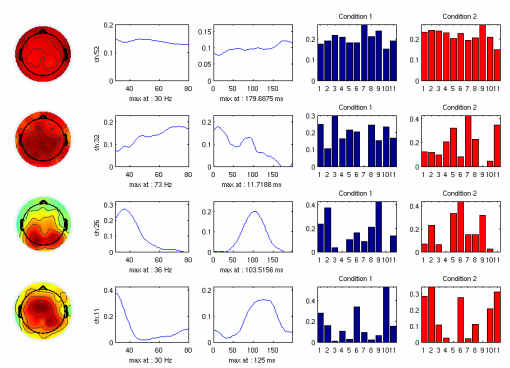
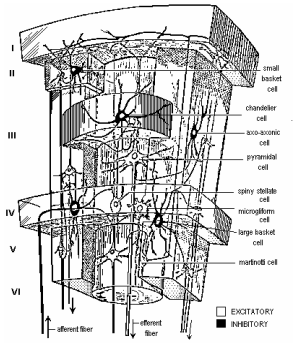
Analysis of Brain Data

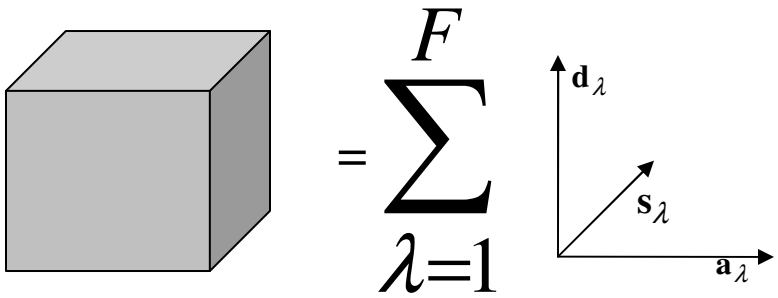
Using Multi-Way Array Models on the EEG

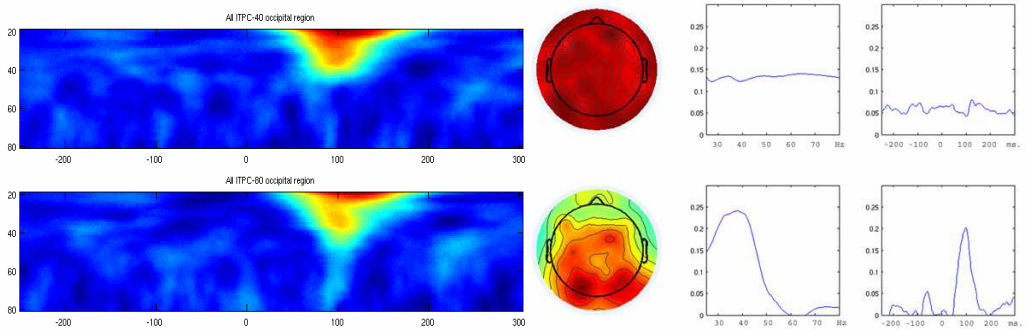
Written by

Morten Mørup









Preface

This master thesis serves as documentation for the final assignment in the requirements to achieve the degree Master of Science in Engineering. The work has been carried out in the period from the 1st of September 2004 to the 14th of February 2005 at the Intelligent Signal Processing group at the Institute of Informatics and Mathematical Modeling, Technical University of Denmark. The work has been supervised by Professor Lars Kai Hansen and part of the work has been done in corporation with Sidse Arnfred at Cognitive Research Unit, Department of Psychiatry, Hvidovre Hospital. I wish to thank both for great discussions, enthusiasm and guidance during the project.

Kgs. Lyngby, February 14, 2005

_____

Morten Mørup, s012198

Abstract

In this thesis the multi-way array model Parallel Factors (PARAFAC) also known as Canonical Decomposition (CANDECOMP) was applied to the event related potential (ERP) of electroencephalographic (EEG) recordings. Previous work done analyzing the ERP by PARAFAC had encountered great problems of degeneracy in the factors. However, in this thesis it is shown that the problem of degeneracy can be effectively circumvented by imposing non-negativity. Furthermore, the PARAFAC analysis was, to my knowledge, for the first time used to analyze the wavelet transformed data of the ERP. Through this analysis, it was shown that PARAFAC is able to access the correct components of the data. Finally, a novel PARAFAC algorithm based on independent component analysis on data having the concept of Combined Independence was proposed. This algorithm proved both fast and efficient in accessing the correct components of simulated as well as real data. In dealing with noise, the algorithm performed even better than the popular PARAFAC algorithm based on alternating least squares.

Keywords: *PARAFAC, CANDECOMP, ERP, EEG, coherence, ITPC, multi-way arrays, tensors, Independent Component Analysis, gamma activity, wavelet analysis, Combined Independence, HOSVD, TUCKER, Core Consistency Diagnostic.*

Abstrakt

I denne afhandling blev multi-way array modellen Parallel Factors (PARAFAC) også kendt som Canonical Decomposition (CANDECOMP) brugt til at analysere event relaterede potentialer (ERP) fra elektroencefalografiske optagelser (EEG). Tidligere forsøg på at analysere ERP'et ved hjælp af PARAFAC havde stødt på problemer med degeneration i faktorerne. I denne afhandling blev det vist, at dette problem kan løses ved at indføre ikke-negativitets begrænsninger. Derudover blev PARAFAC, så vidt jeg ved, for første gang brugt til at analysere wavelet-transformeret ERP-data. Det viste sig, at PARAFAC også her var i stand til at finde de rigtige komponenter i data. Endelig blev en ny PARAFAC algoritme baseret på independent component analysis foreslået til brug på data havende begrebet Combined Independence. Denne algoritme viste sig både hurtig og effektiv til at finde de korrekte komponenter i simuleret såvel som rigtig data. Algoritmen var endda bedre til at håndtere støj end den populære PARAFAC algoritme baseret på alternating least squares.

Nøgleord: *PARAFAC, CANDECOMP, ERP, EEG, coherence, ITPC, multi-way arrays, tensors, Independent Component Analysis, Gamma activity, wavelet analysis, Combined Independence, HOSVD, TUCKER, Core Consistency Diagnostic.*

Table of Contents

Notation

$\mathbf{x}, \vec{\mathbf{X}}$	Vector
$\mathbf{X}$	Matrix
$\boldsymbol{\mathcal{X}}$	Multi-way array in the literature also referred to as tensors, higher-order tensors or multidimensional matrices
x_{i_1} , $x_{i_1 i_2}$, $x_{i_1 i_2 \ldots i_N}$	Denotes the element at the subscribed indices for the corresponding vector, matrix and multi-way array.
$a \underset{\|}{} b$	Same as a=b
$\underset{nr}{=}, \underset{nr}{\Leftrightarrow}, \underset{nr}{\Rightarrow}$	nr refers to the index of an explanation given below the equation.
$\mathbf{X}_{(:,i)}$	The MATLAB notation in this case for the vector consisting of the i[th] column of **X.**
$\mathbf{x}_i$	Vector containing the i[th] column of **X,** i.e. $\mathbf{x}_i = \mathbf{X}_{(:,i)}$
$\mathbf{X}^{(i)}$	Denotes a vector, but where size $\mathbf{X}^{(i)}$ not necessarily equals $\mathbf{X}^{(j)}$
$\mathbf{X}^{+}$	The pseudo inverse of $\mathbf{X}$
$vec(\mathbf{X})$	The vectorization of $\mathbf{X}$ given by: $vec\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$
$diag(\cdot)$	On a matrix: Sets off diagonal elements of a matrix to zero. On a vector: Creates the diagonal matrix where the vector elements are along the diagonal.
$\|\cdot\|$	The Frobenius-norm, see Definition 3, page 111
$\mathbf{A} \otimes \mathbf{B}$	The tensor or Kronecker product, i.e. $\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & \cdots & a_{mm}\mathbf{B} \end{bmatrix}$
$\mathbf{A} \otimes \mathbf{B}$	The Khatri-Rao product: $\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 & \ldots & \mathbf{a}_f \otimes \mathbf{b}_f \end{bmatrix}$ Requires that **A** and **B** have same number of columns.
$\mathbf{a} \circ \mathbf{b}$	The outer product, i.e. $\mathbf{C} = \mathbf{a} \circ \mathbf{b} \Leftrightarrow c_{ij} = a_i b_j$
$\mathbf{X}^{IJ \times K}$	Unfolding $\boldsymbol{\mathcal{X}}$ by the j[th]-way onto the i[th] way, i.e. for a 3-way array $\mathbf{X}^{IJ \times K} = \begin{bmatrix} \mathbf{X}_{(:,1,:)} \\ \vdots \\ \mathbf{X}_{(:,J,:)} \end{bmatrix}$
$\mathbf{X}^{I \times JK}$	Unfolding $\boldsymbol{\mathcal{X}}$ by the K[th]-way onto the J[th] way, i.e. for a 3-way array

	$\mathbf{X}^{I\times JK} = \left[\mathbf{X}_{(:,:,1)} \quad \cdots \quad \mathbf{X}_{(:,:,K)}\right]$
$\mathbf{X}_{(n)}$	The n-mode-matricizing of the multi-way array $\boldsymbol{\mathcal{X}} \in \mathcal{R}^{I_1\times I_2\times\cdots\times I_N}$ to the matrix $\mathbf{X}_{(n)} \in \mathcal{R}^{I_n\times I_1\ldots I_{n-1}I_{n+1}\ldots I_N}$, the inverse operation is denoted $\mathbf{X}_{(n)^{-1}}$
$\mathbf{A}\times_n \mathbf{B}$	The n-mode Multiplication, see also Definition 1, page 111.
$\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle$	The scalar product of two tensors, see also Definition 2, page 111.
$rank(\boldsymbol{\mathcal{A}})$	The rank of the multi-way array $\boldsymbol{\mathcal{A}}$, see also Definition 4, page 111.
$k_{\mathbf{A}}$	The k-rank of the matrix **A**, see also Definition 6, page 112.
$\mathcal{N}(\mathbf{s}_i \mid \mathbf{0}, \mathbf{I})$	The vector $\mathbf{s}_i$ is normal distributed with mean $\mathbf{0}$ and covariance $\mathbf{I}$.
$\dim(\boldsymbol{\mathcal{X}})$	The number of dimensions of $\boldsymbol{\mathcal{X}}$,i.e. if $\boldsymbol{\mathcal{X}} \in \mathcal{R}^{I_1\times I_2\times\cdots\times I_N}$ then $\dim(\boldsymbol{\mathcal{X}}) = N$
$\langle\cdot\rangle$	The expected value of x, i.e. $\langle x\rangle = \int xp(x)dx$
$tr(\cdot)$	The sum of the diagonal elements of a matrix
$\mathrm{cov}(\mathbf{x},\mathbf{y})$	The covariance of **x** and **y**.
$\mathbf{A} \bullet \mathbf{B}$	The Hadamard product (element wise product of two matrices)
$\lvert\mathbf{A}\rvert$	The determinant of the matrix **A**.
∇, ∇^2	$\nabla = \left[\frac{\partial}{\partial x} \quad \frac{\partial}{\partial y} \quad \frac{\partial}{\partial z}\right]$ $\nabla f = \left[\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}\right]$ $\nabla \cdot \mathbf{f}$ - the divergence of $\mathbf{f}$: $\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z}$ $\nabla^T \times \mathbf{f}$ - the curl of $\mathbf{f}$ where $\times$ is the cross product. $\nabla^2 f$ - the Laplacian of f : $\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$
$\exists$	Denotes there exists, i.e. $\exists i; \lVert\mathbf{x}_i\rVert = 1$ means that there is at least one column of **X** having the norm 1.
$\forall$	Denotes for all variable, i.e. $\lVert\mathbf{x}_i\rVert = 1 \forall i$ means that the norm of each column of **X** is 1.
ε	Infinitesimal small value
e, **E**, ***E***	The model error

Abbreviations

ALS	Alternating Least Squares
ARD	Automatic Relevance Determination
CANDECOMP	CANonical DECOMPosition (same as PARAFAC)
CCD	Core Consistency Diagnostic
$CI_{m,n}$	Combined Independent in the m^{th} and n^{th} dimension
BIC	Bayesian Information Criterion
ECG	Electro-Cardiogram
EEG	Electroencephalography
EM	Expectation Maximization
EOG	Electro-Oculogram
EP	Evoked Response Potential
EPSP	Excitatory Post Synaptic Potential
ERD	Evoked Response Desynchronization
ERP	Event Related Potentials
ERPCOH	Event Related Cross Coherence
ERSP	Event Related Spectral Perturbation
FFT	Fast Fourier Transform
fMRI	Functional Magnetic Resonance Imaging
HOSVD	Higher Order Singular Value Decomposition
ICA	Independent Component Analysis
i.i.d.	Independent and identically distributed
IPSP	Inhibitory Post Synaptic Potential
ITPC	Inter Trial Phase Coherence
KL	Kullbach-Leibler divergence
LTM	Long Term Memory
MUM	Match and Utilization Model
NMF	Non-negative Matrix Factorization
PARAFAC -ALSPARAFAC -SR1PARAFAC -EMPARAFAC -VBPARAFAC -ICAPARAFAC	Parallel Factor Analysis Based on: -ALS -Sum of Rank One Components -Expectation Maximization -Variational Bayesian Expectation Maximization - PARAFAC model based on ICA
PCA	Principal Component Analysis
RE	Reticular thalamic nucleus
SNR	Signal to Noise Ratio
STFA	Short Time Fourier Analysis
SVD	Singular Value Decomposition
TCR	Thalamic Relay Cells
VBEM	Variational Bayesian Expectation Maximization

Introduction

Electroencephalography (EEG) refers to electrical activity measured at the scalp that arises from neural activity in the brain. EEG signals generated in response to sensory stimuli events are also referred to as event related potentials (ERP). Traditionally the EEG/ERP has been analyzed by looking at trial averages and spectrums. As much of the focus in the interpretation of the EEG/ERP is based on frequency changes in the data, wavelet analysis has become a popular tool. However, wavelet analysis increases the dimensionality as it adds a frequency dimension to the data giving a multi-way array of $channel \times time \times frequency$. Consequently, to be able to effectively interpret the wavelet analyzed data there is a need to decompose these multi-modal EEG/ERP data into easily interpretable components.

