Tensor decomposition for mining the consistent reproducible patterns in neuroimaging data

Morten Mørup
Informatics and Mathematical Modelling
Cognitive Systems
Technical University of Denmark
joint work with Kristoffer Hougaard Madsen, Sidse Marie Arnfred and Lars Kai Hansen

13th September 2010
Univariate statistical analysis in NeuroImaging

Problems:
1) Multiple comparisons, i.e. many voxels tested.
2) What is the true number of independent tests, i.e. voxels are highly correlated.
3) Data extremely noisy, i.e. low SNR rendering tests insignificant.

Need for advanced multivariate methods that can efficiently extract the underlying sources in the data.
This problem is no different than the problems encountered in general in Modern Massive Datasets (MMDS)

Unsupervised Learning attempts to find the hidden causes and underlying structure in the data.
(Multivariate exploratory analysis – driving hypotheses)
Goal of unsupervised Learning
(Ghahramani & Roweis, 1999)

- Perform dimensionality reduction
- Build topographic maps
- Find the hidden causes or sources of the data
- Model the data density
- Cluster data

Purpose of unsupervised learning
(Hinton and Sejnowski, 1999)

- Extract an efficient internal representation of the statistical structure implicit in the inputs
The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson  06.23.08

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't

Analysis of massive amounts of data will be the main driving force of all sciences in the future!!
Outline of the talk

- NeuroImaging data modeled as tensors (CandeComp/PARAFAC(CP), ShiftCP and ConvCP)
- Bayesian methods for estimating the number of components in tensor decomposition (Automatic Relevance Determination)
- Tensor decomposition of complex functional networks (Infinite Relational Modeling)
NeuroImaging data modeled as tensors

Factor Analysis

Spearman ~1900

The Cocktail Party problem (Blind source separation)
The EEG/MEG/fMRI Party problem

Assumption: Data instantaneous mixture of temporal signatures. (PCA/ICA/NMF)

Flaw: \( \mathbf{X}_{\text{Voxel} \times \text{Time}} \approx \sum_d \mathbf{a}_d^{\text{Voxel}} \mathbf{b}_d^{\text{Time}} \)

\[ \mathbf{X} \approx \mathbf{A} \mathbf{S} = (\mathbf{A} \mathbf{Q}^{-1}) \mathbf{Q} \mathbf{S} = \hat{\mathbf{A}} \hat{\mathbf{S}} \]

\[ \Rightarrow \text{Representation not unique!} \]
From 2-way to multi-way analysis
3 common ways of avoiding tensors

Preaverage

Concatenation
(identical time series varying spatial maps)

Separate Analysis

(identical spatial map, varying time series)
Multilinear modelling

**Bilinear Model:**

\[ X_{\text{Voxel} \times \text{Time}} \approx \sum_d a_d^{\text{Voxel}} b_d^{\text{Time}} \]

**Assumption:** Data instantaneous mixture of temporal signatures. (PCA/ICA/NMF)

**Trilinear Model:**

\[ X_{\text{Voxel} \times \text{Time} \times \text{Trial}} \approx \sum_d a_d^{\text{Voxel}} b_d^{\text{Time}} c_d^{\text{Trial}} \]

**Assumption:** Data instantaneous mixture of temporal signatures that are expressed to various degree over the Subjects/trials (Canonical Decomposition, Parallel Factor (CP))

(weighted averages over the trials)

*Mult. Mod. admits non-ambiguous extraction of consistent patterns of activation (see also k-rank criterion by Kruskal 1976,1977)*
Unfortunately, violation of multi-linearity causes degeneracy.

