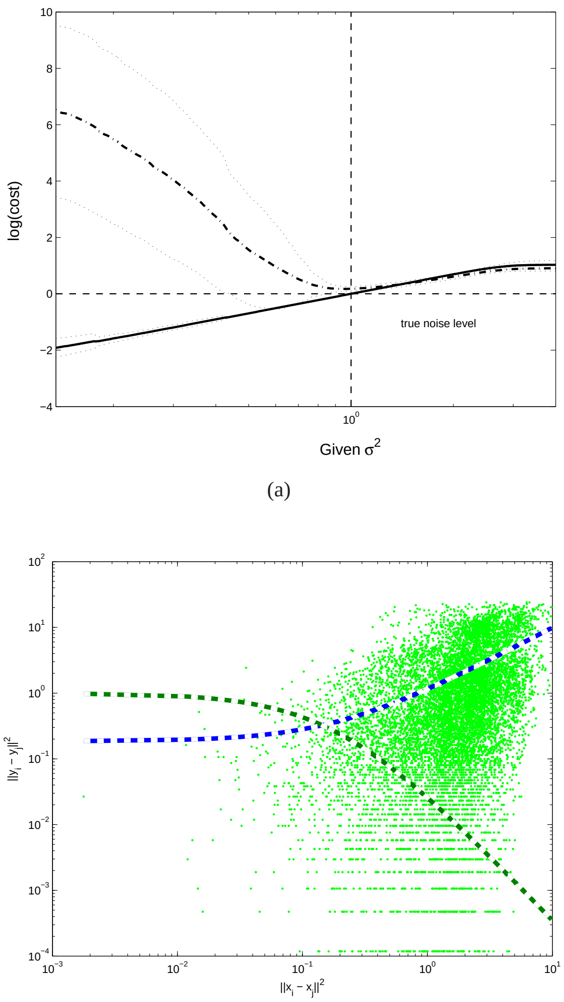
$$\mathscr{M}_p = \sup_{x^{(p)} \in \mathbb{R}^{D_p}} \left| f_p\left(x^{(p)}\right) \right| , \qquad (6.37)$$

for all $x^{(p)}$ sampled from the same distribution underlying the dataset $\mathscr{D}$. belonging to the domain of f_p. The empirical maximal variation can be defined as

$$\hat{\mathscr{M}}_p = \max_{x_i^{(p)} \in \mathscr{D}} \left| f_p\left(x_i^{(p)}\right) \right| , \qquad (6.38)$$

with x_i belonging to the training-set $\mathscr{D}$.

The setting of statistical learning theory may be employed to derive a bound on the deviation of the true maximal variation to the empirical maximal deviation, see also



(a)

(b)

Figure 6.4: *Example of the use of the Morozov discrepancy principle.* **(a)** *Training error (solid line) and validation error (dashed-dotted line) for the LS-SVM regressor with the Morozov scheme as a function of the noise level σ^2 (the dotted lines indicate error-bars by randomizing the experiment). The (dashed lines) denote the true noise level. One can see that imposing small noise levels results in overfitting.* **(b)** *Differogram cloud of the Boston Housing Dataset displaying all differences between two inputs ($\Delta_x = \|x_i - x_j\|_2$) and two corresponding outputs ($\Delta_y = \|y_i - y_j\|_2$). The location of the curve passing the Y-axis given as $E[\Delta_y|\Delta_x = 0]$ results in an estimate of the noise variance.*

example 3.4 in Section 3.5. A main advantage is that this measure is not directly expressed in terms of the parameter vector (which can be infinite dimensional in the case of kernel machines). Moreover, the regularization scheme becomes independent of the normalization and dimensionality of the individual components.

As an example, consider again the linear model (6.1). Furthermore, let $L \in \mathbb{R}$ such that $-L \leq x_d \leq L$ with $L = \max_i(|x_i^p|)$. The following relation holds,

$$|w_d|_1 = \frac{1}{L}|Lw_d|_1 = \frac{\mathscr{M}_d}{L}, \ \ \forall d = 1,\dots,D. \tag{6.39}$$

One then can rewrite (6.38) as follows

$$(\hat{w},\hat{b}) = \arg\min_{w,b} \mathscr{J}_\lambda^{\mathscr{M}}(w,b,\mathscr{M}) = \sum_{p=1}^{P} \mathscr{M}_p + \lambda \sum_{i=1}^{N}(w^T x_i + b - y_i)^2. \tag{6.40}$$

By replacing the maximal variations $\mathscr{M}_d$ by its empirical counterpart, it can be solved efficiently as

$$(\hat{w},\hat{b},\hat{t}) = \arg\min_{w,b,t} \mathscr{J}_\lambda^{\hat{\mathscr{M}}}(w,b,t) = \sum_{d=1}^{D} t_d + \lambda \sum_{i=1}^{N}(w^T x_i + b - y_i)^2$$
$$\text{s.t.} \quad -t_d \leq w_d x_i^d \leq t_d, \ \ \forall d = 1,\dots,D, \ \ \forall i = 1,\dots,N, \tag{6.41}$$

which can be casted as a quadratic programming problem with $2D+1$ unknowns and $2D$ inequalities.

Though this formulation corresponds to a large extents with the methods as LASSO and the SURE formulation, the extension to the kernel version and the way to cope with the missing values will crucially depend on this measure of maximal variation. As the measure of maximal variation depends only on the predicted outputs and not on the parameterized mapping, one may refer to the mechanism of maximal variation as non-parametric regularization principle.

6.4.2 Structure detection in kernel machines

This mechanism is extended towards the setting of primal-dual kernel machines. The formulation of componentwise LS-SVMs suggests the use of a dedicated regularization scheme which is often very useful in practice. In the case where the nonlinear function consists of a sum of components, one may ask oneself which components have no contribution ($f_p(\cdot) = 0$) for prediction. Sparseness amongst the components is often referred to as structure detection. The described method is closely related to the kernel ANOVA decomposition (Vapnik, 1998; Stitson *et al.*, 1999) and the structure detection method of (Gunn and Kandola, 2002). However, the following method as originally described in (Pelckmans *et al.*, 2004, *In press*; Pelckmans *et al.*, 2005*c*) starts from a clear optimality principle, and extends hence the LASSO estimator to a nonlinear kernel setting.

Lemma 6.5. [Primal-Dual Kernel Machine for Structure Detection] *Consider the class of models $\mathscr{F}_\varphi$, see (3.8). The following primal estimator is considered:*

$$
(\hat{w},\hat{b},\hat{t},\hat{e}) = \arg\min_{w,b,t,e} \mathscr{J}^{\hat{\mathscr{M}}}_{\mu,\lambda}(w,b,t,e) = \mu \sum_{p=1}^{P} t_p + \frac{1}{2}\sum_{p=1}^{P} w_p^T w_p + \frac{\lambda}{2}\sum_{i=1}^{N} e_i^2
$$

$$
\text{s.t.} \quad \begin{cases} \sum_{p=1}^{P} w_p^T \varphi_p\left(x_i^{(p)}\right) + b + e_i = y_i & \forall i = 1,\ldots,N \\ -t_p \le w_p \varphi\left(x_i^{(p)}\right) \le t_p & \forall i = 1,\ldots,N, \forall p = 1,\ldots,P. \end{cases} \tag{6.42}
$$

Let $\alpha = (\alpha_1,\ldots,\alpha_N)^T \in \mathbb{R}^N$, $\rho_p^+ = (\rho_{p,1}^+,\ldots,\rho_{p,N}^+)^T \in \mathbb{R}^{+,N}$ and $\rho_p^- = (\rho_{p,1}^-,\ldots,\rho_{p,N}^-)^T \in \mathbb{R}^{+,N}$ be the Lagrange multipliers associated with the corresponding constraints in (6.42). The dual problem is then given as

$$
(\hat{\alpha},\hat{\rho}_p^+,\hat{\rho}_p^-) = \arg\max_{\alpha,\rho_p^+,\rho_p^-} \mathscr{J}_\gamma(\alpha,\rho^+,\rho^-)
$$

$$
-\frac{1}{2}\left(\alpha + \sum_{p=1}^{P}(\rho_p^+ - \rho_p^-)\right)^T \Omega^P \left(\alpha + \sum_{p=1}^{P}(\rho_p^+ - \rho_p^-)\right) - \frac{1}{2\lambda}\alpha^T\alpha + Y^T\alpha
$$

$$
\text{s.t.} \quad \begin{cases} \mu = \sum_{i=1}^{N}(\rho_{ip}^+ + \rho_{ip}^-) & \forall p = 1,\ldots,P \\ \sum_{i=1}^{N}\alpha_i = 0 & \\ \rho_{ip}^+,\rho_{ip}^- \ge 0, & \forall i = 1,\ldots,N, \forall p = 1,\ldots,P \end{cases} \tag{6.43}
$$

where $\Omega_{ij}^P = \sum_{p=1}^{P} K_p\left(x_i^{(p)},x_j^{(p)}\right)$ for all $i,j = 1,\ldots,N$. The estimated predictor can then be evaluated on a new data point $x_* \in \mathbb{R}^D = \left(x_*^{(1)},\ldots,x_*^{(P)}\right)$ as follows

$$
\hat{f}(x_*) = \sum_{p=1}^{P}\sum_{i=1}^{N}\left(\hat{\alpha}_i + \hat{\rho}_{ip}^+ - \hat{\rho}_{ip}^-\right) K_p\left(x_i^{(p)},x_*^{(p)}\right) + \hat{b}, \tag{6.44}
$$

where $\hat{b}$ may be recovered from the complementary slackness conditions associated with the primal-dual derivation.

The proof follows the formulation of the primal-dual kernel machines as in Chapter 3. The main drawback of this approach is the huge number of Lagrange multipliers ($N(2P+1)$) which occur in the dual optimization problem. Note that this number can be reduced readily by only including those constraints of maximal variation belonging to different input values $x_i^{(p)} \ne x_j^{(p)}$. This is especially useful in the case a number of components consist of categorical or binary values. Subsection 8.4.1 describes a computational shortcut.

It is known that the use of 1-norms may lead to a sparse solution which is unnecessarily biased (Fan, 1997). To overcome this drawback, one has proposed the use of norms as the Smoothly Clipped Absolute Deviation (SCAD) penalty function as suggested

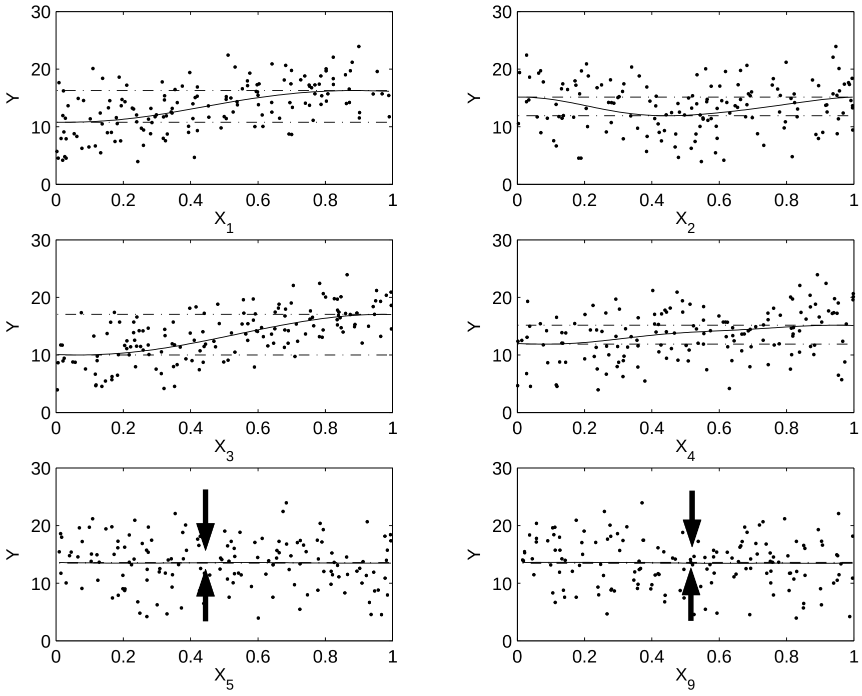


Figure 6.5: *Results from a benchmark study on the dataset as discussed in Example 6.2 with* $N = 100$ *and* $D = 25$. *The four first sub-plots show the contributions of the first* 4 *components, with the dashed line indicating the empirical maximal variation. The last two panels illustrate two components with zero empirical maximal variation.*

by (Fan, 1997) and which have been implemented in a kernel machine in (Pelckmans *et al.*, 2004, *In press*). This text will not pursue this issue as it leads to non-convex optimization problems in general. Instead, the use of the 1-norm is studied in order to detect structure, while the final predictions can be made based on a standard model using only the selected components (compare to basis pursuit, see e.g. (Chen *et al.*, 2001)).

Example 6.2 **[Numerical Example of Structure Detection]** An artificial example is taken from (Vapnik, 1998). Figure 6.5.a and 6.5.b shows results obtained on an artificial dataset consisting of $N = 100$ samples and dimension $D = 25$, uniformly sampled from the interval $[0,1]^{25}$. The underlying function takes the following form:

$$f(x) = 10\sin(X^1) + 20\,(X^2 - 0.5)^2 + 10\,X^3 + 5\,X^4, \tag{6.45}$$

such that $y_i = f(x_i) + e_i$ with $e_i \sim \mathscr{N}(0,1)$ for all $i = 1,\ldots,100$.

Figure 6.5.a gives the nontrivial components ($t_p > 0$) associated with the LS-SVM substrate with μ optimized in validation sense. Here, the hyper-parameters μ and λ were tuned using a 10-fold cross validation criterion. Figure 6.5.b presents the evolution of

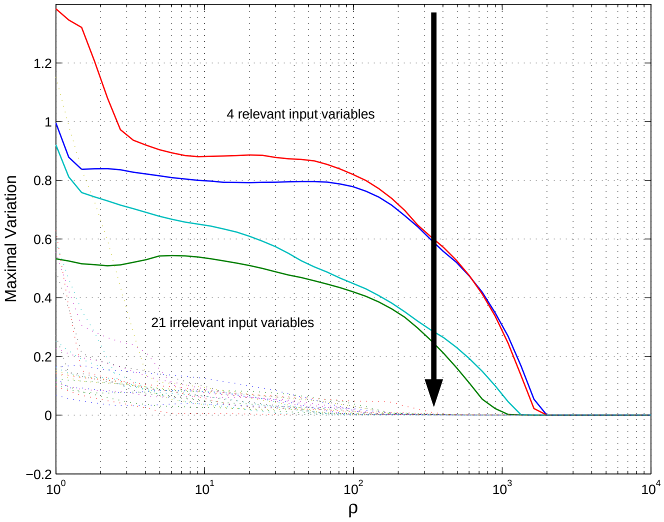


Figure 6.6: *The evolution of the empirical maximal variation of the different components when ranging* μ *from* 1 *to* 10^4*. The black arrow indicates the parameter selected by using* 10*-fold cross-validation, resulting in* 4 *nontrivial contributions of* X^1, X^2, X^3 *and* X^4*.*

values of t when ρ is increased from 1 to 1000 in a maximal variation evolution diagram (similarly as used for LASSO, see Subsection 6.1.2).

Note that an equivalent formulation is obtained by considering the Morozov type of constrained least squares problems. Let $\sigma_\mu \in \mathbb{R}^+$ and $\sigma_\lambda \in \mathbb{R}^+$ be constants. Then one can alternatively write (6.42) as

$$\mathscr{J}^{\hat{\mathscr{M}}}_{\sigma_\mu,\sigma_\lambda}(w,b,t) = \frac{1}{2}\sum_{p=1}^{P} w_p^T w_p$$

$$\text{s.t.} \begin{cases} \frac{1}{P}\sum_{p=1}^{P} t_p \leq \sigma_\mu \\ \frac{1}{N}\sum_{i=1}^{N} e_i^2 \leq \sigma_\lambda \\ \sum_{p=1}^{P} w_p^T \varphi_p\left(x_i^{(p)}\right) + b + e_i = y_i & \forall i = 1,\ldots,N \\ -t_p \leq w_p \varphi\left(x_i^{(p)}\right) \leq t_p, & \forall i = 1,\ldots,N, \forall p = 1,\ldots,P. \end{cases} \tag{6.46}$$

in which case a similar formulation is obtained as in Lemma 6.5 where μ and λ act as multipliers to the two last inequality constraints.

6.4.3 Kernel machines for handling missing values

Black-box techniques as neural networks and SVMs are quite useful in predictive settings but are considered less appropriate for handling missing data (see e.g. (Hastie *et al.*, 2001), Table 10.1). One typically has to resort to preprocessing methods as data imputation, data augmentation (Little and Rubin, 1987) or intractable EM methods, see e.g. (Dempster *et al.*, 1977). The optimization based approach of primal-dual kernel machines however can be employed to approach the problem as proposed in (Pelckmans *et al.*, 2005*b*) for the case of classification. The handling of missing values gives rise to uncertainty in the model's prediction. The use of additive models however can recover still some information in this case associated with components which are not affected.

The following setting is considered in the case of missing values of the input variables where the missing values are complete at random (MCAR) (Rubin, 1976; Little and Rubin, 1987).

Definition 6.2. [Integrated Risk] *An observed input value x_i takes a point distribution X_i at the point x_i, while a missing observation x_m is only known to follow the marginal distribution $x_m \sim P(X)$ with $P(X < x) = \prod_{i=1}^N P(X_i < x)$. Then one may employ the following integrated risk function.*

$$\mathscr{R}(f, P_{XY}) = \int_{x,y} \ell(y - f(x))\, dP_{XY} = \int_x \int_y \ell(y - f(x))\, dP_{Y|X}\, dP_X, \tag{6.47}$$

and the empirical counterpart

$$\hat{\mathscr{R}}(f, \mathscr{D}) = \sum_{i=1}^N \int_y \ell(y_i - f(x))\, dP_{X_i}. \tag{6.48}$$

As such one has to take into account the marginal distribution $P(X)$ only when the observation is missing. In the case of all observed data, (6.48) reduces to the classical risk as in (3.34). The case of building componentwise SVM classifiers in the context of missing values is elaborated based on (Pelckmans *et al.*, 2005*b*). A worst-case counterpart of the integrated empirical risk is studied with the class of models belonging to the componentwise kernel machines $f(x) = \sum_{i=1}^P w_p^T \varphi_p \left(x^{(1)}\right)$.

Definition 6.3. [Worst-case Empirical Risk] *A worst-case upper-bound to the empirical integrated risk of (6.48) is given as follows*

$$\underline{\hat{\mathscr{R}}}_I(f, \mathscr{D}) = \sum_{i=1}^N \max_{u \in [-\mathscr{M}, \mathscr{M}]} \ell(u - y_i), \tag{6.49}$$

which reduces in the case of the Hinge loss function to

$$\underline{\hat{\mathscr{R}}}_h(f, \mathscr{D}) = \sum_{i=1}^N \left[1 - y_i \left(\sum_{p \notin \mathbb{P}_i} w_p^T \varphi_p \left(x_i^{(p)} \right) \right) + \sum_{p \in \mathbb{P}_i} \mathscr{M}_p \right]_+ . \tag{6.50}$$

This can be encoded in a primal-dual kernel machine as follows.

Lemma 6.6. [Primal-Dual Kernel Machine for Handling Missing Values] *Consider the model $f(x) = \sum_{p=1}^P w_p^T \varphi_p\left(x^{(p)}\right) + b$, where the mappings $\varphi_p(\cdot) : \mathbb{R}^{D_p} \rightarrow \mathbb{R}^{n_h}$ denote the potentially infinite dimensional feature map for all $p = 1,\ldots,P$. The following regularized cost-function is considered:*

$$\min_{w,\xi,t} \mathscr{J}_C(w,\xi,t) = \frac{1}{2}\sum_{p=1}^{P} w_p^T w_p + C\sum_{i=1}^{N}\xi_i,$$

$$\text{s.t.} \quad \begin{cases} y_i\left(\sum_{p\notin\mathbb{P}_i} w_p^T\varphi_p\left(x_i^{(p)}\right)+b\right) - \sum_{p\in\mathbb{P}_i} t_p \geq 1-\xi_i \\ \xi_i \geq 0 & \forall i = 1,\ldots,N \\ -t_p \leq w_p^T\varphi_p\left(x_i^{(p)}\right) \leq t_p, & \forall i,p \mid p\in\mathbb{P}_i. \end{cases} \tag{6.51}$$

The dual problem becomes then

$$\max_{\alpha_i,\rho_{ip}^+,\rho_{ip}^-} -\frac{1}{2}\sum_{i,j=1}^{N}\alpha_{y,i}^{(p)}\alpha_{y,j}^{(p)}\tilde{\Omega}_{ij}^P + \sum_{i=1}^{N}\alpha_i$$

$$\text{s.t.} \quad \begin{cases} \alpha_{y,i}^{(p)} = \alpha_i y_i + \rho_{ip}^+ - \rho_{ip}^- & \forall i \mid p\notin\mathbb{P}_i \\ \alpha_{y,i}^{(p)} = \alpha_i y_i & \forall i \mid p\in\mathbb{P}_i \\ \sum_{i=1}^N y_i\alpha_i = 0 \\ \lambda = \sum_{i\mid p\notin\mathbb{P}_i}(\rho_{ip}^+ + \rho_{ip}^-) - \sum_{i\mid p\in\mathbb{P}_i}\alpha_i & \forall p = 1,\ldots,P \\ 0\leq\alpha_i\leq C & \forall i = 1,\ldots,N \\ \rho_{ip}^+,\rho_{ip}^- \geq 0, & \forall i = 1,\ldots,N\ \forall p\in\mathbb{P}_i, \end{cases} \tag{6.52}$$

where $\alpha_i \in \mathbb{R}$ and $\rho_{ip}^+,\rho_{ip}^- \in \mathbb{R}^+$ are the corresponding Lagrange multipliers, $\tilde{\Omega}_{ij}^P = \sum_{p=1}^P \tilde{K}_p\left(x_i^{(p)},x_j^{(p)}\right)$ for all $i,j = 1,\ldots,N$ and where $\tilde{K}_p\left(x_i^{(p)},x_j^{(p)}\right) = K_p\left(x_i^{(p)},x_j^{(p)}\right)$ if $x_i^{(p)}$ nor $x_j^{(p)}$ are missing and zero otherwise. The resulting nonlinear classifier evaluated on a new data point $x_* = \left(x_*^{(1)},\ldots,x_*^{(P)}\right)$ takes the form

$$\operatorname{sign}\left[\sum_{p=1}^{P}\sum_{i=1}^{N}\hat{\alpha}_i^{(p)}K_p\left(x_i^{(p)},x_*^{(p)}\right)+b\right], \tag{6.53}$$

where $\hat{\alpha}_i^{(p)}$ for all $i = 1,\ldots,N$ are solving (6.52).

Proof. The dual problem can be derived in the classical way. The Lagrangian $\mathscr{L}_C$ of the constrained optimization problem becomes

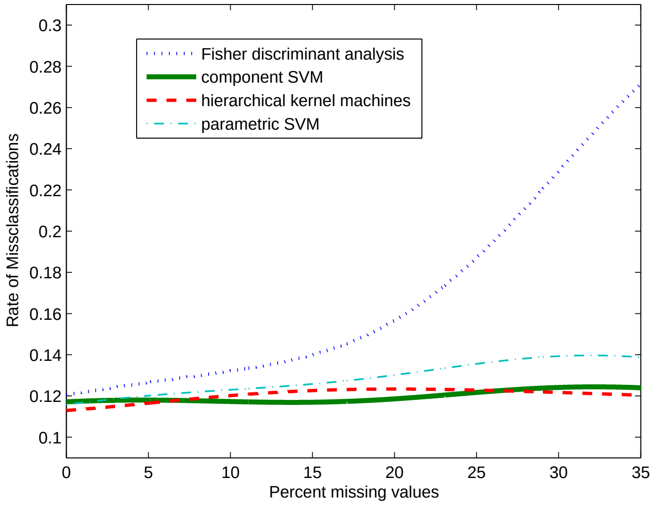
$$\mathscr{L}_C(w_p,\xi_i,t_p;\alpha_i,\nu_i,\rho_{ip}^+,\rho_{ip}^-) = \frac{1}{2}\sum_{p=1}^{P} w_p^T w_p + C\sum_{i=1}^{N}\xi_i$$

$$-\sum_{i=1}^{N}\nu_i\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i\left(\sum_{p\notin\mathbb{P}_i} w_p\varphi_p(x_i^{(p)})+b\right)-\sum_{p\in\mathbb{P}_i}t_p - 1+\xi_i\right]$$
$$-\sum_{ip\in\mathbb{P}}\rho_{ip}^{+}\left(t_p+w_p^T\varphi_p\left(x_i^{(p)}\right)\right)-\sum_{ip\in\mathbb{P}}\rho_{ip}^{-}\left(t_p-w_p^T\varphi_p\left(x_i^{(p)}\right)\right), \quad (6.54)$$

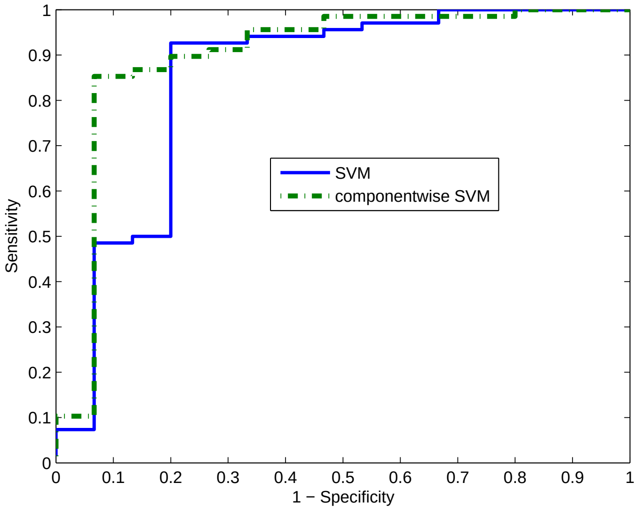
with positive multipliers $0 \leq \alpha_i, \nu_i, \rho_{ip}^{+}, \rho_{ip}^{-}$. The solution is then given as the saddle point of the Lagrangian resulting in the dual problem (6.52). From the condition for optimality $w_p = \sum_{i|p\notin\mathbb{P}_i}\alpha_i y_i\varphi_p\left(x_i^{(p)}\right)$, the result (6.53) follows. □

Example 6.3 **[Numerical Results on Missing Values]** A data set was designed in order to quantify the improvements and the difference of the proposed (linear and kernel) componentwise SVM classificators over standard techniques in the case of missing data and multiple irrelevant inputs. The Ripley dataset ($n = 150$, $d = 2$, binary labels) was extended with three extra (irrelevant) inputs drawn from a normal distribution ($\mathcal{N}(0,1)$). The component consisting of inputs X_1 and X_2 is detected correctly by the hyper-parameter optimizing the validation performance. In a second experiment, a portion of the data was marked as missing data. The performance on a disjoint validation set consisting of 100 points was used to tune hyper-parameters, while the final classifier was trained on all 250 samples. The performance on a fresh test set of size 1000 was used to quantify the generalization performance. For the purpose of comparison, the results of linear Fisher discriminant analysis were computed which cope with the missing values by omitting the corresponding samples, while the other approaches follow the derivations of Subsection 2.3. Figure 6.7.a shows the estimated generalization performance in function of the percentage of missing values.

As a second case, one considered the UCI hepatitis dataset ($n = 80, d = 19$) with approximately 50% of the samples containing at least one missing value. A standard SVM with RBF kernel and the componentwise SVM considering up to second order components were compared. The former replaces the missing values with the sample median of the corresponding variable while the latter follows the described worst-case approach. The respective hyper-parameters were tuned using leave-one-out cross-validation. Figure 6.7.b displays the receiver operating characteristic (ROC) curve of both classifiers on a test-set of size 55. As the componentwise only employed 25 non-sparse components out of the 380 components up to second order ($D_p \leq 2$), the proposed method outperformed the SVM both in interpretability as generalization performance.



(a)



(b)

Figure 6.7: **(a)** *Misclassification rate of the extended Ripley dataset in function of the percentage of missing values. Notice that the worst-case analysis is not breaking down when the percentage of missing values is growing.* **(b)** *ROC curves on the test-set of the UCI hepatitis dataset using an SVM with RBF kernel with imputation of missing values and componentwise SVM employing the measure of maximal variation employing the proposed method for handling missing values. The latter consists of 25 non-sparse out of the approximatively 400 components.*

Chapter 7

Fusion of Training with Model Selection

The amount of regularization is often determined by a set of constants which should be set by the user. The (meta-) problem of setting these is often treated as a problem of model selection and considered as being solved. However, a procedure for the automatic optimization of these hyper-parameters given a certain model selection criterion and model training procedure is highly desirable, at least in practice. This chapter outlines a framework for this purpose based on optimization theory. Section 7.1 introduces the problem and sketches the proposed solution. Various applications of the approach towards model selection problems in linear parametric models are given. Section 7.2 studies the problem of model selection in the case of LS-SVMs and SVMs.

7.1 Fusion of Parametric Models

In order to make intuition on this topic more accessible, the fusion argument for the parametric case is considered first. Unless stated otherwise, the validation performance function is taken as the generic standard for model selection. Let $\mathscr{D}^v = \{(x_j^v, y_j^v)\}_{j=1}^n \subset \mathbb{R}^D \times \mathbb{R}$ be a collection of data-samples i.i.d. sampled from the same distributions as those underlying the training dataset $\mathscr{D}$. Let $X^v = (x_1^v, \ldots, x_n^v)^T \in \mathbb{R}^{n \times D}$ and $Y^v = (y_1^v, \ldots, y_n^v)$, then the validation model selection criterion $\mathsf{Modsel}^v : \mathbb{R}^D \times \mathscr{D}^v \to \mathbb{R}$ becomes

$$\mathscr{J}^v(w) = \sum_{j=1}^{n} (w^T x_j^v - y_j^v)^2. \tag{7.1}$$

Extension to the closely related L-fold and leave-one crossvalidation (Stone, 1974) and to information criteria as Akaikes AIC (Akaike, 1973), C_p (Mallows, 1973) or GCV

(Golub *et al.*, 1979) may follow along the same lines.

7.1.1 Fusion of ridge regression and validation

At first, the task of appropriate selection of the ridge parameter $\gamma \geq 0$ in linear parametric models (see also Section 3.2 and Subsection 6.1.1) is studied. Consider the validation model selection criterion Modsel^v as in (7.1). Necessary and sufficient conditions on a parameter vector w to be the global optimum to (6.2) are given by the normal equations (6.3):

$$\left(X^T X + \gamma I_D\right) w = X^T Y. \tag{7.2}$$

The optimization problem of optimizing the solution-space over the hyper-parameter $\gamma \in \mathbb{R}^+$ with object-function Modsel^v may be formalized as an hierarchical programming problem (see Subsection 2.4.4):

$$\min_{\gamma,w} \mathscr{J}^v(w) = \frac{1}{2}\|X^v w - Y^v\|_2^2 \quad \text{s.t.} \quad \left(X^T X + \gamma I_D\right) w = X^T Y \ \text{ holds and } \ \gamma \geq 0. \tag{7.3}$$

This may be rewritten as the constrained optimization problem

$$(\hat{w}, \hat{\gamma}) = \arg\min_{\gamma,w} \mathscr{J}^v(w) = \frac{1}{2}\|X^v w - Y^v\|_2^2 \quad \text{s.t.} \quad \begin{cases} \left(X^T X\right) w + w_\gamma = X^T Y & (a) \\ \gamma w = w_\gamma & (b) \\ \gamma \geq 0. & (c) \end{cases} \tag{7.4}$$

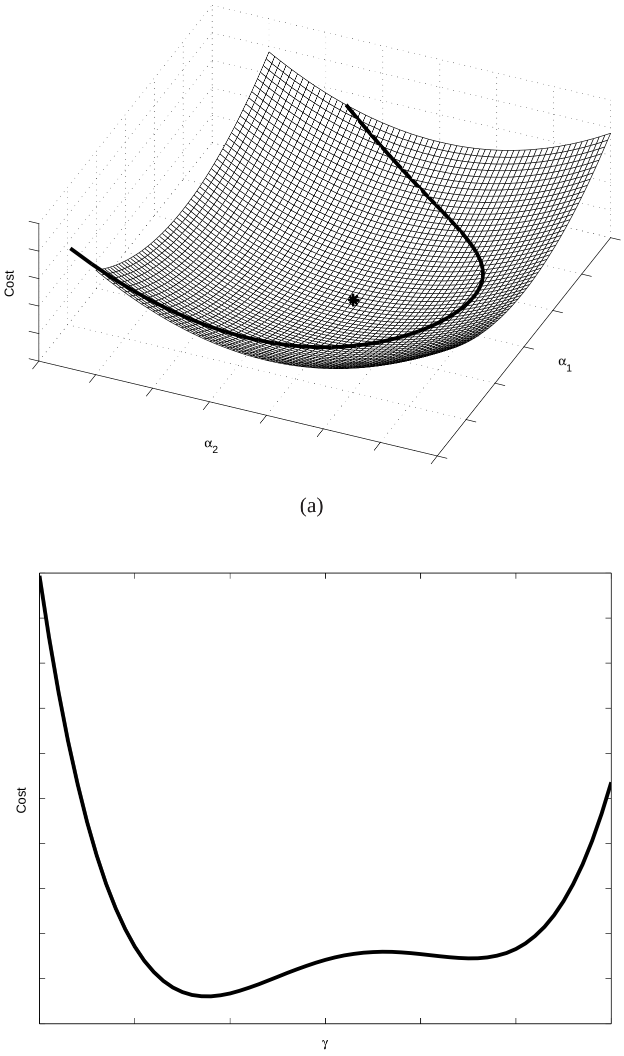
Note that the collinearity constraint (7.4.b) is non-convex. One may refer to this formulation as *Fusion of Ridge regression with model-selection,* or shortly *Fridge regression*. This typical formulation of fusion of training and validation can also be regarded from another perspective.

Definition 7.1 (Solution path). *The solution path of an estimator denotes the set of estimates from the data corresponding to any admissable hyper- or design-parameter.*

The solution path of ridge regression with respect to the regularization constant γ is shown in Figure 7.1.a. Then the task of fusion of an estimator with a (model selection) criterion amounts to minimizing this criterion over the solution path, see Figure 7.1.b. The solution path of the LASSO estimator and the SVM were described and analysed in (Efron *et al.*, 2004) and (Hastie *et al.*, 2004) respectively.

7.1.2 Convex relaxation to fusion of ridge regression

It turns out that in some cases the problem (7.4) can be solved efficiently. Assume that X is orthonormal such that $X^T X = I_D$ as in Lemma 6.1. Then the first order conditions



(a)

(b)

Figure 7.1: *Illustration of the solution path of ridge regression* **(a)** *the costfunction in the parameterspace (surface) and the solution path (solid line) with respect to the regularization constant.* **(b)** *The validation cost of the solutionpath with respect to the regularization constant. This figure illustrates that while the training problem may be convex in the parameters, the subproblem of hyper-parameter tuning may not.*

for optimality become

$$w_d = \lambda X_d^T Y, \ \ \lambda = \frac{1}{1+\gamma}, \ \ \forall d = 1, \ldots, D. \tag{7.5}$$

The fusion problem becomes as such

$$(\hat{w}, \hat{\lambda}) \quad = \quad \underset{\lambda, w}{\arg\min}\, \mathscr{J}^v(w, \lambda) \quad = \quad \frac{1}{2}\|X^v w - Y^v\|_2^2 \quad \text{s.t.} \quad \begin{cases} w = \lambda X^T Y \\ 0 < \lambda < 1, \end{cases} \tag{7.6}$$

which can be solved efficiently as a quadratic programming problem.