In this thesis multi-way array analysis of ERP will be explored. The thesis is inspired by the work of Miwakeichi and Martínez-Montes et al. [24], [25] who applied the multi-way decomposition method Parallel Factor (PARAFAC) to analyze the space-time-frequency components of the EEG. The PARAFAC model used by Miwakeichi and Martínez-Montes et al. will be compared to other PARAFAC models taken from the framework of higher order singular value decomposition (HOSVD) [19] and a more statistical framework using the expectation maximization algorithm (EM) and variational Bayesian expectation maximization (VBEM) described by Beal [3]. Finally, a PARAFAC model based on Independent Component Analysis will be proposed.

The PARAFAC algorithms will be evaluated on real as well as simulated ERP data. The real data was collected by Sidse Arnfred at Cognitive Research Unit, Department of Psychiatry, Hvidovre Hospital. The data reproduces a well known experiment described by Herrmann et al. [15] in which evoked gamma oscillations are found in the posterior regions of the brain. The PARAFAC model will be used both on the ERP as previously done by Field et al.[10], but also for the first time, to my knowledge, to analyze the wavelet transformed ERP-data in terms of the Inter Trial Phase Coherence (ITPC).

1 PARAFAC

"If the only tool you have is a hammer,
you tend to see every problem as a nail."
Abraham H. Maslow

Before addressing the Parallel Factor (PARAFAC) model and algorithms an introduction to wavelet analysis, Bayesian learning and multi-way array algebra will first be given.

1.1 Wavelet Analysis

A wavelet analysis transforms an EEG signal of $channel \times time$ into a multi-way array of $channel \times time \times frequency$.

The spectrum of a signal $x(t)$ is given by its Fourier transform:

$$X(F) = \int_{-\infty}^{\infty} x(t) e^{-i2Ft} dt \qquad \textbf{eq. 1.1}$$

However, the Fourier transform can't reveal frequency changes through the signal. This has lead to the development of the Short-Time Fourier Analysis (STFA). In STFA the signal is Fourier transformed within a finite time-window – giving a temporal resolution of the frequency components of the signal. Unfortunately, the time-window is fixed disabling good temporal resolution for high frequencies. The wavelet transform resolves this problem.

$$C(scale, shift) = \int_{-\infty}^{\infty} x(t) \varphi^{*}(scale, shift, t) dt \qquad \textbf{eq. 1.2}$$

A wavelet is a waveform of effectively limited duration that has an average value of zero. Scaling a wavelet simply means stretching or compressing it, and shifting a wavelet delaying or hastening its onset. The wavelet analysis has grown to become a huge discipline in the analysis of EEG from noise reduction to feature extraction.

Wavelets are separated into continuous and discrete wavelets based on the characteristic of the wavelet rather than the signal's characteristic as is the case for the Fourier transform. A wavelet is called continuous if it can be scaled and shifted to any value. An example of a continuous wavelet is the popular complex Morlet wavelet used in [14],[15], [24]:

$$\tilde{\varphi}(t) = \frac{1}{\sqrt{\pi F_b}} \exp(i2\pi F_c t) \exp\left(-\frac{t^2}{F_b}\right) \qquad \textbf{eq. 1.3}$$

F_c is the center frequency and F_b is a bandwidth parameter. The scaling factor a and shift factor p changes $\tilde{\varphi}$ by:

$$\varphi(a, p, t) = \frac{1}{\sqrt{a}} \tilde{\varphi}\left(\frac{(t-p)}{a}\right) = \frac{1}{\sqrt{\pi F_b a}} \exp\left(i2\pi F_c \frac{(t-p)}{a}\right) \exp\left(-\frac{(t-p)^2}{a^2 F_b}\right) \qquad \textbf{eq. 1.4}$$

$\tilde{\varphi}$ is also called the mother wavelet as it is φ without scaling and shifting. The effect of scaling is illustrated in Figure 1.1.

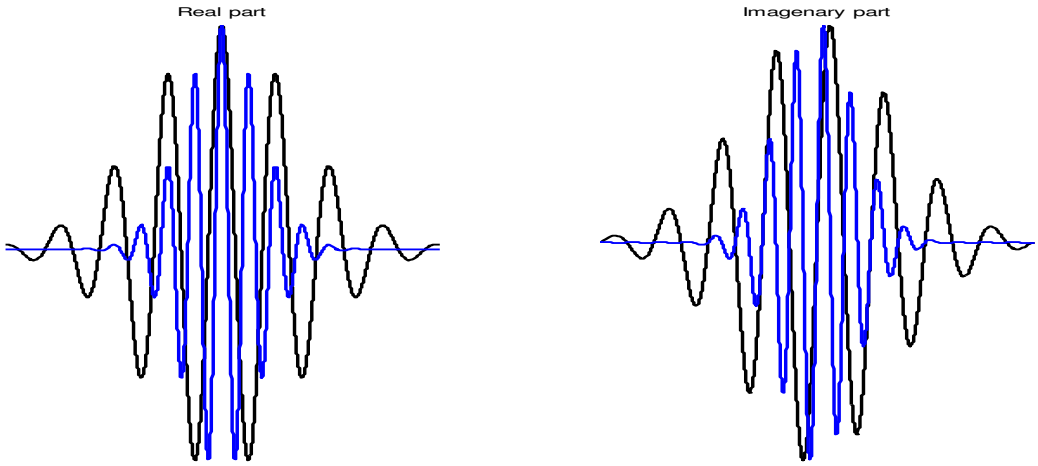


Figure 1.1: The effect of scaling the complex Morlet function. As seen scaling results in a temporal compression of the functions, black has twice the scaling factor of blue.

From the scale of the wavelet transform the frequency of the signal can be estimated as [32] :

$$F = \frac{F_c}{a} \qquad \textbf{eq. 1.5}$$

There is an inherent tradeoff for wavelets between good frequency resolution and good time resolution. This is explained by the Heisenberg-Gabor inequality [17]. As seen from eq. 1.3 a relatively large bandwidth of the wavelet gives a good frequency resolution but the length of the wavelet makes the time point less accurate. Furthermore, there is no simple relation between center frequency and bandwidth as frequency changes with scale according to eq. 1.5 but the bandwidth changes according to eq. 1.4 by $a^2 F_b$.

Consequently, in some literature the bandwidth is denoted ${\sigma_b}^2 = a^2 F_b$ making $\sigma_b \propto \frac{1}{F}$.

Although the wavelet's estimate of the frequency at a given time isn't exact and the whole analysis is slightly influenced by the choice of wavelet, the wavelet analysis is considered a very powerful tool in the analysis of the temporal development of the frequency of the EEG. In the following analysis the complex Morlet wavelet having a bandwidth parameter $F_b = 2$ and a center frequency of $F_c = 1$ will be used, as it has been well accepted in the literature, see also [14],[15].

1.2 Bayesian Learning

Reverend Thomas Bayes (1702-1761) was the first recorded to notice [4]:

$$p(x,y) = p(x|y)p(y) \Rightarrow \left.\begin{array}{l} p(a,b) = p(a|b)p(b) \\ \| \\ p(b,a) = p(b|a)p(a) \end{array}\right\} \Rightarrow \underbrace{p(b|a) = \frac{p(a|b)p(b)}{p(a)}}_{Bayes'Theorem} \qquad \textbf{eq. 1.6}$$

This theorem has become the cornerstone in a probabilistic modeling approach named Bayesian learning.

Given the data **D**, the model m, and the model parameters $\boldsymbol{\theta}$ the posterior probability distribution of the parameters can be expressed using Bayes' theorem as:

$$p(\boldsymbol{\theta}|\mathbf{D},m) = \frac{p(\mathbf{D}|\boldsymbol{\theta},m)p(\boldsymbol{\theta}|m)}{p(\mathbf{D}|m)} \qquad \textbf{eq. 1.7}$$

Where $p(\boldsymbol{\theta}|m)$is the prior probability of the parameters given the model. $p(\mathbf{D}|\boldsymbol{\theta},m)$ is the likelihood of the parameters also called the likelihood function. As $\int p(\boldsymbol{\theta}|\mathbf{D},m)\partial\boldsymbol{\theta} = 1$, $p(\mathbf{D}|m)$ is a normalization constant also denoted the *marginal likelihood*, given by:

$$p(\mathbf{D}|m) = \int p(\mathbf{D}|\boldsymbol{\theta},m)p(\boldsymbol{\theta}|m)\partial\boldsymbol{\theta} \qquad \textbf{eq. 1.8}$$

In probabilistic modeling the goal is to develop models that explain the given data but also generalize well on new data. In Bayesian learning this becomes the two main goals [3]:

1. Approximating the marginal Likelihood of the observed data $p(\mathbf{D} \mid m)$
2. Approximating the posterior distribution over the parameters of a model $p(\boldsymbol{\theta} \mid \mathbf{D}, m)$

Consequently, in probabilistic modeling two main problems must be addressed; finding the right model and the optimal parameters. The solution will be based on the Expectation Maximization algorithm (EM) and the Variational Bayesian Expectation Maximization algorithm (VBEM) based on the analysis given by Beal [3].

The problem of finding the optimal parameters will first be addressed. In maximum likelihood learning equal priors of the parameters given the model is assumed. The maximum likelihood parameter is given as the parameter that is the most probable given the model and the data:

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{D},m) \underset{1}{=} \arg\max_{\boldsymbol{\theta}} \frac{p(\mathbf{D}|\boldsymbol{\theta},m)p(\boldsymbol{\theta}|m)}{p(\mathbf{D}|m)} \underset{2}{=} \arg\max_{\boldsymbol{\theta}} p(\mathbf{D}|\boldsymbol{\theta},m)p(\boldsymbol{\theta}|m)$$
$$\boldsymbol{\theta}_{ML} \underset{3}{=} \arg\max_{\boldsymbol{\theta}} p(\mathbf{D}|\boldsymbol{\theta},m) \qquad \textbf{eq. 1.9}$$

1) *Follows by eq. 1.7.*
2) *Result as the denominator is a constant independent of* $\boldsymbol{\theta}$.
3) *Equality holds as maximum likelihood assumes equal priors of* $\boldsymbol{\theta}$ *given the model. If equal priors can't be assumed, the maximum a posteriori (MAP) estimate is found instead.*

Whereas the EM algorithm is based on the maximum likelihood parameter estimate $\boldsymbol{\theta}_{ML}$, the VBEM algorithm is based on the maximum aposteriori estimate $\boldsymbol{\theta}_{MAP}$.