Common fixes: Impose orthogonality, regularization or non-negativity constraints by analyzing data transformed to a time-frequency domain representation.
Wavelet transformed data

(Mørup et al., NeuroImage 2006)
Features:
- Wavelet analysis
- Data visualization
- Artifact Rejection
- 2-way decomposition
- 3-way decomposition
- Coherence tracking
- Bootstrapping

(Mørup et al, Journ. of Neurosc. Meth. 2007)
(Algorithms described in Mørup et al, Neural Computation 2008)
Degeneracy often a result of multi-linear models being too restrictive

Trilinear model can encompass:
- Variability in strength over repeats

However, other common causes of variation are:
- Delay Variability
  - Trial 1
  - Trial 2
- Shape Variability
  - Trial 1
  - Trial 2
Modelling Delay Variability

Shifted CP:

\[ x_{i,k}(t) \approx \sum_d a_{i,d} b_d(t) \tau_{k,d} c_{k,d} \]
\[ x_{i,d}(t) \approx \sum_{j} a_{i,d} h_{d}(t - \tau_{j,d}) c_{j,d} \]
Delay modelling of fMRI data from retinotopic mapping paradigm

\[ x_{i,k}(t) \approx \sum_d \alpha_{i,d} b_d(t - \tau_{i,d}) c_{k,d} \]

360° rotation in 30 s
8 Hz reversal rate

(Analysis by Kristoffer Hougaard Madsen) (Mørup et al., NeuroImage 2008)
Modeling Shape (and delay) Variability

convolutive CP:

\[ x_{i,k}(t) \approx \sum_{d,\tau} a_{i,d} b_d(t - \tau) c_{k,d}(\tau) \]

(Mørup et al., Nips workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis 2008)
**ConvCP:** Can model arbitrary number of component delays within the trials and account for shape variation within the convolutional model representation. Redundancy between what is coded in C and B resolved by imposing sparsity on C.

(Mørup et al., Nips workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis 2008)
Convolutive Multi-linear decomposition

**Average ERP**

- 96 ms
- 170 ms
- 235 ms

**CP**

- Trials
- Hist. of trial strengths

**convCP**

- Trials
- Hist. of trial strengths
Analysis of fMRI data

Each trial consists of a visual stimulus delivered as an annular full-field checkerboard reversing at 8 Hz.

\[ \lambda' \text{ is } L_1 \text{ sparsity regularization imposed on third mode} \]

(Mørup et al., Nips workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis 2008)
Bayesian Learning and the Principle of Parsimony

The explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory.

Open problem: To get the posterior probability distribution, multiply the prior probability distribution by the likelihood function and then normalize.

Bayesian learning embodies Occam’s razor, i.e. complex models are penalized.

David J.C. MacKay

William of Ockham

Thomas Bayes

EUSIPCO’09  27 August 2009
Many inference paradigms in Bayesian Learning

- **Maximum a posteriori estimation (MAP)**
  
  *seeks optimal solution (admit standard optimization) however, the approach does not take parameter uncertainty into account*

- **Sampling methods**
  
  *Marcov Chain Monte Carlo (MCMC)*

- **Variational methods (VB) and Belief Propagation (BP)**

  *Approximate likelihood $P(\theta)$ by factorized form $Q(\theta)$ that is tractable*

  *VB: minimize the Kulback Leibler divergence $KL(P(\theta)|Q(\theta))$*

  *BP: minimize the Kulback Leibler divergence $KL(Q(\theta)|P(\theta))$*

(Notice: MAP estimation admits direct use of standard optimization tools)
Automatic Relevance Determination (ARD)

- Automatic Relevance Determination (ARD) is a hierarchical Bayesian approach widely used for model selection.
- In ARD hyper-parameters explicitly represents the relevance of different features by defining their range of variation.
  (i.e., Range of variation $\to 0 \Rightarrow$ Feature removed)
A motivating example: A Bayesian formulation of the Lasso /Basis Pursuit Denoising (BPD) problem

\[
\text{LASSO/BPD: } \arg\min_s \frac{1}{2\sigma^2} \|x^I - A^{I \times J}s^J\|_F^2 + \lambda |s|_1
\]

\[
P(x|A, s, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|x-As\|_F^2}{2\sigma^2}}
\]