The inputs are in general not orthonormal at all, especially in the cases where regularization in the form of ridge regression is needed. However, the presented formalism can be used in order to obtain good initial estimates of the regularization constant and the parameters by adopting a suitable preprocessing step. Let USU^T denote the SVD of $X^T X$ with $S = \text{diag}(\sigma_1, \ldots, \sigma_D) \in \mathbb{R}^{D \times D}$ and $U \in \mathbb{R}^{D \times D}$ orthonormal. Then the normal equations (7.2) can be written as follows

$$U\,(S + I_D \lambda)\,U^T w = X^T Y \Leftrightarrow U^T w = \sum_{d=1}^{D} (\sigma_d + \lambda)^{-1} U_d^T X^T Y. \tag{7.7}$$

This can be approximated when the singular values $\{\sigma_d\}_{d=1}^D$ can be clustered in a small numbers around centers $\{\sigma_{\pi_i}\}_{i=1}^I$ where π_i denote disjunct sets of subsets of $1, \ldots, D$ such that $\bigcup_{i=1}^I \pi_i = \{1, \ldots, D\}$. This result in the approximation

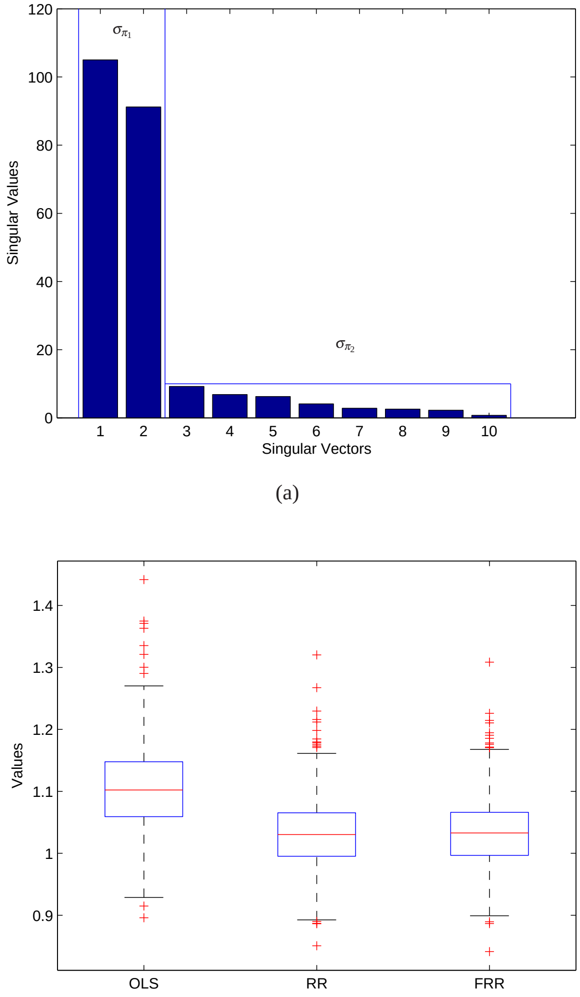
$$U^T w \approx \sum_{i=1}^{I} \lambda_{\pi_i} \sum_{d \in \pi_i} U_d^T X^T Y \quad \text{where} \quad \lambda_{\pi_i} = \frac{1}{\sigma_{\pi_i} + \lambda}. \tag{7.8}$$

A numerical example is constructed with $N = 100$ ten-dimensional $D = 10$ input datapoints which are ill-conditioned (rank larger than 1000), see Figure 7.2.a for a typical spectrum of singular values. The output satisfies the relation $y_i = \omega x_i + e_i$ where ω is a random vector and $e_i \sim \mathscr{N}(0,1)$ and $e_i \sim \mathscr{N}(0,1)$. A separate validationset of size $n = 75$ is used for tuning the regularization trade-off. Results of a Monte Carlo experiment with 1000 iterations are given in Figure 7.2.b. The latter achieves the same performance as the ridge regression but is computationaly much less intensive

7.1.3 A convex relaxation to stepwise selection

Consider the case of input selection for linear models based on model selection criteria. Given a vector of indicators $\imath = (\imath_1, \ldots, \imath_D)^T \in \{0,1\}^D$, the model is given as $f(x) = w_\imath^T I_{(\imath)} x$ where $I_{(\imath)} = \text{diag}(\imath) \in \mathbb{R}^{D \times D}$. The problem of ordinary least squares of this model is given as

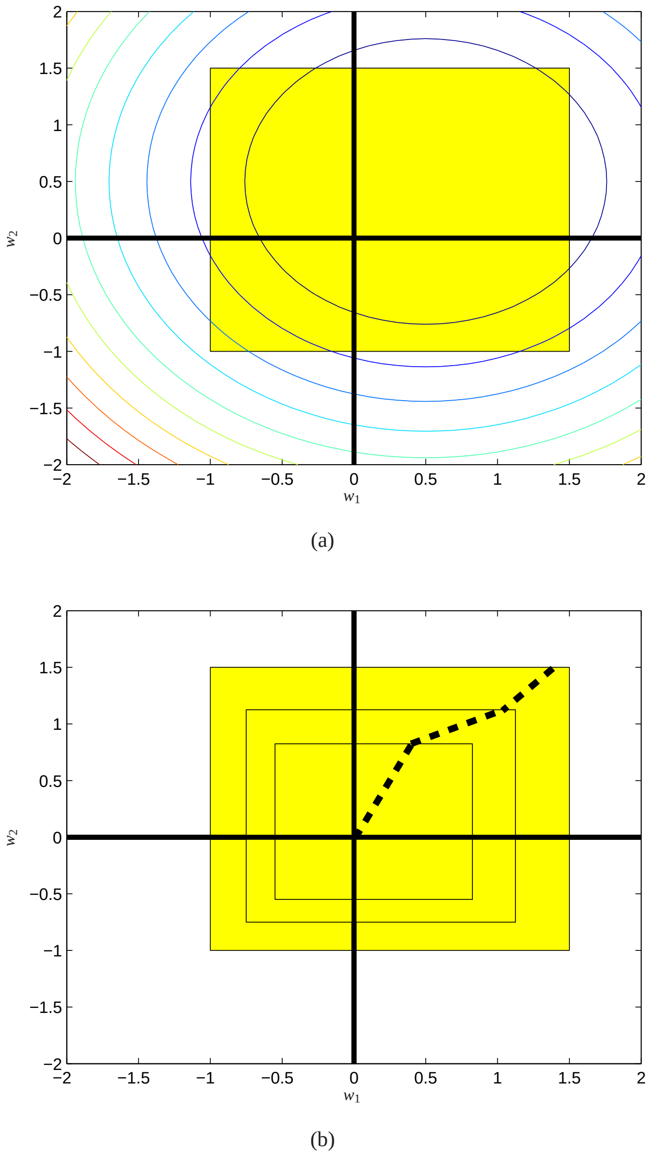
$$\mathscr{J}_\imath(w) = \frac{1}{2} \sum_{i=1}^{N} \left(w^T I_{(\imath)} x_i - y_i\right)^2, \tag{7.9}$$



(a)

(b)

Figure 7.2: *Example of the convex approach to fusion of ridge regression with validation.* **(a)** *A typical spectrum of the covariance matrix in linear parametric regression. The first two singular values and the remaining eight are clustered in two groups with average singular value* 97 *and* 5 *respectively.* **(b)** *Result on a Monte Carlo experiment relating estimqtes of Ordinary Least Squares (OLS) with Ridge Regression estimqtes manually tuned on a validation set and the convex approach to fusion as described in Subsection 7.1.2. The latter achieves the same performance as the ridge regression but is computationaly much less intensive.*



(a)

(b)

Figure 7.3: *Schematic illustration of the hierarchical programming problem approach towards convex stepwise selection.* **(a)** *Contourplot of the least squares costfunction due to the inequality constraints* $|w| \leq W^\iota$ *(square).* **(b)** *Solution path (dashed line) of the global least squares minimizer when varying the constraints* W^ι.

By use of the upper-bound $W^\imath \in \mathbb{R}^D$ such that $W^\imath = |I_{(\imath)} w|$ where $|\cdot|$ denotes the absolute value and $W_{\imath,d} = 0$ if and only if $\imath_d = 0$, for all $d = 1, \ldots, D$, one can write equivalently

$$\mathscr{J}_{W^\imath}(w) = \sum_{i=1}^{N} \left(w^T x_i - y_i\right)^2 \quad \text{s.t.} \quad -W^\imath \leq w \leq W^\imath, \tag{7.10}$$

where the upper-bound $W^\imath$ can now be chosen a-priori when the relevant inputs indicated by the vector $\imath$ are fixed. The Lagrangian becomes

$$\mathscr{L}_{W^\imath}(w; \alpha^+, \alpha^-) = \frac{1}{2} \sum_{i=1}^{N} \left(w^T x_i - y_i\right)^2 + \alpha^{-T}(-w - W^\imath) + \alpha^{+T}(w - W^\imath), \tag{7.11}$$

such that the Lagrange multipliers $\alpha^-, \alpha^+ \in \mathbb{R}^D$ are positive. The necessary and sufficient Karush-Kuhn-Tucker conditions are given as follows:

$$\text{KKT}_{(7.10)}(w; W^\imath, \alpha^+, \alpha^-) \begin{cases} (X^T X)w - X^T Y = \alpha^- - \alpha^+ & & (a) \\ \alpha_d^-, \alpha_d^+ \geq 0 & \forall d = 1, \ldots, D & (b) \\ -W_d^\imath \leq w_d \leq W_d^\imath & \forall d = 1, \ldots, D & (c) \\ \alpha_d^-(W_d^\imath + w_d) = 0 & \forall d = 1, \ldots, D & (d) \\ \alpha_d^+(W_d^\imath - w_d) = 0. & \forall d = 1, \ldots, D & (e) \end{cases} \tag{7.12}$$

Fusion of training and model selection Modsel can be formalized as

$$(\hat{w}; \hat{W}^\imath, \hat{\alpha}^+, \hat{\alpha}^-) = \underset{w; W^\imath, \alpha^+, \alpha^-}{\arg\min} \; \mathscr{J}^{\mathsf{Modsel}}(w) \quad \text{s.t.} \quad \text{KKT}_{(7.10)}(w; W^\imath, \alpha^+, \alpha^-) \quad \text{holds.} \tag{7.13}$$

It is clear that the problem of input selection with respect to a model selection criterion will result into a discrete and non-convex optimization problem. This is often approached with a greedy and somewhat ad hoc stepwise method (see e.g. (Hastie *et al.*, 2001)).

Based on the previous reformulation of the input selection problem in terms of the vector of hyper-parameters $W^\imath$ as in (7.10), a convex relaxation method can be considered. Consider the validation model selection procedure. One can show that the following modification to (7.13) is convex when $\varepsilon \geq 0$ is sufficiently small following the elaboration of hierarchical programming problems given in Subsection 2.4.4:

$$\begin{aligned} (\hat{w}; \hat{W}^\imath, \hat{\alpha}^+, \hat{\alpha}^-) = & \underset{w; W^\imath, \alpha^+, \alpha^-}{\arg\min} \; \mathscr{J}_\varepsilon^v(w; W^\imath, \alpha^+, \alpha^-) \\ & = \|X^v w - Y^v\|_2^2 + \varepsilon \left(\alpha^{+T}(W^\imath - w) + \alpha^{-T}(W^\imath + w)\right) \\ & \text{s.t.} \begin{cases} (X^T X)w - 2X^T Y = \alpha^- - \alpha^+ & \\ \alpha_d^-, \alpha_d^+ \geq 0 & \forall d = 1, \ldots, D \\ -W_d^\imath \leq w_d \leq W_d^\imath & \forall d = 1, \ldots, D. \end{cases} \end{aligned} \tag{7.14}$$

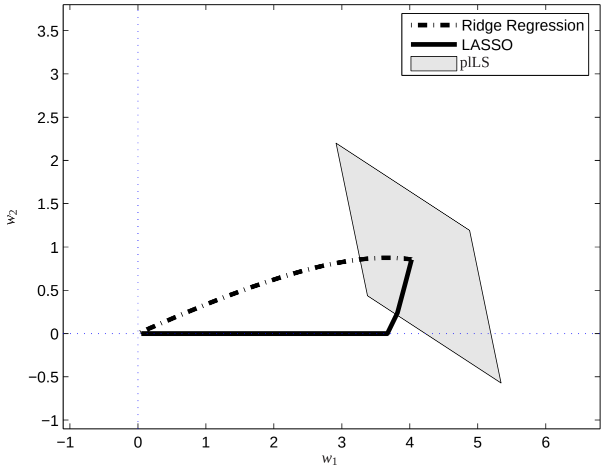


Figure 7.4: *A solution space of ridge regression (RR), LASSO and plausible least squares (pLS) estimators. The parameter space with the solution paths of respectively the Ridge Regressor (dashed-dotted) and the LASSO estimator (solid line) corresponding with different values of their respective hyper-parameters. The rectangle indicates the subspace of solutions which cannot be rejected with a α significance level.*

The following subsection gives an alternative approach based on an entirely different principle and which yields better performances in practice.

7.1.4 Plausible least squares estimates

Another example of fusion of a least squares estimate with a certain criterion is formulated. Here, one does not rely on an explicit parameterization scheme of the solution-space by an hyper-parameter as the regularization constant, but the set of solutions which cannot be rejected by a given significance level is considered instead.

Consider the case of deterministic inputs $x_i \in \mathbb{R}^D$ and stochastic outputs y_i following approximatively a Gaussian distribution $y_i \sim \mathscr{N}(\omega^T x_i, \sigma_e)$. The least squares estimate follows from the normal equations (3.4) where the only stochastic part occurs as $c(X, \mathbf{Y})$. Example 7.1 derives the distribution of the sample covariance estimator $\hat{c}^d(\mathscr{D})$. This can be used to specify a range on the covariance which is plausible given the finite set of samples in the classical way. Let $\hat{c}(\mathscr{D}, \alpha) = (\hat{c}^1(\mathscr{D}, \alpha), \ldots, \hat{c}^D(\mathscr{D}, \alpha))^T \in \mathbb{R}^D$ be the α-quantile of the sample distribution of the sample covariance. Then the solutions w satisfying the following inequalities cannot be rejected with an α_s significance level

$$\hat{c}_D\left(\mathscr{D}, \frac{\alpha_s}{2}\right) \leq (X^T X)w \leq \hat{c}_D\left(\mathscr{D}, 1 - \frac{\alpha_s}{2}\right). \tag{7.15}$$

The set $\{w \quad \text{satisfies eq. (7.15) }\}$ specifies a convex solution set for w, see Figure 7.4.

Example 7.1 **[Sample Covariance Distribution]** Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathbb{R}$ are deterministic points $\forall i = 1, \ldots, N$ and y_i i.i.d. sampled from a random variable $\mathbf{Y}_i$ with $E[\mathbf{Y}_i] = 0$, conditional mean $E[\mathbf{Y}_i|x_i] = wx_1$ and bounded variance $0 < \text{var}(Y_i) < \infty$ and $w \in \mathbb{R}$ fixed but unknown. Different approaches could be taken to derive expressions on the sample distribution of $\hat{c}(\mathscr{D})$.

Consider the sample covariance estimator $\hat{c}(\mathscr{D}) = \frac{1}{n}\sum_{i=1}^N x_i y_i$. It follows from the central limit theorem that $\hat{c}(\mathscr{D}) \rightarrow \mathscr{N}(\mu_c, \sigma_v^2)$ when $N \rightarrow \infty$ where the mean μ_v and σ_v^2 can be computed as follows

$$\begin{cases} \mu_v = E[\hat{c}(\mathscr{D})] = \frac{1}{N}\sum_{i=1}^N x_i E[\mathbf{Y}_i|x_i] = \frac{w}{N}\sum_{i=1}^N x_i^2 \\ \sigma_v^2 = \text{var}[\hat{c}(\mathscr{D})] = \frac{1}{N}\sum_{i=1}^N x_i \,\text{var}(\mathbf{Y}_i|x_i) = \frac{\sigma_e^2}{N}\sum_{i=1}^N x_i. \end{cases} \tag{7.16}$$

When σ_e^2 were not known, $\mathbf{Y}_i$ is approximately Gaussian, the sample variance estimate $\hat{\sigma}_e^2$ can be used. The sample distribution can then be described accurately as a t-distribution with $N-1$ degrees of freedom (see e.g. (Neter *et al.*, 1974)). When also $\mathbf{X}$ can becomes a random variable, the analysis becomes much more cumbersome. Let the random vector $\mathbf{Z}$ be defined as follows $\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{D+1}$ and let $Z \in \mathbb{R}^{N \times D+1}$ contain the N samples $(x_i y_i)$. In the case $\mathbf{Z}$ follows approximatively multivariate Gaussian $\mathbf{Z} \sim \mathscr{N}(0_{D+1}, \Sigma_Z)$, then the covariance matrix $Z^T Z \in \mathbb{R}^{D+1 \times D+1}$ follows a Wishart distribution $\mathscr{W}(\Sigma, N)$ with N degrees of freedom. By definition, the elements C of the Wishart distribution are confined to the positive (semi-) definite cone $S \succeq 0$. In the case $D = 1$ and $\Sigma = \sigma_x^2$, the wishart distribution reduces to the $\sigma_x^2 \chi^2(N)$ (Rao, 1965; Mardia *et al.*, 1979). Details on this approach and its references to the use of the Wishart distribution may be found e.g. in (Letac and Massam, 2004).

From a more practical point of view, the finite sample distribution of $\hat{c}$ may be determined using the bootstrap procedure (Efron, 1979) which results in accurate sample distributions under mild regularity assumptions. Figure 7.5.a gives the sample distribution in the case $\sigma_e^2 = 1$, $\sigma_x^2 = 1$, $N = 100$ and $b_1 = 3.14$ using the bootstrap. Its theoretical counterpart described in (7.16) is given in Figure 7.5.b.

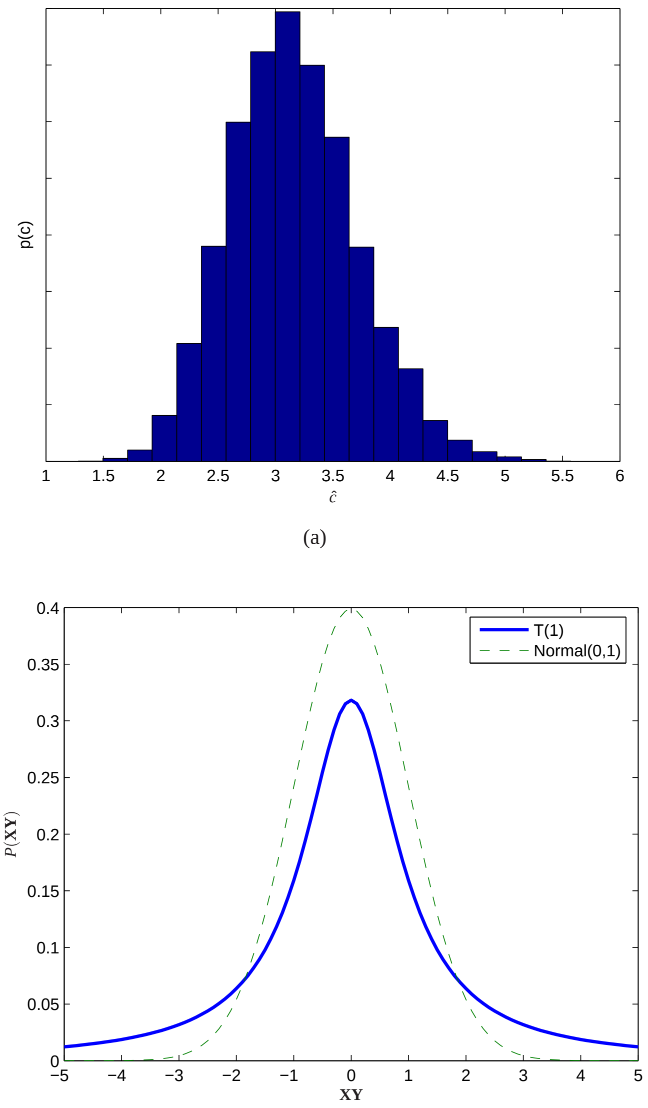
7.1.5 Plausible least squares and subset selection

We proceed by application of this formulation of the plausible solutionset of the least squares estimates towards subset selection. The following question is adressed: *What is the sparsest least squares solution which is still plausible?* As classicaly, the concept of plausibility may be encoded as passing a hypothesis test. A typical test for this simple case is the t-test (see also previous example). Thus one may describe the plausible solutionset of the least squares estimate as in equation (7.15)

$$\text{cond}_{(7.17)}(w, \alpha_s): \quad \hat{c}_D\left(\mathscr{D}, \frac{\alpha_s}{2}\right) \leq (X^T X)w \leq \hat{c}_D\left(\mathscr{D}, 1 - \frac{\alpha_s}{2}\right). \tag{7.17}$$

The desideratum of sparseness is relaxed by the use of the L_1 norm as classicaly. Then this question may be translated as follows

$$\hat{w} = \arg\min_w \mathscr{J}_\alpha(w) = \|w\|_1 \quad \text{s.t.} \quad \text{cond}_{(7.17)}(w, \alpha_s) \ \text{ holds,} \tag{7.18}$$



(a)

(b)

Figure 7.5: **(a)** *Finite sample distribution of the sample covariance estimator using bootstrap.* **(b)** *Limit sample distribution for $N \longrightarrow \infty$ when σ_e were known (normal distribution) and when it were estimated (Student's t-distribution).*

where $\left[\hat{c}_D(\mathscr{D}) - \alpha\sigma_e^2,\ \hat{c}_D(\mathscr{D}) + \alpha\sigma_e^2\right] = S \subset \mathbb{R}^D$ denotes the confidence interval of significance level $0 < \alpha \ll 1$ for the covariance (see previous example). This is another example of the hierarchical programming problem where plausiable model training is fused with a sparsness criterion.

Algorithm 7.1. *(Subset selection using plausible least squares) The algorithm for estimating the most sparse least squares estimate which cannot be rejected with a significance level α_s is found as follows.*

1. *Compute the sample distributions of the covariance of the input X^d with the observed output Y for all $d = 1, \ldots, D$, using either a bootstrap procedure or the sample moments (see example 7.1).*
2. *Given a significance level $0 < \alpha_s < 1$, construct the convex set*
$$\mathscr{S}_{\alpha_s} = \{w \mid \ cond_{(7.17)}(w, \alpha_s) \ \ holds \ \}. \tag{7.19}$$
3. *Find the most sparse solution vector $\hat{w}$ in $\mathscr{S}_{\alpha_s}$ by solving the fusion problem (7.18).*

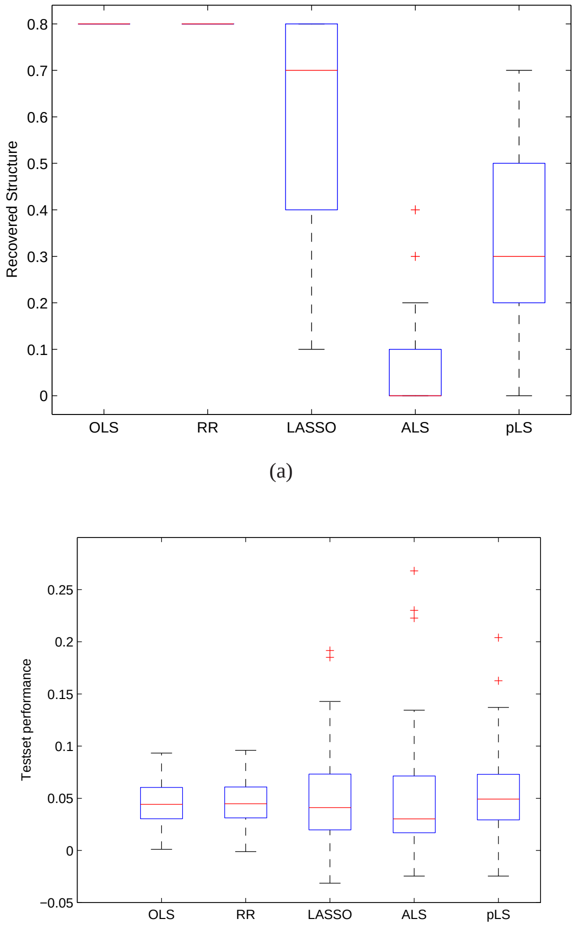
A numerical Monte Carlo experiment relating sparseness and performance of Ordinary Least Squares (OLS), Ridge Regression (RR) (see Subsection 6.1.1), LASSO estimate (see Subsection 6.1.2) Alternative Least Squares (ALS) (see Subsection 6.1.3), and the proposed method (plausible Least Squares or pLS) where the confidence interval was constructed using the quantiles from a simple bootstrap procedure with 10000 iterations. A dataset was constructed as follows, let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$ with $N = 100$, $D = 10$ and the observations generated as $y_i = \omega^T x_i e_i$ with $e_i \sim \mathscr{N}(0,1)$ and $\omega = (\omega_1, \omega_2, 0, \ldots, 0)^T \in \mathbb{R}^D$ where $\omega_1, \omega_2 \sim \mathscr{U}(-5,5)$. The regularization constant of the ridge regression estimate and the LASSO estimate as well as the significance level α of the proposed method are tuned with respect to the performance of the estimate on a separate validation set of size $n = 20$. The final performance is measured using the mean squares error of the estimate on a new testet of size 1000. Panel 7.6.a gives boxplots of the performances, while panel 7.6.b compares the ability to detect structure. Those figures shows that the given approach can have advantage both in performance as in structure detection in this dedicated example.

7.2 Fusion of LS-SVMs and SVMs

7.2.1 Fusion of LS-SVMs with validation

Fusion of the LS-SVM as described in Section 3.3 and the validation criterion Modsel^v as defined in (7.1) can be written as follows

$$\mathscr{J}^v(\alpha, \gamma) = \sum_{j=1}^{n} \left(\Omega_N(x_j^v)^T \alpha - y_j^v\right)^2 \quad \text{s.t.} \quad \begin{cases} \Omega\alpha + \alpha_\gamma = Y & (a) \\ \gamma^{-1}\alpha = \alpha_\gamma & (b) \\ \gamma^{-1} \geq 0 & (c), \end{cases} \tag{7.20}$$



(a)

(b)

Figure 7.6: *Performance of different multivariate estimators based on a least squares cost function. A Monte Carlo experiment relating the Ordinary Least Squares (OLS), Ridge Regression (RR), LASSO, Alternative Least Squares and the proposed method (plausible Least Squares or pLS) on an artificial dataset generated as described below. The regularization constants and the significance level α_S were tuned with respect to the performance on a disjunct validation set.* **(a)** *The recovered structure, while the ALS estimator picks always exactly one significant variable, the plausible least squares outperforms the LASSO method.* **(b)** *This property is traded by a small loss in performance of the estimates.*

where $\Omega_N : \mathbb{R}^D \to \mathbb{R}$ is defined as $\Omega_N(x) = (K(x_1,x),\ldots,K(x_N,x))^T$. As can be seen from this formulation, the constraint set of (7.20) is non-convex because of condition (b) including an unbounded quadratical term $\gamma^{-1}\alpha$. This renders the problem (7.20) non-convex even when the model selection criterion Modsel is convex on its own. A convex approach to the above problem is given in (Pelckmans *et al.*, 2004*b*) based on a matrix A^* leading to an appropriate linearization of the problem. The example below we show an alternative approach.

Example 7.2 **[Convex Approximation of Fusion of LS-SVMs with Validation]** Let K be decomposed as USU^T using a singular value decomposition with $U \in \mathbb{R}^{N\times N}$ orthonormal such that $U^TU = UU^T = I_N$ and $S = diag(\sigma_{(1)},\ldots,\sigma_{(N)}) \in \mathbb{R}^{N\times N}$ with ordered singular values $\sigma_{(1)} \geq \cdots \geq \sigma_{(N)}$. Then problem (7.20) can be rewritten as follows

$$\mathscr{J}^v(\alpha,\gamma) = \sum_{j=1}^{n}\left(\Omega_N(x_j^v)^T\alpha - y_j^v\right)^2 \quad \text{s.t.} \quad \begin{cases} (S+I_N)U^T\alpha = U^TY & (a)\\ \gamma^{-1} \geq 0. & (b)\end{cases} \tag{7.21}$$

Now we define $\lambda_{(i)}$ for all $i = 1,\ldots,N$ as follows

$$\lambda_{(i)} \triangleq \frac{1}{\sigma_{(i)} + 1/\gamma}. \tag{7.22}$$

As the function $f(x) = 1/(x+z)$ is strictly decreasing for $x \in \mathbb{R}^+$ given any fixed value of $z \in \mathbb{R}^+$, the following inequalities are obtained:

$$\begin{cases} \lambda_{(1)} \leq \lambda_{(2)} \leq \cdots \leq \lambda_{(N)} \\ 0 < \lambda_{(i)} \leq \frac{1}{\sigma_{(i)}} \quad \forall i = 1,\ldots,N. \end{cases} \tag{7.23}$$

Now we apply the overparaterization technique by omitting the constraint (7.21.b) and use the linear inequalities (7.23) instead, resulting in the relaxation

$$(\hat{\alpha},\hat{\lambda}) = \arg\min_{\alpha,\lambda} \mathscr{J}^v(\alpha) = \sum_{j=1}^{n}\left(\Omega_N(x_j^v)^T\alpha - y_j^v\right)^2$$

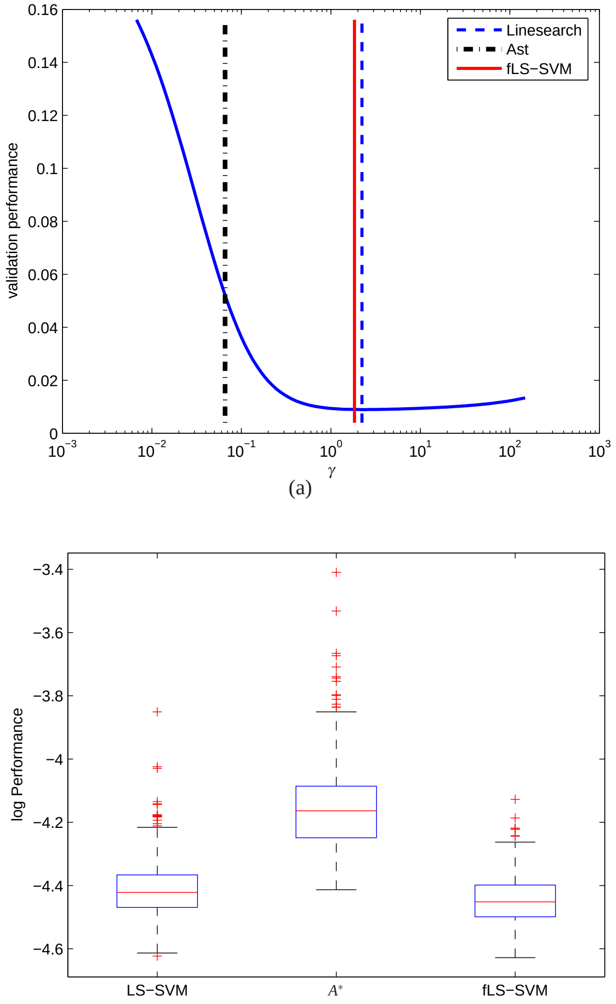
$$\text{s.t.} \quad \begin{cases} U^T\alpha = \sum_{i=1}^N \lambda_{(i)}U_i^TY & (a)\\ \lambda_{(1)} \leq \lambda_{(2)} \leq \cdots \leq \lambda_{(N)} & (b)\\ 0 < \lambda_{(i)} \leq \frac{1}{\sigma_{(i)}} \quad \forall i = 1,\ldots,N & (c), \end{cases} \tag{7.24}$$

where $\lambda = \left(\lambda_{(1)},\ldots,\lambda_{(N)}\right)^T \in \mathbb{R}^N$. Given the estimates, the approximate regularization constant $\hat{\gamma}$ can be recovered from the relation

$$\gamma\alpha = Y - \Omega\alpha, \tag{7.25}$$

and by substituting of the estimate $\hat{\alpha}$.

A monte Carlo study was conducted to assess the practical relevance of the proposed method. Let $\{(x_i,y_i)\}_{i=1}^{100} \subset \mathbb{R}\times\mathbb{R}$ satisfy the relation $y_i = \text{sinc}(x_i) + e_i$ with $\{e_i\}_{i=1}^{100} \sim \mathscr{N}(0,0.1)$. A validation set of size $n = 50$ was used to optimize the regularization constant γ via (a) a linesearch (using 40 evaluations), (b) the method presented in (Pelckmans *et al.*, 2004*b*) using a matrix A^* and (c) the presented method. While the proposed method achieves equivalent performance on a testset, the solution was found a factor 20 faster than the first method. The method proposed in (Pelckmans *et al.*, 2004*b*) gains even a factor 2 in performance, but the loss in performance is significant and the algorithm requires a good choice of the matrix A^*.



(a)

(b)

Figure 7.7: **(a)** *Performance on a validation-set of the estimate with respect to the regularization constant γ in the LS-SVM* estimate. *Vertical lines indicate the minima found by linesearch (dashed), the method based on the matrix A^* (dashed dotted) and the relaxation described in example 7.2 (solid line).* **(b)** *Results of a Monte-Carlo experiment relating the performance of an LS-SVM estimate using linesearch, the method based on a matrix A^* and the presented method. While the first performs as well as the last, the latter is computationally much more attractive.*

7.2.2 Fusion of SVMs with validation

Consider the primal class of classifiers

$$\mathscr{F}_{\text{svm}} = \left\{ f(x) = \text{sign}\left(\omega^T \varphi(x)\right) \;\middle|\; \omega \in \mathbb{R}^{D_\varphi} \right\}. \tag{7.26}$$

By employing the cost-function of the SVM (see Subsection 3.7.1) but using instead the ramp function (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000; Shawe-Taylor and Cristianini, 2004), one may write

$$\begin{aligned}(\hat{w},\hat{e}) = \underset{w,e}{\arg\min}\, \mathscr{J}_C(e) = \frac{1}{2} w^T w + C \sum_{i=1}^N e_i^2 & \\ \text{s.t.} \quad y_i[w^T \varphi(x_i)] \geq 1 - e_i, \;\; e_i \geq 0, \;\; \forall i = 1,\dots,N & \end{aligned} \tag{7.27}$$

Necessary and sufficient conditions are provided by the Karush-Kuhn-Tucker conditions with multipliers $\alpha, \rho \in \mathbb{R}^N$ as in Subsection 3.7.1.

$$\text{KKT}_{(7.27)}(w,e;\alpha,\rho) = \begin{cases} w = \sum_{i=1}^N \alpha_i y_i \varphi(x_i) & & (a) \\ Ce_i = \alpha_i + \rho_i & \forall i = 1,\dots,N & (b) \\ y_i[w^T \varphi(x_i)] \geq 1 - e_i & \forall i = 1,\dots,N & (c) \\ e_i \geq 0 & \forall i = 1,\dots,N & (d) \\ \alpha_i \geq 0, \rho_i \geq 0 & \forall i = 1,\dots,N & (e) \\ \alpha_i \left(y_i[w^T \varphi(x_i)] - 1 + e_i \right) = 0 & \forall i = 1,\dots,N & (f) \\ \rho_i e_i = 0, & \forall i = 1,\dots,N & (g). \end{cases} \tag{7.28}$$

Elimination of the variable w yields the necessary and sufficient conditions for the dual problem. The set of variables $(w, C, \alpha, \rho, e) \in \mathbb{R}^{D+1+3N}$ satisfying those constraints is non-convex due the positive OR constraints (7.28.fg). This solution space was characterized as a piecewise linear set in (Hastie *et al.*, 2004).

Chapter 8

Additive Regularization Trade-off Scheme

This chapter is related to the results of the previous chapter, but rather takes a different approach towards the problem of fusion. Instead of considering existing training procedures, a flexible formulation employing an additive regularization trade-off scheme is taken as the basis for fusion. The resulting substrate is found much easier to proceed with whenever more complex model selection criteria are involved. The basic ingredients are introduced in Section 8.1 and various relations are discussed. Section 8.2 then proceeds with the study of the fusion argument in the context of an LS-SVM regressor with additive regularization trade-off. Furthermore, the concept of an hierarchical kernel machine is introduced, leading to the construction of kernel machines maximizing their own stability (Section 8.3).

8.1 Tikhonov and the Additive Regularization Trade-off

8.1.1 The additive regularization trade-off

A reformulation to the LS-SVM formulation was proposed in (Pelckmans *et al.*, 2003*b*) leading to convex model selection problems. Let $\mathscr{D}$ be as in Chapter 3. Let $c = (c_1, \ldots, c_N)^T \in \mathbb{R}^N$ be a fixed vector of hyper-parameters. The central modification is to consider the following class of cost functions

$$(\hat{w},\hat{e}) = \arg\min_{w,e} \mathscr{J}^c(w,e) = \frac{1}{2}w^T w + \frac{1}{2}\sum_{i=1}^N (e_i - c_i)^2 \quad \text{s.t.} \quad w^T x_i + e_i = y_i. \ \forall i = 1,\ldots,N. \tag{8.1}$$

In the papers (Pelckmans *et al.*, 2003*b*; Pelckmans *et al.*, 2005*c*) this formulation was conceived as a modified trade-off parameterization replacing the classical regularization constant γ in the ridge cost-function (3.9) or (6.2). This is referred to as the Additive regularization trade-off (AReg) scheme. The modified normal equations are given as

$$\left(X^T X + I_D\right) \ w = X^T \left(Y - c\right). \tag{8.2}$$

Once c is fixed, the parameter vector $\hat{w}$ solving (8.2) is the unique global minimizer of (8.1).

8.1.2 A modified loss-function perspective

The parameterization scheme (8.1) can be interpreted as a Modified Loss Function (MLF) scheme. This can be seen most clearly by omitting the regularization term $w^T w$. Let $d = (d_1,\ldots,d_N)^T \in \mathbb{R}^N$ be a fixed vector of terms.

$$\mathscr{J}^b(w,e) = \frac{1}{2}\sum_{i=1}^N (e_i - d_i)^2 \quad \text{s.t.} \quad w^T x_i + e_i = y_i \ \ \forall i = 1,\ldots,N. \tag{8.3}$$

The modified normal equations become

$$\left(X^T X\right) \ w = X^T \left(Y - d\right), \tag{8.4}$$

Note that the formulations (8.2) and (8.4) result in equal solutions w when the following condition on c and d is satisfied:

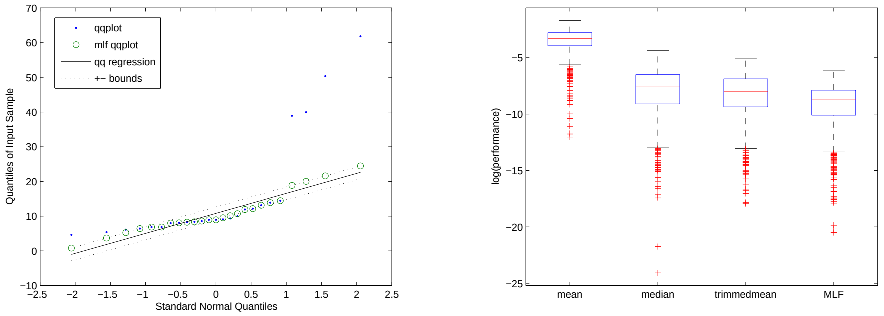
$$X^T c + w = X^T d, \tag{8.5}$$

whenever $X^T X$ is of full rank. This establishes the close connection between the AReg trade-off scheme and the MLF scheme.

Example 8.1 **[Imposing Normal Distribution on the Residuals]** This context of modified loss functions may be used for the formulation of robust estimators as exemplified as follows. Let $\{y_i\}_{i=1}^N$ be an i.i.d. sample from a random variable $\mathbf{Y}$ with fixed but unknown pdf $p_\mathbf{Y}$. Following Example 1.2, the maximum likelihood location parameter of a density with Gaussian distribution corresponds with the least squares estimate.

