Frequency Constrained ShiftCP Modeling of Neuroimaging Data

Morten Mørup
Section for Cognitive Systems
DTU Informatics
Technical University of Denmark
Acknowledgement

Kristoffer Hougaard Madsen
Danish Research Centre for Magnetic Resonance
Copenhagen University Hospital Hvidovre

Lars Kai Hansen
Section for Cognitive Systems
DTU Informatics
Technical University of Denmark

Stefan Haufe
Department of Software Engineering and Theoretical Computer Science
Machine Learning
TU Berlin
Factor Analysis

Spearman ~1900

The Cocktail Party problem  (Blind source separation)

\[
\begin{align*}
X_{\text{tests} \times \text{subjects}} & \approx A_{\text{tests} \times \text{int.}} S_{\text{int.} \times \text{subjects}} \\
X_{\text{microphones} \times \text{time}} & \approx A_{\text{microphones} \times \text{people}} S_{\text{people} \times \text{time}}
\end{align*}
\]
The EEG/MEG/fMRI Cocktail Party/Blind Source Separation problem

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

Flaw: $X \approx AS = (AQ^{-1})(QS) = \hat{S}$ → **Representation not unique!**
Multi-way arrays naturally emerge in NeuroImaging
Unfortunately, multi-way/tensor structure has been widely ignored in many fields of research!
3 common ways of avoiding tensors

- **Preaverage**
  (identical time series varying spatial maps)

- **Concatenation**
  (identical time series varying spatial maps)

- **Separate Analysis**
  (identical spatial map, varying time series)

\[
\sum_{d} \approx \sum_{d}
\]

ASILOMAR 2011
Bilinear Model:

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

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

Trilinear Model:

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

Assumption: Data instantaneous mixture of temporal signatures that are expressed to various degree over the trials.
(Hitchcock 1927: Canonical Polyadic Form)
(Carrol and Chang; Harshman, 1970: Canonical Decomposition, Parallel Factor (CP))
(Möcks, 1988: Topographic Component Analysis TCA (first to analyze ERP of (channel x time x subject))
Model unique under mild conditions (see also Kruskal 1977)
Bilinear decomposition not unique

\[ X \approx AB^\top = AQQ^{-1}B^\top = \tilde{A}\tilde{B}^\top \]

Multi-linear decomposition is in general unique!!

\[ X_{(:,;:,k)} \approx A \text{diag}(C_{k,:})B^T = (AT)(T^{-1}\text{diag}(C_{k,:})Q)(Q^{-1}B^T) = \hat{A}\text{diag}(\hat{C}_{k,:})\hat{B}^T. \]

Kruskal (1976, 1977) derived the following uniqueness criterion generalized to N-ways arrays in (Sidiropoulos and Bro, 2000):

- **3-way array:** \( k_A + k_B + k_C \geq 2D + 2 \)
- **N-way array:** \( \sum_n k_{A(n)} \geq 2D + N - 1 \)

where \( k_A \) is the k-rank denoting the smallest subset of columns of \( \hat{A} \) that is guaranteed to be linearly independent. Thus, \( k_A \leq \text{rank}(A) \).

"A surprising fact is that the nonrotatability characteristic can hold even when the number of factors extracted is greater than every dimension of the three-way array." - Kruskal 1976
Unfortunately, multi-linear models are often too restrictive.

Trilinear model can encompass:
- Variability in strength over repeats

However, a common cause of variability is:
- Delay Variability
Modelling Delay Variability

Shifted CP:

\[ x_{i,k}(t) \approx \sum_{d} a_{i,d} b_{d}(t) \tau_{k,d} c_{k,d} \]

(shiftCP: Harshman, Hong and Lundy 2003)
(Mørup et al., NeuroImage 2008)
Shift CP analysis of event related EEG

\[ x_{i,j,k} \approx \sum_d a_{i,d} b_j - \tau_{k,i,d} c_{k,d} \]

(Mørup et al., NeuroImage 2008)
Two open questions we presently address

- Can we somehow restrict the various components to specific frequency ranges to facilitate component interpretation as well as impose this type of domain knowledge. 
  (The analysis of EEG has historically been constrained to various frequency bands, i.e., Delta, Theta, Alpha, Beta, and Gamma)