\[
P(s|\lambda) = \left(\frac{\lambda}{2}\right)^J e^{\lambda |s|_1}
\]

\[
P(s|A, x, \sigma^2, \lambda) = \frac{P(x|A, s, \sigma^2)P(s|\lambda)}{P(x)}
\]

\[
-\log P(s|A, x, \sigma^2, \lambda) = -\log P(x|A, s, \sigma^2) - \log P(s|\lambda) + \log P(x)
\]

\[
= \frac{\|x-As\|_F^2}{2\sigma^2} + \lambda |s|_1 + \frac{I}{2} \log 2\pi\sigma^2 - J \log\frac{\lambda}{2} + \text{Const.}
\]

\[
\frac{\partial \log P(s|A, x, \sigma^2, \lambda)}{\partial \lambda} = 0 \Rightarrow \lambda = \frac{J}{|s|_1}
\]
ARD in reality a $\ell_0$-norm optimization scheme. As such ARD based on Laplace prior corresponds to $\ell_0$-norm optimization by re-weighted $\ell_1$-norm.

In particular if we define $\lambda$ for each entry in $s$, i.e.

$$\frac{1}{2\sigma^2} \| x^I - A^{I \times J} s^J \|_F^2 + \sum_j \lambda_j |s_j|$$

Corresponding to the Laplace prior $P(s|\lambda) = \prod_j \frac{\lambda_j}{2} e^{-\lambda_j |s_j|}$ optimizing for $\lambda_j$ gives $\lambda_j = \frac{1}{|s_j|}$ such that

$$\frac{1}{2\sigma^2} \| x^I - A^{I \times J} s^J \|_F^2 + \sum_j \frac{|s_j|}{|s_j|}$$

$\ell_0$ norm by re-weighted $\ell_2$ follows by imposing Gaussian prior instead of Laplace.

Notice that we are all the time monotonically decreasing

$$-\log P(s|A,x,\sigma^2,\lambda)$$
Agenda for model order selection

- To use regularization to simplify the Tucker core forming a unique representation as well as enable interpolation between the Tucker (full core) and CP (diagonal core) model.
- To use regularization to turn off excess components in the CP and Tucker model and thereby select the model order.
- To tune the regularization strength from data by Automatic Relevance Determination (ARD) based on Bayesian learning.

(Mørup and Hansen, Journal of Chemometrics 2009)
Sparse Tucker decomposition by ARD

\[
P(\mathcal{X}|\mathcal{R},\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \left| \Lambda - \frac{\|\mathcal{X} - \mathcal{R}\|^2_{F}}{2\sigma^2} \right|_F
\]

\[
P(\mathcal{G}|\alpha^G) = \left( \frac{\alpha^G}{2} \right)^{\frac{\|\mathcal{G}\|_F}{2}} e^{-\frac{\|\mathcal{G}\|_F}{2}}
\]

\[
P(A^{(n)}|\alpha^{(n)}) = \prod_{j_n} \left( \frac{\alpha^{(n)}_{j_n}}{2} \right)^{\frac{I_n}{2}} e^{-\frac{\alpha^{(n)}_{j_n}}{2}|a^{(n)}_{j_n}|_1}
\]

CP follows setting \( \mathcal{G} = \mathcal{I} \)

\[
L = P(\mathcal{G}, A^{(1)}, \ldots, A^{(N)}|\mathcal{X}, \sigma^2, \alpha^G, \alpha^{(1)}, \ldots, \alpha^{(N)})
\]

\[
\propto P(\mathcal{X}|\mathcal{R},\sigma^2) P(\mathcal{G}|\alpha^G) P(A^{(1)}|\alpha^{(1)}) \cdots P(A^{(N)}|\alpha^{(N)}).
\]