Let $p_\mathbf{Y}$ instead follow a contaminated distribution $\mathscr{F}_\varepsilon(\mathscr{N},\mathscr{U})$ defined in (3.58). Let $d \in \mathbb{R}^N$ be fixed such that $\mathscr{D}_d = \{y_i - d_i\}_{i=1}^N \sim \mathscr{N}$, then the MLF argument leads to the following estimator

$$\hat{\mu} = \arg\min_{\mu} \mathscr{J}_d(\mu) = \sum_{i=1}^N (y_i - d_i - \mu)^2 \Leftrightarrow \mu N = 1_N^T (Y - d). \tag{8.6}$$



(a) (b)

Figure 8.1: *Illustration of a use of the MLF mechanism in the case of a sample of a contaminated model.* **(a)** *A Quantile-Quantile plot ('.') of the original sample $\{y_i\}_{i=1}^N$ and of the modified samples $\{y_i - d_i\}_{i=1}^N$ ('o') versus the quantiles of the standard normal distribution. The coefficients of the regression (solid line) equal the estimated location and scale parameter of the nominal model. The figure illustrates the difference in which outliers (at the tails) and samples form the nominal model (at the center) are treated by the MLF mechanism.* **(b)** *Boxplots representing the results of a Monte-Carlo study comparing the mean, median, trimmed mean ($\beta = 25\%$) and the proposed method based on MLF for estimating the location. The performance is expressed as the mean squared error of the estimate and the true location parameter, $N = 50$ and the contamination factor was set to 25%. While the trimmed mean, the median and the MLF based method achieve comparable performance, the latter yields additionally estimates of the scale and quantiles of the nominal model.*

Employing the fusion argument, the question which vector d makes a maximal likelihood estimate $\hat{\mu}$ may be formalized as follows

$$\mathscr{J}^o(\mu,d) = \|d\|_1 \quad \text{s.t.} \quad \begin{cases} (y_i - d_i) \sim \mathscr{N} \\ \mu N = 1_N^T(Y-d). \end{cases} \tag{8.7}$$

The first constraint may be approached by imposing small higher (> 2) moments on the distribution of $\mathscr{D}_d$, see e.g. (Boyd and Vandenberghe, 2004)

An approach may be used using the Quantile-Quantile method comparing two distributions based on the ordered dataset. Let therefor $y_{(i)} \leq y_{(i+1)}$ for all $i = 1,\ldots,N-1$ denote the ordered samples. As the order is retained by translating the samples with a constant μ. By comparison of this ordered samples with Let $\mathscr{D}_z = \left\{z_{(i)}\right\}_{i=1}^N$ be an ordered sample from the standard normal $\mathscr{N}(0,1)$ such that $z_{(i)} \leq z_{(i+1)}$ for all $i = 1,\ldots,N-1$. The deviation of the sample $\mathscr{D}_c$ of the normal distribution $\mathscr{D}_z$ may then be quantified by the maximal deviation $d = \sup_i \left| \left(y_{(i)} - d_{(i)}\right) - \left(\mu + z_{(i)}\right) \right|$ as follows. Let $\sigma_Y \in \mathbb{R}^+$ be the slope of the QQ-plot, see Figure 8.1

$$(\hat{\mu}, \hat{\sigma}_Y, r) = \arg\min_{\mu,\sigma_Y,r} r \quad \text{s.t.} \quad -r \leq \left(y_{(i)} - d_{(i)}\right) - \left(\mu + \sigma_Y z_{(i)}\right) \leq r. \tag{8.8}$$

Let $g = (g_1, \ldots, g_N)^T \in \mathbb{R}^{N,+}$ be a vector of positive slack variables. Using the Pareto approach to multi-criterion optimization results in the following problem

$$\min_{\mu,\sigma_y,r,d,g} \mathscr{J}^o_\lambda(\mu, \sigma_y, r, d, g) = \lambda r + \frac{1}{N}\sum_{i=1}^N g_i$$
$$\text{s.t.} \quad \begin{cases} y_{(i)} - d_{(i)} \leq y_{(i+1)} - d_{(i+1)} & \forall i = 1, \ldots, N-1 \\ -r \leq \left(y_{(i)} - d_i\right) - \left(\mu + \sigma_y z_{(i)}\right) \leq r & \forall i = 1, \ldots, N \\ -g_i \leq d_i \leq g_i & \forall i = 1, \ldots, N \\ \mu N = 1_N^T (Y - d). \end{cases} \tag{8.9}$$

From this problem formulation not only follows an estimate of the location $\hat{\mu}$, but also of the scale parameter $\hat{\sigma}_y$ of the nominal model behind the sample. Moreover, quantile intervals of the nominal model follow from the estimate. The non-sparse elements of d may indicate the outliers in the model, Figure 8.1.a shows an example of a quantile-quantile plot (QQ-plot) of the original samples and of the modified samples using the mechanism as described. Panel 8.1.b reports results of a Monte-Carlo study comparing the mean, median, trimmed mean ($\beta = 25\%$) and the proposed method based on MLF for estimating the location. The performance is expressed as the mean squared error of the estimate and the true location parameter, $N = 50$ and the contamination factor was set to 25%.

8.1.3 LS-SVM substrates

The extension of the AReg scheme to primal-dual kernel machines was studied in (Pelckmans *et al.*, 2003*b*; Pelckmans *et al.*, 2005*c*). Consider the modified cost-function to (3.9) with given values $c \in \mathbb{R}^N$:

$$\mathscr{J}^c(w, e) = \frac{1}{2} w^T w + \frac{1}{2}\sum_{i=1}^N (e_i - c_i)^2 \quad \text{s.t.} \quad w^T \varphi(x_i) + e_i = y_i. \ \forall i = 1, \ldots, N \tag{8.10}$$

The dual solution is then uniquely determined by the following equations

$$\text{KKT}_{(8.10)}(\alpha, e; c) = \begin{cases} (\Omega + I_N)\,\alpha + c = Y & (a) \\ e = \alpha + c, & (b) \end{cases} \tag{8.11}$$

where $\alpha \in \mathbb{R}^N$ are the Lagrange multipliers. The resulting predictor $\hat{f}$ may be evaluated in any point $x_* \in \mathbb{R}^D$ as $\hat{f}(x_*) = \Omega_N(x_*)^T \hat{\alpha}$ where $\Omega_N : \mathbb{R}^D \to \mathbb{R}^N$ is defined as $\Omega_N(x) = (K(x_1, x), \ldots, K(x_N, x))^T$. Note that the vector of residuals e is not eliminated as in Section 3.3 as it will be often needed later-on. We refer to this dual characterization of the solution space to the AReg cost-function as the *LS-SVM substrate*. Note that the LS-SVM formulation (3.9) is taken as a starting point as this lead to the simplest characterization, see also Section 3.3.

Remark that by relating condition (8.11.a) to (3.15.a), one can derive the condition on c and γ for which the solutions equal as follows

$$(\gamma^{-1}-1)\alpha = c, \quad \gamma^{-1} > 0, \tag{8.12}$$

which is clearly non-convex if both γ, c and α are unknown.

8.2 Fusion of LS-SVM substrates

Fusion of the LS-SVM substrate with a model selection criterion $\mathsf{Modsel}(f,\mathscr{D})$ with respect to the regularization constants $c \in \mathbb{R}^N$ may be written as a hierarchical programming problem

$$(\hat{e},\hat{\alpha};\hat{c}) = \arg\min_{e,\alpha;c} \mathscr{J}_{\mathsf{Modsel}}(\alpha) \quad \text{s.t.} \quad \text{KKT}_{(8.10)}(e,\alpha;c) \text{ holds.} \tag{8.13}$$

A crucial property of (8.11) and (8.13) is that the regularization vector $c \in \mathbb{R}^N$ occurs linearly in the constraints. The price one has to pay for this advantage is the increased number of regularization constants $c \in \mathbb{R}^N$ absorbing the non-convex constraints. The remainder of this section will mostly be concerned with the appropriate restriction of the effective degree of freedom of the constants $c \in \mathbb{R}^N$ by imposing a-priori knowledge or model selection criteria on the solution space $\text{KKT}_c(\alpha,e)$ for all $c \in \mathbb{R}^N$.

8.2.1 Fusion of LS-SVM substrates with validation

At first, the case where Modsel is the validation performance Modsel^v on a disjunct validation dataset $\mathscr{D}^v$ is studied.

$$\mathscr{J}_{c,\mathsf{Modsel}^v}(\alpha,c) = \sum_{j=1}^{n} \left(\Omega_N(x_j^v)^T\alpha - y_j^v\right)^2 \quad \text{s.t.} \quad (\Omega+I_N)\alpha + c = Y. \tag{8.14}$$

As was shown in (Pelckmans *et al.*, 2003*b*), the size of the validation-set $\mathscr{D}^v$ should be significantly larger than N in order to obtain stable solutions. This may be seen informally as n samples need to determine N degrees of freedom parameterized by the regularization constant.

In order to approach this disadvantage, the solution α (and thus c) was restricted to the convex hull of the quadratic constraint (8.12). To compute an approximative convex hull of the constraint (8.12), was constructed using a discrete set of regularization constants $\Gamma = \{\gamma_1\}_{q=1}^{Q}$, leading to a convex set

$$\mathscr{S}_\Gamma = \left\{ \alpha = \sum_{q=1}^{Q} g_q \alpha_{\gamma_q} \in \mathbb{R}^N \;\middle|\; \left(\Omega + \gamma_q^{-1} I_N\right)\alpha_{\gamma_q} = Y, \right.$$

$$g_q \geq 0 \ \ \forall q, \ \ \sum_{q=1}^{Q} g_q = 0 \Bigg\}. \quad (8.15)$$

Figure 8.2.a illustrates the solutionset spanned by three Thikonov nodes. Figure 8.2.b gives the results of a numerical comparison of the evolution of the generalization performance in terms of the number of nodes with respect to the generalization ability of the original solution to problem (7.20) using a naive line-search with the same number of evaluations. This formulation is closely related to the marginalization over the noise constant as described in Example 6.1. As can be derived from the set description, the following algorithm may be used:

Algorithm 8.1. [Ensemble Approach to the Fusion of LS-SVMs with Validation] *Let $\Gamma = \{\gamma_q\}_{q=1}^{Q}$ be a set of possible regularization parameters for $1 < Q \in \mathbb{N}$ denoting the vertices of the hull.*

1. *For each γ_q, compute the solution α_{γ_q} to the LS-SVM regressor (3.12).*

2. *Let $g = (g_1 \ldots, g_Q)^T \in \mathbb{R}^Q$ be a vector. Solve the problem*

$$(\hat{f}_g, \hat{g}) = \arg\min_{f_g, g} \mathscr{J}_{\alpha_{\gamma_q}}^{\Gamma,\mathsf{Modsel}}(f_g, g) \quad s.t. \quad \begin{cases} f_g(x) = \Omega_N(x)^T \sum_{q=1}^{Q} g_q \alpha_{\gamma_q} & (a) \\ \sum_{q=1}^{Q} g_q = 1 & (b) \\ g_q \geq 0, \ \ \forall q = 1, \ldots, Q & (c) \end{cases} \quad (8.16)$$

which is convex when $\mathsf{Modsel}(f)$ is a convex measure on $f = w^T \varphi$.

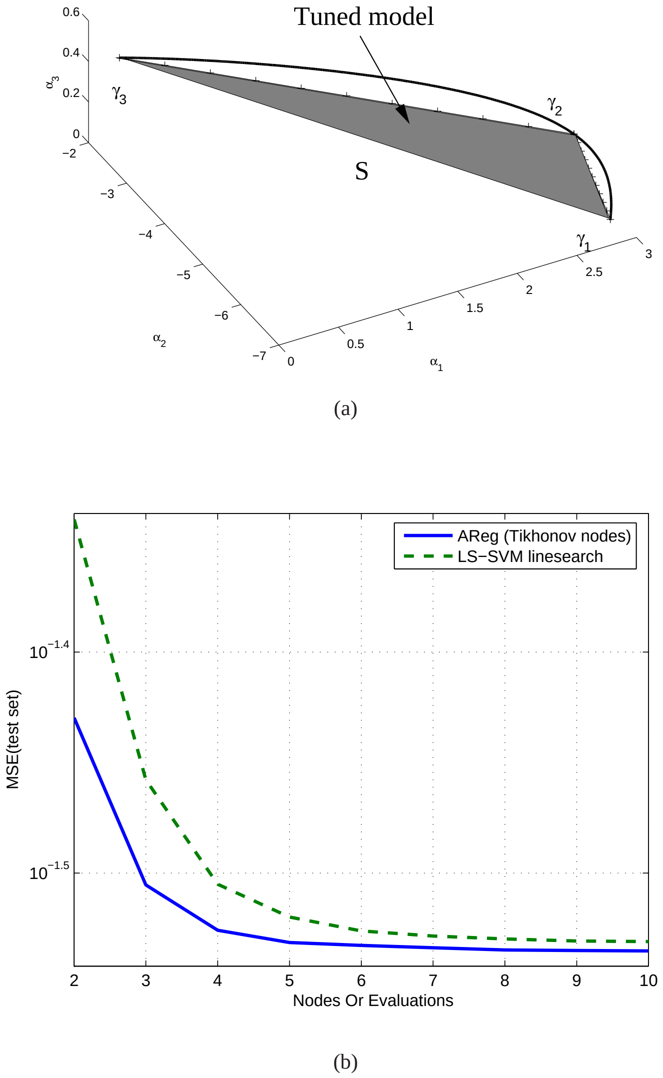
A new point $x \in \mathbb{R}^D$ may be evaluated as $\hat{f}_\Gamma(x_*) = \sum_{q=1}^{Q} g_q \left(\Omega_N(x_*)^T \alpha_{\gamma_q}\right)$.

This algorithm is related to the ensemble approach as elaborated e.g. in (Perrone and Cooper, 1993; Bishop, 1995; Breiman, 1996) and surveyed in (Hamers, 2004).

8.2.2 Fusion of LS-SVM substrates with cross-validation

In order to avoid the non-trivial process of dividing valuable data into a separate training and validation set, Cross-Validation (CV) (Stone, 1974) has been introduced. The following is based on the L-fold CV (where Leave-One-Out CV is a special case with $L = N$). Let $\mathscr{T}$ denote the set of indices of the dataset $\mathscr{D}$ and $\mathscr{V}_l$ denote the set of indices of the lth fold. Then the set $\mathscr{T}$ is repeatedly divided into a training set $\mathscr{T}_l$ and a corresponding disjoint validation set $\mathscr{V}_l$, $\ \forall l = 1, \ldots, L$ such that $\mathscr{T} = \mathscr{T}_l \cup \mathscr{V}_l = \cup_{l=1}^{L} \mathscr{V}_l$ and $\mathscr{V}_l \cap \mathscr{V}_k = \varnothing, \quad \forall l \neq k = 1, \ldots, L$. In the following, $N_{(l)}$ denotes the number of training points and $n_{(l)}$ the number of validation points of the lth fold. Figure 8.3 illustrates this repeated training and validation process.

All L training and validation steps can be solved simultaneously but independently by stacking them into a block diagonal linear system. For notational convenience, the



(a)

(b)

Figure 8.2: **(a)** *Illustration of the convex solution-space according to three Tikhonov nodes.* **(b)** *Evolution of the generalization performance when increasing the number of nodes n compared to the result of a naive line-search using n evaluations. The proposed method is seen to outperform the line-search approach for n small.*

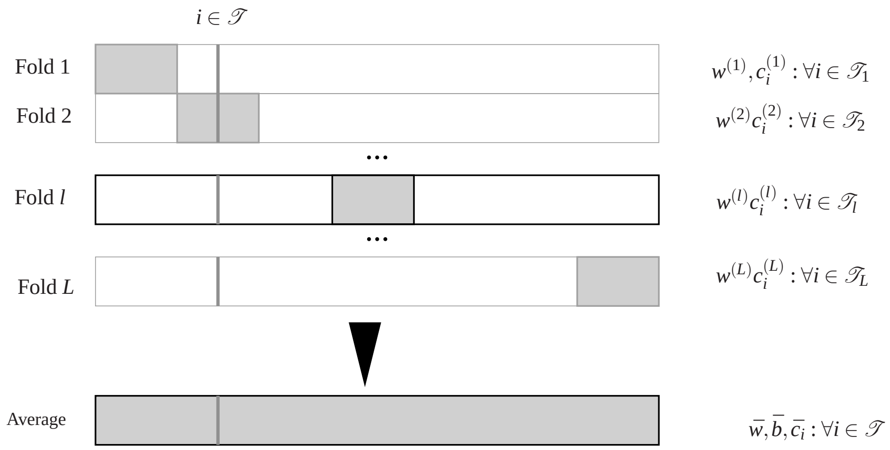

$i \in \mathscr{T}$
Fold 1
Fold 2
Fold l
Fold L
Average
$w^{(1)}, c_i^{(1)} : \forall i \in \mathscr{T}_1$
$w^{(2)} c_i^{(2)} : \forall i \in \mathscr{T}_2$
$w^{(l)} c_i^{(l)} : \forall i \in \mathscr{T}_l$
$w^{(L)} c_i^{(L)} : \forall i \in \mathscr{T}_L$
$\bar{w}, \bar{b}, \bar{c_i} : \forall i \in \mathscr{T}$

Figure 8.3: *Schematical representation of the L-fold cross-validation procedure.*

indicator matrix $I_{(\mathscr{S}_1,\mathscr{S}_2)}$ is introduced denoting a sparse matrix with (i,j)th entry 1 if $\mathscr{S}_1(i) = \mathscr{S}_2(j)$ and 0 otherwise for sets $\mathscr{S}_1$ and $\mathscr{S}_2$, e.g.:

$$I_{(\mathscr{S}_1,\mathscr{S}_2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{where} \quad \mathscr{S}_1 = \{a,b,d\} \quad \text{and} \quad \mathscr{S}_2 = \{a,b,c,d\}. \tag{8.17}$$

As argued in the previous subsection, in each fold the number of validation data may not be smaller than the number of training data. To avoid this difficulty in the cross-validation setting, there is an opportunity to restrict in a natural way the degrees of freedom of the additive regularization constants $c^{(l)}$ for all $l = 1,\dots,N_{(l)}$. As in classical cross-validation practice, the (additive) regularization constants should be held constant over the different folds, i.e.

$$c^{(l)} = I_{\mathscr{T}_l,\mathscr{T}}\, c, \ \forall l = 1,\dots,L. \tag{8.18}$$

This reduces the freedom of the regularization constants from $(L-1)N$ to N. Embedding this in a single linear system results in the following problem. Let $\ell^{cv} : \mathbb{R}^{2LN} \to \mathbb{R}$ be a convex loss function of the training residuals $e^{(l)}$ and the validation errors $e^{(l)v}$ of the L folds.

$$\left(\hat{\alpha}^{(l)}, \hat{c}, \hat{e}^{(l)}, \hat{e}^{(l)v}\right) = \underset{\alpha^{(l)},c,e^{(l)},e^{(l)v}}{\arg\min} \ \ell^{cv}\left(e^{(l)v}, e^{(l)}\right)$$
$$\text{s.t.} \quad \mathrm{KKT}^{l}_{(8.20)}\left(\alpha^{(l)}, c, e^{(l)}, e^{(l)v}\right) \forall l = 1,\dots,L, \tag{8.19}$$

where the L sets of constraints are the Karush-Kuhn-Tucker conditions for the individual folds (8.11) and

$$\mathrm{KKT}^{l}_{(8.20)}\left(\alpha^{(l)},c,e^{(l)},e^{(l)v}\right):\begin{cases} I_{(\mathscr{T}_l,\mathscr{T})}\left(\Omega+I_N\right)I_{(\mathscr{T},\mathscr{T}_l)}\alpha^{(l)}+I_{(\mathscr{T}_l,\mathscr{T})}c=I_{(\mathscr{T}_l,\mathscr{T})}Y & (a)\\ \alpha^{(l)}+I_{(\mathscr{T}_l,\mathscr{T})}c=e^{(l)} & (b)\\ I_{(\mathscr{V}_l,\mathscr{T})}\Omega I_{(\mathscr{T},\mathscr{T}_l)}\alpha^{(l)}+e^{(l)v}=I_{(\mathscr{V}_l,\mathscr{T})}Y, & (c)\end{cases} \tag{8.20}$$

for all $l=1,\ldots,L$. This problem formulation has $2LN$ unknowns with $2LN-N$ different constraints leading to large scale problems already when $N>100$. In (Pelckmans *et al.*, 2003*b*), the following choice for the cost-function ℓ^{cv} was considered.

$$\min_{\alpha^{(l)},c,e^{(l)},e^{(l)v}} \frac{1}{2L}\sum_{l=1}^{L} e^{(l)v^T}e^{(l)v}+\frac{1}{2L}\sum_{l=1}^{L} e^{(l)^T}e^{(l)} \quad \text{s.t.}$$
$$\mathrm{KKT}^{l}_{(8.20)}\left(\alpha^{(l)},c,e^{(l)},e^{(l)v}\right)\forall l=1,\ldots,L, \tag{8.21}$$

A big disadvantage of this approach is the rapid growth of the number of parameters when $N>100$.

8.2.3 A fast approach to fusion with CV

In order to reduce the computational complexity of the approach, a slightly different approach may be formulated leading to a convex problem of $2N$ variables and N constraints. Therefor, the level 1 training of the different folds is written as a multi-criterion optimization problem:

$$\left(w^{(l)},e_k^{(l)},b\right)=\arg\min_{w^{(l)},e_k^{(l)},b}\begin{bmatrix}\frac{1}{2}w^{(1)^T}w^{(1)}+\frac{1}{2}\sum_{k\in\mathscr{T}^{(1)}}\left(e_k^{(1)}-c_k\right)^2\\ \ldots\quad\ldots\\ \frac{1}{2}w^{(L)^T}w^{(L)}+\frac{1}{2}\sum_{k\in\mathscr{T}^{(L)}}\left(e_k^{(L)}-c_k\right)^2\end{bmatrix}$$
$$\text{s.t.}\begin{cases} w^{(1)^T}\varphi(x_k)+b+e_k^{(1)}=y_k & \forall k\in\mathscr{T}^{(1)}\\ \ldots & \\ w^{(L)^T}\varphi(x_k)+b+e_k^{(L)}=y_k. & \forall k\in\mathscr{T}^{(L)}.\end{cases} \tag{8.22}$$

Although the criteria of (8.22) can be solved individually but with coupled regularization constants (8.18), one can relax the problem by trying to find one Pareto-optimal solution (Boyd and Vandenberghe, 2004). The scalarization technique with weights $1_N = (1,\dots,1)^T \in \mathbb{R}^L$ in the objective function is used leading to a much compacter problem than the original formulation (Pelckmans *et al.*, 2003*b*).

$$\sum_{l=1}^{L}\sum_{k\in\mathbb{T}_l}\left(e_k^{(l)}-c_k\right)^2=\sum_{i=1}^{N}\sum_{l|i\in\mathscr{T}_l}\left(e_i^{(l)}-c_i\right)^2=\sum_{i=1}^{N}\left(\sum_{l|i\in\mathscr{T}_l}e_i^{(l)}-(L-1)\tilde{c}_i\right)^2$$

$$\text{s.t.}\quad \tilde{c}_i=\sum_{l|i\in\mathscr{T}_l}e_i^{(l)}+\sqrt{\sum_{l|i\in\mathscr{T}_l}\left(e_i^{(l)}-c_i\right)^2}. \tag{8.23}$$

Eliminating the residuals $e^{(l)}$ and the original regularization term c, the following constrained optimization approach to the cross-validation based AReg LS-SVM is obtained

$$\min_{w^{(l)},b,e_k}\mathscr{J}^{(cv)}=\frac{1}{2}\sum_{l=1}^{L}\frac{{w^{(l)}}^T w^{(l)}}{(L-1)}+\frac{1}{2}\sum_{k=1}^{N}(e_k-\tilde{c}_k)^2$$

$$\text{s.t.}\quad \frac{1}{L-1}\sum_{l|k\in\mathscr{T}_l}{w^{(l)}}^T\varphi(x_k)+b+e_k=y_k,\ \ \forall k=1,\dots N. \tag{8.24}$$

The Lagrangian of this constrained optimization problem becomes

$$\mathscr{L}^{(cv)}(w^{(l)},b,e_k;\alpha_k)=\frac{1}{2}\sum_{k=1}^{N}(e_k-\tilde{c}_k)^2+\frac{1}{2}\sum_{l=1}^{L}\frac{{w^{(l)}}^T w^{(l)}}{L-1}$$

$$-\sum_{k=1}^{N}\alpha_k\left(\frac{1}{L-1}\sum_{l|i\in\mathscr{T}_l}{w^{(l)}}^T\varphi(x_k)+b+e_k-y_k\right). \tag{8.25}$$

The conditions for optimality w.r.t. $w^{(l)},b,e_k,\alpha_k$ for all i,l for the training become:

$$\begin{cases}\partial\mathscr{L}^{(cv)}/\partial e_k=0 & \rightarrow\quad e_k=\tilde{c}_k+\alpha_k \qquad (a)\\ \partial\mathscr{L}^{(cv)}/\partial w^{(l)}=0 & \rightarrow\quad w^{(l)}=\sum_{i\in\mathscr{T}_l}\alpha_k\varphi(x_k) \qquad (b)\\ \partial\mathscr{L}^{(cv)}/\partial b=0 & \rightarrow\quad \sum_{k=1}^{N}\alpha_k=0 \qquad (c)\\ \partial\mathscr{L}^{(cv)}/\partial \alpha_k=0 & \rightarrow\quad \sum_{l|i\in\mathscr{T}_l}{w^{(l)}}^T\varphi(x_k)+b+e_k=y_k. \qquad (d)\end{cases} \tag{8.26}$$

From (8.26.b) one can recover

$$\sum_{l|i\in\mathscr{T}_l}w^{(l)}=\sum_{l|i\in\mathscr{T}_l}\sum_{i\in\mathscr{T}_l}\alpha_k\varphi(x_k)=(L-1)\sum_{k=1}^{N}\alpha_k\varphi(x_k)+\sum_{j\in\mathscr{V}_l}\alpha_j\varphi(x_j). \tag{8.27}$$

After elimination of the variables $w^{(l)}$ and $\tilde{c}$, the dual problem becomes:

$$\left[\begin{array}{c|c}0 & 1_N^T\\ \hline 1_N & \Omega+\frac{1}{L-1}\Omega^{(cv)}\end{array}\right]\left[\begin{array}{c}b\\ \hline \alpha\end{array}\right]+\left[\begin{array}{c}0\\ \hline e\end{array}\right]=\left[\begin{array}{c}0\\ \hline Y\end{array}\right] \tag{8.28}$$

with

$$\Omega^{(cv)} = \begin{bmatrix} \Omega^{\mathscr{V}_1} & & & \\ & \Omega^{\mathscr{V}_2} & & \\ & & \ddots & \\ & & & \Omega^{\mathscr{V}_L} \end{bmatrix} \tag{8.29}$$

and $\Omega^{\mathscr{V}_l} \in \mathbb{R}^{n^{(l)} \times n^{(l)}}$ is the kernel matrix between elements of the validation set of the lth fold $\Omega^{\mathscr{V}_l}_{i,j} = K(x_i, x_j), \forall i, j \in \mathscr{V}_l$. From (8.26.b) one can recover an expression for the individual models of the different folds such that the l-th model can be evaluated in point x^v_j for $j \in \mathscr{V}_l$ as

$$y^v_j = \left(\hat{w}^{(l)}\right)^T \varphi(x^v_j) + \hat{b} + e^v_j = \sum_{k \in \mathscr{T}_l} \hat{\alpha}_k K(x_k, x^v_j) + \hat{b} + e^v_j \tag{8.30}$$

with residual $\hat{f}^{(l)}(x^v_j) - y^v_j$ denoted as e^v_j and $\hat{\alpha}$ and $\hat{b}$ solve (8.28). In matrix notation, conditions (8.28) and (8.30) can be written as

$$\mathrm{KKT}_{(8.24)}(\alpha, b, e, e^v) = \left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \frac{L-1}{L}\Omega + \frac{1}{L}\Omega^{(cv)} \\ \hline 0_N & \frac{L+1}{L}\Omega^{(cv)} - \frac{1}{L}\Omega \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] + \left[\begin{array}{c} 0 \\ \hline e \\ \hline e - e^v \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline Y \\ \hline 0_N \end{array}\right]. \tag{8.31}$$

The fusion of the training equations (8.28) and the validation set of equations (8.30) results in the following constrained optimization problem

$$\textbf{Fusion}: \ (\hat{c}, \hat{\alpha}, \hat{b}) = \arg\min_{c,\alpha,b} \sum_{k=1}^N e_k^2 + \sum_{k=1}^N (e_k - e^v_k)^2 \quad \text{s.t.} \quad \mathrm{KKT}_{(8.24)}(\alpha, b, e, e^v) \ \text{ holds.} \tag{8.32}$$

The estimated average model can then be evaluated in a new point x^* as

$$\hat{f}^{(cv)}(x^*) = \sum_{l | k \in \mathscr{T}_l} {w^{(l)}}^T \varphi(x^*) = \sum_{k=1}^N \hat{\alpha}_k \varphi(x_k) + \hat{b}, \tag{8.33}$$

where $\hat{\alpha}$ and $\hat{b}$ are the solutions to (8.32).

Example 8.2 **[Numerical Comparison of Different Kernel based Fusion Schemes]** A numerical comparisons of the different fusion schemes was reported in (Pelckmans *et al.*, 2003*b*). Table 8.1 gives results of numerical experiments on regression benchmark datasets with the Tikhonov regularization based LS-SVMs (tuned for γ using validation (Val) and cross-validation (CV)) and the LS-SVMs with additive regularization trade-off (AReg) (tuned for λ with validation and cross-validation). For the latter, results are given based on the full implementation (Subsection 8.2.2) and the fast implementation (Subsection 8.2.3). Results of two artificial datasets (a two-dimensional linear function and the *sinc* function) are given. The size of the training, validation and noise free test set were 30, 20, 500, respectively. Cross-validation based tuning procedures were provided with the joint training-validation dataset. Data generation, training and testing were repeated 1000 times. Performance is measured in average mean squared error (Mean(MSE)) and standard deviation (Std(MSE)) of the predictions on the test set which is fixed a priori in the different randomizations. Additionally, the techniques were compared on two benchmark data sets from the UCI Machine Learning Repository, the Abalone data ($N = 700, n = 500, n_{test} = 2977$ and $d = 7$) and the Boston housing dataset ($N = 220, n = 120, n_{test} = 166$, $d = 11$). Data division in training and validation set, tuning, training and testing were repeated respectively 100 and 1000 times. The results show also an increased performance in the case of the first two experiments using the full implementation of AReg LS-SVM based on 10-fold cross-validation. According to the Wilcoxon Rank Sum test, the test set performance is even significantly better using the AReg (CV) LS-SVM for the first two toy examples.

8.3 Stable Kernel Machines

Stability analysis in general aims at determining how much a variation of the formulation (data) influences the estimate of an algorithm. This notion is used in many different domains (numerical, robust statistics, control theory) under different denominators (e.g. sensitivity, perturbation, influence or conditioning). The more specific definition of stability of a learning algorithm defined in e.g. (Devroye *et al.*, 1996; Bousquet and Elisseeff, 2002) is used here. Originally, it was proposed for the estimation of the accuracy of learning algorithms itself by revealing the connection between stability and generalization error (Devroye *et al.*, 1996). In particular, one can derive (Bousquet and Elisseeff, 2002) a bound on the generalization error or risk functional based on an observed quantitative measure of stability. Although many subtle differences exist between different definitions (one distinguishes amongst others between (pointwise) hypothesis, error or uniform stability), this section only works with the two concepts of uniform α and β stability as they are most clearly put within an optimization point of view. Uniform stability was used to derive exponential bounds for different algorithms, including techniques for unsupervised learning (*k*-nearest neighbor), classification (soft margin SVMs) and regression (Regularized least squares regression and LS-SVMs). While in previous papers about stability, the object of interest was the learning algorithm itself (Bousquet and Elisseeff, 2002), the context of hierarchical programming problems and LS-SVM substrates may be used to formulate a constructive approach.

LS-SVM	**Tuned**		**AReg**		
	Val	**CV**	**Val**	**CV**	**Fast CV**
linear regression $(30, 20, 500)$					
Mean(MSE):	0.5887	0.5931	0.5887	**0.3796**	0.5858
Std(MSE):	0.5108	0.5125	0.5108	0.4069	0.5074
sinc $(30, 20, 500)$					
Mean(MSE):	0.0289	0.0269	0.0286	**0.0174**	0.0240
Std(MSE):	0.0217	0.0185	0.0210	0.0086	0.0145
Abalone $(700, 500, 2977)$					
Mean(MSE):	4.6609	4.8502	4.6622	5.0258	4.6216
Std(MSE):	0.1188	0.2311	0.1164	0.1808	0.0952
Computation time (s):	67.81	126.39	10.672	1401.6	19.28
Boston Housing $(220, 120, 166)$					
Mean(MSE):	0.1815	0.1883	0.1814	0.1874	**0.1260**
Std(MSE):	0.0491	0.0523	0.0500	0.0446	0.0262
Computation time (s):	0.1199	9.1834	0.0732	10.0728	1.0195

Table 8.1: *Numerical results from the experiment described in Example 8.2. The mean and the standard deviation of the test-set performance of* 100 *randomizations of the respective datasets are given. These results suggest that the fusion argument does not affect the generalization performance while avoiding the need for non-convex and time consuming line searches.*

8.3.1 Stable regressors

While most stability criteria of learning machines take a form based on the difference in loss between the training and leave-one-out error, a common relaxed version called α-stability can be taken

$$\left| e_k - e_k^{(v)} \right| \leq \alpha_{\mathscr{S}} \quad \forall k = 1, \ldots, N. \tag{8.34}$$

This is considered as a measure for measuring the performance of learning machines and used to derive bounds on the generalization abilities. Here, we use it as a special form of regularization. Imposing $\alpha_{\mathscr{S}}$-stability on additively regularized (AReg) LS-SVMs boils down to a quadratic programming problem

$$\left(\hat{\alpha}^{(l)}, \hat{c}, \hat{e}^{(l)}, \hat{e}^{(l)v}\right) = \underset{\alpha^{(l)}, c, e^{(l)}, e^{(l)v}}{\arg\min} \mathscr{J}_{\alpha_{\mathscr{S}}} \frac{1}{2L} \sum_{l=1}^{L} e^{(l)v^T} e^{(l)v}$$
$$\text{s.t.} \quad \begin{cases} \max\limits_{l \neq h} \max\limits_{j \in \mathscr{V}_h} \left| e_j^{(l)} - e_j^{(h)v} \right| \leq \alpha_{\mathscr{S}} & \forall h = 1, \ldots, L \\ \mathrm{KKT}_{(8.20)} \left(\alpha^{(l)}, c, e^{(l)}, e^{(l)v}\right). & \forall l = 1, \ldots, L \end{cases} \tag{8.35}$$

Note the huge number of unknowns into the formulation which occur already when N has a moderate size. To cure this disadvantage, the fast CV formulation may be used instead

$$\min_{\alpha^{(l)}, c, e^{(l)}, e^{(l)v}} J_{\alpha_{\mathscr{S}}} = \frac{1}{2L} \sum_{l=1}^{L} e^{(l)v^T} e^{(l)v} \quad \text{s.t.} \quad \begin{cases} \mathrm{KKT}^{l}_{(8.20)} \left(\alpha^{(l)}, c, e^{(l)}, e^{(l)v}\right) & \forall l = 1, \ldots, L, \\ \max\limits_{l} \max\limits_{i \in \mathscr{V}_l} \left| e^{(l)} - e^{(l)v} \right| \leq \alpha_{\mathscr{S}}. \end{cases} \tag{8.36}$$

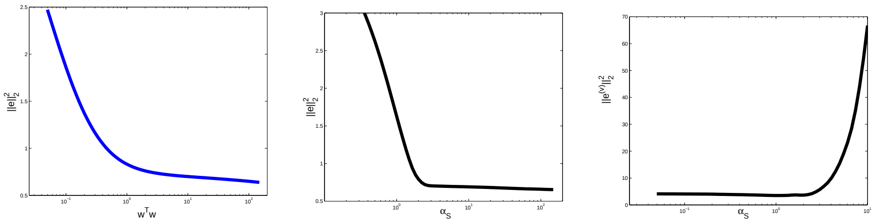
8.3.2 Stability L-curves

One can visualize the trade-off between stability and loss in a graph by exploring the solutions for a range of values of $\alpha_{\mathscr{S}}$. We shall refer to this graph as the L_α-curve, analogously to the L-curve (Hansen, 1992; Neumaier, 1998; Golub and van Loan, 1989) displaying the trade-off between bias and variance (see Figure 8.4).

Example 8.3 This experiments focus on the choice of the regularization scheme in kernel based models. For the design of a Monte-Carlo experiment, the choice of the kernel and kernel-parameter should not be of critical importance. To randomize the design of the underlying functions in the experiment with known kernel-parameter, the following class of functions is considered

$$f(\cdot) = \sum_{k=1}^{N} \bar{\alpha}_k K(x_k, \cdot) \tag{8.37}$$

where the input points x_k are equidistantly taken between 0 and 5 for all $k = 1, \ldots, N$ with $N = 75$ and $\bar{\alpha}_k$ is an i.i.d. uniformly randomly generated term. The kernel is fixed



(a) (b) (c)

Figure 8.4: *The toy problem as described in Section 4 was used to generate the following figures:* **(a)** *Classical L-curve of the regularization parameter γ in (3.12) with respect to the training error;* **(b)** *The L_α curve visualizing the trade-off between fitting error $\|e\|_2^2$ and the α upper bound of the stability measure;* **(c)** *The curve visualizing a typical relationship between the performance of the leave-one-out performance and the α upper bound of the stability measure.*

as $K(x_k, x_j) = \exp(-\|x_k - x_j\|_2^2)$ for all $i, j = 1, \ldots, N$. Output data points points were generated as $y_k = f(x_k) + e_k$ for $k = 1, \ldots, N$ where e_k are N i.i.d. samples of a Gaussian distribution.

Given this method to generate datasets with a prefixed kernel, a Monte Carlo study was conducted to relate the designed algorithms in a practical way as reported in Figure 8.5.