1.2.1 The Expectation Maximization (EM) Algorithm

We consider a model having the hidden variables **S** and the observed data **D**. The parameters describing the (potentially) stochastic dependencies between the hidden and observed variables are given by $\boldsymbol{\theta}$. We assume further that the data $\mathbf{D} = \{\mathbf{d}_1,...,\mathbf{d}_n\}$ consist of n independent and identically distributed (i.i.d.) items, generated using a set of hidden variables $\mathbf{S} = \{\mathbf{s}_1,...,\mathbf{s}_n\}$ such that the likelihood can be written as a function of $\boldsymbol{\theta}$ in the following way [3]:

$$p(\mathbf{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{d}_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} \int p(\mathbf{d}_i,\mathbf{s}_i|\boldsymbol{\theta}) d\mathbf{s}_i \qquad \textbf{eq. 1.10}$$

The logarithm of the likelihood $\mathcal{L}(\boldsymbol{\theta})$ is defined as:

$$\mathcal{L}(\boldsymbol{\theta}) \equiv \ln p(\mathbf{D}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{d}_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \int p(\mathbf{s}_i,\mathbf{d}_i|\boldsymbol{\theta}) d\mathbf{s}_i \qquad \textbf{eq. 1.11}$$

By introducing an auxiliary distribution of the hidden variables given by $q_{\mathbf{s}}(\mathbf{s})$ we can find a lower bound of $\mathcal{L}(\boldsymbol{\theta})$:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \int p(\mathbf{s_i}, \mathbf{d_i}|\boldsymbol{\theta}) d\mathbf{s_i} = \sum_{i=1}^{n} \ln \int q_{\mathbf{s_i}}(\mathbf{s_i}) \frac{p(\mathbf{s_i}, \mathbf{d_i}|\boldsymbol{\theta})}{q_{\mathbf{s_i}}(\mathbf{s_i})} d\mathbf{s_i} \geq_{1}$$

$$\sum_{i=1}^{n} \int q_{\mathbf{s_i}}(\mathbf{s_i}) \ln \frac{p(\mathbf{s_i}, \mathbf{d_i}|\boldsymbol{\theta})}{q_{\mathbf{s_i}}(\mathbf{s_i})} d\mathbf{s_i} = \qquad \textbf{eq. 1.12}$$

$$\sum_{i=1}^{n} \int q_{\mathbf{s_i}}(\mathbf{s_i}) \ln p(\mathbf{d_i}|\boldsymbol{\theta}) d\mathbf{s_i} - \int q_{\mathbf{s_i}}(\mathbf{s_i}) \ln \frac{q_{\mathbf{s_i}}(\mathbf{s_i})}{p(\mathbf{s_i}|\mathbf{d_i}, \boldsymbol{\theta})} d\mathbf{s_i} \equiv$$

$$\mathcal{F}(q_{\mathbf{s}_1}(\mathbf{s}_1), \ldots, q_{\mathbf{s}_n}(\mathbf{s}_n), \boldsymbol{\theta})$$

1) Result of Jensens inequality, see Theorem 1.

It's worth noticing:

$$\int q_{\mathbf{s_i}}(\mathbf{s_i}) \ln \frac{q_{\mathbf{s_i}}(\mathbf{s_i})}{p(\mathbf{s_i}|\mathbf{d_i}, \boldsymbol{\theta})} d\mathbf{s_i} \equiv KL\left[q_{\mathbf{s}_i}(\mathbf{s}_i) \| p(\mathbf{s}_i \mid \mathbf{d}_i, \boldsymbol{\theta}\right] \geq 0 \qquad \textbf{eq. 1.13}$$

Is the Kullbach-Leibler (KL) divergence - a measure of distance between two distributions.

The Expectation-Maximization (EM) algorithm alternates between an E step, which infers posterior distributions over hidden variables given a current parameter setting, and an M step, which maximizes $\mathcal{L}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ given the statistics gathered from the E step [3]:

$$\textbf{E step:} \quad q_{\mathbf{s_i}}^{t+1}(\mathbf{s_i}) \leftarrow \arg\max_{q_{\mathbf{s_i}}} \mathcal{F}(q_{\mathbf{s_i}}(\mathbf{s}_i), \boldsymbol{\theta}^t), \forall i \in \{1, \ldots, n\}$$

$$\textbf{M step:} \quad \boldsymbol{\theta}^{t+1} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q_{\mathbf{s_i}}^{t+1}(\mathbf{s}_i), \boldsymbol{\theta})$$

The E step

To find the optimal distribution $q_{\mathbf{s_i}}(\mathbf{s_i})$ we differentiate $\mathcal{F}\left(q_{\mathbf{s_i}}(\mathbf{s}_i),\boldsymbol{\theta}^t\right)$ with respect to $q_{\mathbf{s_i}}(\mathbf{s_i})$ subject to the constraint $\int q_{\mathbf{s}_i}(\mathbf{s}_i)d_{\mathbf{s}_i}=1$ implemented by the Lagrange multipliers $\{\lambda_i\}_{i=1}^n$, we find:

$$\frac{d\left(\mathcal{F}\left(q_{\mathbf{s_i}}(\mathbf{s}_i),\boldsymbol{\theta}^t\right)+\sum_{i=1}^{n}\lambda_i\left[\int q_{\mathbf{s}_i}(\mathbf{s}_i)d_{\mathbf{s}_i}-1\right]\right)}{d\lambda_i}=\int q_{\mathbf{s}_i}(\mathbf{s}_i)d_{\mathbf{s}_i}-1=0\Leftrightarrow\int q_{\mathbf{s}_i}(\mathbf{s}_i)d_{\mathbf{s}_i}=1$$

$$\left.\begin{aligned}&\frac{d\left(\mathcal{F}\left(q_{\mathbf{s_i}}(\mathbf{s}_i),\boldsymbol{\theta}^t\right)+\sum_{i=1}^{n}\lambda_i\left[\int q_{\mathbf{s}_i}(\mathbf{s}_i)d_{\mathbf{s}_i}-1\right]\right)}{dq_{\mathbf{s_i}}(\mathbf{s}_i)}=\ln p\left(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta}^t\right)-\ln q_{\mathbf{s_i}}(\mathbf{s_i})+\lambda_i-1=0\\&\qquad\Updownarrow\\&q_{\mathbf{s_i}}^{t+1}(\mathbf{s_i})=\exp(\lambda_i-1)p\left(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta}^t\right)\\&\int q_{\mathbf{s}_i}(\mathbf{s}_i)d_{\mathbf{s}_i}=1\Rightarrow\int\exp(\lambda_i-1)p\left(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta}^t\right)d_{\mathbf{s}_i}=1\\&\qquad\Updownarrow\\&\exp(\lambda_i-1)=\frac{1}{\int p\left(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta}^t\right)d_{\mathbf{s}_i}}\end{aligned}\right\}\quad\begin{aligned}&q_{\mathbf{s_i}}^{t+1}(\mathbf{s_i})\\&\|\\&\exp(\lambda_i-1)p\left(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta}^t\right)\\&\|\\&\frac{p\left(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta}^t\right)}{\int p\left(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta}^t\right)d_{\mathbf{s}_i}}\\&\|\\&p\left(\mathbf{s}_i|\mathbf{d}_i,\boldsymbol{\theta}^t\right)\end{aligned}\qquad\textbf{eq. 1.14}$$

The optimal choice of the distribution $q_{\mathbf{s_i}}^{t+1}(\mathbf{s_i})$ is the posterior distribution of $\mathbf{s}_\mathrm{i}$ given by the data and model parameters $p\left(\mathbf{s}_i|\mathbf{d}_i,\boldsymbol{\theta}^t\right)$. This choice of $q_{\mathbf{s_i}}^{t+1}(\mathbf{s_i})$ fulfills according to eq. 1.13 that the KL-divergence is zero.

The M step

To find the optimal parameters we make use of the result from the E step, i.e. $q_{\mathbf{s_i}}^{t+1}(\mathbf{s_i}) = p(\mathbf{s}_i|\mathbf{d}_i,\boldsymbol{\theta})$.

$$\mathcal{F}\left(q_{\mathbf{s}_1}^{t+1}(\mathbf{s}_1),\ldots,q_{\mathbf{s}_n}^{t+1}(\mathbf{s}_n),\boldsymbol{\theta}\right) = \sum_{i=1}^{N}\int q_{\mathbf{s_i}}(\mathbf{s_i})\ln\frac{p(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta})}{q_{\mathbf{s_i}}(\mathbf{s_i})}d\mathbf{s} =$$

$$\sum_{i=1}^{N}\int p(\mathbf{s}_i|\mathbf{d}_i,\boldsymbol{\theta})\ln\frac{p(\mathbf{s_i},\mathbf{d_i}|\boldsymbol{\theta})}{p(\mathbf{s_i}|\mathbf{d_i},\boldsymbol{\theta})}d\mathbf{s_i} = \sum_{i=1}^{N}\int p(\mathbf{s}_i|\mathbf{d}_i,\boldsymbol{\theta})\ln p(\mathbf{d_i}|\boldsymbol{\theta})d\mathbf{s_i} = \sum_{i=1}^{N}\ln p(\mathbf{d_i}|\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) \qquad \textbf{eq. 1.15}$$

The M step defined by $\boldsymbol{\theta}^{t+1} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{F}\left(q_{\mathbf{s_i}}^{t+1}(\mathbf{s}_i),\boldsymbol{\theta}\right)$ therefore becomes a matter of maximizing the likelihood $\mathcal{L}(\boldsymbol{\theta})$.

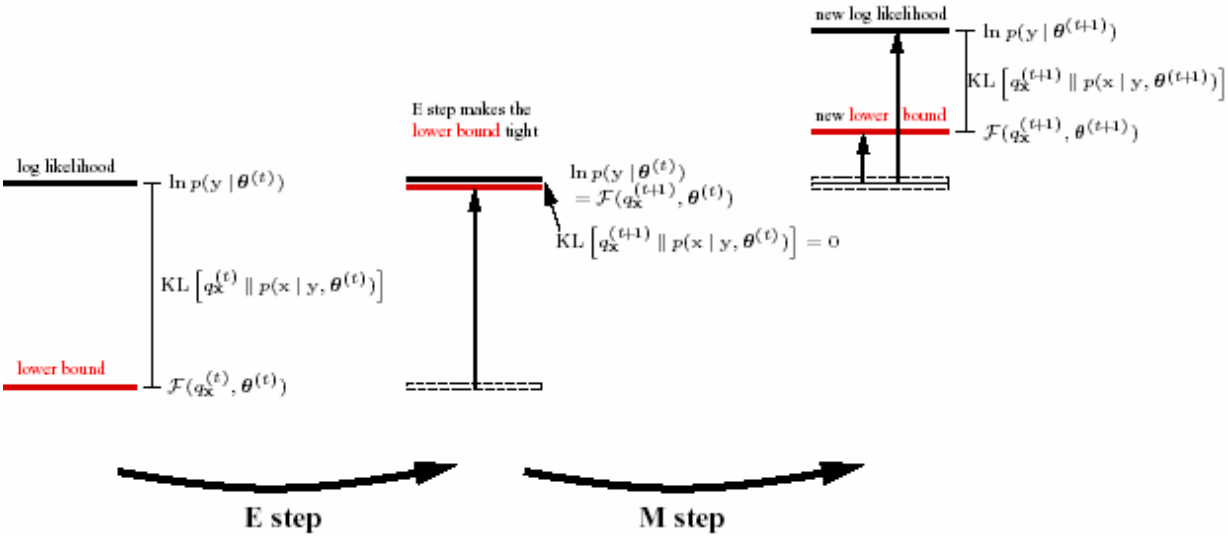


Figure 1.2: The EM algorithm for maximum likelihood learning. In the E step the hidden variable posterior is set to the exact model posterior, making the KL-divergence zero. In the M step, the lower bound of the likelihood of the parameters $\mathcal{F}\left(q_{\mathbf{s}_1}^{t+1}(\mathbf{s}_1),\ldots,q_{\mathbf{s}_n}^{t+1}(\mathbf{s}_n),\boldsymbol{\theta}\right)$ is maximized. (Taken from [3]).