Thus the negative log likelihood based on Laplace priors is proportional to

\[
- \log L \propto c + \frac{1}{2\sigma^2} \|\mathcal{X} - \mathcal{R}\|^2_{F} + \sum_{n} \sum_{j_n} \alpha^{(n)}_{j_n} |a^{(n)}_{j_n}|_1 + \alpha^G |\mathcal{G}|_1
\]

\[
+ \frac{1}{2} I_1 I_2 \cdots I_N \log \sigma^2 - \sum_{n} \sum_{j_n} I_n \log \alpha^{(n)}_{j_n} - J_1 J_2 \cdots J_n \log \alpha^G.
\]

Maximum a posteriori (MAP) estimation

\[
\arg\min_{A^{(n)}} \frac{1}{2\sigma^2} \|\mathcal{X}^{(n)} - A^{(n)} Z^{(n)}\|^2_{F} + \sum_{j} \lambda_j |a^{(n)}_{j}|_1
\]

Update of regularization parameters by ARD

\[
\alpha^G = \frac{J_1 J_2 \cdots J_N}{|\mathcal{G}|_1}, \quad \alpha^{(n)}_{j} = \frac{J_n}{|A^{(n)}_{j}|_1}
\]
Tucker(10,10,10) models were fitted to the data, given are below the extracted cores.
Reversible jump Markov Chain Monte Carlo - a fully Bayesian approach to estimate parameter uncertainty and model order.

(For details see: Schmidt and Mørup, Infinite Non-negative Matrix Factorization, 2010)
Tensor models for complex networks

The Infinite Relational Model
(A Bayesian generative model for graphs)

Learning Systems of Concepts with an Infinite Relational Model (AAAI2006)

Charles Kemp  Josh Tenenbaum  Thomas Griffith  Takeshi Yamada  Naonori Ueda

See also: Infinite Hidden Relational Model (UAI 2006)

Zhao Xu  Kai Yu  Volker Tresp  Hans-Peter Kriegel
The relational model for various types of graphs:

UnDirected

Directed

Bipartite

Multi-graph

\[ \pi_{ij} = z_i \eta z_j \]
\[ \eta = \eta^\top \]

\[ \pi_{ij} = z_i^{(1)} \eta z_j^{(2)} \]

The IRM statistical generative model for graphs:

\[ Z \sim CRP(\alpha) \]
\[ \eta_{ab} \sim Beta(\gamma, \gamma) \]
\[ A_{ij} \sim Bernoulli(\pi_{ij}) \]

Tucker

\[ \pi_{ijk} = \sum_{lmn} \eta_{lmn} z_{ij}^{(1)} z_{mi}^{(2)} z_{nk}^{(3)} \]

Tucker2

\[ \pi_{ij} = z_i^{(1)} \eta^{(k)} z_j^{(2)} \]

Potential symmetry constraints, i.e. \( z^{(1)} = z^{(2)} \)

13th September 2010
Modeling the consistent functional connectivity of the brain

Pairwise Mutual Information (MI) between 2x2x2 voxel groups

$$I(i,j) = \sum_{uv} P_{ij}(u,v) \log \frac{P_{ij}(u,v)}{P_i(u)P_j(v)}$$

Top 100'000 MI links

72 subjects: 42 multiple sclerosis and 30 normal subjects

(Mørup et al., to appear NIPS 2010)
Summary

Multi-linear modeling offers the ability to explicitly extract the most consistent activity of neuroimaging data across repeats/subjects/conditions.

Common causes of variability in neuroimaging data are latency and shape changes—> **shiftCP and convCP**

Important problem in tensor decomposition is to adequately selected the number of components. **Bayesian learning** admits a general framework for model order selection and regularization tuning.

From neuroimaging data complex networks of functional connectivity can be derived. The **Infinite Relational Model** forms an efficient modeling framework for exploring consistent structures in these networks.

**AIM of all the described analyses**

- Extract an efficient internal representation of the statistical structure implicit in the data
- Drive novel hypothesis for formal statistical testing
**Relevant papers**


M.N. Schmidt, M. Mørup, Infinite Non-negative Matrix Factorization, EUSIPCO 2010