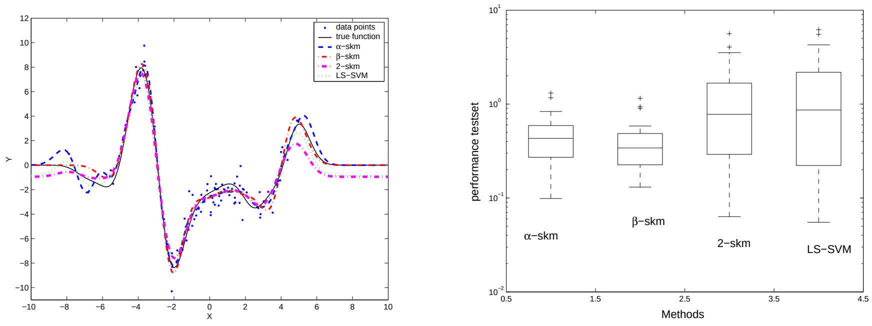
8.4 Hierarchical Kernel Machines

The idea of hierarchical programming and fusion of training and model selection levels was used to formalize an hierarchical modeling strategy.

8.4.1 Alternative training criteria

Sometimes the designers assumptions and optimality criteria do not allow for straightforward primal-dual derivations or do result in a number of unknowns (Lagrange multipliers) which makes the approach less practical. Consider e.g. the case of structure detection as elaborated in Section 6.4.

Sparseness is often regarded as good practice in the machine learning community (Vapnik, 1998; von Luxburg *et al.*, 2004) as it gives an optimal solution with a minimal representation (from the viewpoint of VC theory and compression). The primal-dual framework also provides another motivation for trying to sparsify the support values based on sensitivity analysis. The optimal Lagrange multipliers $\hat{\alpha}$ contain



(a)

(b)

Figure 8.5: *Results from numerical experiments with the data generating mechanism as described in Section 4.* **(a)** *Result of the α-stable, β-stable, 2-norm (8.32) and standard LS-SVM on a particular realization of the dataset.* **(b)** *Boxplot of the obtained accuracy obtained on a testset on a Monte-carlo study of the different methods for randomly generated functions according to equation (8.37).*

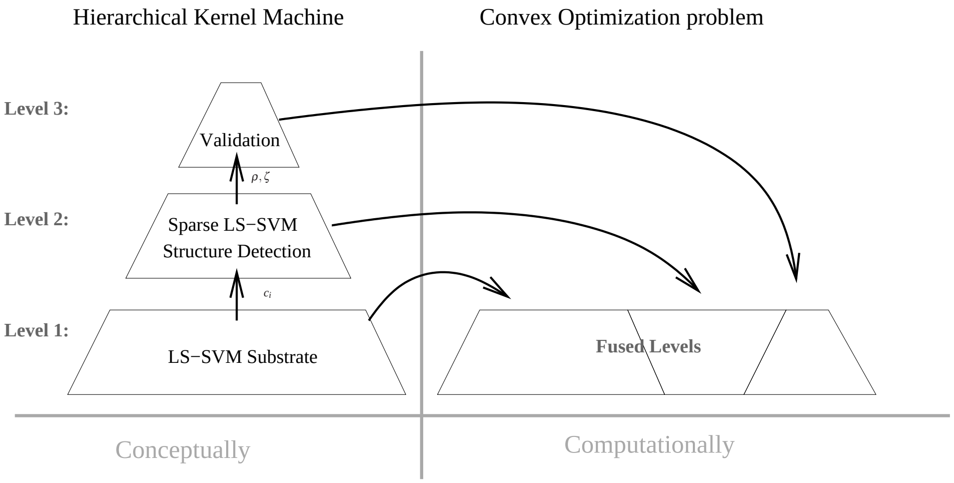


Figure 8.6: *Schematic representation of an hierarchical kernel machine. Conceptually, one formulates the problem of substrates (level 1), modeling (level 2) and model selection (level 3) on different levels. Interaction of the levels is guided by a proper set of hyper-parameters. Computationally, the different levels are treated as an hierarchical programming problem employing the KKT conditions to impose the conceptual structure.*

information of how much the (dual) optimal solution changes when the corresponding constraints are perturbed, see Subsection 3.3.3. In this respect, one can design a kernel machine that minimizes its own sensitivity to model mis-specifications or atypical data observations by minimizing an appropriate norm on the Lagrange multipliers. Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex and differentiable loss-function. The 1-norm is considered

$$\min_{e,\alpha,b,c} \sum_{i=1}^{N} \ell(e_i) + \zeta \|\alpha\|_1 \quad \text{s.t.} \quad \text{KKT}_{(8.10)}(\alpha, e; c) \ \text{ hold}, \tag{8.38}$$

where $0 < \zeta \in \mathbb{R}$ acts as a hyper-parameter. This criterion leads to sparseness (Vapnik, 1998) and was studied in (Pelckmans *et al.*, 2004*e*).

As already hinted at in Subsection 6.4.2, the current framework may be used to obtain a much more practical formulation to the problem of structure detection for componentwise kernel models using the measure of maximal variation. The kernel machine for structure detection minimizes the following criterion for a given tuning constant $0 < \rho \in \mathbb{R}$:

$$\min_{e,t_p,\alpha,b,c} \sum_{i=1}^{N} \ell(e_i) + \rho \sum_{p=1}^{P} t_p \quad \text{s.t.} \quad \begin{cases} \text{KKT}_{(8.10)}(\alpha, e; c) \ \text{ hold} & \text{with } \ \Omega = \sum_{p=1}^{P} \Omega^{(p)} \\ -t_p 1_N \leq \Omega^{(p)} \alpha \leq 1_N t_p, & \forall p = 1, \ldots, P \end{cases} \tag{8.39}$$

which has a unique minimum and can be solved efficiently when ℓ is convex.

8.4.2 Finishing it all up: fusion with validation

As argued in Chapter 7, the automatic tuning of the hyper-parameter ρ in (8.39) or ζ in (8.38) of the second level with respect to an appropriate model selection criterion is highly desirable, at least in practice. A similar approach with respect to a validation criterion using a third level of inference. This three level architecture constitutes the hierarchical kernel machine. The LS-SVM substrate constitute the first level, while the sparse LS-SVM and the LS-SVM for structure detection makes up the second level. The validation performance is used to tune the hyper-parameters ζ (or ρ) on a third level.

A third level is added to the LS-SVM for structure detection in order to tune the hyper-parameter ζ of the second level where one chooses $\ell(e) = e^2$. Figure 8.8 summarizes the derivation below and points out the hierarchical approach. Reconsider the problem (8.39) where ρ acts as a hyper-parameter. One can eliminate e and c from this optimization problem leading to

$$\min_{t,\alpha} \mathscr{J}_\rho(\alpha, t) = \frac{1}{2} \|\Omega^P \alpha - y\|_2^2 + \rho \sum_{p=1}^{P} t_p \quad \text{s.t.} \quad -t_p 1_N \leq \Omega^{(p)} \alpha \leq t_p 1_N, \ \forall p = 1, \ldots, P. \tag{8.40}$$

Let ξ^{+p} and $\xi^{-p} \in \mathbb{R}^{+,N}$ for all $p = 1, \ldots, P$ be multipliers of the Lagrangian. The

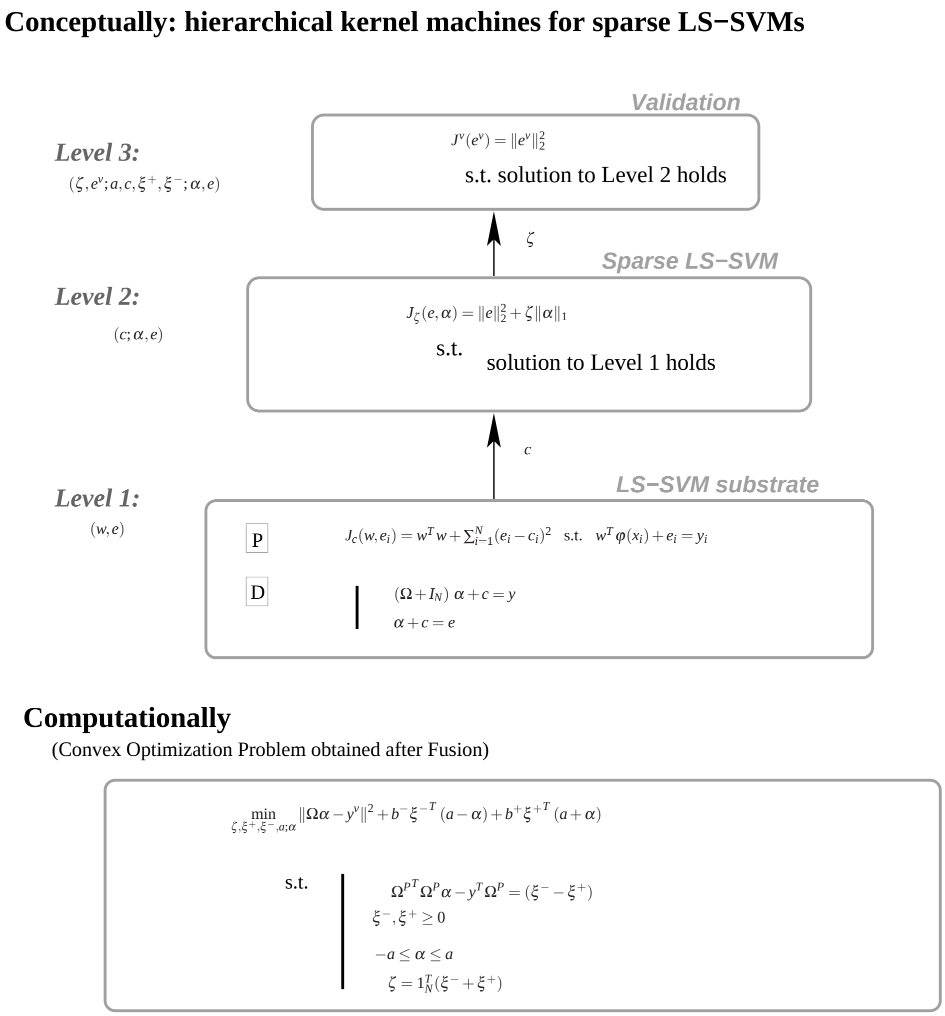

Conceptually: hierarchical kernel machines for sparse LS−SVMs
Validation
Level 3:
$(\zeta, e^v; a, c, \xi^+, \xi^-; \alpha, e)$
$J^v(e^v) = \|e^v\|_2^2$
s.t. solution to Level 2 holds
ζ
Sparse LS−SVM
Level 2:
$(c; \alpha, e)$
$J_\zeta(e, \alpha) = \|e\|_2^2 + \zeta\|\alpha\|_1$
s.t.
solution to Level 1 holds
c
LS−SVM substrate
Level 1:
(w, e)
P
$J_c(w, e_i) = w^T w + \sum_{i=1}^N (e_i - c_i)^2$ s.t. $w^T \varphi(x_i) + e_i = y_i$
D
$(\Omega + I_N)\ \alpha + c = y$
$\alpha + c = e$
Computationally
(Convex Optimization Problem obtained after Fusion)
$\min_{\zeta, \xi^+, \xi^-, a; \alpha} \|\Omega\alpha - y^v\|^2 + b^- \xi^{-T} (a - \alpha) + b^+ \xi^{+T} (a + \alpha)$
s.t.
$\Omega^{P^T} \Omega^P \alpha - y^T \Omega^P = (\xi^- - \xi^+)$
$\xi^-, \xi^+ \geq 0$
$-a \leq \alpha \leq a$
$\zeta = 1_N^T (\xi^- + \xi^+)$

Figure 8.7: *Schematical representation of the hierarchical kernel machine for sparse representations. From a conceptual point of view, inference is done at different levels and interaction is guided via a set of hyper-parameters. The first level constitutes of an LS-SVM substrate. On the second level, inference of the c_i is defined in terms of a cost function inducing sparseness, while ζ is optimized on a third level using a validation criterion.*

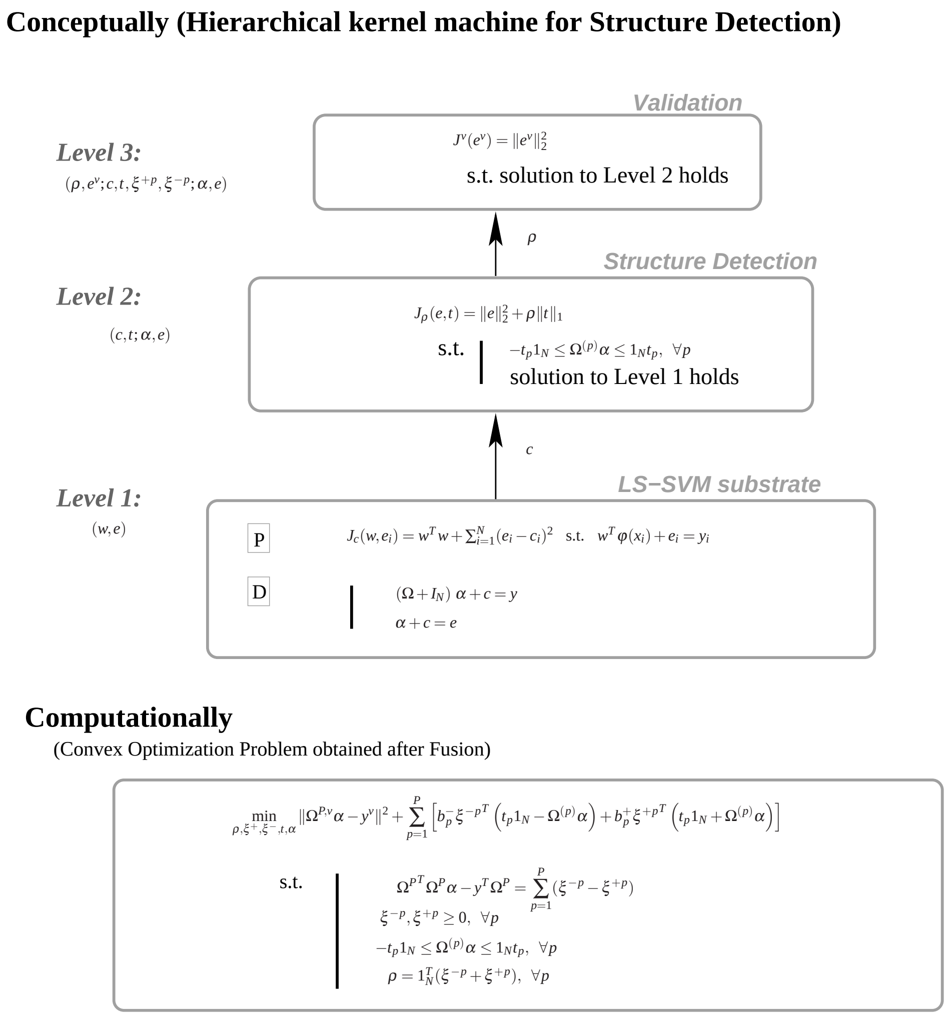

Conceptually (Hierarchical kernel machine for Structure Detection)
Validation
Level 3:
$(\rho, e^v; c, t, \xi^{+p}, \xi^{-p}; \alpha, e)$
$J^v(e^v) = \|e^v\|_2^2$
s.t. solution to Level 2 holds
ρ
Structure Detection
Level 2:
$(c, t; \alpha, e)$
$J_\rho(e,t) = \|e\|_2^2 + \rho\|t\|_1$
s.t.
$-t_p 1_N \leq \Omega^{(p)}\alpha \leq 1_N t_p, \ \forall p$
solution to Level 1 holds
c
LS−SVM substrate
Level 1:
(w, e)
P
$J_c(w, e_i) = w^T w + \sum_{i=1}^N (e_i - c_i)^2 \quad \text{s.t.} \quad w^T \varphi(x_i) + e_i = y_i$
D
$(\Omega + I_N)\ \alpha + c = y$
$\alpha + c = e$
Computationally
(Convex Optimization Problem obtained after Fusion)
$\min_{\rho, \xi^+, \xi^-, t, \alpha} \|\Omega^{P,v}\alpha - y^v\|^2 + \sum_{p=1}^P \left[b_p^- \xi^{-p^T} \left(t_p 1_N - \Omega^{(p)}\alpha \right) + b_p^+ \xi^{+p^T} \left(t_p 1_N + \Omega^{(p)}\alpha \right) \right]$
s.t.
$\Omega^{P^T}\Omega^P\alpha - y^T\Omega^P = \sum_{p=1}^P (\xi^{-p} - \xi^{+p})$
$\xi^{-p}, \xi^{+p} \geq 0, \ \forall p$
$-t_p 1_N \leq \Omega^{(p)}\alpha \leq 1_N t_p, \ \forall p$
$\rho = 1_N^T(\xi^{-p} + \xi^{+p}), \ \forall p$

Figure 8.8: *Schematical representation of the hierarchical kernel machine for structure detection. On the second level, inference of the c_i is expressed in terms of a least squares cost function with a minimal amount of maximal variation, while ρ is optimized on a third level using a validation criterion.*

corresponding Karush-Kuhn-Tucker conditions then become

$$\mathrm{KKT}_\rho(\alpha, t; \xi^+, \xi^-) = \left\{ \begin{array}{lll} \Omega^{P^T}\Omega^P\alpha - y^T\Omega^P = \sum_{p=1}^P(\xi^{-p} - \xi^{+p}) & & (a) \\ \rho = 1_N^T(\xi^{-p} + \xi^{+p}) & \forall p = 1, \ldots, P & (b) \\ \xi^{+p}, \xi^{-p} \geq 0 & \forall p = 1, \ldots, P & (c) \\ -t_p 1_N \leq \Omega^{(p)}\alpha \leq t_p 1_N & \forall p = 1, \ldots, P & (d) \\ \xi_i^{-p}(t_p + \Omega_i^{(p)}\alpha) = 0, \ \forall i = 1, \ldots, N & \forall p = 1, \ldots, P & (e) \\ \xi_i^{+p}(t_p - \Omega_i^{(p)}\alpha) = 0, \ \forall i = 1, \ldots, N, & \forall p = 1, \ldots, P & (f) \end{array} \right. \tag{8.41}$$

The problem of fusion then becomes

$$\textbf{Fusion:} \quad \min_{\rho, t, \alpha, \xi^-, \xi^+} \mathscr{J}^v = \frac{1}{2}\|\Omega^{P,v}\alpha - y^v\|_2^2 \quad \text{s.t.} \quad \mathrm{KKT}_\rho(\alpha, t; \xi^+, \xi^-) \tag{8.42}$$

where $\Omega^{P,v} \in \mathbb{R}^{n \times N} = \sum_{p=1}^P \Omega^{(p),v}$ and $\Omega_{ij}^{(p),v} = K^p\left(x_i^{(p)}, x_j^{(p),v}\right)$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, N$. The problem (8.42) is convex up to the complementary slackness constraints (8.41.ef) which belong to the class of positive OR constraints, see also Subsection 2.4.3.

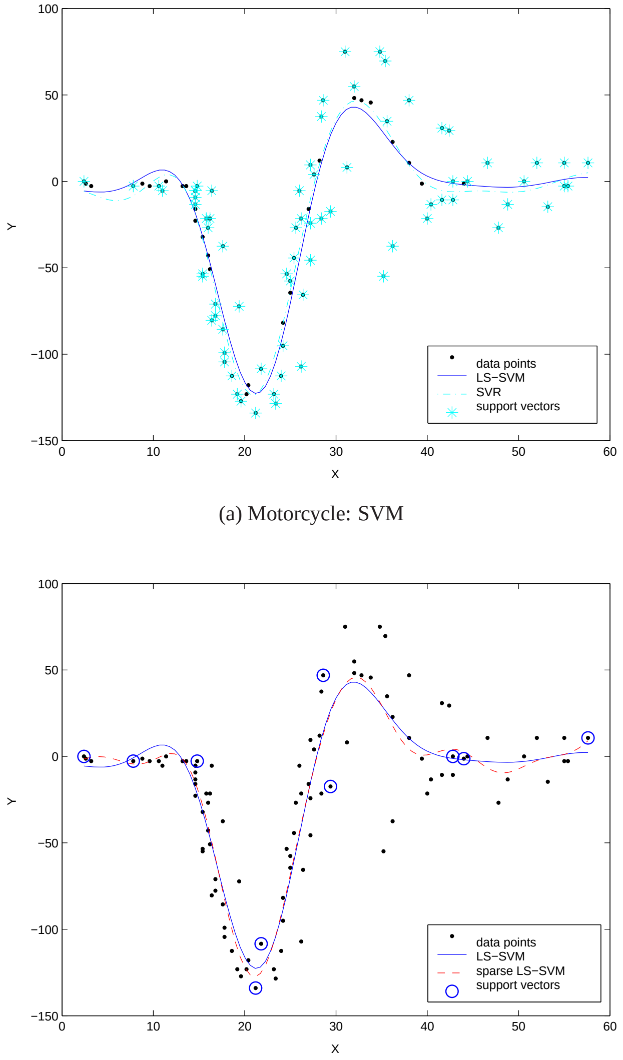
The estimated model can be evaluated at new data points $x_* \in \mathbb{R}^d$ as

$$\hat{f}(x^*) = \hat{w}^T\varphi(x^*) = \sum_{i=1}^N \hat{\alpha}_i \sum_{t_p \neq 0} K^p\left(x_i^{(p)}, x_*^{(p)}\right), \tag{8.43}$$

where $\hat{\alpha}$ and $\hat{t}_p$ are solutions to (8.42).

Example 8.4 **[Numerical Results of Sparse LS-SVMs]** The performance of the proposed sparse LS-SVM substrate was measured on a number of regression and classification datasets, respectively an artificial dataset sinc (generated as $Y = \mathrm{sinc}(X) + e$ with $e \sim \mathscr{N}(0, 0.1)$ and $N = 100$, $d = 1$) and the motorcycle dataset (Eubank, 1999) ($N = 100$, $d = 1$) for regression (see Figure 8.9), the artificial Ripley dataset ($N = 250$, $d = 2$) (see Figure 8.10) and the PIMA dataset ($N = 468$, $d = 8$) from UCI at classification problems. The models resulting from sparse LS-SVM substrates were tested against the standard SVMs and LS-SVMs where the kernel parameters and the other tuning-parameters (respectively C, ε for the SVM, γ for the LS-SVM and ξ for sparse LS-SVM substrates) were obtained from 10-fold cross-validation (see Table 8.2).

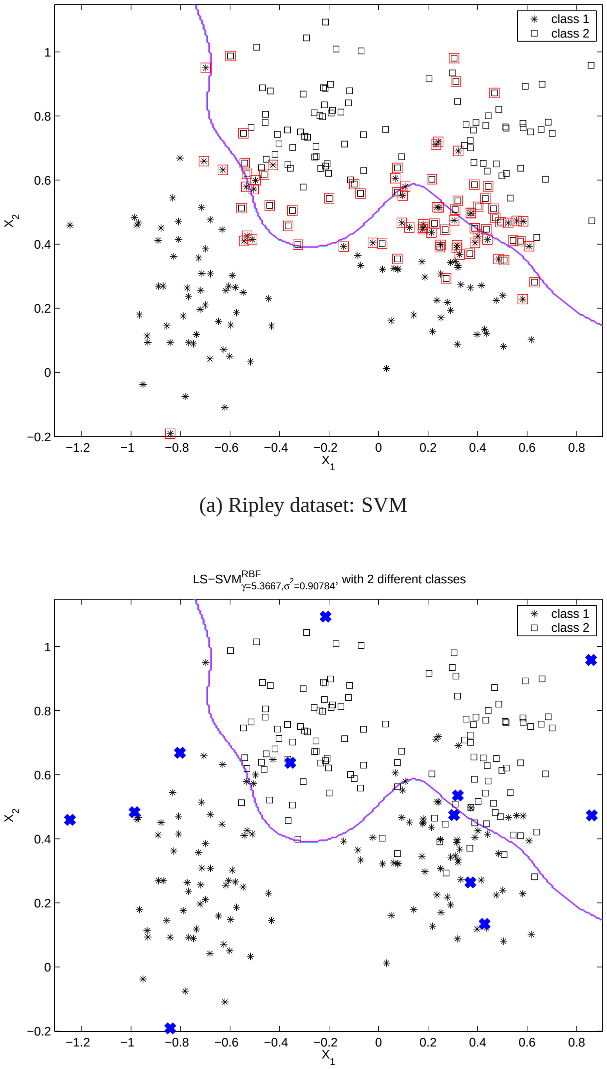
Example 8.5 **[Numerical Results of Structure Detection]** An artificial example is taken from (Vapnik, 1998) and the Boston housing dataset from the UCI benchmark repository was used for analyzing the practical relevance of the structure detection mechanism. This



(a) Motorcycle: SVM

(b) Motorcycle: sparse LS-SVM substrate

Figure 8.9: *Comparison of the SVM, LS-SVM and sparse LS-SVM substrate of subsection 8.4.1 on the Motorcycle regression dataset. One sees the difference in selected support vectors of* **(a)** *a standard SVM and* **(b)** *a sparse hierarchical kernel machine.*



(a) Ripley dataset: SVM

(b) Ripley dataset: sparse LS-SVM substrate

Figure 8.10: *Comparison of the SVM, LS-SVM and sparse LS-SVM substrate of subsection 8.4.1 on the Ripley classification dataset. One can see the difference in selected support vectors of* **(a)** *a standard SVM and* **(b)** *a sparse hierarchical kernel machine. The support vectors of the former concentrate around the margin while the sparse hierarchical kernel machine will provide a more global support.*

	SVM		**LS-SVM**	**Sparse LS-SVM substr.**	
	MSE	**Sparse**	**MSE**	**MSE**	**Sparse**
Sinc	0.0052	68%	0.0045	0.0034	9%
Motorcycle	516.41	83%	444.64	469.93	11%
	PCC	**Sparse**	**PCC**	**PCC**	**Sparse**
Ripley	90.10%	33.60%	90.40%	90.50%	4.80%
Pima	73.33%	43%	72.33%	74%	9%

Table 8.2: *Performances of SVMs, LS-SVMs and the sparse LS-SVM substrates of Subsection 8.4.1 expressed in Mean Squared Error (MSE) on a test set in the case of regression or Percentage Correctly Classified (PCC) in the case of classification. Sparseness is expressed in percentage of support vectors w.r.t. number of training data. The kernel machines were tuned for the kernel parameter and the respective hyper-parameters C,ε; γ and ζ with 10-fold cross-validation. These results indicate that sparse LS-SVM substrates are at least comparable in generalization performance with existing methods, but are often more effective in achieving sparseness.*

subsection considers the formulation from Subsection 8.4.1, where sparseness amongst the components is obtained by use of the sum of maximal variation. The performance on a validation set was used to tune the parameter ρ both via a naive line-search as well as using the method which is described in Subsection 8.4.2.

Figure 8.11 shows results obtained on an artificial dataset consisting of 100 samples and dimension 25, uniformly sampled from the interval $[0,1]^{25}$. The underlying function takes the following form:

$$f(x) = 10\,\sin(X^1) + 20\,(X^2 - 0.5)^2 + 10\,X^3 + 5\,X^4 \tag{8.44}$$

such that $y_i = f(x_i) + e_i$ with $e_i \sim \mathcal{N}(0,1)$ for all $i = 1,\ldots,100$. Figure 8.11 gives the nontrivial components ($t_p > 0$) associated with the LS-SVM substrate with ρ optimized in validation sense. Figure 8.12 presents the evolution of values of t when ρ is increased from 1 to 1000 in a maximal variation evolution diagram (similarly as used for LASSO (Hastie *et al.*, 2001)).

The Boston housing dataset was taken from the UCI benchmark repository. This dataset concerns the housing values in suburbs of Boston. The dependent continuous variable expresses the median value of owner-occupied homes. From 13 given inputs, an additive model was build using the mechanism of maximal variation for detection of which input variables have a non-trivial contribution. 250 data-points were used for training purposes and 100 were randomly selected for validation. The analysis works with standardized data (zero mean and unit variance), while results are expressed in the original scale. The structure detection algorithm as proposed in Subsection 8.4.1 was used to construct

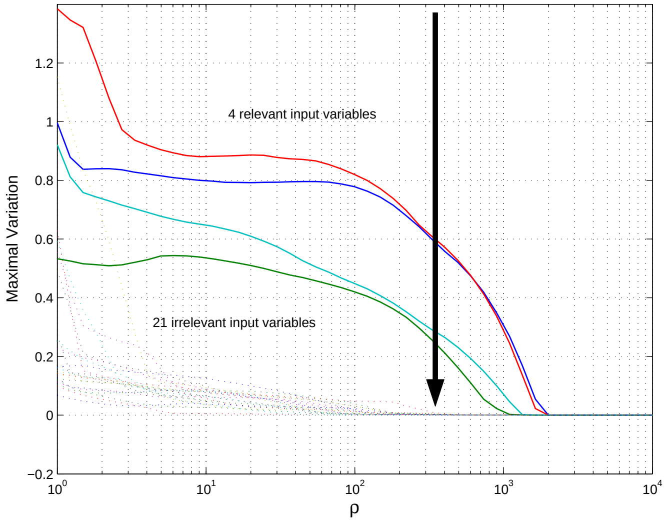


Figure 8.11: *Results of structure detection on an artificial dataset as used in (Vapnik, 1998), consisting of 100 data samples generated by four componentwise non-zero functions of the first 4 inputs and 21 irrelevant inputs and perturbed by i.i.d. unit variance Gaussian noise. This diagram shows the evolution of the maximal variations per component when increasing the hyper-parameter ρ from 1 to 10000. The black arrow indicates a value ρ corresponding with a minimal cross-validation performance. Note that for the corresponding value of ρ, the underlying structure is indeed detected successfully.*

the maximal variation evolution diagram. Figure 8.13 displays the contributions of the individual components. The performance on the validation dataset was used to tune the kernel parameter and ρ. The latter was determined both manually (by a line-search) as automatically by fusion as described in Subsection 8.4.2. For the optimal parameter ρ, the following inputs have a maximal variation of zero:

1 CRIM: per capita crime rate by town,

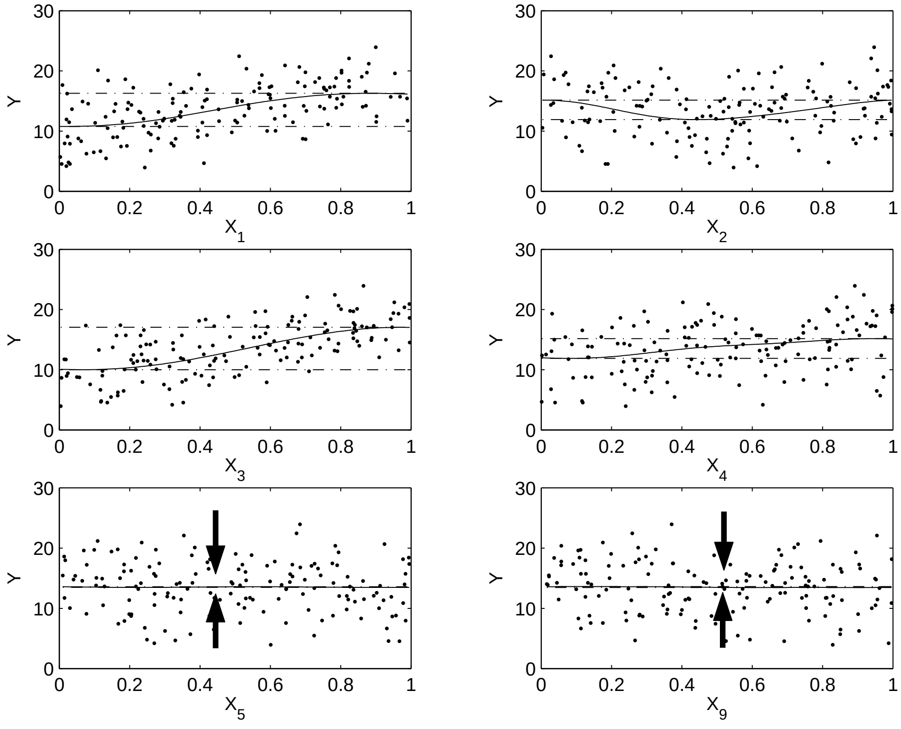
2 ZN: proportion of residential land zoned for lots over $25,000$ sq.ft.,

4 CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise),

10 TAX: full-value property-tax rate per $10,000$,

12 B: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks.

Testing was done by retraining a componentwise LS-SVM based on only the selected inputs. The resulting additive model increases in performance expressed in MSE on an independent test-set with 22%. The improvement is even more significant (32%) with respect to a standard nonlinear LS-SVM model with an RBF-kernel.



(a)

Figure 8.12: *Results of structure detection on an artificial dataset as used in (Vapnik, 1998), consisting of 100 data samples generated by four componentwise non-zero functions of the first 4 inputs and 21 irrelevant inputs and perturbed by i.i.d. unit variance Gaussian noise. The resulting nontrivial components ($t_p > 0$) associated with the LS-SVM substrate with ρ optimized in validation sense.*

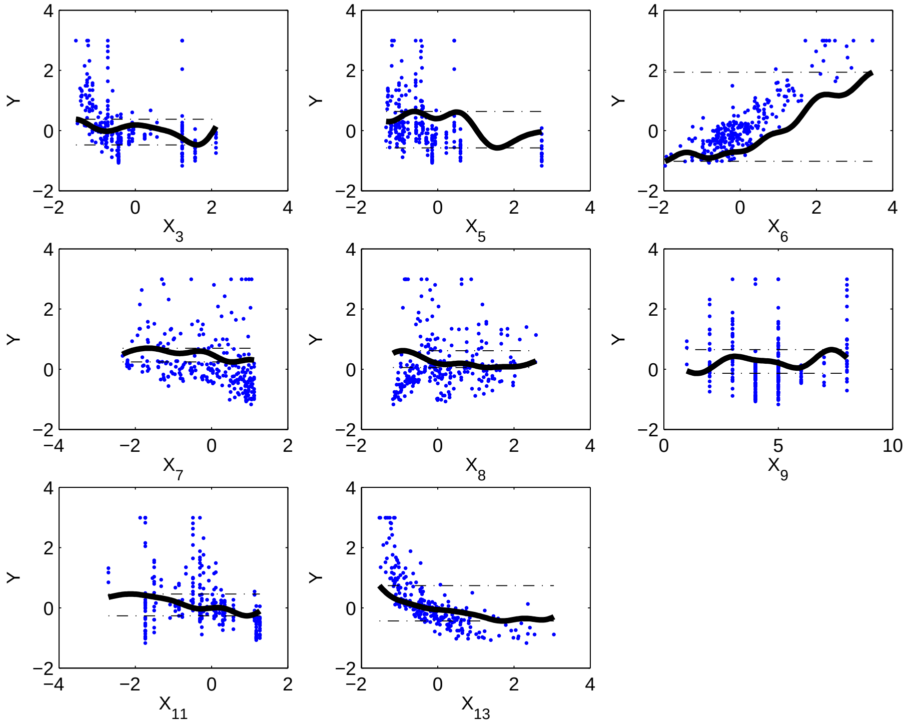


Figure 8.13: *Results of structure detection on the Boston housing dataset consisting of 250 training, 100 validation and 156 randomly selected testing samples. The contributions of the variables which have a non-zero maximal variation are shown. The fusion argument as described in Subsection 8.4.2 was used to tune the parameter* ρ.

Part III

σ

Chapter 9

Kernel Representations & Decompositions

The generalization performance of kernel machines in general often depends crucially on the choice of the (shape of the) kernel and its parameters. The following chapter shows the relationship between the issue of regularization and the choice of the kernel. Furthermore, the idea of kernel decompositions is proposed to approach the problem of the choice of the kernel. Finally, relations with techniques from the field of system identification are elaborated. Given observed second moments, the task of stochastic realization amounts to finding those internal (kernel) structures effectively realizing this empirical characterization. This results in a tool which can assist the user in the decision for a good (shape of the) kernel. Section 9.1 introduces a formal argument relating the regularization scheme and a weighting term in the loss function respectively with the form of the kernel using a primal-dual argument. Then Section 9.2 proceeds with the elaboration of a method for searching compact kernel decompositions based on the method of maximal variation. Section 9.4 then discusses a method for recovering the shape of the kernel from the observed second order moments in the univariate case and is also extended to the multivariate case.

9.1 Duality between regularization and kernel design

9.1.1 Duality between kernels and regularization scheme

A classical result in the theory of smoothing splines (Wahba, 1990) can be cast in the more general context of kernels using a primal-dual argument.

Theorem 9.1. [Duality between Regularization and Kernel Design] *Let $\varphi : \mathbb{R}^D \to \mathbb{R}^{D_\varphi}$ be a fixed mapping where $D_\varphi \in \{\mathbb{N}, +\infty\}$. Consider the class of models (3.8) given as $\mathscr{F}_\varphi = \left\{ f(x) = \omega^T \varphi(x) \;\middle|\; \omega \in \mathbb{R}^{D_\varphi} \right\}$. Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex and differential loss function and let $G \in \mathbb{R}^{D_\varphi \times D_\varphi}$ be a positive semi-definite matrix. Consider the class of estimation methods optimizing the following L_2 regularized cost function on the training dataset $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$*

$$(\hat{w}, \hat{e}) = \arg\min_{w,e} \mathscr{J}_G(w,e) = \frac{1}{2}\sum_{i=1}^N \ell(e_i) + \frac{1}{2} w^T G w$$
$$\text{s.t.} \quad w^T \varphi(x_i) + e_i = y_i, \;\; \forall i = 1, \ldots, N. \quad (9.1)$$

Let $\{\phi_d : \mathbb{R}^D \to \mathbb{R}^R\}$ be a set of functions spanning the null-space of $G\varphi$. Let $\phi \in \mathbb{R}^{N \times R}$ be defined as $\phi_{ir} = \phi_r(x_i)$ for all $i = 1, \ldots, N$ and $r = 1, \ldots, R$ be of full rank. Then $G\phi_d = 0$ for all $d = 1, \ldots, D$. The resulting estimate can be evaluated as follows

$$\hat{f}(x) = \sum_{i=1}^N \hat{\alpha}_i K_G(x_i, x) + \sum_{r=1}^R \hat{\beta}_r \phi(x), \quad (9.2)$$

where $K_G(x_i, x) = \varphi(x_i) G^\dagger \varphi(x)$ with $G^\dagger \in \mathbb{R}^{D_\varphi \times D_\varphi}$ the pseudo-inverse to G. Furthermore the unknowns $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)^T \in \mathbb{R}^N$ and $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_R\right)^T \in \mathbb{R}^R$ are unique for the given loss function ℓ and dataset $\mathscr{D}$.