- Can we automatically infer the number of components casting the shiftCP in a Bayesian formulation and apply automatic relevance determination as in (Mørup et al. 2010).
  (Estimating the number of components has traditionally been carried out by estimating the models for all considered model orders and evaluated the estimated models by some heuristic)
The frequency constrained ShiftCP model

**Time Domain:**
\[
    x_{i,t,k} \approx \sum_{d=1}^{D} a_{i,d} b_{t-\tau_{i,d},d} c_{k,d}
\]

**Frequency Domain:**
\[
    \tilde{x}_{i,f,k} \approx \sum_{d=1}^{D} a_{i,d} \tilde{b}_{f,d} c_{k,d} \exp\left(-i 2\pi \frac{f-1}{T} \tau_{k,d}\right) = \sum_{d=1}^{D} a_{i,d} \tilde{b}_{f,d} \tilde{c}_{k,d}^{(f)}
\]

s.t. \( \tilde{b}_{f,d} = 0 \) if \( w_{f,d} = 0 \)

\( W \) is a binary indicator matrix such that the element \( w_{fd} = 1 \) indicate that frequency \( f \) is included in component \( d \) and \( w_{fd} = 0 \) that it is not.
A Bayesian formulation of the frequency constrained shift CP model

\[
P(\lambda | \alpha, \beta) = \prod_d \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_d^{\alpha - 1} \exp(-\beta \lambda_d)
\]

\[
P(A|\lambda) = \prod_d \left( \frac{\lambda_d}{2\pi} \right)^{1/2} \exp\left(-\frac{\lambda_d \| a_d \|^2}{2}\right)
\]

\[
P(B|\lambda) = \prod_d \left( \frac{\lambda_d}{2\pi} \right)^{T/2} \exp\left(-\frac{\lambda_d \| b_d \|^2}{2}\right)
\]

\[
P(C|\lambda) = \prod_d \left( \frac{\lambda_d}{2\pi} \right)^{K/2} \exp\left(-\frac{\lambda_d \| c_d \|^2}{2}\right)
\]

\[
P(\tau|T) = \begin{cases} \frac{1}{T} & \text{for } -T/2 < \tau \leq T/2 \\ 0 & \text{otherwise} \end{cases}
\]

\[
P(\chi|A, B, C, \tau, \sigma^2) = \prod_{itk} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_{i,t,k} - \sum_d a_{id} b_{t-\tau_{k,d},d} c_{k,d})^2}{2\sigma^2}\right)
\]

The joint log-likelihood is given by:

\[
\log P(\chi, A, B, C, \tau, \lambda|\sigma^2, \alpha, \beta) = -\frac{1}{2\sigma^2} \sum_{itk} (x_{i,t,k} - \sum_d a_{id} b_{t-\tau_{k,d},d} c_{k,d})^2
\]

\[
+ \sum_d (I + T + K + \alpha - 1) \log \lambda_d - \frac{\lambda_d}{2} (\| a_d \|^2_F + \| b_d \|^2_F + \| c_d \|^2_F + \beta) + \text{const}
\]

s.t. \quad b_{f,d} = 0 \text{ if } W_{f,d} = 0
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.

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. The horizontal axis represents the space of possible data sets $D$. Bayes rule rewards models in proportion to how much they predicted the data that occurred. These predictions are quantified by a normalized probability distribution on $D$. 