Bayesian Information Criterion (BIC)

Finding the right model for the EM algorithm, i.e. choosing the number of source signals amounts to finding the optimal model M that according to the Bayesian Information Criterion satisfies:

$$M_{opt} = \arg\max_{M} p(M \mid \mathbf{D}) \qquad \textbf{eq. 1.16}$$

Where $p(M \mid \mathbf{D})$ is given by Bayes' theorem:

$$p(M \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid M)\, p(M)}{p(\mathbf{D})} \qquad \textbf{eq. 1.17}$$

Where:

$$p(\mathbf{D}) = \sum_{M} p(\mathbf{D} \mid M)\, p(M) \qquad \textbf{eq. 1.18}$$

$p(M)$ is the prior of the model and assumed to be uniform. For a particular choice of model, the probability of finding the observed data $\mathbf{D}$ is given by the integral over all model parameters:

$$\begin{aligned} p(\mathbf{D} \mid M) &= \int p(\mathbf{D}, \boldsymbol{\theta} \mid M)\, d\boldsymbol{\theta} = \int p(\mathbf{D} \mid M, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid M)\, d\boldsymbol{\theta} \\ &= \int e^{\log p(\mathbf{D} \mid M, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta} \mid M)}\, d\boldsymbol{\theta} = \int e^{-f(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \end{aligned} \qquad \textbf{eq. 1.19}$$

Where

$$f(\boldsymbol{\theta}) = -\log p(\mathbf{D} \mid M, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta} \mid M) \qquad \textbf{eq. 1.20}$$

As equal priors are assumed in maximum likelihood learning the optimum of $f(\boldsymbol{\theta})$ is given by $\boldsymbol{\theta}_{ML}$. Making a second order Taylor expansion around the optimum given by $\boldsymbol{\theta}_{ML}$ yields:

$$f(\boldsymbol{\theta}) \approx f(\boldsymbol{\theta}_{ML}) + \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{ML})^{T}\, \mathbf{H}\, (\boldsymbol{\theta} - \boldsymbol{\theta}_{ML}) \qquad \textbf{eq. 1.21}$$

Which gives:

$$p(\mathbf{D} \mid M) = \int e^{-f(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \approx e^{-f(\boldsymbol{\theta}_{ML})} \int e^{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{ML})^{T}\, \mathbf{H}\, (\boldsymbol{\theta} - \boldsymbol{\theta}_{ML})}\, d\boldsymbol{\theta} \qquad \textbf{eq. 1.22}$$

As the integral has a Gaussian form it can be written by:

$$p(\mathbf{D} \mid M) \approx p(\mathbf{D} \mid \boldsymbol{\theta}_{ML}, M)\, p(\boldsymbol{\theta}_{ML} \mid M)\, (2\pi)^{\frac{D}{2}} \left| H \right|^{-\frac{1}{2}} \qquad \textbf{eq. 1.23}$$

Where D is the number of free parameters. Neither the prior $p(\boldsymbol{\theta}_{ML} \mid M)$ nor $(2\pi)^{\frac{D}{2}}$ depends on the number of samples N. The Hessian H holds a $D \times D$ product over samples that can be factored out as $|H| = N^D |\tilde{H}|$. Neglecting $|\tilde{H}|$ gives:

$$p(\mathbf{D} \mid M) \approx p(\mathbf{D} \mid \boldsymbol{\theta}_{ML}, M) N^{-\frac{D}{2}} \qquad \textbf{eq. 1.24}$$

The probability of the observed data given the model is therefore the probability of the observed data given the optimal parameters of the model weighted by a function of the number of observations N and free parameters D.

1.2.2 Variational Bayesian EM (VBEM) algorithm

Once more we consider a model having the hidden variables **S** and the observed data **D**. The parameters describing the (potentially) stochastic dependencies between the hidden and observed variables are given by $\boldsymbol{\theta}$. We again assume that the data $\mathbf{D} = \{\mathbf{d}_1, ..., \mathbf{d}_n\}$ consist of n independent and identically distributed (i.i.d.) items, generated using a set of hidden variables $\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$ such that the likelihood can be written as a function of **S** and $\boldsymbol{\theta}$ in the following way [3]:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \mathbf{S}) \equiv \ln p(\mathbf{D}|m) = \ln \int p(\mathbf{D}, \mathbf{S}, \boldsymbol{\theta}|m) d\boldsymbol{\theta} d\mathbf{S} = \ln \int q(\boldsymbol{\theta}, \mathbf{S}) \frac{p(\mathbf{D}, \mathbf{S}, \boldsymbol{\theta}|m)}{q(\boldsymbol{\theta}, \mathbf{S})} d\boldsymbol{\theta} d\mathbf{S} \underset{1}{\geq} \\ \int q(\boldsymbol{\theta}, \mathbf{S}) \ln \frac{p(\mathbf{D}, \mathbf{S}, \boldsymbol{\theta}|m)}{q(\boldsymbol{\theta}, \mathbf{S})} d\boldsymbol{\theta} d\mathbf{S} \underset{2}{=} \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S}) \ln \frac{p(\mathbf{D}, \mathbf{S}, \boldsymbol{\theta}|m)}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S})} d\boldsymbol{\theta} d\mathbf{S} = \\ = \mathcal{F}(q_{\mathbf{S}}(\mathbf{S}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \end{aligned} \qquad \textbf{eq. 1.25}$$

1. Result of Jensens inequality, see Theorem 1.
2. Comes from the assumption that $\boldsymbol{\theta}$ and **S** are mutually independent.

We now have:

$$\begin{aligned} \ln p(\mathbf{D} \mid m) - \mathcal{F}(q_{\mathbf{S}}(\mathbf{S}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \ln p(\mathbf{D} \mid m) - \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S}) \ln \frac{p(\mathbf{D}, \mathbf{S}, \boldsymbol{\theta}|m)}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S})} d\boldsymbol{\theta} d\mathbf{S} = \\ \ln p(\mathbf{D} \mid m) - \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S}) \ln \frac{p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{D}, m) p(\mathbf{D} \mid m)}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S})} d\boldsymbol{\theta} d\mathbf{S} = \\ \ln p(\mathbf{D} \mid m) - \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S}) \ln p(\mathbf{D} \mid m) - q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S}) \ln \frac{q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S})}{p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{D}, m)} d\boldsymbol{\theta} d\mathbf{S} = \\ \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S}) \ln \frac{q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S})}{p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{D}, m)} d\boldsymbol{\theta} d\mathbf{S} = KL(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{S}}(\mathbf{S}) \| p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{D}, m)) \end{aligned} \qquad \textbf{eq. 1.26}$$

The E and M step

From eq. 1.26 it is seen that improving the likelihood corresponds to setting the hidden variables as well as $\boldsymbol{\theta}$ equal their posteriors, as this minimizes the Kullbach-Leibler divergence $KL(q_{\boldsymbol{\theta}}(\boldsymbol{\theta})q_{\mathbf{S}}(\mathbf{S})\|p(\mathbf{S},\boldsymbol{\theta}|\mathbf{D},m))$.

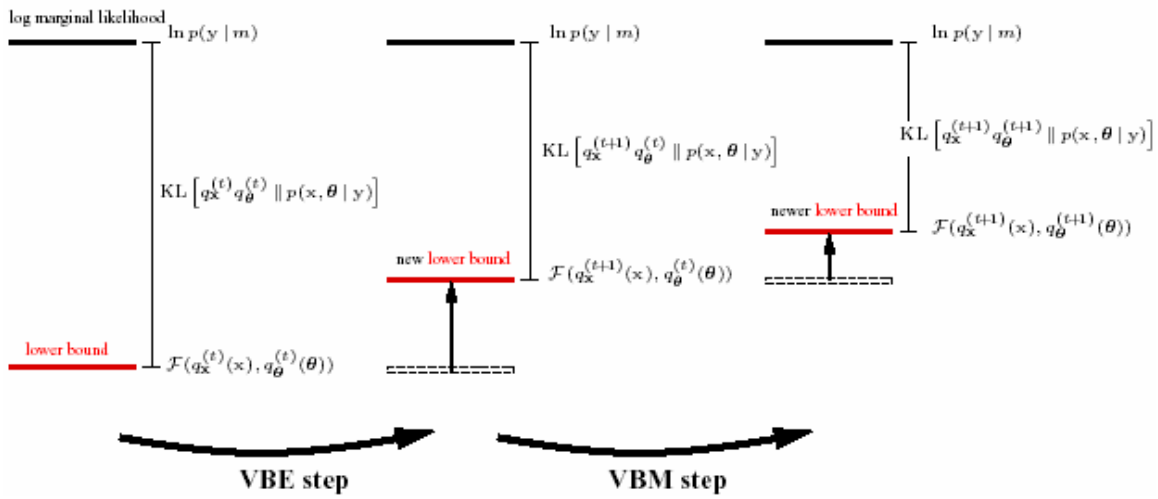


Figure 1.3: The VBEM algorithm. In the VBE step the variational posterior over hidden variable is no longer set to the exact model posterior. However, each VBE and VBM step is assured to improve the lower bound of the likelihood (Figure taken from [3]).

As the marginal likelihood doesn't change, choosing the right model amounts to finding the model having the largest lower bound of the marginal likelihood. Furthermore, hyper parameters can be used to model the various parameters of the model. The hyper parameters can then indicate how many factors to include by how certain the underlying parameter is zero. Estimating the number of factors to include in the model by hyper parameters is called Automatic Relevance Determination (ARD).

1.3 Multi-way arrays

Multi-way arrays in the literature also referred to as tensors, higher-order tensors or multidimensional matrices [19] are simply any set of data for which the elements can be arranged as [5]:

$$x_{ijk\ldots}\ \text{i=1.I, j=1…J, k=1…K, ... i.e. } \boldsymbol{\mathcal{X}} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n}$$

Notice that vectors and matrices are two special cases of multi-way arrays; a 1-way array and a 2-way array. In the following the various ways of a multi-way array will also be referred to as modalities. For a description of the different aspects of multi-way arrays see Appendix B: Multi-way array algebra.

1.3.1 Unfolding

The unfolding operation folds one of the ways of the multi-way data onto another. Consider for example the three-way array $\boldsymbol{\mathcal{X}}$ defined by x_{ijk}, i=1.I, j=1…J, k=1…K. Unfolding the third way of $\boldsymbol{\mathcal{X}}$ onto the second way gives:

$$\boldsymbol{\mathcal{X}}^{I \times J \times K} \overset{unfolding}{\rightarrow} \mathbf{X}^{I \times JK}$$

While unfolding the second way of $\boldsymbol{\mathcal{X}}$ onto the third way gives:

$$\boldsymbol{\mathcal{X}}^{I \times J \times K} \overset{unfolding}{\rightarrow} \mathbf{X}^{I \times KJ}$$

For a three-way array there are 6 different options of unfolding $\boldsymbol{\mathcal{X}}$ into a matrix as revealed in Figure 1.4. The unfolding can be performed consecutively turning for instance a four-way array into a vector by three unfolding operations:

$$\boldsymbol{\mathcal{X}}^{I \times J \times K \times L} \overset{unfolding}{\rightarrow} \boldsymbol{\mathcal{X}}^{IL \times J \times K} \overset{unfolding}{\rightarrow} \mathbf{X}^{IL \times JK} \overset{unfolding}{\rightarrow} \mathbf{x}^{ILJK}$$

Unfolding multi-way data enables manipulation of the data using normal vector and matrix calculation.

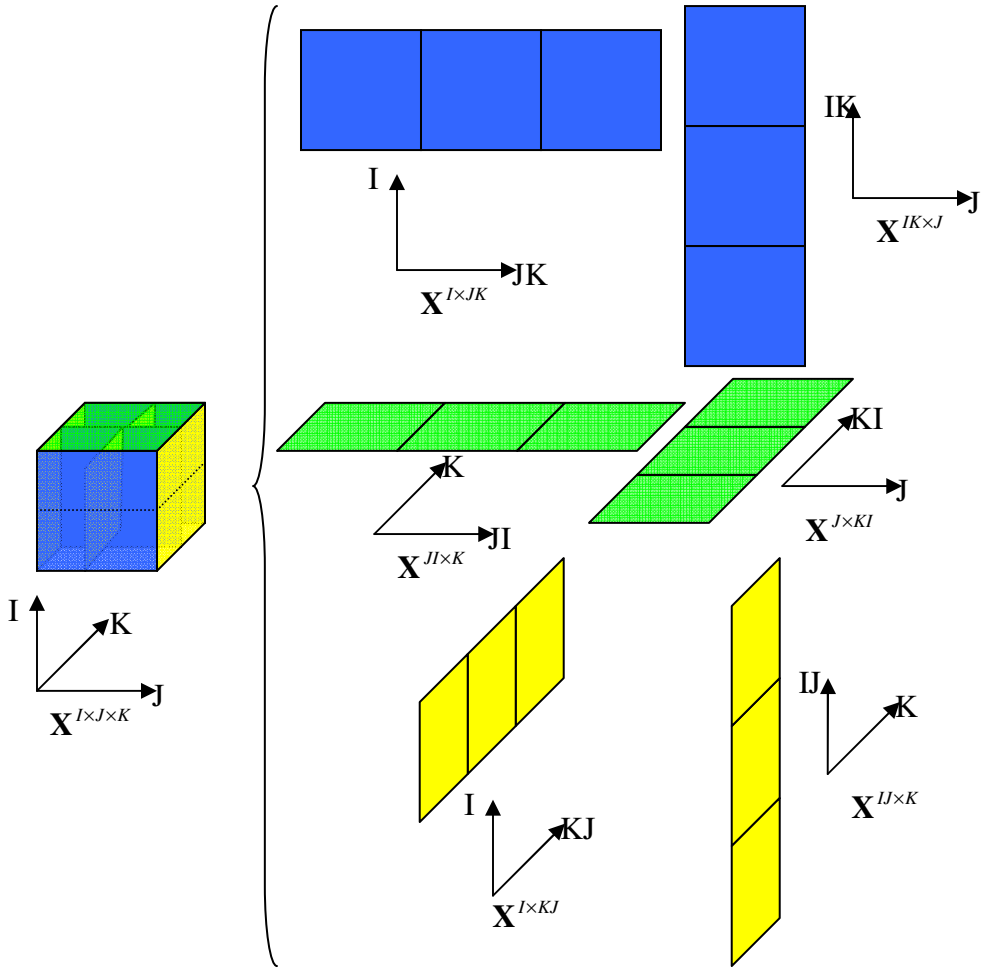


Figure 1.4: The six ways of unfolding the three-way array $\boldsymbol{\mathcal{X}}$ into a matrix.