Proof. The proof starts with the primal-dual characterization of the global optimum to the constrained optimization problem (9.1), see condition (a) of Subsection 3.3.2. Let $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ be the Lagrange multipliers in the corresponding Lagrangian $\mathscr{L}_G$. An invariant condition for optimality independently for the choice of ℓ is

$$\frac{\partial \mathscr{L}_G}{\partial w} = 0 \to G w = \Phi_N^T \alpha, \quad (9.3)$$

where $\Phi_N = (\varphi(x_1), \ldots, \varphi(x_N))^T \in \mathbb{R}^{N \times D_\varphi}$ which holds in the optimum. If the inverse G^{-1} to G exists such that $G^T G^{-1} = G^{-T} G = I_{D_\varphi}$, then the solution takes the form

$$\hat{f}(x) = \hat{\alpha}^T \Phi_N G^{-1} \varphi(x)^T = \sum_{i=1}^N \hat{\alpha}_i K_G(x_i, x), \quad (9.4)$$

where the modified kernel $K_G : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is defined as $K_G(x_i, x) = \varphi(x_i)^T G^{-1} \varphi(x)$ and the vector $\hat{\alpha}$ contains the unique Lagrange multipliers following from the problem (9.1).

In the case the matrix is not invertible, the proof is a little bit more involved. Let $s \in \mathbb{N}_0$ denote the rank of the matrix G. Let $G = USU^T$ be the SVD of the matrix G such that $U^T U = I_{D_\varphi}$ and $S = \operatorname{diag}(\sigma_{(1)}, \sigma_{(2)}, \ldots, \sigma_{(s)}, 0, \ldots, 0) \in \mathbb{R}^{D_\varphi \times D_\varphi}$. Let $G^\dagger$ be the pseudo-inverse of G such that $G^\dagger = US^\dagger U^T$ with $S^\dagger = \operatorname{diag}(\sigma_{(1)}^{-1}, \sigma_{(2)}^{-1}, \ldots, \sigma_{(s)}^{-1}, 0, \ldots, 0) \in$

$\mathbb{R}^{D_\varphi \times D_\varphi}$. Let $Q \in \mathbb{R}^{D_\varphi \times D_\varphi}$ be span the null-space, e.g. $Q = U\,\mathrm{diag}(0_s^T, 1, \ldots, 1)U^T$. Then condition (9.3) can be rewritten as follows

$$Gw = \Phi_N^T \alpha \Leftrightarrow w = G^\dagger \Phi_N^T \alpha + Qw. \tag{9.5}$$

If the rank of the null-space of Q defined $R = D_\varphi - s$ is finite, a finite set of functions $\{\phi_r : \mathbb{R}^D \to \mathbb{R}\}_{r=1}^R$ can be constructed as follows. Let $U^0 \in \mathbb{R}^{D_\varphi \times R}$ contain the R eigenvectors corresponding with the zero singular values.

$$\phi_r = \varphi(x)^T U_r^0, \;\; \forall r = 1, \ldots, R, \tag{9.6}$$

then this set is a minimal set. From this it follows that the matrix $\phi \in \mathbb{R}^{N \times R}$ defined as $\phi_{ir} = \phi_r(x_i)$ for all $i = 1, \ldots, N$ and $r = 1, \ldots, R$ must be full rank. Thus, the solution to (9.1) can then be written as (9.2) where uniqueness follows from the convexity properties.

Moreover, from condition (9.5) it follows that $\alpha\Phi_N^T$ cannot be contained in the null-space $Q\varphi$ or in the span of $\{\phi_r\}_{r=1}^R$ such that the condition

$$0_{D_\varphi} = \alpha \Phi_N^T Q \Leftrightarrow \phi^T \alpha = 0_R. \tag{9.7}$$

is necessary and sufficient for uniqueness. □

This result also holds in the case of SVMs (Section 3.4) and SVTs (Section 3.5) which both employ a related formulation based on slackness variables.

The semi-parametric primal-dual kernel machines as elaborated in Section 4.1 may be seen as a direct application of this result. Let $\{\phi_d : \mathbb{R}^D \to \mathbb{R}\}_{r=1}^R$ be a set of parametric basis functions such that $\phi \in \mathbb{R}^{N \times R}$ (where $\phi_{ir} = \phi_r(x_i)$) is of full rank. Let φ_ϕ be an extended version of the mapping φ such that

$$\varphi_\phi(x) = (\phi_1(x), \ldots, \phi_R(x), \varphi)^T \in \mathbb{R}^{R+D_\varphi}. \tag{9.8}$$

Let $G = \mathrm{diag}(0_R^T, 1_{D_\varphi}^T) \in \mathbb{R}^{R+D_\varphi}$ be a diagonal matrix with zero weights to the parametric components. Then consider the estimator minimizing the regularized squared loss

$$(\hat{w}, \hat{e}) = \arg\min_{w,e} \mathscr{J}_{\gamma,G}(w,e) = \frac{\gamma}{2}\sum_{i=1}^N e_i^2 + \frac{1}{2} w^T G w \quad \text{s.t.} \quad w^T \varphi(x_i) + e_i = y_i, \;\; \forall i = 1, \ldots, N. \tag{9.9}$$

The pseudo inverse $G^\dagger$ and Q have then a particular easy form such that the solution is characterized by the following set of linear equations

$$\left[\begin{array}{c|c} 0_{R\times R} & \phi^T \\ \hline \phi & \Omega_G + \frac{1}{\gamma} I_N \end{array}\right] \left[\begin{array}{c} \beta \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0_R \\ \hline Y \end{array}\right], \tag{9.10}$$

following conditions (9.5) and (9.7) and where $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T G^\dagger \varphi(x_j)$. This set of linear equations is equivalent to equation (4.3).

9.1.2 Kernels as smoothing filters

Theorem 9.1 not only relates the quest of regularization with the research on learning the kernel but also supports the interpretation of kernel machines as smoothing filters as discussed in the following example.

Example 9.1 [Learning Machine based on a Fourier Decomposition, II] The setting of example 3.2 is studied in some more detail. Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N$ contain a sample with univariate inputs x_i uniformly sampled from a finite interval. Let $\varphi_F : \mathbb{R} \to \mathbb{R}^\infty$ be a mapping of a point x to its Fourier coefficients defined as follows

$$\varphi_F(x)_\lambda = \exp(i\lambda x) \tag{9.11}$$

where $\lambda = -\infty, \ldots, \infty$ acts similarly as an index. the inner product with any $\omega \in \mathbb{R}^\infty$ is then defined as

$$< \omega, \varphi_F(x) >= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega_\lambda \exp(i\lambda x) d\lambda \triangleq \omega^T \varphi_F(x). \tag{9.12}$$

which amounts to the classical inverse Fourier transform where λ plays the role of the frequency parameter. Let $(\mathscr{F} f) : \mathbb{R} \to \mathbb{R}$ denote the Fourier transform of the function f. The previous elaboration proves that one works with a kernel machine which implicitly works with a Fourier representation $\omega : \mathbb{R} \to \mathbb{R}$ if the following kernel is used

$$K_f(x_i, x_j) =< \varphi_F(x_i), \varphi_F(x_j) >= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(i\lambda (x_j - x_i)\right) d\lambda, \tag{9.13}$$

which equals a generalized function in the form of a Dirac function $\Delta(x_j - x_i)$ which integrates to one.

Given this Fourier interpretation, a plausible choice is to impose a decreasing weighting term penalizing for higher frequencies leading to less smooth solutions. This corresponds with a complexity measure corresponding with a high-pass filter on the estimated model, see e.g. (Wahba, 1990; Girosi *et al.*, 1995). Let the function $g : \mathbb{R} \to \mathbb{R}$ be defined as

$$g(\lambda) = \begin{cases} \exp\left(-\frac{\lambda^2}{h}\right) & \lambda \neq 0 \\ 0, & \lambda = 0 \end{cases} \tag{9.14}$$

where $h < c \in \mathbb{R}$ is an appropriate constant. Then the regularization term with weighting matrix can be formalized as

$$\omega^T G \omega \triangleq \int_{-\infty}^{\infty} g(\lambda) \ \ \omega^2(\lambda) d\lambda. \tag{9.15}$$

Following the previous theorem, this would coincide with the use of a parametric intercept term (lying in the null space of G) and the use of the kernel

$$K_G(x_i, x_j) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\lambda) \exp\left(i\lambda (x_i - x_j)\right) d\lambda = \exp\left(-\frac{(x_i - x_j)^2}{h}\right), \tag{9.16}$$

following from the invariance property of the function $f(x) = \exp(-x^2)$ with respect to the Fourier transform such that $(\mathscr{F} f)(x) = f(x)$ for all $x \in \mathbb{R}$. which results in the classical RBF kernel with bandwidth h, see e.g. Appendix A in (Girosi *et al.*, 1995).

9.1.3 Duality between error weighting schemes and kernel design

A similar argument can be used to explicify the relationship between the a weighted least squares scheme and the dual representations in terms of kernels.

Theorem 9.2. [Weighted Least Squares Primal-Dual Kernel Machines] *Consider the same setting as in the previous theorem. Let $H \in \mathbb{R}^{N\times N}$ be the known positive definite weighting matrix of the errors.*

$$(\hat{w},\hat{e}) = \arg\min_{w,e} \mathscr{J}_H(w,e) = \frac{1}{2}e^T He + \frac{1}{2}w^T w$$

$$\text{s.t.} \quad w^T \varphi(x_i) + e_i = y_i. \ \forall i = 1,\ldots,N \quad (9.17)$$

The global optimum follows from the set of linear equations

$$(\Omega H + I_N)\ e = Y, \qquad (9.18)$$

The solution then may be evaluated in any point $x_* \in \mathbb{R}^D$ as follows

$$\hat{f}(x_*) = \Omega_N(x_*)^T H\hat{e}, \qquad (9.19)$$

where $\hat{e} = (\hat{e}_1,\ldots,\hat{e}_N)^T \in \mathbb{R}^N$ solves (9.18) and $\Omega_N : \mathbb{R}^D \to \mathbb{R}^N$ is defined as $\Omega_N(x) = (K(x_1,x),\ldots,K(x_N,x))^T \in \mathbb{R}^N$.

Proof. The proof again starts with the primal-dual derivations as in Section 3.3. Let $\alpha = (\alpha_1,\ldots,\alpha_N)^T \in \mathbb{R}^N$ be a vector containing Lagrange multipliers. The Lagrangian becomes

$$\mathscr{L}_H(w,e;\alpha) = \frac{1}{2}e^T He + \frac{1}{2}w^T w - \sum_{i=1}^{N} \alpha_i \left(w^T \varphi(x_i) + e_i - y_i\right). \qquad (9.20)$$

Necessary and sufficient first order conditions for optimality then characterize uniquely the global optimum as follows

$$\begin{cases} \dfrac{\partial \mathscr{L}_H}{\partial w} = 0 \to \quad w = \Phi_N^T \alpha \\ \dfrac{\partial \mathscr{L}_H}{\partial e} = 0 \to \quad He = \alpha \\ \dfrac{\partial \mathscr{L}_H}{\partial \alpha_i} = 0 \to \quad w^T \Phi_N + e = Y, \end{cases} \qquad (9.21)$$

where $\Phi_N \in \mathbb{R}^N$ is defined as $\Phi_N = (\varphi(x_1),\ldots,\varphi(x_N))^T$. Let $H^\dagger$ denote the pseudo-inverse to H, then after eliminating w and α, the dual set of equations becomes as in (9.18). Remark that this time the result is not expressed in the Lagrange multipliers α but in the vector of residuals e as the latter contains more information (as e is not restricted to the image of H). □

Remark 9.1. Note that if an inverse H^{-1} to H exists, the solution can be expressed alternatively as follows

$$\left(\Omega + H^{-1}\right)\ \alpha = Y, \tag{9.22}$$

where the relation $wH = \Phi_N^T \alpha$ is used.

This result enables the construction of models consisting of a deterministic component modeled by a primal-dual kernel machine and a stochastic component modeled by a Gaussian process. Let $\{\mathbf{Y}_i\}_{i=1}^N$ be a Gaussian process with a non-parametric function for the mean $f(x)$ and fixed covariance function $\rho : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, then the probabilistic rules governing the observations may be written as

$$\begin{cases} E[\mathbf{Y}|x] = w^T \varphi(x) & \forall x \in \mathbb{R}^D \\ \\ \operatorname{cov}(\mathbf{Y}_i, \mathbf{Y}_j) = E[(\mathbf{Y}_i - f(x_i))(\mathbf{Y}_j - f(x_j))] = \rho(x_i, x_j), & \forall x_i, x_j \in \mathbb{R}^D. \end{cases} \tag{9.23}$$

Let $C \in \mathbb{R}^{N\times N}$ be the covariance matrix such that $C_{ij} = \rho(x_i, x_j)$ for all $i, j = 1, \ldots, N$ which is strictly positive definite. Define the random variables $\mathbf{Z}$ as follows $\mathbf{Z}_i = \mathbf{Y}_i - f(x_i)$, then $\{\mathbf{Z}_i\}_{i=1}^N$ is a Gaussian process. The log likelihood of a realization $Z = (z_1, \ldots, z_N)^T \in \mathbb{R}^N$ of this non i.i.d. process is given as

$$\ell(Z) = \log\left(Z^T C^\dagger Z\right), \tag{9.24}$$

as in e.g. (Whittle, 1954; Box and Jenkins, 1979; Brockwell and Davis, 1987). This motivates the following penalized likelihood cost-function

$$(\hat{w}, \hat{Z}) = \arg\min_{w,Z} \mathscr{J}_{\gamma,\rho}(w, Z) = \frac{\gamma}{2} Z^T C^{-1} Z + \frac{1}{2} w^T w$$
$$\text{s.t.} \quad w^T \varphi(x_i) + z_i = y_i, \ \ \forall i = 1, \ldots, N, \tag{9.25}$$

with C^{-1} the inverse of the covariance matrix C such that $C^{-T}C = C^T C^{-1} = I_N$. The output value corresponding to a new datapoint $x_* \in \mathbb{R}^D$ can be estimated as follows

$$\hat{f}(x_i) = \Omega_N(x_*)^T \hat{\alpha}, \tag{9.26}$$

where $\hat{\alpha}$ solve the dual system (9.18). One may refer to $\hat{f}$ as the (deterministic) mean function of the process. Following a similar argument as standard in Gaussian Processes based on the matrix inversion Lemma (see also Section 5.2), the expected response at position $x_* \in \mathbb{R}^D$ is given as

$$E[\mathbf{Y}_* \mid x_*, \mathbf{Y}_1 = y_1, \ldots, \mathbf{Y}_N = y_N] = \left(\Omega_N(x_*) + \rho_N(x_*)\right)^T \hat{\alpha}, \tag{9.27}$$

where the function $\rho_N : \mathbb{R}^D \to \mathbb{R}^N$ is defined as $\rho_N(x) = (\rho(x_1, x), \ldots, \rho(x_N, x))^T \in \mathbb{R}^N$. From this expression and (9.18), it can be seen that the difference between the covariance (and the weighting scheme) on the one hand and the kernel on the other is indistinguishable in the formulations. In the extremal case of the same functional form of the kernel and the covariance function, the difference dissolves completely. A similar result was obtained in the theory of smoothing splines (Wahba, 1990).

Example 9.2 **[Colored Noise Scheme]** A classical example is considered where the noise scheme can be modeled by a first order Auto-Regressive (AR) process

$$\mathscr{F}_{\varphi,a} = \left\{ f(x) = w^T \varphi(x),\ \ y_t = f(x_t) + (1+aq)e_t \ \middle|\ w \in \mathbb{R}^{D_\varphi}, \right\},\ \ |a| < 1, \tag{9.28}$$

where q denotes the backshift operator $qe_t = e_{t-1}$. Define $qe_1 = e_0$ where $e_0 \in \mathbb{R}$ is an appropriate initial condition, to setup a proper initial condition. This type of models was elaborated by (Engle *et al.*, 1986) in the case of modeling the electricity load as a function of amongst others the temperature. Let $\{(x_t, y_t)\}_{t=1}^T$ be a set of observations recorded at a finite sequence of equal time intervals corresponding with $t = 1, \ldots, T$. In this case the following cost-function may be written

$$(\hat{w}, \hat{e}) = \arg\min_{w,e} \mathscr{J}_{a,\gamma}(w,e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{t=1}^{N} e_t^2 \quad \text{s.t.} \quad w^T \varphi(x_t) + (1+aq)e_t = y_t \ \ \forall t = 2, \ldots, T, \tag{9.29}$$

where $z_t = (1+aq)e_t$ for all $t = 2, \ldots, T$ and $z_t = e_t$ defines a Gaussian process $\{z_t\}_{t=1}^T$ with covariance matrix $C \in \mathbb{R}^T$ defined as follows

$$C_{kl} = \operatorname{cov}(z_k, z_l) = \begin{cases} \sigma_e^2 & \text{if } \ k = l \\ a\sigma_e^2 & \text{if } \ |k-l| = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{9.30}$$

After constructing the Lagrangian $\mathscr{L}_{a,\gamma}$ with multipliers $\alpha = (\alpha_2, \ldots, \alpha_T)^T \in \mathbb{R}^T$, one obtains the following conditions for optimality

$$\begin{cases} \dfrac{\partial \mathscr{L}_{a,\gamma}}{\partial w} = 0 \to & w = \sum_{t=1}^T \alpha_t \varphi(x_t) \\ \dfrac{\partial \mathscr{L}_{a,\gamma}}{\partial e} = 0 \to & \gamma e_t = (1+aq)\alpha_t \qquad \forall t = 1, \ldots, T \\ \dfrac{\partial \mathscr{L}_{a,\gamma}}{\partial \alpha_i} = 0 \to & w^T \varphi(x_t) + (1+aq)e_t = y_t, \ \ \forall t = 1, \ldots, T. \end{cases} \tag{9.31}$$

Let the matrix $T_a \in \mathbb{R}^{T \times T}$ be defined as follows

$$T_a = \begin{bmatrix} 1 & 2a & a^2 & 0 & \ldots & 0 \\ 0 & 1 & 2a & a^2 & & \\ \vdots & & \ddots & \ddots & & \\ & & & & 1 & a \\ 0 & & & & 0 & 1 \end{bmatrix}. \tag{9.32}$$

As this matrix has all eigenvalues one (Golub and van Loan, 1989), the variables e_t and w may be eliminated from the set of equations (9.31) resulting in the following set of linear equations

$$\left(\Omega + \frac{1}{\gamma} T_a \right) \alpha = Y, \tag{9.33}$$

The resulting mean function $\hat{f}$ may be evaluated in a new point $x_* \in \mathbb{R}^D$ as follows

$$\hat{f}(x_*) = \Omega_N(x_*)^T \hat{\alpha}, \tag{9.34}$$

where $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_T)^T \in \mathbb{R}^T$ solves the system (9.33). In this example, the parameter a was considered to be known. It becomes apparent (from an optimization point of view) that the determination of the regularization constant and the auto-regressive parameter amounts to non-convex model selection problems, as also regarded in this way in (Engle *et al.*, 1986).

From the dual system (9.33), it may be concluded that the problem is equivalent to the weighted problem as follows. Define $T_a^{-1} = \operatorname{diag}\left((1+aq)^{-1}, \ldots, (1+aq)^{-1}\right)$, then

$$(\hat{w}, \hat{Z}) = \arg\min_{w,Z} \mathscr{J}^T_{\gamma,a}(w,Z) = \frac{1}{2} w^T w + \frac{\gamma}{2} Z^T T_a^{-1} Z \quad \text{s.t.} \quad w^T \varphi(x_t) + z_t, \;\; \forall t = 2, \ldots, T, \tag{9.35}$$

where $\{z_t\}_{t=1}^T$ is a non-white process with covariance ρ,

9.1.4 Duality of linear structure and kernel design

This subsection shows how imposed structure in the form of symmetric functions reflect in the design of the kernel matrix. Specificly, consider the task of estimating even functions f from data such that $f(x) = f(-x)$ for all $x \in \mathbb{R}^D$. Consider the following model

$$f(x) = \frac{1}{2}\left(w^T \varphi(x) + w^T \varphi(-x)\right), \tag{9.36}$$

which should be even by construction. Consider the primal problem:

$$\begin{aligned} (\hat{w}, \hat{e}) = \arg\min_{w,e} \mathscr{J}_\gamma(w,e) &= \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^N e_i^2 \\ \text{s.t.} \quad & \frac{1}{2}\left(w^T \varphi(x) + w^T \varphi(-x)\right) + e_i = y_i \;\; \forall i = 1, \ldots, N \end{aligned} \tag{9.37}$$

Eliminating the latter infinite constraint results in the following problem

$$\begin{aligned} (\hat{w}, \hat{e}) = \arg\min_{w,e,f} \mathscr{J}_\gamma(w,e) &= \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^N e_i^2 \\ \text{s.t.} \quad & \frac{1}{2} w^T \varphi(x_i) + \frac{1}{2} w^T \varphi(-x_i) + e_i = y_i, \;\; \forall i = 1, \ldots, N. \end{aligned} \tag{9.38}$$

Using a primal-dual argument, the corresponding dual problem can be summarized as follows

$$\left(\Omega^{(2)} + \frac{1}{\gamma} I_N\right) \alpha = Y, \tag{9.39}$$

where the modified kernel matrix becomes $\Omega^{(2)} = \frac{1}{4}\left(\Omega^{-,-} + 2\Omega^- + \Omega\right)$ and the matrices $\Omega^-, \Omega^{-,-} \in \mathbb{R}^{N \times N}$ are defined as $\Omega^-_{ij} = K(x_i, -x_j)$ and $\Omega^{-,-}_{ij} = K(-x_i, -x_j)$ respectively. The function can be evaluated in a new point as

$$\hat{f}(x_*) = \Omega_N^{(2)}(x_*)^T \hat{\alpha}, \tag{9.40}$$

where $\hat{\alpha}$ solve the dual set of equations (9.39) and $\Omega_N^{(2)} : \mathbb{R}^D \to \mathbb{R}^N$ is defined as $\Omega_N^{(2)}(x_*) = \frac{1}{4}\left(K(x_1,x_*) + 2K(-x_1,x_*) + K(-x_1,-x_*),\ldots\right)^T \in \mathbb{R}^N$.

Remark 9.2. This structural approach should be contrasted with the approach sketched in Section 4.3 where structure was imposed pointwise. The present technique also guarantees that future prediction on (yet unknown) testpoints will satisfy the constraints. It is however more difficult to apply than the pointwise approach as an appropriate model definition (9.36) is not easily found e.g. in the case of inequality (monotonicity) constraints. Note finally that this form of structural constraints also translates into the use of an appropriate kernel.

9.2 Kernel decompositions and Structure Detection

9.2.1 Kernel decompositions

The problem of choosing an appropriate kernel may be approached in correspondence with the following principle *"If nothing were known a priori on the choice of the kernel, then let the data decide"*, which situates this issue closely to a Bayesian interpretation as in (MacKay, 1992) and was elaborated in the case of LS-SVM models in (Van Gestel *et al.*, 2002). The motivation for the concept of kernel decompositions is summarized in the following lemma.

Lemma 9.1. [Kernel Decomposition] *Let $D_p = \sum_{p=1}^P D_{\varphi_p} \in \mathbb{N}_0$ be a fixed nonzero positive integer. Let $\varphi_P^* : \mathbb{R}^D \to \mathbb{R}^{D_P}$ denote the extended feature space mapping defined as*

$$\varphi_{(P)}(x) = \left(\varphi_1(x)^T, \ldots, \varphi_P(x)^T\right)^T \in \mathbb{R}^{D_P}. \tag{9.41}$$

Let $c = (c_1, \ldots, c_P)^T \in \mathbb{R}^{+,P}$ be a vector of positive constants. Consider the modified regularized least squares cost-function of the LS-SVM regressor given as

$$\mathscr{J}_c = \frac{1}{2}\sum_{p=1}^P c_p \left(w_p^T w_p\right) + \frac{1}{2}\sum_{i=1}^N e_i^2 \quad s.t. \quad \sum_{p=1}^P w_p^T \varphi_p(x_i) + e_i = y_i,\ \ \forall i = 1,\ldots,N, \tag{9.42}$$

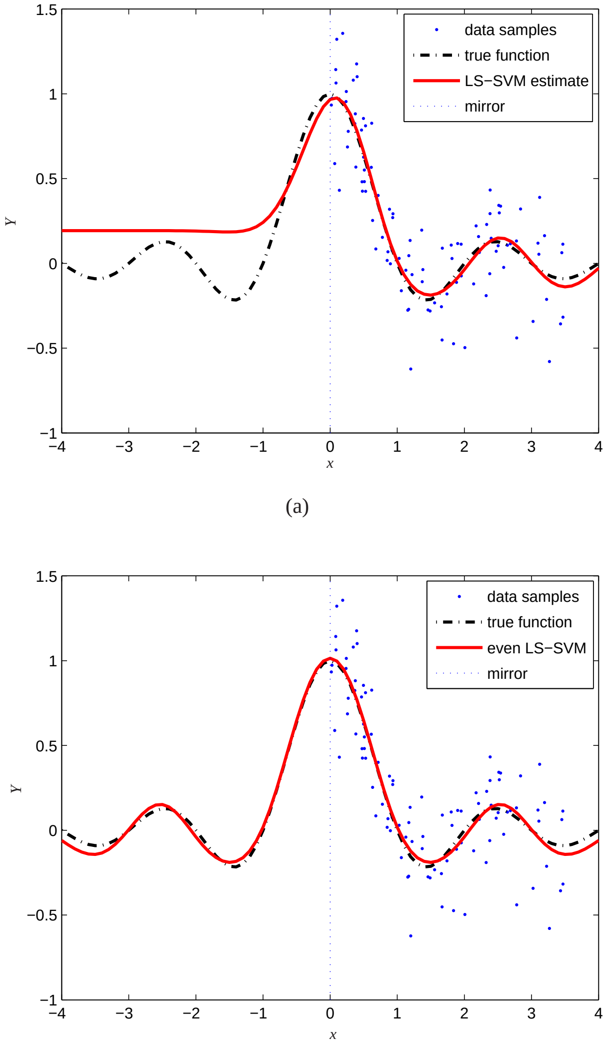
where the vector $c \in \mathbb{R}^{+,P}$ determines the regularization trade-off. Let $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)^T \in \mathbb{R}^N$ denote the unique solution to the dual problem of (9.42). Then the solution takes the form

$$\hat{f}(x) = \sum_{i=1}^N \hat{\alpha}_i K_{(P)}(x_i, x), \tag{9.43}$$

where $K_{(P)} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is defined as

$$K_{(P)}(x_i,x_j) = \sum_{p=1}^P c_p K_p(x_i,x_j),\ \ \forall x_i, x_j \in \mathbb{R}^D, \tag{9.44}$$

and K_p is the kernel corresponding with the pth feature map such that $K_p(x_i,x_j) = \varphi_p(x_i)^T \varphi_p(x_j)$. We refer to the kernel $K_{(P)}$ as a kernel decomposition.



(a)

(b)

Figure 9.1: *Illustratic example showing the benefits of imposing structural constraints on the estimate of a function (dashed-dotted line) with noisy observations (dots).* **(a)** *estimate of standard LS-SVM without imposing the structure.* **(b)** *estimate using the presented method imposing the even structure of the data. This latter has improved generalization on the left-half plane. This approach is especially usefull as a modular approach for semi-parametric tasks (see Section 4.1).*

This result is easily proven by using a primal-dual argument and is closely related to Theorem 9.1. A special case is encountered when the vector of constants c is taken constant, say $c_p = 1/\gamma$ for all $p = 1,\dots,P$ in which case the formulation reduces to the componentwise kernel machines formulation as elaborated in Section 4.2. However, the present result has a slightly different focus.

9.2.2 Structure detection using kernel decompositions

Let $K_{(P)} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ denote a kernel decomposition consisting of $P \in \mathbb{N}_0$ components $K_{(P)}(x_i,x_j) = \sum_{p=1}^{P} K_p(x_i,x_j)$. From the close relationships between componentwise kernel machines (4.2) and kernel decompositions (9.1), one can consider methods for obtaining models that contain sparse in the components, which would lead to a sparse kernel decomposition. The approach towards structure detection using the measure of maximal variation as described in Subsection 6.4.2 may be employed to let the data decide on which specific kernel and parametric terms to use.

Example 9.3 **[Modeling discontinuities]** An example is elaborated in the case one knows that the underlying function may contain a number of discontinuities of Kth order. Let the set $\{x_q \in \mathbb{R}\}_{q=1}^{Q}$ denote the set of knots at which place a discontinuity may occur of the kth derivative. A conveniently broad class of discontinuities is obtained when this set correspond with the data samples $\{x_i\}_{i=1}^{N}$. Let $\{\varsigma_q^{(k)} : \mathbb{R} \to \mathbb{R}\}_{q,k}$ denote the set of basis functions modeling the discontinuities as follows

$$\varsigma^{(k)}(x;x_q) = \int \dots \int I^*(x > x_q) dx^k, \tag{9.45}$$

where $I^*(x > 0)$ equals $+1$ if $x > 0$ and -1 otherwise. Then the primal model takes the form

$$\textbf{Model:} \quad f(x) = w^T \varphi(x) + \sum_{k=0}^{K} \sum_{i=1}^{N} w_{ik} \varsigma_i^{(k)}(x;x_i). \tag{9.46}$$

Using the regularized least squares cost-function (9.42) with the weights $c = 1_R/\gamma$, the estimated model takes the form

$$\textbf{Result:} \quad \hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i K_{(P)}(x_i,x), \tag{9.47}$$

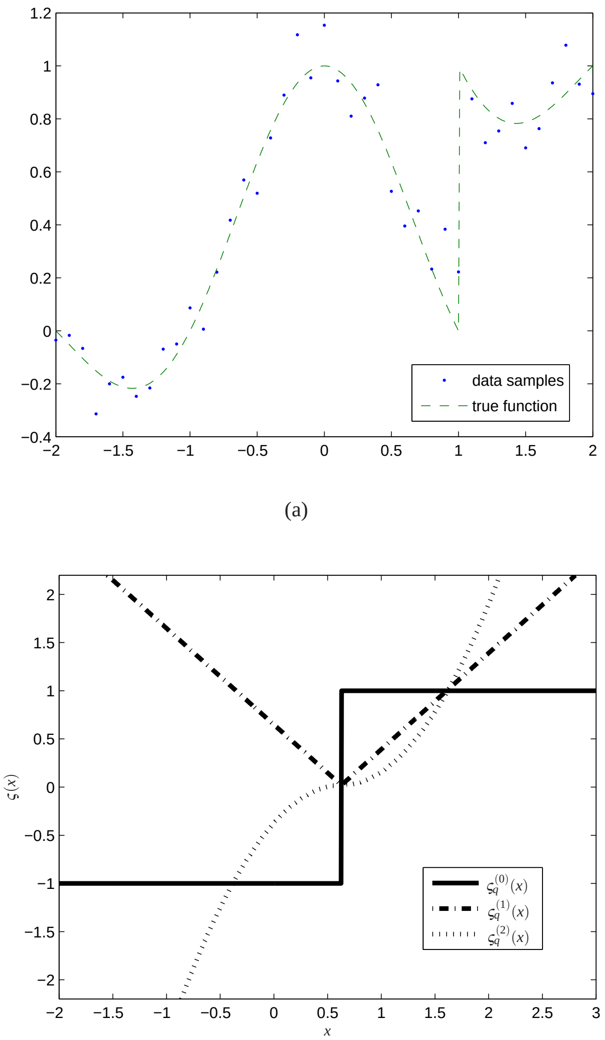
where $K_{(P)} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is defined as

$$\textbf{Kernel:} \quad K^*(x_i,x_j) = K(x_i,x_j) + \sum_{i=1}^{N} \sum_{k=1}^{K} K_{\varsigma}^{x_i,k}(x_i,x_j), \;\; \forall x_i,x_j \in \mathbb{R}^D, \tag{9.48}$$

and the kernel $K_{\varsigma}^{x_i,k}$ is defined as

$$K_{\varsigma}^{x_q,k}(x_i,x_j) = \varsigma^{(k)}(x_i;x_q)\varsigma^{(k)}(x_j;x_q), \quad \forall x_i,x_j \in \mathbb{R}. \tag{9.49}$$

Note that the discontinuities land up into the kernel as regularization is applied to it. This was necessary in order to avoid ill-posedness due to the large set of basis functions $\{\varsigma^{(k)} : \mathbb{R} \to \mathbb{R}\}_{q,k}$.



(a)

(b)

Figure 9.2: *Illustration of the technique for the modeling of data with underlying function containing discontinuities at the observed points.* **(a)** *Given a function including a discontinuity (dashed line) and $N = 40$ noisy observations (dots).* **(b)** *Example of the basis functions $\varsigma_q^{(0)}$ and $\varsigma_q^{(1)}$ at the knot $x_q = 0.6283$.*

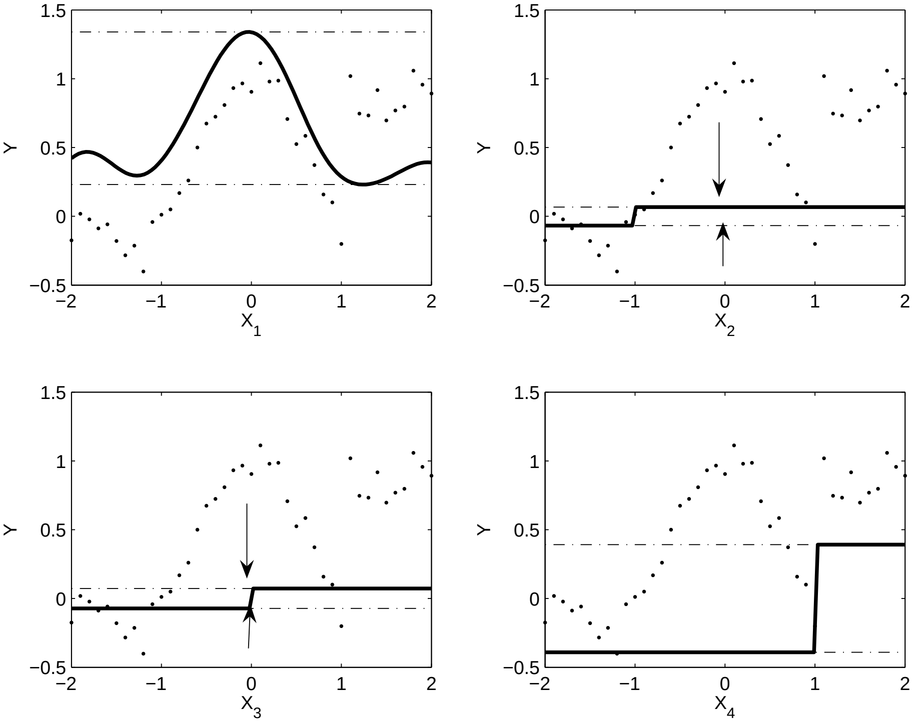


Figure 9.3: *A toy example using $N = 40$ datapoints. The contributions of the second and the third discontinuity tends to zero as the impact of the maximal variations are increased in the loss function as indicated by the arrows.*

Now the stage is set for application of the structure detection approach based on maximal variation as elaborated in Subsection 6.4.2. This is particular relevant here for a number of reasons, including (1) knowing the location and number of discontinuities is important for understanding and analysis of the result, (2) the measure of maximal variation is suited for this type of basis functions as a zero maximal variation does imply a zero weighting of the term, (3) the scale-independence of the measure of maximal variation decreases the impact of the scale of the basis functions on the prediction. As the number of basis functions grows in the number of datapoints, the hierarchical modeling strategy is advisable.

Figure 9.3 illustrates this application. The first panel shows basis functions modeling discontinuities of order $k = 0, 1, 2$, while the second panel shows the contributions of a simple toy example. This example is based on a set of $N = 40$ observations generated as $y_i = \text{sinc}(x_i) + I(x_i > 1.11) + e_i$ with $e_i \sim \mathscr{N}(0, 0.1)$. Only first-order discontinuities are considered, while they only can occur at a finite number of places $\{x_q\}_{q=1}^{Q} = \{-1, 0, 1\}$. The contributions of the bases $\varphi^{(0)}(\cdot; -1)$ and $\varphi^{(0)}(\cdot; 0)$ will tend to zero by increasing the impact of the maximal variation term in the cost function indicating that no effective discontinuity is present in the data on the knots -1 and 0. This example was loosely motivated on the research on modeling discontinuities as described by (Ansley and Wecker, 1981) and mentioned in (Wahba, 1990).

9.3 One-sided Representations

9.3.1 Time series analysis and signal processing

As was already touched upon in example 3.2 and in Section 5.1, there is a close relation between harmonic analysis and smoothing functions (Vapnik, 1998; Girosi *et al.*, 1995). However, there is a conceptual difference between this field with the subject of signal processing and time series analysis, quoting (Wiener, 1949):

> "While the past of a time series is accessible for examination, its future is not. That means that the involved operators (for time series analysis) must have an inherent certain one-sidedness."

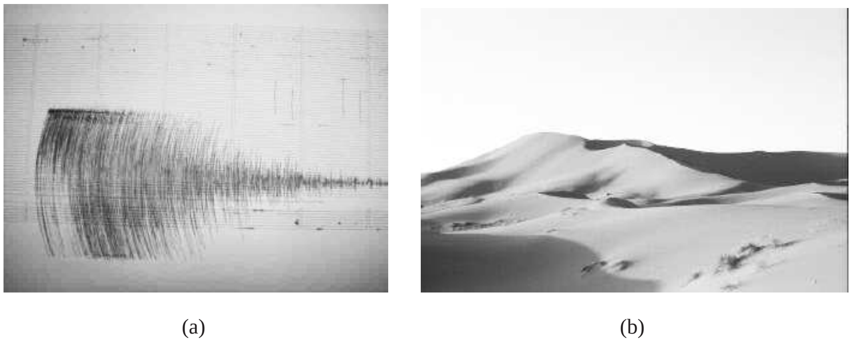
which is not valid in the case of the mentioned methods. This principle will constitute the main difference between Gaussian processes as reviewed in Section 5.2 and stochastic processes with a time index set $\mathbb{T}$. This difference becomes apparent by studying the Wiener-Hopf equation for the causal filtering problem.