William of Ockham

Thomas Bayes

David J.C. MacKay

Evidence

$P(D|H_0)$

$P(D|H_1)$

$D$

$H_0$

$H_1$
Model Inference by MAP
(i.e., Maximizing log P with respect to the model parameters)

**Updating $A$, $B$, $C$:** solve for \( \frac{\partial \log P}{\partial A} = 0 \), \( \frac{\partial \log P}{\partial B} = 0 \), \( \frac{\partial \log P}{\partial C} = 0 \) to obtain:

\[
A \leftarrow X_{(1)} Z (\sigma^2 \text{diag}(\lambda) + Z^T Z)^{-1}
\]

\[
\dot{B}_{f,w_f:} \leftarrow X_{(2)} (\dot{C}_{w_{f,:}}^{(f)} \odot A_{w_{f,:}}) (\sigma^2 \text{diag}(\lambda_c) + \dot{C}_{w_{f,:}}^{(f)T} \dot{C}_{w_{f,:}}^{(f)} \odot A_{w_{f,:}}^T A_{w_{f,:}})^{-1}
\]

\[
C_{k,:} \leftarrow X_{(3)} (\dot{B}_{k,,:}^{(k)} \odot \dot{A}) (\sigma^2 \text{diag}(\lambda) + \dot{B}_{k,,:}^{(k)T} \dot{B}_{k,,:}^{(k)} \odot \dot{A}^T \dot{A})^{-1}
\]

($B^{(k)}$ correspond to time-delayed version of $B$ with respect to $\tau_{k,:}$)

**Updating $\lambda$:** Solving for \( \frac{\partial \log P}{\partial \lambda_d} = 0 \) we obtain

\[
\lambda_d = \frac{I + T + K + \alpha - 1}{\|a_d\|_F^2 + \|b_d\|_F^2 + \|c_d\|_F^2 + \beta}.
\]

**Setting $\sigma^2$:** Rather than estimating $\sigma^2$ from data we set $\sigma^2$ according to a predefined signal to noise ratio (SNR) of 0db. As a result

\[
\sigma^2 = \frac{\|\chi\|_F^2}{(1 + 10^{\text{SNR}/10})ITK} = \frac{\|\chi\|_F^2}{2ITK}
\]
Model Inference by MAP
(i.e., Maximizing log P with respect to the model parameters)

Updating τ: Let

\[ R_{(3)_{k,:}^d'} = X_{(3)_{k,:}} - \sum_{d \neq d'} C_{k,d'} (B_{d}^{(k)} \odot A_{d'})^T, \]

i.e. \( R_{(3)_{k,:}^d'} \) is the remaining signal at the \( k \)th row when projecting all but the \( d' \)th source out of \( X_{(3)} \). Let,

\[ S_{k,d'}(t) = \sum_{i} R_{i,i,d}^d A_{i,d'} \]
\[ \tilde{V}_{k,d'}(f) = S_{k,d'}^r(f) \tilde{B}_{f,d'}. \]

\( \tau_{k,d'} \) is estimated according to

\[ \hat{\tau}_{k,d'} = \arg\max_{\tau} |V_{k,d'}(\tau)| \]
\[ \tau_{k,d'} = \hat{\tau}_{k,d'} - (T + 1). \]

I.e. as the delay corresponding to maximum absolute cross-correlation between \( S_{k,d'}(j) \) – the time profile of the residual for the \( d' \) component and \( B_{d} \) – the component time profile.
Analysis of synthetic data
True data contains 4 components; drift, 12 Hz, 24 Hz and 50 Hz

True and estimated temporal signatures
Analysis of real event related EEG
40 components, no frequency constraints

Most components mixtures of multiple frequency ranges. No components pruned by ARD.
Analysis of real event related EEG
40 components, four different frequency constraints applied to groups of 10 components

Extracted components constrained to specified frequency regions.
No components pruned by ARD.

ASILOMAR 2011
Conclusion

- We successfully derived the frequency constrained shift CP model.
  *motivation:* frequency constraints can facilitate interpretation and component identification.

- We stated the model within a Bayesian formulation.
  This admitted automatic estimation of the number of components by automatic relevance determination (ARD).

- We demonstrated on synthetic data the success of the proposed model
  Both in terms of constraining the components to given frequencies as well as in terms of how the ARD can be used to learn the adequate number of components at the cost of fitting one conventional model.

- Applied the model to real event related EEG data
  It seems the Shift CP model is very constrained and is supported by a large number of components. The frequency constrained shift CP was able to extract prominent frequency specific components.

- The proposed frequency constrained framework readily generalize to a large variete of other models
  such as PCA/ICA, TUCKER, PARAFAC2
References