The n-mode-matricizing $\mathbf{X}_{(n)}$ of the tensor $\boldsymbol{\mathcal{X}} \in \mathcal{R}^{I_1\times I_2\times\cdots\times I_N}$ is defined as the $\dim(\boldsymbol{\mathcal{X}})$-2 unfolding giving [19] ,[23]:

$$\mathbf{X}_{(n)} \in \mathcal{R}^{I_n\times I_1\ldots I_{n-1}I_{n+1}\ldots I_N}$$

The inverse operation is denoted:

$$\mathbf{X}_{(n)^{-1}} = \boldsymbol{\mathcal{X}} \in \mathcal{R}^{I_1\times I_2\times\cdots\times I_N}.$$

For MATLAB implementation of the described multi-way array manipulation as well as the algebra given in Appendix B: Multi-way array algebra, confer Appendix C: MATLAB implementation of multi-way array manipulations.

1.4 Models

The two most used forms of decomposition of multi-way arrays are the PARAFAC and the TUCKER model [23]. Where the PARAFAC decomposition gives easily interpretable components, the TUCKER model is a convincing multilinear generalization of the SVD concept to higher order [19] . Furthermore, the TUCKER model enables evaluation of the PARAFAC model using the so-called Core Consistency Diagnostic, see also paragraph 1.5.1.

1.4.1 Parallel Factor Analysis (PARAFAC)

The PARAFAC model is intrinsically related to the principle of parallel proportional profiles [5]. Suppose that the matrix $\mathbf{X}^{(1)}$ can be adequately modeled as $\mathbf{AS}^T$ where the number of columns of $\mathbf{A}$ and $\mathbf{S}$ is the same.

$$\mathbf{X}^{(1)} = \mathbf{AS}^T = \mathbf{a}_1\mathbf{s}_1^T + \mathbf{a}_2\mathbf{s}_2^T + \ldots + \mathbf{a}_F\mathbf{s}_F^T = \mathbf{a}_1\mathbf{s}_1^T d_{11}^{(1)} + \mathbf{a}_2\mathbf{s}_2^T d_{22}^{(1)} + \ldots + \mathbf{a}_F\mathbf{s}_F^T d_{FF}^{(1)} = \mathbf{AD}^{(1)}\mathbf{S}^T \quad \text{where } \mathbf{D}^{(1)} = \mathbf{I} \qquad \textbf{eq. 1.27}$$

Suppose another matrix $\mathbf{X}^{(2)}$ can be described by the same matrices $\mathbf{A}$ and $\mathbf{S}$ only in different proportions:

$$\mathbf{X}^{(2)} = \mathbf{a}_1\mathbf{s}_1^T d_{11}^{(2)} + \mathbf{a}_2\mathbf{s}_2^T d_{22}^{(2)} + \ldots + \mathbf{a}_F\mathbf{s}_F^T d_{FF}^{(2)} = \mathbf{AD}^{(2)}\mathbf{S}^T \quad \text{where } \mathbf{D}^{(2)} \text{ is a diagonal matrix} \qquad \textbf{eq. 1.28}$$

The two models consist of the same (parallel) profiles only in different proportions. Cattell was the first to prove that the presence of parallel proportional profiles would lead to an unambiguous decomposition [5].

The Parallel Factor, PARAFAC, model was independently proposed by Harshman [13] and by Carrol and Chang [5] in 1970. The latter naming it Canonical Decomposition, CANDECOMP. The model can be expressed in several ways:

$$x_{ijk} = \sum_{\lambda=1}^{F} a_{i\lambda} b_{j\lambda} s_{k\lambda} + e_{ijk} \quad \text{, where } F \text{ is the number of factors.} \qquad \textbf{eq. 1.29}$$

Due to the symmetry of the components in eq. 1.29 the index order of the components doesn't matter. Another formulation of the model is given by:

$$\mathbf{X}^{(i)} = \mathbf{AD}^{(i)}\mathbf{S} + \mathbf{E}^{(i)} \quad \text{, where } \mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(M)} \end{bmatrix}, \text{and } \mathbf{D}^{(i)} \text{ is a diagonal matrix} \qquad \textbf{eq. 1.30}$$

From the formulation of the model in eq. 1.30 the relation of PARAFAC to parallel proportional profiles is evident. Finally, the model can be expressed more compact by the Khatri-Rao product.

$$\mathbf{X}^{I\times JK} = \mathbf{A}\left(\mathbf{S}\middle|\otimes\middle|\mathbf{B}\right)^T \qquad \textbf{eq. 1.31}$$

Where the i[th] row of **B** corresponds to the diagonal of $\mathbf{D}^{(i)}$. The Khatri-Rao product is given by [5]:

$$\mathbf{A}\middle|\otimes\middle|\mathbf{B} = \left[\mathbf{a}_1 \otimes \mathbf{b}_1 \quad \ldots \quad \mathbf{a}_f \otimes \mathbf{b}_f\right], \text{ where } \begin{matrix}\mathbf{A} = \left[\mathbf{a}_1 & \cdots & \mathbf{a}_f\right] \\ \mathbf{B} = \left[\mathbf{b}_1 & \cdots & \mathbf{b}_f\right]\end{matrix}$$

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \ldots & a_{1m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & \ldots & a_{nm}\mathbf{B} \end{bmatrix} \qquad \textbf{eq. 1.32}$$

and the Kronecker product

The PARAFAC model is easily generalized to higher orders. The higher order equivalents are given by:

$$x_{i_1 i_2 \ldots i_N} = \sum_{\lambda=1}^{F} a_{i_1\lambda}^{(1)} a_{i_2\lambda}^{(2)} \cdots a_{i_N\lambda}^{(N)} + e_{i_1 i_2 \ldots i_N} \quad \text{, where } F \text{ is the number of factors.} \qquad \textbf{eq. 1.33}$$

Expressing the PARAFAC for arrays of more than three dimensions in terms of eq. 1.29 and eq. 1.30 yields:

$$\mathbf{X}^{(I_1\times I_2 i_3 \ldots i_N)} = \mathbf{A}\mathbf{D}^{(i_3)}\mathbf{D}^{(i_4)}\ldots\mathbf{D}^{(i_N)}\mathbf{S} + \mathbf{E}^{(I_1\times I_2 i_3 \ldots i_N)} \qquad \textbf{eq. 1.34}$$

where $\mathbf{D}^{(j)}$ are diagonal matrices

$$\mathbf{X} = \mathbf{A}\left(\mathbf{B}^{(N-1)}\middle|\otimes\middle|\mathbf{B}^{(N-2)}\middle|\otimes\middle|\ldots\middle|\otimes\middle|\mathbf{B}^{(1)}\right)^T + \mathbf{E} \qquad \textbf{eq. 1.35}$$

From eq. 1.29 and eq. 1.33 it is seen that PARAFAC decomposes the multi-way array into a sum of effects pertaining to each dimension. Each factor consists of one vector from each dimension. Consequently, each factor's relation to each dimension can easily be read from the factor vector corresponding to the dimension, see also Figure 1.5.

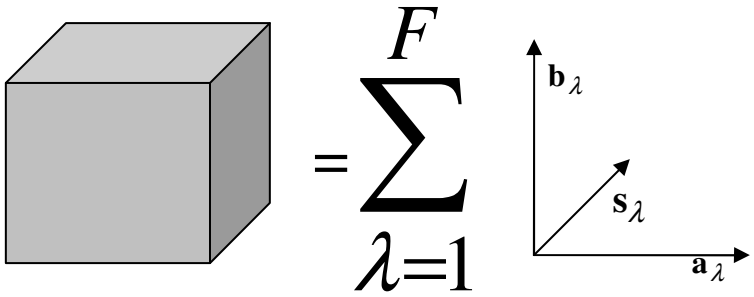


Figure 1.5: Graphical representation of the PARAFAC model as formulated in eq. 1.29. The model decomposes the multi-way array into a sum over factor effects pertaining to each dimension.

The PARAFAC model is however very restricted as the number of free parameters, D, is given by:

$$D = F\sum_{j=1}^{N} I_j \ll \prod_{j=1}^{N} I_j \text{ as } F \text{ in general is less than } \max(I_i \forall i) \qquad \textbf{eq. 1.36}$$

Uniqueness

From the formulation of the PARAFAC model given in eq. 1.30 PARAFAC doesn't hold the rotational freedom other factor models such as independent component analysis, ICA and principal component analysis, PCA have.

$$\mathbf{X}^{(i)} = \mathbf{A}\mathbf{D}^{(i)}\mathbf{S} = \mathbf{A}\mathbf{P}\mathbf{P}^{-1}\mathbf{D}^{(i)}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{S} = (\mathbf{A}\mathbf{P})\left(\mathbf{P}^{-1}\mathbf{D}^{(i)}\mathbf{Q}\right)\left(\mathbf{Q}^{-1}\mathbf{S}\right)$$
$$\Downarrow \qquad \textbf{eq. 1.37}$$
$$\left(\mathbf{P}^{-1}\mathbf{D}^{(i)}\mathbf{Q}\right) \text{ must be a diagonal matrix}$$

According to eq. 1.37 the rotational freedom of PARAFAC requires that the product $\left(\mathbf{P}^{-1}\mathbf{D}_i\mathbf{Q}\right)$ must be a diagonal matrix. In practice, this means that **P** and **Q** can only be scaling and permutation matrices. Consequently, the only indeterminacies are the order of the components and the magnitudes of the loading vectors [5].

The PARAFAC model seems a logic extension of the factor analysis as the generalization to any dimension given in eq. 1.33 yields the well known factor analysis model in the 2-way array case: $x_{ij} = \sum_{\lambda=1}^{F} a_{i\lambda} b_{j\lambda}$. It is, however, much more restricted than the normal factor analysis, as a matrix of $N_1 \times N_2 \cdot N_3$ in a factor analysis with no restrictions would give $F \cdot (N_1 + N_2 \cdot N_3)$ free parameters while the PARAFAC model of the corresponding $N_1 \times N_2 \times N_3$ multi-way array only yields $F(N_1 + N_2 + N_3)$ free parameters.

Sidiropoulos and Bro have extended J. B. Kruskal's result of uniqueness from 1977 to higher orders, for the proof see [31]. The result makes use of the k-rank which is given by the least amount of columns of a matrix that are linearly independent, see also Definition 6 page 112. Let the PARAFAC model be defined as in eq. 1.33. The model is insured to be unique apart from permutations and scaling if:

$$\sum_{i=1}^{N} k_{\mathbf{A}^{(i)}} \geq 2F + (N-1) \qquad \textbf{Eq. 1.38}$$

1.4.2 TUCKER and Higher Order Singular Value Decomposition

The generalization of singular value decomposition to multidimensional data has not yet come to one "ideal" form. For a discussion on what might define the "ideal" HOSVD see [23]. However, higher order singular value decomposition, HOSVD, of the multi-way array $\boldsymbol{X}$ will follow the definition of Lathauwer, Moor and Vandewalle [19] .