Let the two time series $\{u_t\}_{t=1}^N$ (input) and $\{y_t\}_{t=1}^N$ (output) be equidistantly sampled and let $U = (u_1, \ldots, u_N)^T \in \mathbb{R}^N$ and $Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$. Let $K^f \in \mathbb{R}^{N\times N}$ be a lower diagonal matrix such that $K^f_{ij} = 0$ if $j > i$. This will represent the linear operator filtering the input as to mimic the output signal, or informally $K_f U \approx Y$. Note that the lower diagonal form of the linear filter K^f represents the one-sided character of the operator, see (Kailath *et al.*, 2000) and also the literature on Volterra equations of the first kind (Press *et al.*, 1988). Under the assumption of stationarity, the covariance matrices $E[UY^T] \in \mathbb{R}^{N\times N}$ and $\Omega = E[YY^T] \in \mathbb{R}^{N\times N}$ are Toeplitz. Let $[\cdot]_{\text{lower}} : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ denote an operation mapping a matrix $A \in \mathbb{R}^{N\times N}$ to its upper-diagonal counterpart $B \in \mathbb{R}^{N\times N}$ such that $B_{ij} = A_{ij}$ if $j \leq i$ and zero otherwise. Then the Wiener-Hopf technique for finding the optimal predictive filter is summarized as follows.

$$\min_{K^f} \sum_i (K^f_i U - y_i)^2 \Leftrightarrow \left[E[UY^T] - K^f E[YY^T]\right]_{\text{lower}} = 0_{N\times N}$$

$$\Leftrightarrow K_f = \left[E[UY^T] L^{-T} D^{-1}\right]_{\text{lower}} L^{-1}, \quad (9.50)$$

where the *LDL* transformation of the covariance matrix is used such that $\Omega = LDL^T$ with $L \in \mathbb{R}^{N\times N}$ lower triangular and $D \in \mathbb{R}^{N\times N}$ diagonal (Golub and van Loan, 1989), see e.g. (Kailath *et al.*, 2000). It is interesting to relate this central derivation to the smoothing problem (Kailath *et al.*, 2000), the LS-SVM modeling approach (Section 3.3) and the realization approach discussed in Chapter 9.2.2.



(a) (b)

Figure 9.4: *Illustration of a one-sided and non-causative process occurring in nature.* **(a)** *Seismograms measuring the strength of earthquakes have an inherent one-sidedness as they do present oscillatory behavior caused by the main quake.* **(b)** *Sand dunes in the desert do not present an inherent time order but consists of a spatial process as the hill peaks depend smoothly on the neighboring slopes.*

Another crucial assumption for statistical analysis of time series is that operators which come into consideration are not tied down to an origin in space as any statistical distribution may not be affected by a shift in origin (Wiener, 1949). This assumption is described readily by the ergodic theorems, see (Birkhoff, 1931), which relevance in the static smoothing problem is yet latent.

9.3.2 One-sided representations

One-sided representations for univariate time-series include the popular Auto-Regressive (AR) model of order $K \in \mathbb{N}_0$

$$\hat{y}_{t+1} = \sum_{k=0}^{K} a_k y_{t-k}, \ \ \forall t = K, \ldots, T. \tag{9.51}$$

A non-causative counterpart was formulated in the context of spatial data analysis named as the Spatial Auto-Regressive (SAR) models (Ripley, 1988). Consider the univariate process $\mathbf{Z}$ sampled at equidistant points enumerated by $i = 1, \ldots, N$. The simplified SAR model of order K takes the form

$$E[Z_i | Z_j, i \neq Z_j] = \sum_{k=1}^{K} a_k \left(Z_{i-k} + Z_{i+k} \right), \ \ \forall i = k+1, \ldots, N-k, \tag{9.52}$$

where $a = (a_1, \ldots, a_K)^T \in \mathbb{R}^K$ is the vector with parameters. The difference between the one-sided representation (9.51) and the spatial (9.52) can be clearly seen, although their theoretical properties coincide to large extents (Cressie, 1993).

We define here the phrase “a certain one-sidedness” as in the previous quote in the definition of one-sidedness and spatial representation.

Definition 9.1. [One-sided Representations] *A model with a one-sided representation does only describe relationships of the outcome with previous variates. A model with a spatial representation is violating this constraint.*

Note that the literature on time-series and systems theory define causality of a model estimate in a different way, see e.g. (Brockwell and Davis, 1987; Kailath *et al.*, 2000).

System theory and identification have a slightly different focus as they study the behavior and modeling of a one-sided dynamical system from input-output measurements typically denoted as $\{(u_t, y_t)\}_{t=1}^T \in \mathbb{R}^{D_u} \times \mathbb{R}^{D_y}$. A linear one-sided input-output relation is characterized by its so-called impulse response $h = (h_0, \ldots, h_\infty)^T$ defined as follows

$$E[\, y_t \,\big|\, (u_{-\infty}, \ldots, u_t)\,] = \sum_{\tau=0}^{\infty} h_\tau u_{t-\tau}, \quad \forall t = -\infty, \ldots, \infty, \tag{9.53}$$

where one also refers to h as the Markov parameters. As this representation involves a possibly infinite vector of parameters h, identification often employs more parsimonious system representations. Important examples are the rational polynomial representations as the Box-Jenkins class of models (see e.g. (Box and Jenkins, 1979; Ljung, 1987)), and the state-space models. Let again $K \in \mathbb{N}_0$ be the order of the system and let $A \in \mathbb{R}^{K\times K}$, $B \in \mathbb{R}^{K\times D_u}$, $C \in \mathbb{R}^{D_y\times K}$ and $D \in \mathbb{R}^{D_y\times D_u}$ be the system matrices. Then a state-space model can be written as follows (Kalman, 1960), see e.g. (Kailath *et al.*, 2000)

$$\begin{cases} x_{t+1} = Ax_t + Bu_t & \forall t = 1, \ldots, T \\ y_t = Cx_t + Du_t, & \forall t = 1, \ldots, T, \end{cases} \tag{9.54}$$

where the sequence x_t is called the state of the system at time instants $t = 2, \ldots, T$ and represent (informally) the memory of the system at a time instant t. The goal of one-sided models as (9.51) and (9.54) is prediction, explanation and control as well as smoothing. It then comes as no surprise that the issue of determining the required amount of smoothing in static tasks have inherent relations to the mentioned approaches as illustrated in the next example.

Example 9.4 [One-sided auto-regressive representation and the convolution] Consider the sequence $\{y_t\}_{t=1}^T$ which constitutes of a convolution of an unobserved indexed array $\{e_t\}_{t=1}^T$ (the index set denotes typically the time) with a given convolution vector $h \in \mathbb{R}^T$

$$y_t = \sum_{\tau=0}^{T-t} h_\tau e_{t-\tau}, \;\; \forall t = 1, \ldots, T. \tag{9.55}$$

Let h be defined as follows

$$h_\tau = \exp\left(-\frac{\tau}{\sigma}\right), \;\; \forall \tau = 0, \ldots, \tag{9.56}$$

where $0 < \sigma \in \mathbb{R}$ denotes a bandwidth parameter. The task of optimizing this bandwidth parameter such that two given series $\{e_t\}$ and $\{y_t\}$ are related optimally as (9.55) amounts to solving

$$\min_{\sigma,e} \quad = \quad \sum_{t=1}^{T} e_t^2 \quad \text{s.t.} \quad y_t \quad = \quad \sum_{\tau=0}^{t} \exp\left(-\frac{\tau}{\sigma}\right)(e_{t-\tau}), \qquad \forall t \quad = \quad 1,\ldots,T. \tag{9.57}$$

In order to tackle the problem the following analytical property is used

$$\sum_{\tau=0}^{\infty} a^{\tau} q = \frac{1}{1-aq}, \ \text{ if } \ |a| < 1, \tag{9.58}$$

where q is a linear operator (more specific, q is the backshift operator $qx_t = x_{t-1}$). Using this equation, it follows that

$$\begin{aligned} \sum_{\tau=0}^{\infty} a^{\tau} q \quad &= \quad \sum_{\tau=0}^{\infty} \exp(\tau \ln(a))q = \sum_{\tau=0}^{\infty} \exp\left(-\tau \ln\left(\frac{1}{a}\right)\right) q \\ &= \quad \frac{1}{(1-aq)}, \end{aligned} \tag{9.59}$$

such that (9.57) and (9.59) are equivalent if $\sigma = 1/\ln(\frac{1}{a})$. Problem (9.57) can be written equivalently as

$$\min_{a,e} \mathscr{J}(a,e) = \sum_{t=2}^{T} e_t^2 \quad \text{s.t.} \quad y_t = ay_{t-1} + x_t + e_t, \ \ -1 \leq a \leq 1, \tag{9.60}$$

where $e = (e_1,\ldots,e_t)^T \in \mathbb{R}^T$. This amounts to solving a convex constrained least squares problem.

A cornerstone of the research on system identification is given by realization theory which establishes the relation between the system matrices and the Markov parameters parameterizing the impulse response (9.53) of the system under study. In the case of stochastic state-space models without external inputs u_t, stochastic realization theory provides a related approach based on the auto-covariances of the model (Kung, 1978).

9.4 Stochastic Realization for LS-SVM Regressors

A numerical method is proposed to access the shape of the underlying kernel under the assumption of stationarityy of the data (the covariance measure underlying the data is only a function of the displacement between two measurements).

9.4.1 Univariate and equidistantly sampled data

In order to fix the ideas, let us consider here the case of univariate and equidistantly sampled data. In this case the kernel matrix takes a particularly simple form

$$\Omega_T = \begin{bmatrix} k_0 & k_1 & \dots & k_{N-1} \\ k_1 & k_0 & \dots & \\ & \ddots & \ddots & \\ k_{N-1} & & k_1 & k_0 \end{bmatrix} \quad \text{s.t.} \quad k_\tau = K(x_i, x_{i+\tau}) = K(x_i, x_{i-\tau}),\ \forall \tau = 0, \dots, N-1, \tag{9.61}$$

which is known as a symmetric Toeplitz matrix (Golub and van Loan, 1989) and plays a central role in the research on system identification, see e.g. (Kailath *et al.*, 2000). As such, the admissible class of kernel matrices Ω_T may be described as follows

$$\mathscr{K}_T = \left\{ \Omega_T \;\middle|\; \Omega_T \succeq 0,\ \Omega_T = \Omega_T^T,\ \Omega_T \text{ Toeplitz} \right\}, \tag{9.62}$$

which is a proper pointed cone, see e.g. (Alizadeh and Goldfarb, 2003; Genin *et al.*, 2003; Boyd and Vandenberghe, 2004).

Definition 9.2. [Admissible LS-SVM models] *The set of optimal LS-SVM models for any admissible kernel and constant regularization term may be described as*

$$\mathscr{F}_{\mathscr{K}_T} = \Big\{ f : \mathbb{R}^D \to \mathbb{R},\ \alpha \in \mathbb{R}^N,\ \gamma \in \mathbb{R}_0^+,\ \Omega_T \in \mathscr{K}_T,\ e \in \mathbb{R}^N \quad s.t. \quad \begin{cases} \left(\Omega_T + \frac{1}{\gamma} I_N\right) \alpha = Y & (a) \\ \gamma e = \alpha & (b) \\ f(x_i) = \Omega_{T,i}\, \alpha & (c) \\ \gamma > 0 & (d) \end{cases} \Big\}. \tag{9.63}$$

The subset of optimal LS-SVM smoothers then can be written after elimination of γ, α and f as follows

$$\mathscr{Y}^Y_{\mathscr{K}_T} = \Big\{ Y_s \in \mathbb{R}^N, \tilde{\Omega}_T \in \mathbb{R}^{N \times N}, e \in \mathbb{R}^N \;\Big|\; \left(\tilde{\Omega}_T + I_N\right) e = Y,\quad Y_s = \tilde{\Omega}\, e,\ \ \tilde{\Omega}_T \in \mathscr{K}_T \Big\}. \tag{9.64}$$

where $\tilde{\Omega}_T = \gamma \Omega_T$ is still in the scale invariant set $\mathscr{F}_{\mathscr{K}_T}$.

Note that these sets are non-convex by the occurrence of quadratic terms. Consider model selection criteria based on the smoothing abilities on the training output observations denoted as $\mathsf{Modsel}_s : \mathbb{R}^N \times \mathbb{R}^{N \times N} \times \mathscr{A} \to \mathbb{R}$, such as e.g. C_p's statistic (Mallows, 1973) or the generalized cross-validation criterion (Golub *et al.*, 1979). The model selection problem may then be formulated as follows

$$(\hat{Y}_s, \hat{\Omega}_T, \hat{e}) = \arg\min_{Y_S, \Omega_T, e} \ \mathscr{J}_{\mathsf{Modsel}}(Y_s, \Omega_T, e) \qquad \text{s.t.} \quad (Y_s, \Omega_T, e) \in \mathscr{Y}^Y_{\mathscr{K}_T}, \tag{9.65}$$

which formalizes again the fusion argument as introduced in Chapter 7. This type of problems is in general non-convex even if the function Modsel is convex. However, one can find numerically efficient methods to solve the problem exactly in a number of cases where one is described explicitly.

Example 9.5 One can frame the recent literature on learning (Lanckriet *et al.*, 2004) the kernel in the presented framework. Especially the kernel characterization (9.61) seems appropriate to study the transductive setting where the input points of points which need evaluation are known beforehand.

9.4.2 A realization approach

The method of moments estimates parameters by finding expressions of those in terms of the lowest possible order moments and then substituting sample moments in the expression, see e.g. (Rice, 1988). In the case of second order moments for Gaussian processes, a generalization of this principle was formulated under the denominator of stochastic realization theory. Although the original formulation was described towards the identification of the system matrices of the one-sided state-space model from the observed sample auto-covariances (Kung, 1978), the same approach may be employed in order to approach the problem (9.65). Reconsider definition 5.2 of a Gaussian process.

Definition 9.3. [Second Order Moments of a Univariate Gaussian Process] *The second-order moments of a stationary Gaussian process $\{\mathbf{Y}_x\}_{x\in\mathscr{D}}$ with zero mean are defined as*

$$\rho_{\mathbf{Y}}(\tau) = E\left[\mathbf{Y}_i\mathbf{Y}_j \;\middle|\; \|x_i - x_j\| = \tau\right]. \tag{9.66}$$

Then let $C \in \mathbb{R}^{N\times N}$ be the positive semi-definite covariance matrix which is Toeplitz such that $C_{ij} = \rho_{\mathbf{Y}}(x_i - x_j) = \rho_{\mathbf{Y}}(\tau)$. Let $Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$ be a vector containing the observations corresponding with the equidistantly sampled data-points, e.g. $x_i = \frac{i-1}{N-1}$. The sample covariance may then be written as follows

$$C_Y \in \mathbb{R}^{N\times N}, \quad s.t. \quad C_{Y,ij} = \frac{1}{N}\sum_{k,l|\ |k-l|=|i-j|} y_k y_l, \tag{9.67}$$

which is positive semi-definite and Toeplitz. The choice for the term $\frac{1}{N}$ over the familiar $\frac{1}{N-\tau}$ ensures the property of positive-definiteness although it introduces a small bias (Brockwell and Davis, 1987).

Here the assumption of stationarity is essential as it guides the process of averaging out the effect of the i.i.d errors in the observations.

An expression of the theoretical second order moments of the LS-SVM smoother is now derived. Under the assumption that the errors e are i.i.d and conditional independent on f such that $E[e_i|f(x_i)] = 0$, the following equalities hold as

$$\begin{aligned} C_s = E\left[(Y_s+e)(Y_s+e)^T\right] &= \gamma^2 E\left[\Omega e e^T \Omega^T\right] + \sigma_e^2 I_N \\ &= \gamma^2 \Omega E[ee^T]\Omega^T + \sigma_e^2 I_N = \gamma^2 \sigma_e^2 \Omega\Omega^T + \sigma_e^2 I_N. \end{aligned} \tag{9.68}$$

Then substituting the sample covariance matrix $\hat{C}_Y$ into the expression will result into the equalities

$$\hat{C}_Y = C_s \;\; \Rightarrow \;\; \hat{C}_Y - \sigma_e^2 I_N = \gamma^2 \sigma_e^2 \Omega\Omega^T, \tag{9.69}$$

where the constant $\sigma \in \mathbb{R}_0^+$ may dissolve into the kernel matrix. If $N \to \infty$ and $\hat{C}_Y$ is Toeplitz, then also Ω will be Toeplitz (Kailath *et al.*, 2000). This expression leads to the following algorithm

Algorithm 9.1. *Let $\hat{C}_Y$ denote the sample covariance matrix (9.69).*

1. *Determine an appropriate estimate of the noise level underlying the data using model-free techniques as described in e.g. (Pelckmans* et al.*, 2004*a*) such that* $\left(\hat{C}_Y - \hat{\sigma}_e^2 I_N\right) \succeq 0$.

2. *Take the square root of the resulting positive definite Toeplitz matrix. Let $USU^T = \hat{C}_Y - \hat{\sigma}_e^2$ be the singular value decomposition (SVD) such that $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_N) \in \mathbb{R}^{N\times N}$ and $U^TU = UU^T = I_N$, then*

$$\hat{C}_Y - \hat{\sigma}_e^2 I_N = \tilde{\Omega}\tilde{\Omega}^T \;\; \Leftrightarrow \;\; \tilde{\Omega}_T = U \operatorname{diag}(\sqrt{\sigma}_1, \ldots, \sqrt{\sigma}_N)\; U^T. \tag{9.70}$$

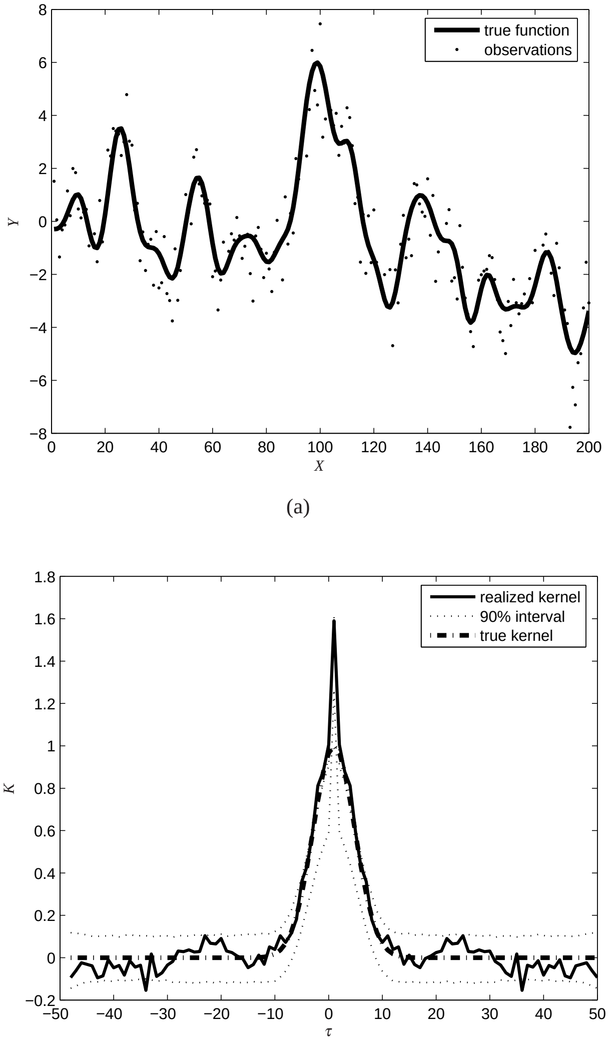
3. *Proper normalization of the resulting matrix $\tilde{\Omega}_T$ leads to a kernel matrix $\hat{\Omega}_T$ and a regularization term $\gamma > 0$. The form of the kernel may be accessed by plotting x_i against the first row of $\hat{\Omega}$.*

The obtained kernel can only be evaluated at the same sampling rate as the original data, which is a severe restriction for most learning tasks. Nevertheless, the plot of the discrete kernel may be used as a tool suggesting the form of the kernel. As in the stochastic realization algorithm (Kung, 1978), realization would amount to look for a parsimonious model description of the kernel (impulse response).

Example 9.6 [Monte Carlo Example of the Realization Approach to Kernel Design] A simple toy dataset is considered to illustrate the realization algorithm. In order to generate an appropriate dataset, the assumptions of the method must be incorporated carefully.

For an optimal trade-off between accuracy and clarity of exposition, the size of the training set is taken $N = 200$. Let the collection $\{(x_i, z_i, \varepsilon_i)\}_{i=1}^N \subset \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ be a set consisting of univariate point locations $x_i \in \mathbb{R}$ which are equidistantly sampled and and two corresponding i.i.d. samples of the standard distribution such that $z_i \sim \mathcal{N}(0,1)$ and $\varepsilon_i \sim \mathcal{N}(0,1)$. A dataset with underlying stationary covariance measure $h : \mathbb{R} \to \mathbb{R}$ is then generated as follows

$$y_i = \sum_{j=1}^N h(x_i - x_j) z_i + \varepsilon_i, \;\; \forall i = 1, \ldots, N. \tag{9.71}$$



(a)

(b)

Figure 9.5: *An example of a kernel realization.* **(a)** *Given* $N = 200$ *noisy data-samples of a nonlinear stationary function generated as a convolution of a white noise sequence with a two-sided function.* **(b)** *The kernel estimate (solid line) resulting from the realization algorithm versus the two-sided convolution function (dashed line) used to generate the data and the* 90% *quantile interval of a Monte Carlo experiment (dotted line). The peek at* $\tau = 0$ *of the kernel estimate is to be attributed to the noise level.*

Following Herglotz's theorem, see e.g. (Brockwell and Davis, 1987), the generated process is stationary if h is a positive definite function. Let in this example h be defined as the familiar mapping

$$h(x_i - x_j) = \exp\left(\frac{-\|x_i - x_j\|_2^2}{\sigma^2}\right), \tag{9.72}$$

with the constant $\sigma = 1$.

Figure 9.5 gives the results of a Monte Carlo experiment. The dataset generated in one specific iteration is given in Figure 9.5.a where the solid line gives the true stationary function and the dots give the actual observations. Panel 9.5.b then gives the realization of the kernel from this data using algorithm 9.1 (solid line). The dashed line indicates the function h employed to generate the data as in (9.72). The dotted lines denote the 90% quantile interval of the Monte Carlo experiment after 1000 iterations. This example shows that one can successfully recover the shape of the kernel from the sample covariances in the data. The peak of the realizations at $\tau = 0$ corresponds with the impact of the noise level on the estimators and is to be attributed to the regularization parameter γ.

9.4.3 The differogram: non-equidistant and multivariate data

The classical case of stochastic realization proceeds by imposing a parametric model on the derived decomposition. This subsection approaches the case of non-equidistantly sampled and higher dimensional data within the same spirit. The main difference is that no discrete state-space model is imposed, but an appropriate element of a parametric class of kernels is identified instead. Hereto, the same tool is used as presented in Appendix A in the context of estimating the noise level.

Definition 9.4. [The differogram] *Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}$ be a dataset. Then define the sample differences as follows*

$$\begin{cases} \Delta_{x,ij} = \|x_i - x_j\|_2 \in \mathbb{R}^+ & \forall i, j = 1, \ldots, N \\ \Delta_{y,ij} = \|y_i - y_j\|_2 \in \mathbb{R}^+ & \forall i, j = 1, \ldots, N \end{cases} \tag{9.73}$$

which are samples of the random variable $\Delta_{\mathbf{X}}$ and $\Delta_{\mathbf{Y}}$ respectively. The differogram is then defined as the

$$\Upsilon(\delta_x) = \frac{1}{2} E\left[\Delta_Y \mid \Delta_X = \delta_x\right]. \tag{9.74}$$

The graphical presentation of all sample differences $\{(\Delta_{x,ij}, \Delta_{y,ij})\}_{i<j}$ is called the differogram cloud.

This definition is closely related to the concept of the variogram (Cressie, 1993) in the concept of spatial data analysis and to the U-statistics as studied in statistics (Lee, 1990). The definition was coined in (Pelckmans *et al.*, 2003*a*) and (Pelckmans *et al.*, 2004*a*) for the purpose of the model free estimation of the noise level. This was based on the following result

$$\sigma_e^2 = \lim_{\Delta_x \to 0} \Upsilon(\Delta_x). \tag{9.75}$$

which was proven in (Pelckmans *et al.*, 2004*a*). Appendix A surveys the main results of this research focussed towards the estimation of the noise level.

Now a simple result extends the use of the differogram to the estimation of auto-covariances in the case of univariate data with stationary covariance $\mathrm{cov}(x_i, x_j) = \rho\left(\|x_i - x_j\|_2\right)$. From the differogram, an expression for the covariance function can be computed as follows

$$\begin{aligned}\Upsilon(\delta_x)^2 &= \frac{1}{2}E\left[(\mathbf{Y}_i - \mathbf{Y}_j)^2 \mid \Delta_{\mathbf{X}} = \Delta_x\right] \\ &= \frac{1}{2}E\left[\mathbf{Y}_i^2 + \mathbf{Y}_j^2 - 2\mathbf{Y}_i\mathbf{Y}_j \mid \Delta_{\mathbf{X}} = \Delta_x\right] \\ &= \sigma_{\mathbf{Y}}^2 - E[\mathbf{Y}_i\mathbf{Y}_j \mid \Delta_{\mathbf{X}} = \Delta_x] \\ &= \sigma_{\mathbf{Y}}^2 - \rho\left(\|x_i - x_j\|_2^2\right). \end{aligned} \tag{9.76}$$

This results in the estimate $\hat{\rho} : \mathbb{R}^+ \to \mathbb{R}$ from the estimated differogram $\hat{\Upsilon} : \mathbb{R}^+ \to \mathbb{R}$

$$\hat{\rho}\left(\|x_i - x_j\|_2^2\right) = \sigma_{\mathbf{Y}}^2 - \hat{\Upsilon}(\delta_x)^2 \tag{9.77}$$

Consider e.g. the parametric differogram model

$$\Upsilon_{h,v,s}(\Delta_x) = v - \exp\left(\frac{-\Delta_x}{h}\right), \quad h, s > 0,\ v > s. \tag{9.78}$$

The use of the following estimator of the model Υ_h was motivated in (Pelckmans *et al.*, 2004*a*).

$$(\hat{h}, \hat{v}, \hat{s}) = \arg\min_{h,v,s} \sum_{i<j} \frac{\left(\Upsilon_{h,v,s}(\Delta_{x,ij}) - \Delta_{y,ij}\right)^2}{\Upsilon_{h,v,s}(\Delta_x)} \quad \text{s.t.} \quad h, s > 0,\ v > s. \tag{9.79}$$

which can be efficiently solved using an iterative approach.

The following result motivates then the a continuous counterpart to the realization context.

Lemma 9.2. [A Stochastic Realization Approach in the case of Non-equidistant Samples] *Let* $\rho : \mathbb{R}^+ \to \mathbb{R}$ *be a stationary covariance function. Then its Fourier transform is positive*

$$\mathscr{F}\rho(\lambda) = \int_{-\infty}^{\infty} \rho(\Delta_x)\exp(-i\Delta_x\lambda)d\Delta_x, \tag{9.80}$$

following from the Hertzglotz theorem, see e.g. (Doob, 1953; Brockwell and Davis, 1987). The square-root decomposition of this function can theoretically be formulated as the pointwise square root of the Fourier transform $\mathscr{F}\rho$ *such that*

$$\begin{aligned}\rho(\|x_i - x_j\|) = \int_{-\infty}^{\infty} k(x_i, z)k(z, x_j)dz \\ \Longleftrightarrow (\mathscr{F}k)(\lambda) = \sqrt{(\mathscr{F}\rho)(\lambda)}, \ \forall -\infty < \lambda < \infty, \end{aligned} \tag{9.81}$$

following Parseval's theorem (Doob, 1953).

Chapter 10

Conclusions

This chapter reviews the most important results of this text and formulates some general conclusive remarks on the discussed methodology. Furthermore, some interesting prospects of the research track are summarized and the general ideas for some paths suitable for future investigation are described.

10.1 Concluding Remarks

The main goal of this dissertation was twofold. At first, we argued that the tasks of design of an appropriate learning algorithm, the determination of the regularization trade-off and the design of an appropriate kernel are interrelated in different ways and should be considered jointly. Secondly we centralized the primal-dual argument originating from the theory on convex optimization in the research on the design of learning machines. To support both conclusions, different new results were studied and reported, including (1) new learning machines as the SVT and kernel machines handling missing values, non i.i.d errors, censored observations and others; (2) incorporating model structure and prior knowledge in the learning algorithm itself and its close relation to the design of kernels. (3) the issue of complexity control or regularization was investigated in some detail and new formulations of such mechanisms are discussed; (4) the notion of hierarchical programming and fusion of training with model selection resulting in an automatic procedure for tuning the global characterization of a variety of learning machines and model selection problems; (5) the relation between techniques in system identification and signal processing on the one hand, and kernel design on the other hand led to new approaches in the task of kernel design.

The text is organized as follows. The introduction surveyed the current state-of-the-art of machine learning and primal-dual kernel machines biased towards the further exposition. Chapter 2 discussed the important backbone for the methodology of primal-dual kernel machines as found in convex optimization theory.

The first part studied the design and analysis of learning machines employing the primal-dual argument in some detail. While the stage was set by the elaboration of the simplest case in the form of the LS-SVM regressor, extensions towards L_1, L_∞ and robust counterparts were formulated in Chapter 3. The following chapter then discussed extensions of those learning machines towards the handling of structure in the form of parametric components, additive model structures and pointwise inequalities. Chapter 5 studied the relationship of the present methodology towards different established approaches in some detail.

The second part discusses the impact and the different forms of complexity control and regularization. Chapter 6 surveyed different forms of regularization methods as found in the literature. A number of extensions made by the author were discussed in the the setting of primal-dual kernel machines. An important contribution in this respect was the formulation of the non-parametric measure of maximal variation. Various consequences of this scheme were elaborated e.g. towards the problem of handling missing values. Chapter 7 then discusses the hierarchical programming argument towards the fusion of training and model selection in a number of parametric and non-parametric cases. Chapter 8 took the argument a step further in the formulation of the additive regularization scheme. This framework was then used for the formulation of fusing training and cross-validation and making stable kernel machines.

Chapters 9 initiated the research on learning the kernel in the context of primal-dual kernel machines. The first three sections discussed some results establishing the relationship between regularization schemes, weighted least squares based primal-dual kernel machines and the design of kernels. The final sections studied a tool which can play a crucial role into the design and learning of kernels by exploiting results in system identification.

10.2 Directions towards Further Work

Mining for invariances and functional relationships

The task of machine learning may be summarized as follows

> "Given a dataset, which patterns and relations are invariantly present?"

The meaning of (statistical) invariance can be formalized as classically in terms of frequency or belief, see e.g. (Fisher, 1922; O'Hagen, 1988; Jaynes, 2003) and (Shawe-Taylor and Cristianini, 2004). An alternative translation may be defined as *"invariant under different realizations"* meaning that any collection of the same variables under different situations should preserve the invariant patterns. The present text follows in many derivations this spirit. For example, *fusion of training and cross-validation* (Part γ) can be alternatively presented as *identify the functional relationship between input* *and output which is mostly invariant over the different folds*. A simple abstract example explains this reasoning: *Given a number of digitized images of writings of the digit "7" collected from different writers, what is the invariant structure over all realizations?*

Although the setting is in some way natural to the unsupervised learning problem, counterparts can be formulated to the supervised case. Consider for example the case of regression. Given a set of measurements of variables, one could ask oneself which set of variables can be explained using a deterministic mapping using which variables:

> "Given a collection of observed variables, which subset can be explained optimally given the remaining variables?"

While this problem of mining for functional relations encompasses classical statistical inference, it can have a high relevance in case studies where even the assumption about which variable acts as output and which as covariate cannot be made a priori. Applications can be found in automatic compression methods and various detection algorithms.

Errors-in-variables and Nonlinear System Identification

The main body of derivations in the text assumed input data which can be considered as deterministic. In the case only perturbed versions of the inputs are observed, the learning problem becomes much more complex. In the case the learning task has no prior assignment of the labels "input" and "outputs", the problem of stochastic components in the variables becomes even more prominent as neglecting of this perturbation cannot be characterized as an assumption. Typical examples include the case of unsupervised learning and time-series prediction. In the case of the latter, a NARX model for example is known to be often inferior in prediction performance compared to nonlinear output error models.

However, a major problem is inherently connected to the setting of stochastic inputs: the errors on the input variables are to be propagated through the (unknown) model. Even in the linear case, this will lead to quadratical constrained (non-convex) optimization problems, which eventually can be solved efficiently using a Singular Value Decomposition or a worst-case analysis. In the setting of nonlinear models, the errors on the inputs have to propagate through the unknown nonlinearity which result in complex global optimization problems. Desiderata in this case would be to formulate efficient optimization problems for solving the described problem approximatively.

Interval Estimation

Most classical learning algorithms focus on point estimators. Inference of the uncertainty of the model is usually obtained via computer intensive sampling methods as bootstrap or Gibbs sampling schemes, or by exploiting sufficient assumptions or approximations as Normality of all involved distributions. However, those approaches digress in spirit from Vapnik's main principle as described in Subsection 1.2.5.

Section 3.5 and Subsection 6.4.3 initiate a direction towards the construction of models for interval estimation based on tolerance intervals. The elaboration of those issues and the analysis of the strategy makes up a new interesting area of research in statistical learning and kernel machines. The relevance is not only given by the frequent need of the users to assess the quality and uncertainty of the prediction, but is also a necessary tool for approaches towards the study of design of experiments (Fisher, 1935) which is also closely related to the next directive.

Interactive Learning and Design of Experiments

The learning task as described may be labeled as *passive* as the analysis draws conclusions (hypothesis) $\mathbb{H}$ based on given data $\mathscr{D}_N$:

$$\mathscr{D} \Rightarrow \mathbb{H}.$$

At least in the social sciences, one more often looks for optimal strategies to investigate a certain phenomenon. A strategy depends amongst others in the way one samples the different outcomes. In the statistical design of experiments one investigates which future data samples are most likely to increase the amount of knowledge of the phenomenon under study. The amount of knowledge is often translated mathematically as the inverse of the variance of the corresponding inferred model. This approach towards the task of learning can be described as *active*. Schematically

$$\mathscr{D}_1 \Rightarrow \mathbb{H}_1 \Rightarrow \mathscr{D}_2 \Rightarrow \mathbb{H}_2 \Rightarrow \cdots \Rightarrow \mathbb{H}_N.$$

Bibliography

Abramowitz, M. and I.G. Stegun (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 9th printing ed.. Dover. New York.

Aizerman, M., E. Braverman and L. Rozonoer (1964). Theoretical foundations of the potential function method in pattern recognition. *Automation and Remote Control* **25**, 821–837.

Akaike, H. (1973). Statistical predictor identification. *Annuals Institute of Statistical Mathematics* **22**, 203–217.

Alizadeh, F. and D. Goldfarb (2003). Second-order cone programming. *Math. Program.* **95**(1), 3–51.

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125–127.

Amato, U., A. Antoniadis and M. Pensky (2004). Wavelet kernel penalized estimation for non-equispaced design regression. Technical report. IAP Statistics Network.

Andrews, D. F., P.J. Bickel, F.R. Hampel, P.J. Huber, W. Roger and J.W. Tukey (1972). *Robust estimation of location*. Princeton University Press.

Anguita, D., A. Boni and S. Ridella (2003). A digital architecture for support vector machines: Theory, algorithm, and FPGA implementation. *IEEE Transactions on Neural Networks* **14**(5), 993–1009.

Anguita, D., A.Boni and A.Zorat (2004). Mapping LSSVM on digital hardware. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN2004)*.

Ansley, C. and W. Wecker (1981). Extensions and applications of the signal extraction approach to regression. In: *Proceedings of ASA-CENSUS-NBER conference on applied time-series*. Washington, D.C.

Antoniadis, A. and I. Gijbels (2002). Detecting abrupt changes by wavelet methods. *Journal of Non-parametric Statistics* **14**(1-2), 7–29.

Antoniadis, A. and J. Fan (2001). Regularized wavelet approximations (with discussion). *Jour. of the Am. Stat. Ass.* **96**, 939–967.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**, 337–404.

Bach, F.R. and M.I. Jordan (2004). Learning graphical models for stationary time series. *IEEE Transactions on Signal Processing* **52**(8), 2189– 2199.

Backus, G. and F. Gilbert (1970). Uniqueness in the inversion of inaccurate gross earth data. *Philos. Trans. Royal Society London* **266**, 123–192.

Baudat, G. and F. Anouar (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation* **12**, 2385–2404.

Bellman, R. and R. Kalaba (1965). *Dynamic Programming and Modern Control Theory*. Academic Press. New York.

Bertero, M., T. Poggio and V. Torre (1988). Ill-posed problems in early vision. *Proceedings of the IEEE* **76**(8), 869–889.

Bertino, E., G. Piero Zarri and B. Catania (2001). *Intelligent Database Systems*. ACM Press. Addison Wesley Professional.

Bhattacharya, C. (2004). Second order cone programming formulations for feature selection. *Journal of Machine Learning Research* **5**, 1417–1433.

Billingsley, P. (1986). *Probability and Measure*. Wiley & Sons.

Birkhoff, G.D. (1931). Proof of the ergodic theorem. *Proceedings Natl. Acad. Sci. U.S.A.* **17**, 656–660.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Boor, C. De and B. Schwartz (1977). Piecewise monotone interpolation. *Journal of Approximation Theory* **21**, 411–416.

Boser, B., I. Guyon and V. Vapnik (1992). A training algorithm for optim margin classifier. In: *In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*. ACM. pp. 144–52.

Bousquet, O. and A. Elisseeff (2002). Stability and generalization. *Journal of Machine Learning Research* **2**, 499–526.

Bousquet, O., S. Boucheron and G. Lugosi (2004). Introduction to statistical learning theory. *in* Advanced Lectures on Machine Learning Lecture Notes in Artificial Intelligence, *eds. O. Bousquet and U. von Luxburg and G. Rätsch.* Springer.

Box, G.E.P. and G.M. Jenkins (1979). *Time Series Analysis, Forecasting and Control*. series, Time Series analysis and digital processing. Holden-Day. Oakland, California.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Boyd, S., C. Crusius and A. Hansson (1998). Control applications of nonlinear convex programming. *Journal of Process Control* **8**(5-6), 313–324.

Boyd, S., L. El Ghaoui, E. Feron and V. Balakrishnan (1993). Linear matrix inequalities in system and control theory. In: *Proceedings Annual Allerton Conf. on Communication, Control and Computing*. Allerton House, Monticello, Illinois.

Boyd, S., L. El Ghaoui, E. Feron and V. Balakrishnan (1994). *Linear Matrix Inequalities in System and Control Theory*. SIAM.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**(2), 123–140.

Brockwell, P. J. and A. D. Davis (1987). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer-Verlag.

Buckley, M.J. and G.K. Eagleson (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75**(2), 189–199.

Burman, P. (1989). A comparative study of ordinary cross-validation, *v*-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**(3), 123–140.

Cawley, G.C. and N.L.C. Talbot (2003). Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition* **36**(11), 2585–2592.

Chapelle, O., V. Vapnik, O. Bousquet and S. Mukherjee (2002). Choosing multiple parameters for support vector machines. *Machine Learning* **46**(1-3), 131–159.

Chebyshev, P. L. (1859). Sur les questions de minima qui se rattachent la reprsentation approximative des fonctions. *Mmoires Academie des Science Petersburg* **7**, 199–291. *Oeuvres de P. L. Tchebychef*, 1, 273-378, Chelsea, New York, 1961.

Chen, S.S., D.L. Donoho and M.A. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM Review* **43**(1), 129–159.

Cherkassky, V. (submitted, 2002). Practical selection of svm parameters and noise estimation for svm regression. *Neurocomputing, Special Issue on SVM*.

Conover, W.J. (1999). *Practical Nonparameteric Statistics*. Wiley.

Cortes, C. and V. Vapnik (1995). Support vector networks. *Machine Learning* **20**(3), 273–297.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistics Society B* (34), 187–220.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press. Princeton, N.J.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–390.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley.

Cristianini, N. and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

Cucker, F. and S. Smale (2002). Best choices for regularization parameters in learning theory: On the bias-variance problem. *Foundations of Computational Mathematics* **2**(4), 413–428.

Dantzig, G.B. (1963). *Linear Programming and Extensions*. Princeton University Press. Princeton, NJ.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41**, 909–996.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

De Brabanter, J. (2004). LS-SVM regression modelling and its applications. PhD thesis. Faculty of Engineering, K.U.Leuven. Leuven, Belgium. 243 pages.

De Brabanter, J., K. Pelckmans, J.A.K. Suykens and B. De Moor (2002*a*). Robust cross-validation score function for LS-SVM non-linear function estimation. Internal Report 02-94. KULeuven - ESAT. Leuven, Belgium.

De Brabanter, J., K. Pelckmans, J.A.K. Suykens and B. De Moor (2003). Robust complexity criteria for nonlinear regression in NARX models. In: *Proceedings of the 13th System Identification Symposium (SYSID2003)*. Rotterdam, Netherlands. pp. 79–84.

De Brabanter, J., K. Pelckmans, J.A.K. Suykens and B. De Moor (2004). Robust statistics for kernel based NARX modeling. Internal Report 04-38. KULeuven - ESAT. Leuven, Belgium.

De Brabanter, J., K. Pelckmans, J.A.K. Suykens, J. Vandewalle and B. De Moor (2002*b*). Robust cross-validation score function for non-linear function estimation. In: *International Conference on Artificial Neural Networks (ICANN 2002)*. Madrid, Spain. pp. 713–719.

De Cock, K., B. De Moor and B. Hanzon (2003). On a cepstral norm for an ARMA model and the polar plot of the logarithm of its transfer function. *Signal Processing* **83**(2), 439–443.

De Moor, B., Pelckmans K., Hoegaert L. and Barrero O. (2002). Linear and non-linear modeling in soft4s. Technical report. ESAT-SISTA, K.U.Leuven. Leuven, Belgium.

De Smet, F. (2004). Microarrays : algorithms for knowledge discovery in oncology and molecular biology. PhD thesis. Faculty of Engineering, K.U.Leuven. Leuven, Belgium.

Decoste, D. and B. Schölkopf (2002). Training invariant support vector machines. *Machine Learning* **46**(1-3), 161 – 190.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Jour. of the Royal Stat. Soc. series B* **39**, 1–38.

Devroye, L. and L. Györfi (1985). *Nonparametric Density Estimation: The L_1 View*. John Wiley. New York.

Devroye, L., L. Györfi and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag.

Devroye, L., L. Györfi, D. Schäfer and H. Walk (2003). The estimation problem of minimum mean squared error. *Statistics and Decisions* **21**, 15–28.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Oxford University Press.

Dietterich, T.G. and G. Bakiri (1995). Solving multi-class learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research* **2**, 263–286.

Diggle, P.J. (1990). *Time-series: A biostatistical introduction*. Oxford University Press, Oxford.

Dodd, T.J. and C.J. Harris (2002). Identification of nonlinear time series via kernels. *International Journal of Systems Science* **33**(9), 737–750.

Donoho, D. and I. Johnstone (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika* **81**, 425–455.

Doob, J.L. (1953). *Stochastic Processes*. Wiley Publications in Statistics. John Wiley & Sons.

Duchon, J. (1977). *Spline Minimizing Rotation Invariant Semi-norms in Sobolev Spaces*. Vol. 571 of *Lecture Notes in Mathematics*. Springer-Verlag. Berlin.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**(1), 1–26.

Efron, B., T. Hastie, I. Johnstone and R. Tibshirani (2004). Least angle regression. *Annals of Statistics* **32**(2), 407–499.

El Ghaoui, L. and H. Lebret (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal Matrix Analysis and Applications* **18**(4), 1035–1064.

Engle, R., C. Granger, J. Rice and A. Weiss (1986). Semiparameteric estimates of the relation between weather and electricity sales. *Journal of the American Statistics Society* **81**, 310–320.

Espinoza, M., K. Pelckmans, L. Hoegaerts, J.A.K. Suykens and B. De Moor (2004). A comparative study of LS-SVMs applied to the silverbox identification problem. In: *Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS 2004)*. Stuttgart, Germany.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. Vol. 157. Marcel Dekker, New York.

Fan, J. (1997). Comments on wavelets in statistics: A review. *Journal of the Italian Statistical Association* (6), 131–138.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag. New York. 570pp.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London, A* **222**, 309–368.

Fisher, R.A (1935). *The Design of Experiments*. Oliver and Boyd. Edinburgh.

Frank, L.E. and J.H. Friedman (1993). A statistical view of some chemometric regression tools. *Technometrics* **35**, 109–148.

Friedman, J. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* **84**, 165–175.

Friedman, J. and J.W. Tukey (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **23**, 881–890.

Friedmann, J. H. and W. Stuetzle (1981). Projection pursuit regression. *Jour. of the Am. Stat. Assoc.* **76**, 817–823.

Fu, W.J. (1998). Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Fukumizu, K., F. R. Bach and M. I. Jordan (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning* **5**, 73–99.

Fung, G. and O.L. Mangasarian (2001). Proximal support vector machine classifiers. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD 2001)* (Association for Computing Machinery, Ed.). San Francisco. pp. 77–86.

Gasser, T., L. Sroka and C. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

Gaylord, C.K. and D.E. Ramirez (1991). Monotone regression splines for smoothed bootstrapping. *Computational Statistics Quarterly* **6**(2), 85–97.

Genin, Y., Y. Hachez, Yu. Nesterov and P. van Dooren (2003). Optimization problems over positive pseudopolynomial matrices. *SIAM Journal on Matrix Analysis and Applications* **25**(1), 57–79.

Genton, M. (2001). Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research* **2**, 299–312.

Girolami, M. (2002). Orthogonal series density and the kernel eigenvalue problem. *Neural Computation* **14**(3), 669–688.

Girosi, F., M. Jones and T. Poggio (1995). Regularization theory and neural networks architectures. *Neural Computation* **7**, 219–269.

Goethals, I. (2005). Subspace Identification for linear, Hammerstein and Hammerstein-Wiener systems. PhD thesis. ESAT, KULeuven. Leuven, Belgium.

Goethals, I., K. Pelckmans, J.A.K. Suykens and B. De Moor (2004*a*). NARX identification of Hammerstein models using least squares support vector machines. In: *Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS 2004)*. Stuttgart, Germany. pp. 507–512.

Goethals, I., K. Pelckmans, J.A.K. Suykens and B. De Moor (2004*b*). Subspace identification of Hammerstein systems using least squares support vector machines. Technical report. ESAT-SISTA, K.U.Leuven. Leuven, Belgium.

Goethals, I., K. Pelckmans, J.A.K. Suykens and B. De Moor (2005*a*). Identification of MIMO Hammerstein models using least squares support vector machines. *Automatica*.

Goethals, I., K. Pelckmans, L. Hoegaerts, J.A.K. Suykens and B. De Moor (2005*b*). Subspace intersection algorithm for Hammerstein-Wiener systems. Technical report. ESAT-SISTA, K.U.Leuven.

Goethals, I., L. Hoegaerts, J.A.K. Suykens and B. De Moor (2004*c*). Kernel canonical correlation analysis for the identification of Hammerstein-Wiener models. Technical report. ESAT-SISTA, K.U.Leuven. Leuven, Belgium.

Goldfarb, D. and G. IYengar (2003). Robust convex quadratically constrained programs. *Mathematical Programming Series B* **97**, 495–515.

Golub, G. H. and C. F. van Loan (1989). *Matrix Computations*. John Hopkins University Press.

Golub, G. H., M. Heath and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

Grant, M.C. (2004). Disciplined Convex Programming. PhD thesis. Stanford, Electrical Engineering.

Grenander, U. and M. Rosenblatt (1957). *Statistical Analysis of Stationary Time Series*. John Wiley and Sons. New York.

Grötschel, M., L. Lovasz and A. Schrijver (1988). *Geometric Algorithms and Combinatorial Optimization*. Springer.

Gunn, S. R. and J. S. Kandola (2002). Structural modelling with sparse kernels. *Machine Learning* **48**(1), 137–163.

Haar, A. (1910). Zur theorie der orthogonalen funktionen-systeme. *Mat. Ann.* **69**, 331–371.

Hamers, B. (2004). Kernel Models for Large Scale Applications. PhD thesis. Faculty of Engineering, K.U.Leuven. Leuven. 218 p., 04-105.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel (1986). *Robust statistics, the approach based on influence functions*. Wiley & Sons, New York.

Hanley, J.A. and B.J. McNeil (1982). The meaning and use of the area under a receiver operating characteristics. *Radiology* **143**, 29–36.

Hansen, P.C. (1992). Analysis of discrete ill-posed problems by means of the l-curve. *SIAM Review* **34**(4), 561–580.

Hansen, P.C. (1998). *Rank-deficient and Discrete Ill-posed Problems*. SIAM.

Hardle, W. (1990). *Applied Nonparameteric Regression*. Vol. 19 of *Econometric Society Monographs*. Cambridge University Press.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models*. Chapman and Hall.

Hastie, T., R. Tibshirani and J. Friedman (2001). *The Elements of Statistical Learning*. Springer-Verlag. Heidelberg.

Hastie, T., S. Rosset and R. Tibshirani (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* **5**, 1391–1415.

Herbrich, R. (2001). *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press.

Herrmann, D.J.L. and O. Bousquet (2003). *Advances in Neural Information Processing Systems*. Chap. On the Complexity of Learning the Kernel Matrix, pp. 399–406. Vol. 15. MIT Press. Cambridge, MA.

Hettmansperger, T.P. and J.W. McKean (1994). *Robust Nonparametric Statistical Methods*. Vol. 5 of *Kendall's Library of Statistics*. Arnold.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325.

Hoegaerts, L. (2005). Eigenspace Methods and Subset Selection in Kernel based Learning. PhD thesis. SCD - ESAT - KULeuven.

Hoerl, A. E., R. W. Kennard and K. F. Baldwin (1975). Ridge regression: Some simulations. *Communications in Statistics, Part A - Theory and Methods* **4**, 105–123.

Hoerl, A.E. and R.W. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–82.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.

Ivanov, V.V. (1976). *The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations*. Nordhoff International.

Jaakkola, T.S. and D. Haussler (1999). Probabilistic kernel regression models. In: *Proceedings of the 1999 Conference on AI and Statistics*. Morgan Kaufmann.

Jaynes, E. T. (2003). *Probability Theory, The Logic of Science*. Cambridge University Press.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Kluwer.

Jollife, I.T. (1986). *Principal Component Analysis*. Springer-Verlag.

Kailath, T., A.H. Sayed and B. Hassibi (2000). *Linear Estimation*. Prentice Hall.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*.

Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. *Combinatorica* **4**(4), 373–395.

Kearns, M. (1997). A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split. *Neural Computation* (4), 698–714.

Klein, J.P., M.L. Moeschberger and L. Melvin (1997). *Survival Analysis, Techniques for Censored and Truncated Data*. Springer.

Kocijan, J., A. Girard, B. Banko and R. Murray-Smith (2003). Dynamic systems identification with gaussian processes. In: *Proceedings of the 4th Mathmod conference*. Int. Association. for Mathematics and Computers in Simulation. Vienna.

Kolmogorov, A.N. (1933). *Foundations of the Theory of Probability*. second english edition ed.. Chelsea Publishing Company. New York.

Kung, S.Y. (1978). A new identification method and model reduction algorithm via singular value decomposition. In: *Proceedings of the 12th Asilomar Conference on Circuits, Sytems and Computation*. pp. 705–714.

Lanckriet, G.R.G., L. El Ghaoui, C. Bhattacharyya and M.I. Jordan (2002). A robust minimax approach to classification. *Journal of Machine Learning Research* **3**, 555–582.

Lanckriet, G.R.G., N. Cristianini, P. Bartlett and M.I. Jordan L. El Ghaoui (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research* **5**, 27–72.

Lee, A.J. (1990). *U-statistics, Theory and Practice*. Marcel Dekker. New York.

Letac, G. and H. Massam (2004). All invariant moments of the Wishart distribution. *Scandinavian Journal of Statistics* **31**, 295–318.

Linton, O. B. and J. P. Nielsen (1995). A kernel method for estimating structured nonparameteric regression based on marginal integration. *Biometrika* **82**, 93–100.

Little, R.J.A. and D.B. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley.

Ljung, L. (1987). *System Identification, Theory for the User*. Prentice Hall.

Lobo, M.S., L. Vandenberghe, S. Boyd and H. Lebret (1998). Applications of second order programming. *Linear Algebra and its Applications* **284**, 193–228.

Loeve, M. (1955). *Probability Theory*. D. van Nostrand. New York.

Luenberger, D.G. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.

MacKay, D. J. C. (1992). The evidence framework applied to classification networks. *Neural Computation* **4**, 698–714.

MacKay, D.J. (1998). *Introduction to Gaussian processes*. Vol. 168 of *NATO Asi Series. Series F, Computer and Systems Sciences*. Springer Verlag.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

Mangasarian, O.L. and D.R. Musicant (1999). Succesive overrelexations for support vector machines. *Neural Networks* **10**, 1032–1037.

Mardia, K.V., J.T. Kent and J.M. Bibby (1979). *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press.

Markowitz, H. (1956). The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly* **3**, 111–133.

Mattera, D. and S. Haykin (2001). Support vector machines for dynamic reconstruction of a chaotic system. *in* Advances in Kernel Methods, *eds. B. Schölkopf and C. Burges and A. Smola* pp. 211–241. MIT Press.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London, A* **209**, 415–446.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Molenberghs, G., M.G. Kenward and E. Lesaffre (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika*.

Mood, A.M., F.A. Graybill and D.C. Boes (1963). *Introduction to the Theory of Statistics*. Series in Probability and Statistics. McGraw-Hill.

Morozov, V.A. (1984). *Methods for Solving Incorrectly Posed Problems*. Springer-Verlag.

Mukherjee, S., E. Osuna and F. Girosi (1997). Nonlinear prediction of chaotic time series using a support vector machine. In: *Proceedings of the 1997 IEEE Workshop on Neural Networks for Signal Processing* (J. Principe, L. Gile, N. Morgan and E. Wilson, Eds.). Vol. VII.

Müller, K.-R., A.J. Smola, Gunnar Rätsch, B. Schölkopf, J. Kohlmorgen and V. Vapnik (1999). Using support vector machines for time series prediction. *in* Advances in Kernel Methods - Support Vector Learning, *eds. B. Schölkopf and C. Burges and A. Smola*. MIT Press, Cambridge, MA.

Müller, U.U., A. Schick and W. Wefelmeyer (2003 (to appear)). Estimating the error variance in nonparametric regression by a covariate-matched u-statistic. *Statistics*.

Neal, R.N. (1994). Bayesian Learning for Neural Networks. PhD thesis. Dept. of Computer Science, University of Toronto.

Nesterov, Y. (1998). Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods & Software* **9**(1-3), 141–160.

Nesterov, Y. and A. Nemirovski (1994). *Interior-Point Polynomial Methods in Convex Programming: Theory and Applications*. Society for Industrial and Applied Mathematics (SIAM). Philadelphia.

Nesterov, Y. and M.J. Todd (1997). Self-scaled barriers and interior point methods for convex programming. *Mathematics of Operational Research* **22**(1), 1–42.

Neter, J., W. Wasserman and M.H. Kutner (1974). *Applied Linear Models, Regression, Analysis of Variance and Experimental Designs*. third ed.. Irwin.

Neumaier, A. (1998). Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review* **40**(3), 636–666.

Neyman, J. and E.S. Pearson (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, part I.. *Biometrika* **15**, 175–240.

Nocedal, J. and S.J. Wright (1999). *Numerical Optimization*. Springer Series in Operational Research. Springer. New York.

O'Hagen, A. (1978). On curve fitting and optimal design for regression. *Journal of the Royal Statistical Society B* **40**, 1–42.

O'Hagen, A. (1988). *Probability: Methods and Measurements*. Chapmann & Hall. London.

Osborne, M.R., B. Presnell and B.A. Turlach (2000). On the LASSO and its dual. *Journal of Computational & Graphical Statistics*.

Pareto, V. (1971). *Manual of Political Economy*. A.M. Kelley Publishers. First Published in Italian in 1906.

Parzen, E. (1961). An approach to time series analysis. *The Annals of Statistics* **32**(4), 951–989.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–1076.

Parzen, E. (1970). Statistical inference on time series by RKHS methods. In: *Proceedings 12th Biennial Seminar on Time Series Analysis*, ed. R. Pyke. Montreal, Canada.

Pearson, K. (1902). On the systematic fitting of curves to observations and measurements. *Biometrika* **1**, 265–303.

Pelckmans, K., I. Goethals, J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*a*). Componentwise least squares support vector machines. Chapter in *Support Vector Machines: Theory and Applications*, L. Wang (Ed.), Springer pp. 77–98.

Pelckmans, K., I. Goethals, J.A.K Suykens and B. De Moor (2005*a*). On model complexity control in identification of hammerstein systems. Technical report. ESAT-SISTA, K.U.Leuven. Leuven, Belgium.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2003*a*). Variogram based noise variance estimation and its use in kernel based regression. In: *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing (NNSP 2003)*. Toulouse, France. pp. 199–208.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*b*). The differogram: Nonparametric noise variance estimation and its use for model selection. *Accepted for publication in Neurocomputing*.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*c*). Regularization constants in LS-SVMs : a fast estimate via convex optimization. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2004)*. Budapest, Hungary. pp. 699–704.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2005*b*). Maximal variation and missing values for componentwise support vector machines. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2005)*. IEEE. Montreal, Canada.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2003*b*). Additive regularization trade-off: Fusion of training and validation levels in kernel methods. *Internal Report 03-184, ESAT-SISTA, K.U.Leuven, Belgium, submitted*.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2004*d*). Alpha and beta stability for additively regularized LS-SVMs via convex optimization. In: *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2004)*. Leuven, Belgium.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2004*e*). Morozov, ivanov and tikhonov regularization based LS-SVMs. In: *Proceedings of the11th International Conference on Neural Information Processing (ICONIP 2004)*. Kolkata, India.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2004*f*). Sparse LS-SVMs using additive regularization with a penalized validation criterion. In: *Proceedings of the 12e European Symposium on Artificial Neural Networks*. pp. 435–440.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2005*c*). Building sparse representations and structure determination on LS-SVM substrates. *Neurocomputing* **64**, 137–159.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2005*d*). Componentwise support vector machines for structure detection. Technical report. ESAT-SISTA, K.U.Leuven, accepted on International Conference on Artificial Neural Networks (ICANN 2005). Leuven, Belgium.

Pelckmans, K., J.A.K. Suykens, T. van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor and J. Vandewalle (2002*a*). LS-SVMlab : a matlab/c toolbox for least squares support vector machines. Tutorial. KULeuven - ESAT. Leuven, Belgium.

Pelckmans, K., J.A.K. Suykens, T. van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor and J. Vandewalle (2002*b*). LS-SVMlab : a matlab/c toolbox for least squares support vector machines. Demonstration at NIPS 2002 02-44. KULeuven - ESAT.

Pelckmans, K., M. Espinoza, J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*g*). Primal-dual monotone kernel machines. *Accepted for publication in Neural Processing Letters*.

Perrone, M.P. and L.N. Cooper (1993). When networks disagree: Ensemble method for neural networks. *in* eural Networks for Speech and Image Processing *ed. R.J. Mammone*. Chapman-Hall.

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. *in* Advances in Kernel Methods - Support Vector Learning, *eds. B. Schölkopf and C. Burges and A. Smola* pp. 185–208. MIT Press.

Pochet, N., F. De Smet, J.A.K. Suykens and B. De Moor (2004). Systematic benchmarking of micorarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics* **20**(17), 3185–3195.

Powell, M.J.D. (1981). *Approximation Theory and Methods*. Cambridge University Press. Cambridge.

Press, W.H., A.A. Teukolsky, W.T. Vetterling and B.P. Flannery (1988). *Numerical recipes in C*. The art of scientific computing. Cambridge University Press.

Rao, C.R. (1965). *Linear Statistical Inference and Its Applications*. Wiley Series in Probablity and Mathematical Statistics. John Wiley & Sons.

Rao, P. (1983). *Nonparameteric Function Estimation*. Probability and Mathematical Statistics. Academic press.

Rasmussen, C.E. (1996). Evaluation of Gaussian Processes and other Methods for Non-Linear Regression. PhD thesis. graduate department of Computer Science, University of Toronto.

Rätsch, G. (2001). Robust Boosting via convex Optimization: Theory and Applications. PhD thesis. University of Potsdam.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. of Statist* **12**, 1215–1230.

Rice, J.A. (1988). *Mathematical statistics and data analysis*. Duxbury Press. Pacific Grove, California.

Rifkin, R. (2002). Everything Old Is New Again: A Fresh Look at Historical Approaches in Machine Learning. PhD thesis. MIT.

Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica* **14**, 465–471.

Rockafellar, R.T. (1970). *Convex Analysis*. Princeton University Press.

Rockafellar, R.T. (1993). Lagrange multipliers and optimality. *SIAM Review* **35**, 183–283.

Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

Rubinstein, R.Y. (1981). *Simulation and the Monte Carlo Method*. Wiley.

Saunders, C., A. Gammerman and V. Vovk (1998). Ridge regression learning algorithm in dual variables. In: *Proceedings of the 15th Int. Conf. on Machine learning(ICML'98)*. Morgan Kaufmann. pp. 515–521.

Schoenberg, I.J. (1946). Contribution to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics* **4**(2), 45–99 & 112–141.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. MIT Press. Cambridge, MA.

Schölkopf, B., R. Herbrich and A.J. Smola (2001). A generalized representer theorem. In: *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*. pp. 416–426.

Schölkopf, B., Tsuda, K. and Vert, J.-P., Eds.) (2004). *Kernel Methods in Computational Biology*. MIT Press.

Schoukens, J., J.G. Nemeth, P. Crama, Y. Rolain and R. Pintelon (2003). Fast approximate identification of nonlinear systems. *Automatica* **39**(7), 1267–1274.

Schumaker, L.L. (1981). *Spline functions: basic theory*. John Wiley & Sons. New York.

Schwartz, G. (1979). Estimating the dimension of a model. *Annuals of Statistics* **6**, 461–464.

Scott, D.W. (1992). *Multivariate Density Estimation, theory, practice and visualization*. Wiley series in probability and mathematical statistics. Wiley.

Sen, A. and M. Srivastava (1990). *Regression Analysis, Theory, Methods, and Applications*. Springer.

Shawe-Taylor, J. and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Singer, Y. (2003). Multiclass learning with output codes. *in* Advances in Learning Theory: Methods, Models and Applications, *eds. Suykens, J.A.K., G. Horvath, B. Sankar, C. Michelli and J. Vandewalle* **190**, 251–266. IOS Press.

Spanos, A. (1999). *Probability Theory and Statistical Inference*. Econometric Modeling with Observational Data. Cambridge University Press.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: *Proceedings of the third Berkeley Symposium on Mathematical Probability* (University of California Press, Ed.). Berkeley. pp. 197–206.

Stitson, M., A. Gammerman, V. Vapnik, V. Vovk, C. Watkins and J. Weston (1999). Support vector regression with ANOVA decomposition kernels. *in* Advanced in Kernel methods: Support Vector Learning, *eds. B. Schölkoph, B. Burges and A. Smola*. The MIT Press, Cambridge Massachusetts.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* **13**, 1040–1053.

Stone, C.J. (1985). Additive regression and other nonparameteric models. *Annals of Statistics* **13**, 685–705.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistics Society Series* **B**(36), 111–147.

Sturm, J.F. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software* **11-12**, 625–653.

Suykens, J.A.K. and J. Vandewalle (1999). Least squares support vector machine classifiers. *Neural Processing Letters* **9**(3), 293–300.

Suykens, J.A.K., Horvath G., Basu S., Micchelli C. and Vandewalle J. (eds.) (2003*a*). *Advances in Learning Theory: Methods, Models and Applications*. Vol. 190 of *NATO Science Series III: Computer & Systems Sciences*. IOS Press Amsterdam.

Suykens, J.A.K., J. De Brabanter, L. Lukas and J. Vandewalle (2002*a*). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing* **48**(1-4), 85–105.

Suykens, J.A.K., J. Vandewalle and B. De Moor (2001). Optimal control by least squares support vector machines. *Neural Networks* **14**(1), 23–35.

Suykens, J.A.K., L. Lukas, P. van Dooren, B. De Moor and J. Vandewalle (1999). Least squares support vector machine classifiers: a large scale algorithm. In: *Proceedings of the European Conference on Circuit Theory and Design (ECCTD'99)*. Stresa, Italy. pp. 839–842.

Suykens, J.A.K., T. van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle (2002*b*). *Least Squares Support Vector Machines*. World Scientific, Singapore.

Suykens, J.A.K., T. van Gestel, J. Vandewalle and B. De Moor (2003*b*). A support vector machine formulation to PCA analysis and its kernel version. *IEEE Transactions on Neural Networks* **14**(2), 447–450.

Tax, D.M.J and R.P.W. Duin (1999). Support vector domain description. *Pattern Recognition Letters* **20**(11-13), 1191–1199.

Tibshirani, R.J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society* **58**, 267–288.

Tikhonov, A. N. and V. Y. Arsenin (1977). *Solution of Ill-Posed Problems*. Winston. Washington DC.

Todd, M.J. (2002). The many facets of linear programming. *Mathematical Programming Series B* **91**, 417–436.

Trafalis, T.B. and S.A. Alwazzi (2003). Robust support vector regression and applications. *in* Intelligent Engineering Systems Through Artificial Neural Networks, *eds. C.H. Dagli and A.L. Buczak and J. Ghosh and M.J. Embrechts and O. Ersoy and S.W. Kercel* **13**, 181–186. ASME Press.

Tukey, J.W. (1958). Bias and confidence in not quite large samples. *Abstract. Ann. Math. Statist* **29**, 614.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley. Reading, MA.

Valentini, G. and T.G. Dietterich (2004). Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. *Journal of Machine Learning Research* **5**, 725–775.

Van Dooren, P. (2004). The basics of developing numerical algorithms. *Control Systems Magazine* pp. 18–27.

Van Gestel, T. (2002). From linear to Kernel Based Methods in Classification, Modelling and Prediction. PhD thesis. Faculty of Engineering, K.U.Leuven. Leuven. 286 p.,02-104.

Van Gestel, T., J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor and J. Vandewalle (2002). A bayesian framework for least squares support vector machine classifiers, gaussian processes and kernel fisher discriminant analysis. *Neural Computation* **14**(5), 1115–1147.

Vanoverschee, P. and B. De Moor (1996). *Subspace Identification for Linear System: Theory, Implementation, Applications*. Kluwer Academic Publishers.

Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag. New York.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag. New York.

Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley and Sons.

Verleysen, M. (2003). *Limitations and future trends in neural computation*. Chap. Learning high-dimensional data, pp. 141–162. IOS Press. Amsterdam (The Netherlands).

von Luxburg, U., O. Bousquet and B. Schölkopf (2004). A compression approach to support vector model selection. *Journal of Machine Learning Research* (5), 293–323.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya A* **26**, 359–372.

Weinert, H.L. (1982). *Reproducing Kernel Hilbert Spaces*. Hutchinson Ross Publishing Company. New York.

Weston, J., A. Elisseeff, B. Schölkopf and M. Tipping (2003). Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Methods* **3**, 1439–1461.

Wetherill, G.B. (1986). *Regression Analysis with Applications*. Monographs on Statistics and Applied Probability. Chapman and Hall.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434–449.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series. With Engineering Applications*. Classics Series. The MIT Press.

Yu, Y., W. Lawton, S.L. Lee, S. Tan and J. Vandewalle (1998). Wavelet based modeling of nonlinear systems. *in* Nonlinear Modeling, Advanced Blackbox Techniques, *eds. J.A.K. Suykens and J. Vandewalle* pp. 119–148. Kluwer Academic Publisher.

Zhang, S. (2000). Quadratic maximization and semidefinite relaxation. *Mathematical Programming*.

Biography

Kristiaan Pelckmans was born at 3 november 1978 in Merksplas, Belgium. He received a M.Sc. degree ("Licentiaat") in Computer Science in 2000 from the Katholieke Universiteit Leuven. After a projectwork for an implementation of kernel machines and LS-SVMs (LS-SVMlab), he currently pursues a Ph.D. at the KULeuven in the faculty of engineering, department of Electrical Engineering in the SCD/ SISTA laboratory. His research mainly focusses on machine learning and statistical inference using primal-dual kernel machines.

List of Publications

Book Chapter

Pelckmans, K., I. Goethals, J. De Brabanter, J.A.K. Suykens and B. De Moor (2004). Componentwise least squares support vector machines. Chapter in *Support Vector Machines: Theory and Applications*, L. Wang (Ed.), Springer, pp. 77-98.

Accepted Journal Papers

Pelckmans, K., J.A.K. Suykens and B. De Moor (2005*c*). Building sparse representations and structure determination on LS-SVM substrates. *Neurocomputing*, Special Issue, vol. 64, Mar. 2005, pp. 137-159.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*a*). The differogram: Nonparametric noise variance estimation and its use for model selection. *Neurocomputing,* in press.

Goethals, I., K. Pelckmans, J.A.K. Suykens and B. De Moor (2005*a*). Identification of MIMO Hammerstein models using least squares support vector machines. *Automatica*.

Pelckmans, K., M. Espinoza, J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*g*). Primal-dual monotone kernel machines. *Neural Processing Letters, accepted*.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2005*b*). Handling missing values in support vector machine classifiers. *Accepted for publication in Neural Networks*.

Accepted Papers at International Conferences

De Brabanter, J., K. Pelckmans, J.A.K. Suykens, J. Vandewalle and B. De Moor (2002*b*). Robust cross-validation score function for non-linear function estima-

tion. In: *International Conference on Artificial Neural Networks (ICANN 2002)*. Madrid, Spain. pp. 713–719.

De Brabanter, J., K. Pelckmans, J.A.K. Suykens and B. De Moor (2003). Robust complexity criteria for nonlinear regression in NARX models. In: *Proceedings of the 13th System Identification Symposium (SYSID2003)*. Rotterdam, Netherlands. pp. 79–84.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2003*a*). Variogram based noise variance estimation and its use in kernel based regression. In: *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing (NNSP 2003)*. Toulouse, France. pp. 199–208.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2004*e*). Sparse LS-SVMs using additive regularization with a penalized validation criterion. In: *Proceedings of the 12e European Symposium on Artificial Neural Networks*. pp. 435–440.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2004*c*). Alpha and beta stability for additively regularized LS-SVMs via convex optimization. In: *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2004)*. Leuven, Belgium.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*b*). Regularization constants in LS-SVMs : a fast estimate via convex optimization. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2004)*. Budapest, Hungary. pp. 699–704.

Espinoza, M., K. Pelckmans, L. Hoegaerts, J.A.K. Suykens and B. De Moor (2004). A comparative study of LS-SVMs applied to the silverbox identification problem. In: *Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS 2004)*. Stuttgart, Germany.

Goethals, I., K. Pelckmans, J.A.K. Suykens and B. De Moor (2004*a*). NARX identification of Hammerstein models using least squares support vector machines. In: *Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS 2004)*. Stuttgart, Germany. pp. 507–512.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2004*d*). Morozov, ivanov and tikhonov regularization based LS-SVMs. In: *Proceedings of the11th International Conference on Neural Information Processing (ICONIP 2004)*. Kolkata, India.

Pelckmans, K., J. De Brabanter, J.A.K. Suykens and B. De Moor (2005*b*). Maximal variation and missing values for componentwise support vector machines. In: *in Proceedings International Joint Conference on Neural Networks (IJCNN 2005)*.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2005*e*). Componentwise support vector machines for structure detection. Technical report. Accepted on International Conference on Artificial Neural Networks, ICANN 2005.

Submitted Papers

Submitted Journal Papers

De Brabanter, J., K. Pelckmans, J.A.K. Suykens and B. De Moor (2002*a*). Robust cross-validation score function for LS-SVM non-linear function estimation. Internal Report 02-94. KULeuven - ESAT. Leuven, Belgium.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2003*b*). Additive regularization trade-off: Fusion of training and validation levels in kernel methods. *Internal Report 03-184, ESAT-SISTA, K.U.Leuven, Belgium, submitted*.

De Brabanter, J., K. Pelckmans, J.A.K. Suykens and B. De Moor (2004). Robust statistics for kernel based NARX modeling. Internal Report 04-38. KULeuven - ESAT. Leuven, Belgium.

Goethals, I., K. Pelckmans, J.A.K. Suykens and B. De Moor (2004*b*). Subspace identification of Hammerstein systems using least squares support vector machines. Technical report. ESAT-SISTA, K.U.Leuven. Leuven, Belgium, *Submitted to IEEE Transactions on Automatic Control, conditionally accepted*.

Submitted Conference Papers

Goethals, I., K. Pelckmans, L. Hoegaerts, J.A.K. Suykens and B. De Moor (2005*b*). Subspace intersection algorithm for Hammerstein-Wiener systems. Technical report. ESAT-SISTA, K.U.Leuven.

Pelckmans, K., I. Goethals, J.A.K Suykens and B. De Moor (2005*a*). On model complexity control in identification of hammerstein systems. Technical report. ESAT-SISTA, K.U.Leuven. Leuven, Belgium.

Internal Reports

Pelckmans, K., J.A.K. Suykens, T. Van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor and J. Vandewalle (2002*a*). LS-SVMlab : a matlab/c toolbox for least squares support vector machines. Tutorial. KULeuven - ESAT. Leuven, Belgium.

Pelckmans, K., J.A.K. Suykens, T. Van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor and J. Vandewalle (2002*b*). LS-SVMlab : a matlab/c toolbox for least squares support vector machines. Demonstration at NIPS 2002 02-44. KULeuven - ESAT.

De Moor, B., Pelckmans K., Hoegaert L. and Barrero O. (2002). Linear and non-linear modeling in soft4s. Technical report. ESAT-SISTA, K.U.Leuven. Leuven, Belgium.

Appendix A

The Differogram

This appendix reviews the result of the differogram for estimating the noise level without relying exlicitly on an estimated model. The differogram cloud constitutes of a representation of the data in terms of the mutual distances amongst input- and output samples respectively. The behaviour of this representation towards the origin is then proven to be closely related with the noise level. The use of a parametric differogram model is used to estimate the noise level accurately. The main difference with existing methods is that there is no need for an extra hyperparameter whatever.

A.1 Estimating the Variance of the Noise

A.1.1 Model based estimators

Given a random vector (X,Y) where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, let $\{(x_i,y_i)\}_{i=1}^N$ be samples of the random vector satisfying the relation

$$y_i = f(x_i) + e_i, \quad i = 1,\ldots,N. \tag{A.1}$$

The error terms e_i are assumed to be uncorrelated random variables with zero mean and variance $\sigma^2 < \infty$ (independent and identically distributed, i.i.d.), and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ a smooth function. The same setting was adopted e.g. in (Devroye *et al.*, 2003). An estimate $\hat{f}$ of the underlying function can be used to estimate the noise variance by suitably normalizing the sums of squares of its associated residuals, see e.g. (Wahba, 1990). A broad class of model based variance estimators can be written as

$$\hat{\sigma}_e^2 = \frac{y^T Q y}{\mathrm{tr}[Q]}$$

with $y = (y_1, \ldots, y_N)^T$ (Buckley and Eagleson, 1988), $\mathrm{tr}(\cdot)$ denotes the trace of the matrix and $Q = (I_N - S)^2$ a symmetric $N \times N$ positive definite matrix. Let $\hat{y}_i = \hat{f}(x_i)$ and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_N)^T \in \mathbb{R}^N$. For most modeling methods, one can determine a smoother matrix $S \in \mathbb{R}^{N \times N}$ with $\hat{y} = Sy$ such as e.g. in the cases of ridge regression, smoothing splines (Eubank, 1999) or Least Squares Support Vector Machines (LS-SVMs) (Suykens *et al.*, 2002*b*).