The TUCKER model is defined by:

$$x_{i_1 i_2 \ldots i_N} = \sum_{j_1}^{J_1} \sum_{j_2}^{J_2} \cdots \sum_{j_N}^{J_N} s_{j_1 j_2 \ldots j_N} U^{(1)}_{i_1 j_1} u^{(2)}_{i_2 j_2} \cdots u^{(N)}_{i_N j_N} \qquad \textbf{eq. 1.39}$$

In the 3-Way case the Tucker model can be formulated the following way using the Kronecker product:

$$\mathbf{X}^{I \times JK} = \mathbf{A}\mathbf{G}(\mathbf{B} \otimes \mathbf{S})^T \qquad \textbf{eq. 1.40}$$

eq. 1.39 can equivalently be expressed by

$$\boldsymbol{X} = \boldsymbol{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)} \qquad \textbf{eq. 1.41}$$

Where $\boldsymbol{S}$ is denoted the core multi-way array, and $\times_n$ is the n-mode multiplication see also Definition 1 page 111.

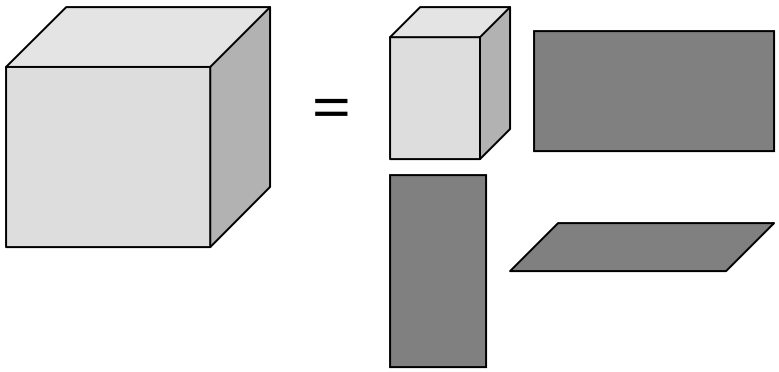


Figure 1.6: Graphical representation of the TUCKER model of a 3-way array. The model decomposes the multi-way array into matrices (dark grey) pertaining to each modality, while the core array relates each modality.

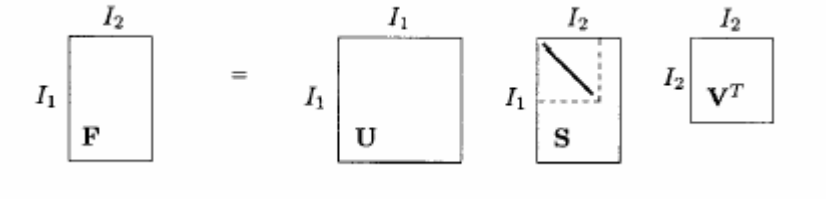
Although the TUCKER model doesn't impose any constraints, to obtain the HOSVD $\mathbf{U}^{(i)}$ has to be an orthonormal $(I_i \times I_i)$ matrix, and $\boldsymbol{\mathcal{S}}$ a multi-way array of same size as $\boldsymbol{\mathcal{X}}$ subject to:

$$\begin{aligned} \left\langle \boldsymbol{\mathcal{S}}_{I_n=\alpha}, \boldsymbol{\mathcal{S}}_{I_n=\beta} \right\rangle &= 0 \text{ if } \alpha \neq \beta \qquad (\text{all-orthogonality}) \\ \left\| \boldsymbol{\mathcal{S}}_{I_n=1} \right\| \geq \left\| \boldsymbol{\mathcal{S}}_{I_n=2} \right\| &\geq \ldots \geq \left\| \boldsymbol{\mathcal{S}}_{I_n=I_n} \right\| \geq 0 \qquad (\text{ordering}) \end{aligned} \qquad \textbf{eq. 1.42}$$

eq. 1.39 and eq. 1.41 can be equivalently expressed in matrix notation as:

$$\mathbf{X}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{S}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \otimes \cdots \otimes \mathbf{U}^{(n-1)} \right) \qquad \textbf{eq. 1.43}$$

Where $\mathbf{X}_{(n)}$ is the n-mode-matricizing of $\boldsymbol{\mathcal{X}}$ and $\mathbf{S}_{(n)}$ the n-mode-matricizing of $\boldsymbol{\mathcal{S}}$. The resemblance of HOSVD to SVD becomes evident as the singular value decomposition of a matrix $\mathbf{F}$ can be expressed by [19] :



$$\mathbf{F} = \underbrace{\mathbf{U}^{(1)} \mathbf{S} \mathbf{U}^{(2)}}_{\text{eq. 1.43}} = \underbrace{\mathbf{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}}_{\text{eq. 1.41}} \qquad \textbf{eq. 1.44}$$

Here $\mathbf{S} \in \mathcal{R}^{I_1 \times I_2}$ is an ordered pseudo-diagonal matrix, i.e.

$$\mathbf{S} = diag\left(\sigma_1, \sigma_1 \ldots \sigma_{\min(I_1, I_2)}\right)$$
$$\sigma_1 \geq \sigma \geq \ldots \geq \sigma_{\min(I_1, I_2)} \geq 0$$

Evidently, HOSVD becomes SVD for the 2-way array.

The number of free parameters for the HOSVD is given by:

$$D = \underbrace{\prod_{i=1}^{N} I_i - \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{I_i-1} j}_{All\ orthogonal}}_{\boldsymbol{S}} + \underbrace{\sum_{i=1}^{N}\left(I_i^{\,2} - \underbrace{\left(\sum_{j=1}^{I_i-1} j\right)}_{Orthogonality} - \underbrace{I_i}_{norm=1}\right)}_{\mathbf{U}^{(i)}} = \quad \textbf{eq. 1.45}$$

$$\underbrace{\prod_{i=1}^{N} I_i - \sum_{i=1}^{N}\frac{I_i^{\,2}-I_i}{2}}_{\boldsymbol{S}} + \underbrace{\sum_{i=1}^{N}\left(I_i^{\,2} - \frac{I_i^{\,2}-I_i}{2} - I_i\right)}_{\mathbf{U}^{(i)}} = \prod_{i=1}^{N} I_i$$

1. Follows as $\sum_{i=1}^{n} \frac{n(n+1)}{2}$

As can be seen from eq. 1.45 the free parameters of HOSVD exactly match the number of parameters of the multi-way array. Consequently, the HOSVD as a decomposition will reconstruct the multi-way array exact.

Calculating the HOSVD

According to eq. 1.43:

$$\mathbf{X}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{S}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \otimes \cdots \otimes \mathbf{U}^{(n-1)}\right)^T$$

Defining the following set of matrices [19] :

$$\mathbf{V}^{(n)} = \tilde{\mathbf{S}}_{(n)} \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \otimes \cdots \otimes \mathbf{U}^{(n-1)}\right)^T$$
$$\mathbf{S}_{(n)} = \boldsymbol{\Sigma}^{(n)} \tilde{\mathbf{S}}_{(n)}$$
$$\boldsymbol{\Sigma}^{(n)} = diag(\left[\sigma_1, \sigma_1 \ldots \sigma_N\right])$$

Where $\boldsymbol{\Sigma}^{(n)}$ is selected so that $\tilde{\mathbf{S}}_{(n)}$ is a normalized version of $\mathbf{S}_{(n)}$ with the rows scaled to unit length. This gives:

$$\mathbf{X}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{S}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \otimes \cdots \otimes \mathbf{U}^{(n-1)}\right)^T = \quad \textbf{eq. 1.46}$$
$$\mathbf{U}^{(n)} \boldsymbol{\Sigma}^{(n)} \mathbf{V}^{(n)}$$

From eq. 1.46 it is seen that $\mathbf{U}^{(n)}$ can be calculated by the normal singular value decomposition (SVD) of $\mathbf{X}_{(n)}$.

From eq. 1.41 we had:

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{S}} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}$$

This gives:

$$\begin{aligned}
&\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{S}} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)} \underset{1}{\Leftrightarrow} \\
&\boldsymbol{\mathcal{X}} \times_N \mathbf{U}^{(N)T} = \boldsymbol{\mathcal{S}} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_{N-1} \mathbf{U}^{(N-1)} \underset{1}{\Leftrightarrow} \\
&\boldsymbol{\mathcal{X}} \times_N \mathbf{U}^{(N)T} \times_{N-1} \mathbf{U}^{(N-1)T} \cdots \times_1 \mathbf{U}^{(1)T} = \boldsymbol{\mathcal{S}} \underset{2}{\Leftrightarrow} \\
&\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{X}} \times_1 \mathbf{U}^{(1)T} \times_2 \mathbf{U}^{(2)T} \cdots \times_N \mathbf{U}^{(N)T}
\end{aligned} \qquad \textbf{eq. 1.47}$$

1. Follows as $\mathbf{U}^{(i)}$ is an orthonormal $(I_i \times I_i)$ matrix.
2. Result of the end of Definition 1 page 111.

$\boldsymbol{\mathcal{S}}$ can equivalently be calculated by the Kronecker product:

$$\mathbf{S}_{(n)} = \mathbf{U}^{(n)T} \mathbf{X}_{(n)} \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \otimes \cdots \otimes \mathbf{U}^{(n-1)} \right) \qquad \textbf{eq. 1.48}$$

From eq. 1.42 $\boldsymbol{\mathcal{S}}$ had to fulfill $\left\| \boldsymbol{\mathcal{S}}_{I_n=1} \right\| \geq \left\| \boldsymbol{\mathcal{S}}_{I_n=2} \right\| \geq \ldots \geq \left\| \boldsymbol{\mathcal{S}}_{I_n=I_n} \right\| \geq 0$ this ordering is ensured by the fact that $\mathbf{U}^{(n)} \forall n$ is ordered. $\left\langle \boldsymbol{\mathcal{S}}_{I_n=\alpha}, \boldsymbol{\mathcal{S}}_{I_n=\beta} \right\rangle = 0 \text{ if } \alpha \neq \beta$ can be proven by:

$$\mathbf{S}_{(n)} = \mathbf{U}^{(n)T} \mathbf{X}_{(n)} \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \cdots \otimes \mathbf{U}^{(n-1)} \right)$$

$$\mathbf{S}_{(n)} \mathbf{S}_{(n)}{}^T = \mathbf{U}^{(n)T} \mathbf{X}_{(n)} \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \cdots \otimes \mathbf{U}^{(n-1)} \right) \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \cdots \otimes \mathbf{U}^{(n-1)} \right)^T \mathbf{X}_{(n)}{}^T \mathbf{U}^{(n)} \underset{1}{=}$$

$$\mathbf{U}^{(n)T} \mathbf{U}^{(n)} \boldsymbol{\Sigma}^{(n)} \mathbf{V}^{(n)} \mathbf{V}^{(n)T} \boldsymbol{\Sigma}^{(n)T} \mathbf{U}^{(n)T} \mathbf{U}^{(n)} = \boldsymbol{\Sigma}^{(n)} \mathbf{V}^{(n)} \mathbf{V}^{(n)T} \boldsymbol{\Sigma}^{(n)T} = \boldsymbol{\Sigma}^{(n)} \boldsymbol{\Sigma}^{(n)T}$$

1. Follows as
   $\left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \cdots \otimes \mathbf{U}^{(n-1)} \right) \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \cdots \otimes \mathbf{U}^{(n-1)} \right)^T = \mathbf{I}$ and
   $\mathbf{X}_{(n)} = \mathbf{U}^{(n)} \boldsymbol{\Sigma}^{(n)} \mathbf{V}^{(n)}$

As $\boldsymbol{\Sigma}^{(n)} \boldsymbol{\Sigma}^{(n)T}$ is a diagonal matrix the assumption of $\left\langle \boldsymbol{\mathcal{S}}_{I_n=\alpha}, \boldsymbol{\mathcal{S}}_{I_n=\beta} \right\rangle = 0 \text{ if } \alpha \neq \beta$ holds.

$$
\begin{aligned}
&N = \dim(\boldsymbol{\mathcal{X}}) \\
&\mathbf{U}^{(1)} \leftarrow SVD\left(\mathbf{X}_{(1)}\right) \\
&\mathbf{U}^{(2)} \leftarrow SVD\left(\mathbf{X}_{(2)}\right) \\
&\vdots \qquad\qquad \vdots \\
&\mathbf{U}^{(N)} \leftarrow SVD\left(\mathbf{X}_{(N)}\right) \\
&\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{X}} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \cdots \times_N \mathbf{U}^{(N)^T}
\end{aligned}
$$

Figure 1.7: The calculation of HOSVD

As shown in Figure 1.7, HOSVD as for SVD doesn't require any iterative operations to be estimated. Furthermore, for SVD $\mathbf{U}^{(1)}$ is the eigenvectors of $\mathbf{X}\mathbf{X}^T = \mathbf{X}_{(1)}\mathbf{X}_{(1)}{}^T$ and $\mathbf{U}^{(2)}$ is the eigenvectors of $\mathbf{X}^T\mathbf{X} = \mathbf{X}_{(2)}\mathbf{X}_{(2)}{}^T$. Clearly, HOSVD keeps this characteristic, as $\mathbf{U}^{(n)}$ is the eigenvectors of $\mathbf{X}_{(n)}\mathbf{X}_{(n)}{}^T$. Consequently, HOSVD seems like a parsimonious multi-dimensional generalization of SVD.