A.1.2 Model free estimators

Model-free variance estimators were proposed in the case of equidistantly ordered data. In the work of (Rice, 1984) and (Gasser *et al.*, 1986), such estimators of σ^2 have been proposed based on first- and second-order differences of the values of y_i, respectively. For example Rice suggested estimating σ^2 by

$$\hat{\sigma}^2 = \frac{1}{2(N-1)} \sum_{i=1}^{N-1} (y_{i+1} - y_i)^2 . \tag{A.2}$$

Gasser *et al.* (1986) have suggested a similar idea for removing local trend effects by using

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=2}^{N-1} c_i^2 \hat{\varepsilon}_i^2, \tag{A.3}$$

where $\hat{\varepsilon}_i$ is the difference between y_i and the value at x_i of the line joining (x_{i-1}, y_{i-1}) and (x_{i+1}, y_{i+1}). The values c_i are chosen to ensure that $E\left[c_i^2 \hat{\varepsilon}_i^2\right] = \sigma^2$ for all i when the function f in (A.1) is linear. Note that one assumes that $x_1 < \cdots < x_N$, $x_i \in \mathbb{R}$ in both methods.

In the case of non-equidistant or higher dimensional data an alternative approach is based on a density estimation technique. Consider the regression model as defined in (A.1). Assume that $e_1, \ldots, e_N$ are i.i.d. with a common probability distribution function F belonging to the family

$$\mathscr{F} = \left\{ F : \int x dF(x) = 0,\ 0 < \int |x|^r dF(x) < \infty \right\},\ r \in \mathbb{N}_0 \text{ and } 1 \leq r \leq 4. \tag{A.4}$$

Let $K : \mathbb{R}^d \to \mathbb{R}$ be a function called the kernel function and let $h > 0$ be a bandwidth or smoothing parameter. Then (Müller *et al.*, 2003 (to appear)) suggested an error variance estimator given by

$$\hat{\sigma}_e^2 = \frac{1}{N(N-1)h} \sum_{1 \leq i < j \leq N} \frac{1}{2} (y_i - y_j)^2 \frac{1}{2} \left(\frac{1}{\hat{f}_i} + \frac{1}{\hat{f}_j} \right) K\left(\frac{x_i - x_j}{h} \right), \tag{A.5}$$

where $\hat{f}_i$ is defined as

$$\hat{f}_i = \frac{1}{(N-1)h} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h} \right),\ i = 1, \ldots, N. \tag{A.6}$$

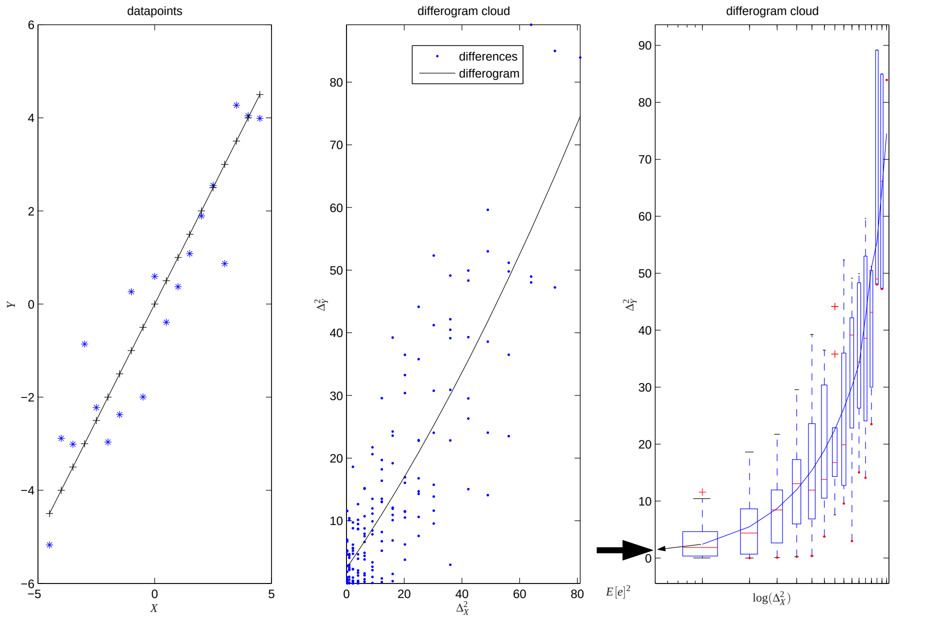


Figure A.1: *Differogram of a linear function.* **(a)** *Data are generated from* $y_i = x_i + e_i$ *with* $e_i \sim \mathcal{N}(0,1)$*, i.i.d and* $i = 1,\ldots,N = 25$*;* **(b)** *All differences* $\Delta^2_{x,ij} = \|x_i - x_j\|_2^2$ *and* $\Delta^2_{y,ij} = \|y_i - y_j\|_2^2$ *for* $i < j = 1,\ldots,N$*. The solid line represents the estimated differogram model;* **(c)** *All differences boxed using a log scale for* $\Delta^2_{x,ij}$*. The intercept of the curve crossing the Y-axis corresponds to twice the estimated noise variance* $2\hat{\sigma}_e^2$*.*

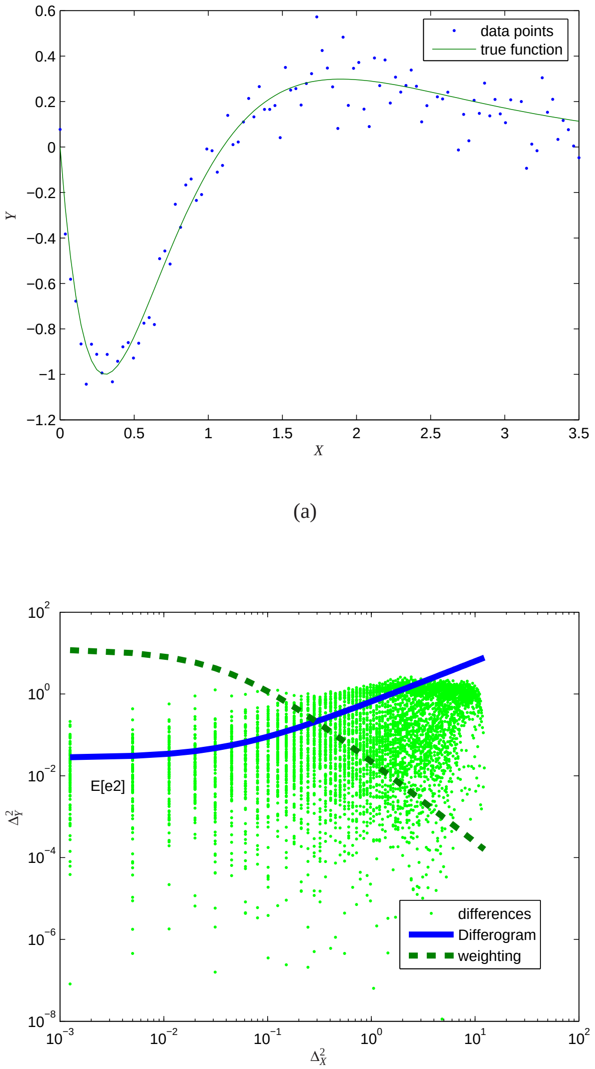
The cross-validation principle can be used to select the bandwidth h. This paper is related to (A.5) and (A.6) but avoids the need for an extra hyper-parameter such as the bandwidth and is naturally extendible to higher dimensional data.

A.2 Variogram and Differogram

The differogram was motivated from a perspective of the semi-variogram cloud employed in spatial statistics and defined as follows

Definition A.1 (Semi-variogram). (Cressie, 1993) *Let* $\{Z(x_i), i \in \mathbb{N}\}$ *be a stationary Gaussian process with mean* $\overline{z}$*,* $\mathrm{Var}[Z(x_i)] < \infty$ *for all* $i \in \mathbb{N}$ *and a correlation function which only depends on* $\Delta^2_{x,ij} = \|x_i - x_j\|_2^2$ *for all* $i,j \in \mathbb{N}$*. It follows from the stationarity of the process* $Z(x_1),\ldots,Z(x_N)$ *that*

$$\begin{aligned} \frac{1}{2} E\left[\left(Z(x_i) - Z(x_j)\right)^2\right] &= \sigma^2 + \tau^2\left(1 - \rho(\Delta^2_{x,ij})\right) \\ &= \eta\left(\Delta^2_{x,ij}\right), \quad \forall i,j \in \mathbb{N}, \end{aligned} \tag{A.7}$$



(a)

(b)

Figure A.2: *Differogram of a nonlinear function.* **(a)** *Data are generated according to the nonlinear dataset described in (Wahba, 1990). with the noise standard deviation of* 0.1 *and* $N = 100$. **(b)** *Differogram cloud of all differences of the inputs and the outputs respectively. The solid line represents the estimated differogram* $\hat{\Upsilon}(\Delta_x^2)$ *and the dashed line denotes the corresponding weighting function* $1/\vartheta(\Delta_x^2)$. *The estimate of the noise variance is* 0.1086.

where σ^2 is the small scale variance (the nugget effect), τ^2 is the variance of the serial correlation component and $\rho : \mathbb{R} \to \mathbb{R}$ is the correlation function (Diggle, 1990; Cressie, 1993). The function $\eta : \mathbb{R} \to \mathbb{R}^+$ is called the semi-variogram.

The prefix *semi-* refers to the constant $\frac{1}{2}$ in the definition. A scatter-plot of the differences is referred to as the variogram cloud. A number of parametric models were proposed to model η (Cressie, 1993). Estimation of the parameters of a variogram model often employs a maximum likelihood criterion (Cressie, 1993) leading (in most cases) to non-convex optimization problems. The variogram can be considered as being complementary to the auto-covariance function of a Gaussian process as $E(Z(x_i) - Z(x_j))^2 = 2E(Z(x_i))^2 - 2E(Z(x_i)Z(x_j))$. The auto-covariance function is often employed in an equidistantly sampled setting in time-series analysis and stochastic system identification, while the variogram allows to handle non-equidistantly sampled data, see also Subection 9.4.3.

Instead of working with a Gaussian process Z, machine learning is concerned (amongst others) with learning an unknown smooth regression function $f : \mathbb{R}^d \to \mathbb{R}$ from observations $\{(x_i, y_i)\}_{i=1}^N$ sampled from the random vector (X, Y). We now define the differogram similar to the semi-variogram as follows:

Definition A.2 (Differogram). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous function such that $y_i = f(x_i) + e_i$. Let $\Delta^2_{x,ij} = \|x_i - x_j\|_2^2$ for all $i, j = 1, \ldots, N$ be samples of the random variable Δ^2_X and let $\Delta^2_{y,ij} = \|y_i - y_j\|_2^2$ be samples from the random variable Δ^2_Y. The differogram function $\Upsilon : \mathbb{R}^+ \to \mathbb{R}^+$ is defined as*

$$\Upsilon(\Delta^2_x) = \frac{1}{2} E[\Delta^2_Y | \Delta^2_X = \Delta^2_x]. \tag{A.8}$$

This function is well-defined as the expectation operator results in a unique value for each different conditioning $\Delta^2_X = \Delta^2_x$ by definition (Mood *et al.*, 1963). A main difference with the semi-variogram is that the differogram does not assume an isotropic structure of the regression function f. A motivation for this choice is that the differogram will be of main interest in the direct region of $\Delta^2_X = 0$ where the isotropic structure emerges because of the Lipschitz assumption. A similar reasoning lies at the basis of the use of RBF-kernels and nearest neighbor methods (Hastie *et al.*, 2001; Devroye *et al.*, 2003).

Although the definition is applicable to the multivariate case, some intuition is given by considering the case of one-dimensional inputs. Let $\Delta^2_{e_{ij}} = (e_i - e_j)^2$ be samples form the random variable Δ^2_e. For one-dimensional linear models $y_i = wx_i + b + e_i$ where $w, b \in \mathbb{R}$ and $\{e_i\}_{i=1}^N$ is an i.i.d. sequence where the inputs are standardized (zero mean and unit variance), the differogram equals
$\Upsilon_w(\Delta^2_x) = \frac{1}{2} w \Delta^2_X + \frac{1}{2} E[\Delta^2_e]$, as illustrated in Figure A.1. Figure A.2 presents the differogram cloud and the (estimated) differogram function of a non-linear regression, while Section 6 reports on some experiments on higher dimensional data.

Equivalently to the nugget effect in the variogram, one can proof the following lemma relating the differogram function to the noise variance.

Lemma A.1. *Assume a Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$ such that $\exists M \in \mathbb{R}^+$ where $\|f(X) - f(X')\|_2^2 \leq M\|X - X'\|_2^2$ with X' a copy of the random variable X. Let $\{(x_i, y_i)\}_{i=1}^N$ be sampled from the random vector (X,Y) and e obeying the relation $Y = f(X) + e$. Assume that the random variable e has bounded moments and is independent of $f(X)$. Under these assumptions, the limit $\lim_{\Delta_x^2 \to 0} \Upsilon(\Delta_x^2)$ exists and equals σ_e^2.*

Proof: Let $\Delta_{e,ij}^2 = (e_i - e_j)^2$ be samples of the random variable $\Delta_e^2 = (e - e')^2$ where e' is a copy of the random variable e. As the residuals are not correlated, it follows that $E[\Delta_e^2] = E\left[e^2\right] + 2E\left[ee'\right] + E\left[e'^2\right] = 2\sigma_e^2$. Substitution of the definition of the Lipschitz continuity into the definition of the differogram gives

$$\begin{aligned}
2\Upsilon(\Delta_x^2) &= E[\Delta_Y^2 \mid \Delta_X^2 = \Delta_x^2] \\
&= E\left[\left(f(X) + e - f(X') - e'\right)^2 \mid \|X - X'\|_2^2 = \Delta_x^2\right] \\
&= E\left[\left(e - e'\right)^2 + \left(f(X) - f(X')\right)^2 \mid \|X - X'\|_2^2 = \Delta_x^2\right] \\
&\leq E\left[\Delta_e^2 + M\|X - X'\|_2^2 \mid \|X - X'\|_2^2 = \Delta_x^2\right] \\
&= 2\sigma_e^2 + E\left[M\|X - X'\|_2^2 \mid \|X - X'\|_2^2 = \Delta_x^2\right] \\
&= 2\sigma_e^2 + M\Delta_x^2, \qquad \text{(A.9)}
\end{aligned}$$

where the independence between the residuals and the function f (and hence between Δ_e^2 and $(f(X) - f(X'))^2$), and the linearity of the expectation operator E is used (Mood *et al.*, 1963). From this result, it follows that $\lim_{\Delta_x^2 \to 0} \Upsilon(\Delta_x^2) \to \sigma_e^2$.

□

The differogram function will only be of interest near the limit $\Delta_x^2 \to 0$ in the sequel. A similar approach was presented in (Devroye *et al.*, 2003) where the nearest neighbor paradigm replaces the conditioning on Δ_X^2 and fast rates of convergence were proved.

A.2.1 Differogram models based on Taylor-series expansions

Consider the Taylor series expansion of order r centered at $m \in \mathbb{R}$ for local approximation in $x_i \in \mathbb{R}$ for all $i = 1, \ldots, N$

$$T_r[f(x_i)](m) = f(m) + \sum_{l=1}^{r} \frac{1}{l!} \nabla^{(l)} f(m)(x_i - m)^l + \mathcal{O}(x_i - m)^{r+1}, \qquad \text{(A.10)}$$

where $\nabla f(x) = \frac{\partial f}{\partial x}$, $\nabla^2 f(x) = \frac{\partial^2 f}{\partial x^2}$, etc. for $l \geq 2$. One may motivate the use of an r-th order Taylor series approximation of the differogram function with center $m = 0$ as a suitable model because one is only interested in the case $\Delta_x^2 \to 0$:

$$\Upsilon_{\mathscr{A}}(\Delta_x^2) = a_0 + \mathscr{A}(\Delta_x^2), \;\text{ where }\; \mathscr{A}(\Delta_x^2) = \sum_{l=1}^{r} a_l (\Delta_x^2)^l, \; a_0, \ldots, a_r \in \mathbb{R}^+, \qquad \text{(A.11)}$$

where the parameter vector $a = (a_0, a_1, \dots, a_r)^T \in \mathbb{R}^{+,r+1}$ is assumed to exist uniquely. The elements of the parameter vector a are enforced to be positive as the (expected) differences should always be strictly positive. The function ϑ of the mean absolute deviation of the estimate can be bounded as follows

$$\begin{aligned}
\vartheta(\Delta_x^2; a) &= E\left[|\Delta_Y^2 - \Upsilon_{\mathscr{A}}(\Delta_X^2; a)| \;\middle|\; \Delta_X^2 = \Delta_x^2\right] \\
&= E\left[|\Delta_Y^2 - a_0 - \sum_{l=1}^{r} a_l (\Delta_X^2)^l| \;\middle|\; \Delta_X^2 = \Delta_x^2\right] \\
&\leq E\left[|a_0 + \sum_{l=1}^{r} a_l (\Delta_x^2)^l|\right] + E\left[|\Delta_Y^2| \middle| \Delta_X^2 = \Delta_x^2\right] \\
&= 3\left(a_0 + \sum_{l=1}^{r} a_l (\Delta_x^2)^l\right) \triangleq \bar{\vartheta}(\Delta_x^2; a),
\end{aligned} \tag{A.12}$$

where respectively the triangle inequality, the property $|\Delta_Y^2| = \Delta_Y^2$ and definition A.2 are used. The function $\bar{\vartheta} : \mathbb{R}^+ \to \mathbb{R}^+$ is defined as an upperbound to the spread of the samples Δ_Y^2 from the function $\Upsilon(\Delta_x^2)$. Instead of deriving the parameter vector a from the (estimated) underlying function f, it is estimated immediately based on the observed differences $\Delta_{x,ij}^2$ and $\Delta_{y,ij}^2$ for $i < j = 1, \dots, N$. The following weighted least squares method can be used

$$a^* = \arg\min_{a \in \mathbb{R}_+^{r+1}} \mathscr{J}(a) = \sum_{i \leq j}^{N} \frac{c}{\vartheta(\Delta_{x,ij}^2; a)} \left(\Delta_{y,ij}^2 - \Upsilon_{\mathscr{A}}(\Delta_{x,ij}^2; a)\right)^2, \tag{A.13}$$

where the constant $c \in \mathbb{R}_0^+$ normalizes the weighting function such that $1 = \sum_{i<j} c / \bar{\vartheta}(\Delta_{x,ij}^2; a)$. The function $\bar{\vartheta}$ corrects for the heteroscedastic variance structure inherent to the differences (see e.g. (Sen and Srivastava, 1990)). As the parameter vector a is positive, the weighting function is monotonically decreasing and as such represents always a local weighting function.

A.3 Differogram for Noise Variance Estimation

A U-statistic is proposed to estimate the variance of the noise from observations.

Definition A.3 (U-statistic). (Hoeffding, 1948) *Let $g : \mathbb{R}^l \to \mathbb{R}$ be a measurable and symmetric function and let $\{u_i\}_{i=1}^N$ be i.i.d. samples drawn from a fixed but unknown distribution. The function*

$$U_N = U(g; u_1, \dots, u_N) = \frac{1}{\binom{N}{l}} \sum_{1 \leq i_1 \leq \cdots \leq i_l \leq N} g(u_{i_1}, \dots, u_{i_l}), \tag{A.14}$$

for $l < N$, is called a U-statistic of degree l with kernel g.

It is shown (Lee, 1990) that for every unbiased estimator based on the same observations, a U-statistic exists with a smaller variance of the corresponding estimator. If the regression function was known, the errors e_i for all $i = 1, \ldots, N$ were observable and the sample variance can be written as a U-statistic of order $l = 2$

$$\hat{\sigma}_e^2 = U(g; e_1, \ldots, e_N) = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} g_1(e_i, e_j)$$
$$\text{and} \quad g_1(e_i, e_j) = \frac{1}{2}(e_i - e_j)^2 = \frac{1}{2}\Delta_{e,ij}^2. \quad \text{(A.15)}$$

However, the true function f is not known in practice. A key step deviating from classical practice is to abandon trying to estimate the global function (Vapnik, 1998) or the global correlation structure (Cressie, 1993). Instead, knowledge of the average local behavior is sufficient for making a distinction between smoothness in the data and unpredictable noise. As an example, consider $r = 0$, the 0th order Taylor polynomial of f centered at x_i evaluated at x_j for all $i, j = 1, \ldots, N$. This approximation scheme is denoted as $T_0[f(x_j)](x_i) = f(x_i)$ such that (A.15) becomes

$$\begin{aligned} \hat{\sigma}_e^2 &= \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} \frac{1}{2}(y_i - y_j)^2 \\ &\approx \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} \frac{1}{2}\left(e_i + f(x_i) - e_j - T_0[f(x_j)](x_i)\right)^2 \\ &= \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} \frac{1}{2}\Delta_{e,ij}^2, \end{aligned} \quad \text{(A.16)}$$

where the approximation improves as $x_i \longrightarrow x_j$. To correct for this, a localized second order isotropic kernel $g_2 : \mathbb{R}^2 \longrightarrow \mathbb{R}$ can be used

$$g_2(y_i, y_j) = \frac{c}{2\vartheta(\Delta_{x,ij}^2)} \Delta_{y,ij}^2, \quad \text{(A.17)}$$

where the decreasing weighting function $1/\bar{\vartheta}(\Delta_x^2)$ is taken from (A.12) in order to favor good (local) estimates. The constant $c \in \mathbb{R}_0^+$ is chosen such that the sum of the weighting terms are constant: $2c(\sum_{i \le j}^N 1/\bar{\vartheta}(\Delta_{x,ij}^2)) = N(N-1)$.

From this derivations one may motivate the following kernel for a U-statistic based on the differogram model (A.11) and weighting function as derived in (A.12):

$$g_3(y_i, y_j) = \frac{c}{2\vartheta(\Delta_{x,ij}^2)} \left(\Delta_{y,ij}^2 - \mathscr{A}(\Delta_{x,ij}^2)\right)$$
$$\text{with} \quad \bar{\vartheta}(\Delta_{x,ij}^2) = 3\left(a_0 + \mathscr{A}(\Delta_{x,ij}^2)\right), \quad \text{(A.18)}$$

where $c \in \mathbb{R}_0^+$ is a normalization constant. The resulting U-statistic becomes

$$\hat{\sigma}_e^2 = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} g_3(y_i, y_j). \quad \text{(A.19)}$$

One can show that this U-estimator equals the estimated intercept of the differogram model (A.11):

Lemma A.2. *Let $x_1,\ldots,x_N \in \mathbb{R}^d$ and $y_1,\ldots,y_N \in \mathbb{R}$ be samples drawn according to the distribution of the random vector (X,Y) with joint distribution F. Consider a U-statistic as in Definition A.3 with kernel g such that $g : \mathbb{R}^l \to \mathbb{R}$ is a measurable and symmetric function. Consider the differogram according to Definition A.2 and the differogram model (A.11). The estimator of the weighted U-statistic (A.18) of the noise variance estimator (A.19) equals the intercept a_0 of the estimated differogram model using the weighted least squares estimate (A.13).*

Proof: This can be readily seen as the expectation can be estimated empirically in two equivalent ways. Consider for example the mean μ of the error terms $e_1,\ldots,e_N$ which can be estimated as $\hat{\mu} = \arg\min_\mu \sum_{i=1}^N (e_i-\mu)^2$ and as $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N e_i$, see e.g. (Hettmansperger and McKean, 1994). As previously, one can write

$$\begin{aligned} 2\hat{\sigma}_e^2 &= \lim_{\Delta_x^2\to 0} E[\Delta_Y^2|\Delta_X^2 = \Delta_x^2] \\ &= \lim_{\Delta_x^2\to 0} E\left[\frac{c}{\vartheta(\Delta_X^2)}\left(\Delta_Y^2 - \mathscr{A}(\Delta_X^2)\right) \;\middle|\; \Delta_X^2 = \Delta_x^2\right], \qquad \text{(A.20)} \end{aligned}$$

if $\lim_{\Delta_x^2\to 0}\mathscr{A}(\Delta_x^2) = 0$. The sample mean estimator becomes

$$\begin{aligned} 2\hat{\sigma}_e^2 &= \frac{2}{N(N-1)}\sum_{k=1}^{N(N-1)/2}\frac{c}{2\vartheta(\Delta_{x,k}^2)}\left(\Delta_{y,k}^2 - \mathscr{A}(\Delta_{x,k}^2)\right) \\ &= \frac{2}{N(N-1)}\sum_{i<j}^{N}\frac{c}{2\bar{\vartheta}(\Delta_{x,ij}^2)}\left(\Delta_{y,ij}^2 - \mathscr{A}(\Delta_{x,ij}^2)\right) \\ &= U(g_3;u_1,\ldots,u_N), \qquad \text{(A.21)} \end{aligned}$$

where a unique index $k = 1,\ldots,N(N-1)/2$ corresponds with every distinct pair $1 \leq i < j \leq N$. Alternatively, using the least squares estimate

$$\begin{aligned} 2\hat{\sigma}_e^2 &= \arg\min_{a_0\geq 0}\sum_{k=1}^{N(N-1)/2}\frac{c}{\bar{\vartheta}(\Delta_{x,k}^2)}\left(\Delta_{y,k}^2 - \mathscr{A}(\Delta_{x,k}^2) - a_0\right)^2 \\ &= \arg\min_{a_0\geq 0}\sum_{i<j}\frac{c}{\bar{\vartheta}(\Delta_{x,ij}^2)}\left(\Delta_{y,ij}^2 - \mathscr{A}(\Delta_{x,ij}^2) - a_0\right)^2. \qquad \text{(A.22)} \end{aligned}$$

In both cases, the function $\mathscr{A} : \mathbb{R}^+ \to \mathbb{R}^+$ of the differogram model and the weighting function $\bar{\vartheta} : \mathbb{R}^+ \to \mathbb{R}^+$ are assumed to be known from (A.13).

□

A.4 Applications

A model-free estimate of the noise variance plays an important role in the practice of model selection and setting tuning parameters. Examples of such applications are given:

1. Well-known complexity criteria (or model selection criteria) such as the Akaike Information Criterion (Akaike, 1973), the Bayesian Information Criterion (Schwartz, 1979) and C_p statistic (Mallows, 1973) take the form of a prediction error criterion which consists of the sum of a training set error (e.g. the residual sum of squares) and a complexity term. In general:

$$J(S) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i;S))^2 + \lambda\left(Q_N(\hat{f})\right)\hat{\sigma}_e^2, \tag{A.23}$$

   where S denotes the smoother matrix, see (De Brabanter *et al.*, 2002*a*). The complexity term $Q_N(\hat{f})$ represents a penalty term which grows proportionally with the number of free parameters (in the linear case) or the effective number of parameters (in the nonlinear case (Wahba, 1990; Suykens *et al.*, 2002*b*)) of the model $\hat{f}$ grows.

2. Consider the linear ridge regression model $y = w^T x + b$ with w and b optimized w.r.t.

$$J_{RR,\gamma}(w,b) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{N}(y_i - w^T x_i - b)^2. \tag{A.24}$$

   Using the Bayesian interpretation (MacKay, 1992; Van Gestel, 2002) of ridge regression and under i.i.d. Gaussian assumptions, the posterior can be written as $p(w,b \mid x_i, y_i, \mu, \zeta) \propto \exp(-\zeta(wx_i + b - y_i)^2)\exp(-\mu(w^T w))$, the estimate of the noise variance $\zeta = 1/\hat{\sigma}_e^2$ and the expected variance of the first derivative $\mu = 1/\sigma_w^2$ can be used to set respectively the expected variance of the likelihood $p(y_i|x_i, w, b)$ and on the prior $p(w,b)$. As such, a good guess for the regularization constant when the input variables are independent becomes $\hat{\gamma} = \hat{a}_1^2/\hat{\sigma}_e^2$.

   Another proposed guess for the regularization constant $\hat{\gamma}$ in ridge regression (A.24) can be derived as in (Hoerl *et al.*, 1975): $\hat{\gamma} = \hat{w}_{LS}^T \hat{w}_{LS}/(\hat{\sigma}_e d)$ where $\hat{\sigma}_e$ is the estimated variance of the noise, d is the number of free parameters and $\hat{w}_{LS}$ are the estimated parameters of the ordinary least squares problem. These guesses can also be used to set the regularization constant in the parametric step in fixed size LS-SVMs (Suykens *et al.*, 2002*b*) where the estimation is done in the primal space instead of the dual via a Nÿstrom approximation of the feature map.

3. Given the non-parametric Nadaraya-Watson estimator $\hat{f}(x) = [\sum_{i=1}^{N}(K((x - x_i)/h)y_i)]/\ [\sum_{i=1}^{N} K((x - x_i)/h)]$, the plugin estimator for the bandwidth h is calculated under the assumption that a Gaussian kernel is to be used and the noise is Gaussian. The derived plugin estimator becomes $h_{opt} = C\hat{\sigma}^2 N^{-\frac{1}{5}}$ where $C \approx 6\sqrt{\pi}/25$, see e.g. (Hardle, 1990).

4. We note that $\hat{\sigma}_e^2$ also plays an important role in setting the tuning parameters of SVMs, see e.g. (Vapnik, 1998; Cherkassky, submitted, 2002).

Appendix B

A Practical Overview: LS-SVMlab

While the presented research is rather methodological in nature, much effort was spent on the practical abilities of the methods and on increasing the userfrinedliness of the tools by elaborating a MATLAB/C toolbox called LS-SVMlab. The content and implementation details of the Matlab/C toolbox are discussed qualitatively and some details are given about the interface.

B.1 LS-SVMlab toolbox

In 2002, a freeware Matlab/C toolbox was released by the same authors for the use of algorithms based on LS-SVM classifiers and regressors, and various extensions (Pelckmans *et al.*, 2002*b*; Pelckmans *et al.*, 2002*a*)

http://www.esat.kuleuven.ac.be/sista/lssvmlab/,

which is freely available for research purposes (for precise conditions, see website).

Two years of experience and feedback were embodied in a new upgrade (LS-SVMlab2). This section reviews and discusses issues concerning the main structure, the newly implemented tools, a new graphical user interface and a number of useful extensions of this software package. Note that a whole range of related software for the estimation of SVMs and other Machine Learning techniques is available on the web (see e.g. http://www.kernel-machines.org). The present approach mainly differs from most approaches as the package focuses not on only one technique but offers a whole spectrum of kernel based methods for the application at hand. Moreover, a graphical interface was designed to ease the application of most described methods. A couple of sometimes conflicting desiderata were put first:

1. The toolbox should provide algorithmic tools as developed recently by the authors and co-workers for the generic user.

2. The use of the toolbox should be highly robust and user-friendly in order to facilitate the application of the methodology to the unexperienced as well as the demanding users.

3. The calls of the core algorithms and the implementation should correspond with the mathematical formulations as well as possible.

4. Functionality should be extendible towards other training and tuning algorithms and other kernels.

Furthermore, the new toolbox should be backwards compatible to the first release.

B.1.1 Software architecture

Somewhat at the core of the software design is the definition of an appropriate Matlab structure containing all information for the inference of a type of kernel machine. A typical example of such a model is represented in Figure B.1, but can be extended with extra fields containing details on the specific method or dataset. We shall refer to such container as a data-structure if at least the substructure with the data definition is present. One can speak of a model structure if the container includes the data definition and the specifications in `method`. With a small abuse in notation, we will refer to the latter as a *model*. As an example, Table B.1 expands the substructure `method` containing details on the involved training methodology. Every substructure contains a `status` flag indicating whether the according stage (preprocessing, training,...) is already processed successfully or will need to be redone.

The software folder (the different .m files) is organized as follows. The root directory of the toolbox contains generic calls (`trainm`, `simm`, `tunem`, `prem` and `dispm`) which support the model interface and redirects the user to the appropriate implementation. On a second level the core functionalities are implemented as close to the formulas as possible. Those are located in a set of subdirectories making the extension and interpretability highly accessible. The implementations are functional and make no use of the model structure interface.

B.1.2 Model selection and generalization

A main advantage of this toolbox is its functionality regarding the task of model selection as it contains a wide range of useful routines and algorithms for measuring and maximizing the generalization performance of specific models. A number of commonly used model selection criteria are implemented in the package. These include the classical procedures for computing different model selection criteria as *L*-fold cross-validation, leave-one-out, Generalized Cross-Validation (GCV), a variety

model –	– data	: Definitions of the data-sample involved in the modeling process
	– pre	: Information on the pre- and post-processing
	– train	: Details on the used implementation
	⌞ type	
	⌞ status	
	⌞ train	
	⌞ sim	
	⌞ reg	
	⌞ kernel	
	– modsel	: Specifications on the model selection procedure
	– disp	: Information on the used visualization technique

Table B.1: *Definition of the model structure at the core of the toolbox*

of information criteria and fast implementations of those. Following the contributions in (De Brabanter *et al.*, 2002*a*), robust counterparts to some of the model selection criteria were implemented. Apart from this estimation methods, different methods for the optimization of a model selection criterion are including, ranging from very generic algorithms as a computer-intensive grid search and local optimization routines to fast initial estimates. Implementation of the fusion argument as elaborated in this thesis are provided. Furthermore, some useful tools assisting the user in the design of an appropriate kernel are encoded.

B.1.3 Building blocks

While the previous discussion describes the general setup of the toolbox, this Subsection gives some details and illuminates some choices of the implementation.

Preprocessing The toolbox contains a set of functions for automatically preprocessing the data before the stage of modeling. While this is often highly dependent on the application at hand, some procedures as normalization and standardization is useful in most application. The standard preprocessing procedure will handle binary, categorical and continuous data in different ways.

Modeling and Estimation Somewhat central to the toolbox is an efficient C implementation for solving standard LS-SVMs. A variety of related parametric techniques as ridge regression are supplied in order to ease comparisons of the method. Furthermore, a set of structured and dedicated primal-dual kernel machines are implemented as described in the text. Special attention is given

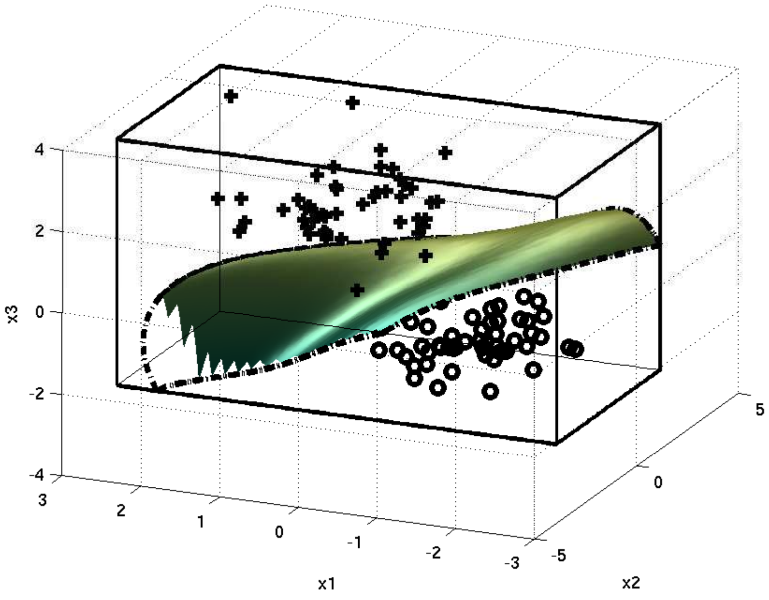


Figure B.1: *Example of a decision hyperplane found by application of a Support Vector Machine.*

to the construction of a user-interface assisting the user in the choice of an appropriate algorithm.

Visualization Techniques Of direct concern to the user is the visual format in which the result is presented on screen. In first instance, every training procedure is engaged for making an appropriate visualization. Furthermore, some visualization tools are implemented for representing the raw data as the differogram technique and others. A final set of visualization tools are involved with the visualization of the model tuning process as evolution diagrams for structure detection and computer-intensive grid-searches for hyper-parameter tuning.

Resampling Schemes and Bayesian Inference Most results in the context of statistical learning and kernel machines focus on the formulation of learning machines for point estimation. However, the user is often also interested in quantitative estimates of the (un)certainty of the provided prediction. This need is approached in two disjunct ways. Classical non-parametric statistics provides a number of results on resampling schemes based on the bootstrap procedure. An entirely different approach emerged from the Bayesian point of view. This implementation mainly builds on results described in (Van Gestel *et al.*, 2002).

Extensions for Classification In the task of classification dedicated tools as the Receiver Operating Characteristic (ROC) curve of a binary classifier (Hanley and McNeil, 1982) is often a useful tool to analyze the learned model. Another useful extension towards the task of classification are the functions for converting multi-class classification problems in sets of binary classification task using different encoding schemes, see e.g. (Singer, 2003). Special attention was paid to efficient calculation of error correcting output codes as presented in (Dietterich and Bakiri, 1995).

Large Scale Methods A number of dedicated functions enable the handling and processing of large scale databases in the toolbox. A principal tool here is the fixed-size LS-SVM as introduced in (Suykens *et al.*, 2002*b*) which is based on a Nÿstrom approximation scheme combined with estimation in the primal space. A problem especially apparent in medium to large scale problems is the problem of hyper-parameter tuning and model selection. Dedicated formulations based on the fusion argument are implemented.

Unsupervised Learning The task of finding patterns in unlabeled data in the context of primal-dual kernel machines is discussed in some detail in (Suykens *et al.*, 2002*b*) and advances are given in (Hoegaerts, 2005). The toolbox contains implementations of kernel PCA, kernel CCA and kernel PLS together with fast approximation schemes to those algorithms capable of handling large datasets.