The uniqueness of HOSVD

As $\mathbf{U}^{(n)}$ can be calculated by the normal singular value decomposition (SVD) of $\mathbf{X}_{(n)}$ it follows that $\mathbf{U}^{(n)}$ share the uniqueness properties of SVD:
If the singular values found in eq. 1.46 all are different then $\mathbf{u}_i^{(n)}$ is unique up to sign. Furthermore, the vectors corresponding to the same n-mode singular value can be replaced by multiplication with an orthogonal matrix.

If $\mathbf{U}^{(n)}$ isn't unique there exist an orthogonal matrix $\mathbf{Q}$ so: $\mathbf{V}^{(n)} = \mathbf{U}^{(n)}\mathbf{Q}$ is also a solution of eq. 1.46. This gives a new core given by $\hat{\boldsymbol{\mathcal{S}}} = \boldsymbol{\mathcal{S}} \times_{(n)} \mathbf{Q}^{-1}$. As $\boldsymbol{\mathcal{S}}$ in eq. 1.48 is found by the knowledge of $\mathbf{U}^{(n)} \forall n$ and $\mathbf{X}_{(n)}$ it follows that $\boldsymbol{\mathcal{S}}$ due to the fact that $\mathbf{u}_i^{(n)}$ only maximally is unique up to sign can't be unique. As revealed in eq. 1.48 the change of sign gives a new value of $\boldsymbol{\mathcal{S}}$. This gives as many different values as there are sign-combinations of the vectors of $\mathbf{U}^{(n)} \forall n$.

1.4.3 Model Relations

According to Definition 7 page 112 a multi-way array is called diagonalizable if the core multi-way array $\boldsymbol{\mathcal{S}}$ of the HOSVD fulfills $\boldsymbol{\mathcal{S}}_{i_1 i_2 \ldots i_N} = 0$ unless $i_1 = i_2 = \ldots = i_N$.
With this definition the TUCKER, HOSVD and PARAFAC models can be related as shown in Figure 1.8.

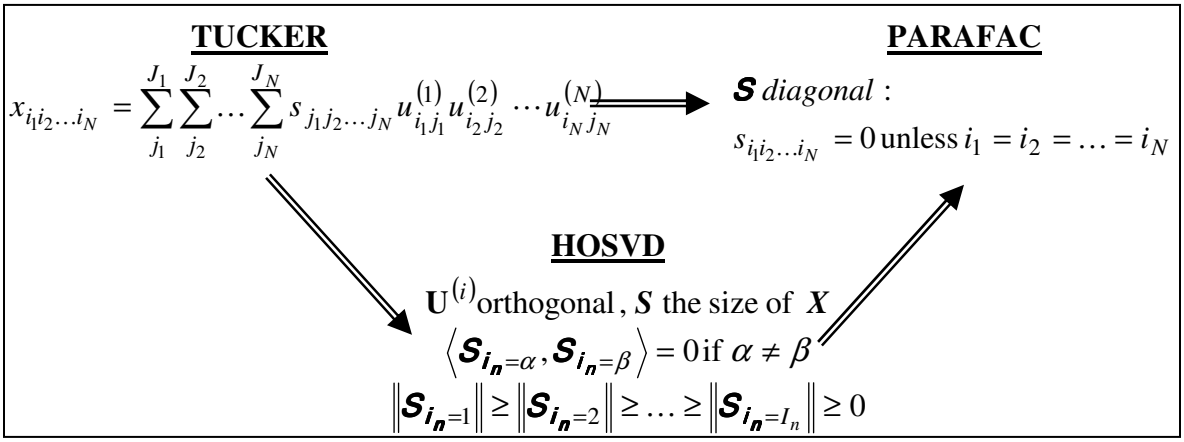


Figure 1.8: The relation between the TUCKER, HOSVD and PARAFAC model.

From Figure 1.8 it follows that if a multi-way array is diagonalizable it can be perfectly represented by the PARAFAC model. Furthermore, if the HOSVD was based on a PARAFAC model this would require the core $\boldsymbol{\mathcal{S}}$ to be diagonal. However, this is too strong a condition greatly reducing the number of free parameters (compare eq. 1.36 to eq. 1.45) disabling the HOSVD to perfectly model the data. As the core $\boldsymbol{\mathcal{S}}$ has same size as $\boldsymbol{X}$ the results of HOSVD contrary to the PARAFAC model is very hard to interpret.

The PARAFAC model is also related to the rank of a multi-way array. An N-way array $\boldsymbol{\mathcal{A}}$ has rank-1 when it equals the outer product of N vectors, i.e. $\boldsymbol{\mathcal{A}} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \ldots \circ \mathbf{u}^{(N)}$. PARAFAC of a multi-way array is the decomposition of the multi-way array into a minimal sum of rank-1 components [21]. Furthermore, by the definition of the rank of multi-way arrays, see the notice in Definition 4 page 111, the PARAFAC decomposition approximately describes the rank of the multi-way array. Consequently, a PARAFAC model based on finding the best sum of rank-1 components as described in [20],[33] will also be implemented in the following section.

1.5 PARAFAC Algorithms

Algorithms are based on an initialization and an iterative optimization. Before addressing the three iterative optimization approaches; Alternating Least Squares, EM-algorithm and VBEM-algorithm, the problem of initialization will first be addressed. This will be followed by a brief description of the handling of non-negativity and a way of evaluating the PARAFAC model by the so called Core Consistency Diagnostic.

Initialization

The choice of initialization can have significant impact on the time it takes an algorithm to converge, but more serious problems arises in situations where the function to optimize has local extremes. In this situation, the algorithm might also converge to different values depending on where it is initialized, see Figure 1.9.

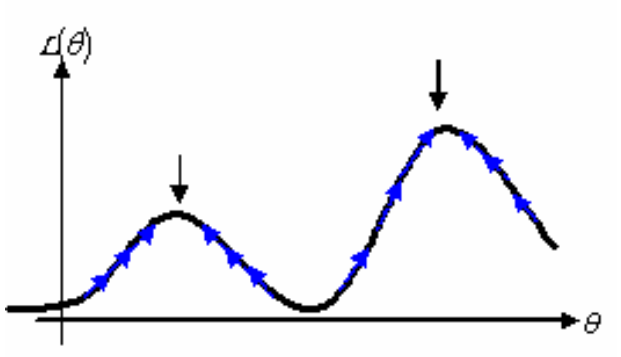


Figure 1.9: A one dimensional optimization situation where convergence depends on initialization. Here, initializing with a low value of θ makes the algorithm converge to a local extreme.

From Figure 1.8 it was seen that if the HOSVD had a diagonal core it could be considered a PARAFAC model. Furthermore, HOSVD has the advantage that $\mathbf{U}^{(i)}$ is ordered making the first F eigenvectors of $\mathbf{U}^{(i)}$ describe the most of the variation of the data. Let **k** be the vector containing the diagonal of $\boldsymbol{\mathcal{S}}$. From the HOSVD an initial guess of the PARAFAC model parameters can be found by taking the F first eigenvectors of $\mathbf{U}^{(i)} \forall i$:

$$x_{i_1 i_2 \ldots i_N} = \sum_{\lambda=1}^{F} k_\lambda u_{i_1 \lambda}^{(1)} u_{i_2 \lambda}^{(2)} \cdots u_{i_N \lambda}^{(N)} = \sum_{\lambda=1}^{F} v_{i_1 \lambda} u_{i_2 \lambda}^{(2)} \cdots u_{i_N \lambda}^{(N)} \text{ where } v_{i_1 \lambda} = k_\lambda u_{i_1 \lambda}^{(1)} \qquad \textbf{eq. 1.49}$$

However, to insure no local optimum is found from the algorithms using different initialization points is recommended.

Non negativity

In the problems at hand, the PARAFAC models are required to yield non-negative results. In the case of the VBEM-algorithm, insuring non-negativity could be achieved by using non-negative distributions such as gamma distributions as priors of the factors,

however only the hidden variables in the EM-algorithm can in this way be assured non negativity. Bro derives in his thesis two simple methods of assuring non-negativity.

One way of assuring non-negativity is simply to find the optimal unconstrained solution and to set all negative elements in every column of the factors to zero, see also Theorem 10 page 109. This method is denoted "Column-wise Non-Negativity" [5].

A more flexible method is to insure the non-negativity constraint using the following algorithm which manipulates iteratively each row of the constrained factors. The algorithm uses the feature that each element in the row of a factor is only affected by the other elements of the same row, consequently the name "Row-wise Non-Negativity" [5].

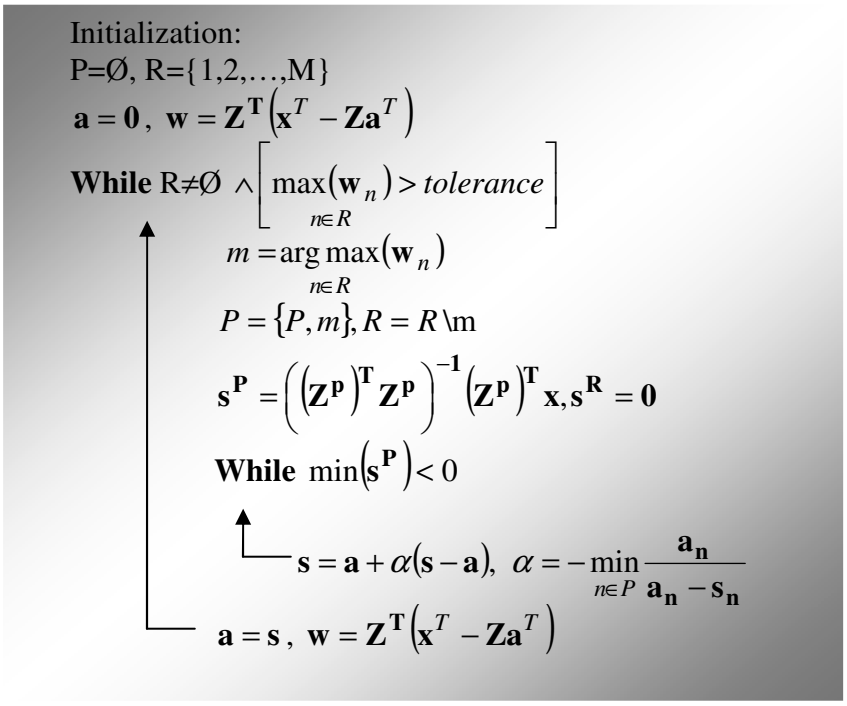


Figure 1.10: Algorithm for "Row-wise Non-Negativity". x corresponds to a row of X, where X and Z is defined as in Figure 1.11.

The algorithm is optimal in a least square sense as it optimizes the parameters according to which variable contributes the most to the squared error, i.e. $\left\|\mathbf{x}^{T}-\mathbf{Z}\mathbf{a}^{T}\right\|^{2}$. The "Row-wise" and "Column-wise" implementation of non-negativity is only valid for algorithms based on Least Square optimization, i.e. minimizing the sum of square error.

Even though the data to model aren't non-negative it can be an advantage to impose non-negativity by adding a constant and fitting the model under the non-negativity constraint. This insures no factor can counteract the effect of any other factor by having opposite sign eliminating the risk of degeneracy in the factors.

1.5.1 Core Consistency Diagnostic

The Core Consistency Diagnostic, CCD, can be applied to any model that can be considered a restricted 3-way TUCKER model [5]. Consider the PARAFAC model given in eq. 1.31. i.e. $\mathbf{X}^{I\times JK} = \mathbf{A}\left(\mathbf{S}\middle|\otimes\middle|\mathbf{B}\right)^T$. According to Figure 1.8 this can be considered a restricted 3-way TUCKER model as given in eq. 1.40, i.e. $\mathbf{X}^{I\times JK} = \mathbf{AG}\left(\mathbf{S}\middle|\otimes\middle|\mathbf{B}\right)^T$ where the core ***G*** is zero apart from along the superdiagonal which has ones. This core is denoted ***T***. ***G*** of the TUCKER model is now found by inserting **A**, **B** and **S** obtained from the PARAFAC model, i.e [5]:

$$\min\left\|\mathbf{X}^{I\times JK} - \mathbf{AG}\left(\mathbf{S}\middle|\otimes\middle|\mathbf{B}\right)^T\right\|^2 = \min\left\|vec\mathbf{X} - \left(\mathbf{B}\otimes\mathbf{S}\middle|\otimes\middle|\mathbf{A}\right)vec\mathbf{G}\right\|^2 \qquad \textbf{eq. 1.50}$$

If the PARAFAC model is valid ***G*** should resemble ***T***. A measure of resemblance is the core consistency[1]:

$$Core\,Consistency = 100\cdot\left(1 - \frac{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}\left(t_{def} - g_{def}\right)^2}{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F} g_{def}{}^2}\right) \qquad \textbf{eq. 1.51}$$

From eq. 1.51 it's seen that if the PARAFAC model is perfect, the nominator becomes zero giving a 100% consistency. If the PARAFAC model isn't correct the percentage of **G** not consistent with **T** reduces the Core Consistency. A core consistency well below 70-90% indicates that either too many components are used or the model otherwise is mis-specified [1].

Although the Core Consistency is an effective measure of how many factors to include, Bro emphasizes that other measures such as sum of squared residuals versus number of factors, inspection of the parameters and cross validation also should be taken into consideration.

[1] Bro has unclearly defined the Core Consistency in his thesis [5] and several other papers, he defines:

$$Core\,Consistency = 100\cdot\left(\frac{1 - \sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}\left(t_{def} - g_{def}\right)^2}{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F} t_{def}{}^2}\right)$$

However, in his implementation he uses the definition given in eq. 1.51.

The core consistency generalized to higher orders yields:

$$Core\,Consistency = 100 \cdot \left(1 - \frac{\sum_{i_1=1}^{F}\sum_{i_2=1}^{F}\cdots\sum_{i_N=1}^{F}\left(t_{i_1 i_2 \ldots i_N} - g_{i_1 i_2 \ldots i_N}\right)^2}{\sum_{i_1=1}^{F}\sum_{i_2=1}^{F}\cdots\sum_{i_N=1}^{F} g_{i_1 i_2 \ldots i_N}{}^2} \right) \qquad \textbf{eq. 1.52}$$

1.5.2 PARAFAC by Alternating Least Squares

Rasmus Bro and Claus Andersson have created a multi-way toolbox for Matlab [1]. Their implementation of PARAFAC is based on the technique of alternating least squares (ALS).

The principle of ALS is quite simple; initialize all model parameters for example randomly. Update each parameter by minimizing a cost-function with respect to the parameter while holding all other parameters fixed.

Consider the PARAFAC model as defined in eq. 1.31. Giving the cost function $\min\left\|\mathbf{X}^{I\times JK} - \mathbf{A}\left(\mathbf{S}\middle|\otimes\middle|\mathbf{B}\right)^T\right\|^2$, the Alternating Least Squares algorithm for PARAFAC is then defined by [5]:

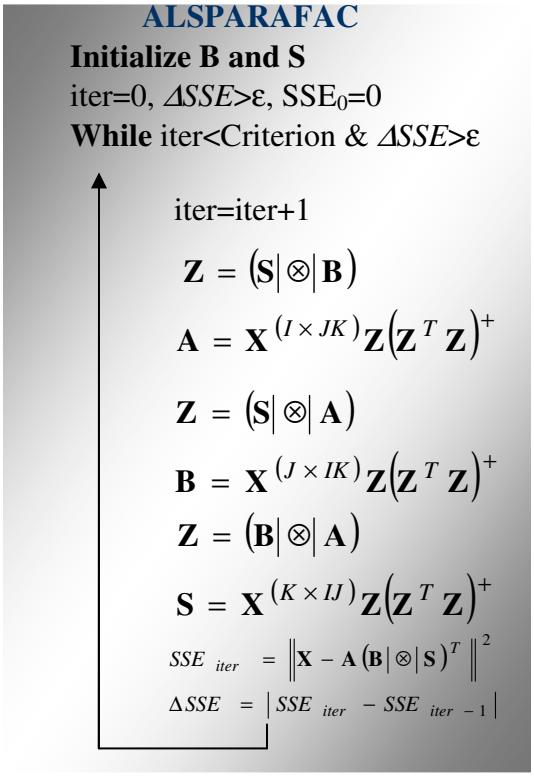


Figure 1.11: The ALSPARAFAC algorithm.

Notice how the algorithm makes use of the interchangeability of the model parameters as revealed in eq. 1.29 changing the order of the parameters by changing the unfolding of $\boldsymbol{X}$.

1.5.3 PARAFAC by multi-way rank one decomposition

As the PARAFAC model can be formulated as a sum of rank one components a PARAFAC algorithm can be defined in the framework of the HOSVD as explained in [19] , [20]. The algorithm finds the best rank one decomposition by an alternating least square approach where each dimensions factor can be directly found from the n-mode multiplication. Each consecutive factor explains the most of the remaining variation in the multi-way array.

SR1PARAFAC

$S_1=X$

For k=1:r (where r is the number of rank-1 decompositions)

Initialize $\mathbf{u}^{(i)}$ by taking the first eigenvectors of the HOSVD of $\boldsymbol{S}_k$

While $\left\|\mathbf{u}_t^{(i)} - \mathbf{u}_{t-1}^{(i)}\right\| > \varepsilon$

$$\mathbf{u}_{k,t+1}^{(1)} = \boldsymbol{S}_k \times_2 \mathbf{u}_{k,t}^{(2)} \times_3 \mathbf{u}_{k,t}^{(3)} \cdots \times_N \mathbf{u}_{k,t}^{(N)}$$

$$\mathbf{u}_{k,t+1}^{(1)} = \frac{\mathbf{u}_{k,t+1}^{(1)}}{\left\|\mathbf{u}_{k,t+1}^{(1)}\right\|}$$

$$\mathbf{u}_{k,t+1}^{(2)} = \boldsymbol{S}_k \times_1 \mathbf{u}_{k,t+1}^{(1)} \times_3 \mathbf{u}_{k,t}^{(3)} \cdots \times_N \mathbf{u}_{k,t}^{(N)}$$

$$\mathbf{u}_{k,t+1}^{(2)} = \frac{\mathbf{u}_{k,t+1}^{(2)}}{\left\|\mathbf{u}_{k,t+1}^{(2)}\right\|}$$

$$\vdots$$

$$\mathbf{u}_{k,t+1}^{(N)} = \boldsymbol{S}_k \times_1 \mathbf{u}_{k,t+1}^{(1)} \times_2 \mathbf{u}_{k,t+1}^{(2)} \cdots \times_{N-1} \mathbf{u}_{k,t+1}^{(N-1)}$$

$$\mathbf{u}_{k,t+1}^{(N)} = \frac{\mathbf{u}_{k,t+1}^{(N)}}{\left\|\mathbf{u}_{k,t+1}^{(N)}\right\|}$$

$$\lambda = \boldsymbol{S}_k \times_1 \mathbf{u}_{k,t+1}^{(1)} \times_2 \mathbf{u}_{k,t+1}^{(2)} \cdots \times_N \mathbf{u}_{k,t+1}^{(N)}$$

$$\boldsymbol{S}_{k+1} = \boldsymbol{S}_k - \lambda \mathbf{u}_{k,t+1}^{(1)} \circ \mathbf{u}_{k,t+1}^{(2)} \circ \cdots \circ \mathbf{u}_{k,t+1}^{(N)}$$

Figure 1.12: PARAFAC based on a sum of multi-way rank one decompositions, SR1PARAFAC.

1.5.4 PARAFAC by EM and VBEM

Frederik Brink Nielsen seems to be the first to derive the PARAFAC model in a statistical framework using the EM and VBEM algorithms [27]. His derivation is based on the assumption of normal distributed factors. Although non-negativity could be insured for example by insuring gamma priors it turns out to be unnecessary in the case of the EM algorithm as the row wise non negativity constraint can be implemented on all but the hidden variable **S**. This practically always insures **S** to be positive. Gamma priors could have been used in the VBEM algorithm to ensure the non-negativity. In order to implement the gamma priors each element of the k^{th} factor would depend on the other elements of its row. The algorithm would have to iterate over each dimensions own factors slowing down the already very slow algorithm. Consequently, the factors are assumed normal distributed as given by [27]. However, if the VBEM algorithm finds the true factors, non-negativity becomes just a matter of choosing the correct sign of each factor.

PARAFAC by EM

The expectation maximization of PARAFAC will be based on the following assumptions:

Model :	$\mathbf{X}^{(i)} = \mathbf{A}\mathbf{D}^{(i)}\mathbf{S} + \mathbf{E}^{(i)}$	
Assumptions :	$\mathcal{N}\left(\mathbf{e}^{(i)} \mid \mathbf{0}, \boldsymbol{\Psi}^{(i)}\right)$	(1)
	$\mathcal{N}\left(\mathbf{s}_j \mid \mathbf{0}, \mathbf{I}\right)$	(2)
	$\boldsymbol{\Psi}^{(i)}, \mathbf{D}_i$ diagonal	(3)

Due to assumption 1 the following is valid:

$$p\left(\left\{\mathbf{x}_i^{(m)}\right\}_{m=1}^{M} \mid \mathbf{s}_i, \mathbf{A}, \left\{\mathbf{D}^{(m)}, \boldsymbol{\Psi}_j^{(m)}\right\}_{m=1}^{M}\right) = \prod_{m=1}^{M} (2\pi)^{-N/2} \left|\boldsymbol{\Psi}^{(m)}\right|^{-\frac{1}{2}} \exp\left[\left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\boldsymbol{\Psi}^{(m)-1}\left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right] \quad \textbf{eq. 1.53}$$

$$\propto \exp\left[\sum_{m=1}^{M}\left(\left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\boldsymbol{\Psi}^{(m)-1}\left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right)\right]$$

E Step

As derived in Theorem 2 page 95 the following holds:

$$p\left(\mathbf{s}_i \middle| \mathbf{A}, \left\{\mathbf{x}_i^{(m)}, \mathbf{D}^{(m)}, \boldsymbol{\Psi}^{(m)}\right\}_{m=1}^{M}\right) \propto \exp\left(\mathbf{s}_i{}^T \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} - \frac{1}{2} \mathbf{s}_t^T \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)}\right) \mathbf{s}_i\right) \qquad \textbf{eq. 1.54}$$

This yields the update rule below, as derived in Theorem 3 page 96 :

$$\boldsymbol{\Sigma}_{\mathbf{S}} = \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)}\right)^{-1}$$

$$\langle \mathbf{S} \rangle = \boldsymbol{\Sigma}_{\mathbf{S}} \left(\sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{X}_m\right)$$

$$\left\langle \mathbf{S}\mathbf{S}^T \right\rangle = N\boldsymbol{\Sigma}_{\mathbf{S}} + \langle \mathbf{S} \rangle \langle \mathbf{S} \rangle^T$$

M Step

In the M Step the likelihood $\mathcal{L}(\boldsymbol{\theta})$ is maximized. From eq. 1.15 and eq. 1.53 the following is derived as revealed in Theorem 4 page 98:

$$\mathcal{L}(\boldsymbol{\theta}) =$$
$$-\frac{N}{2} \sum_{m=1}^{M} \ln\left|\boldsymbol{\Psi}^{(m)}\right| - \frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbf{X}_i^{m\,T} \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} + tr\left(\mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)} \langle \mathbf{s}\mathbf{s}_i \rangle\right) - 2\mathbf{x}_i^{(m)T} \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)} \langle \mathbf{s}_i \rangle + const \qquad \textbf{eq. 1.55}$$

This yields the update rule derived in Theorem 5 page 99:

$$\mathbf{a}_k = \left(\sum_{m=1}^{M} \boldsymbol{\Psi}^{(m)-1} \mathbf{X}^{(m)} \langle \mathbf{S} \rangle^T \mathbf{D}^{(m)}\right)_k \left(\sum_{m=1}^{M} \left(\boldsymbol{\Psi}^{(m)-1}\right)_{kk} \mathbf{D}^{(m)} \left\langle \mathbf{S}\mathbf{S}^T \right\rangle^T \mathbf{D}^{(m)}\right)^{-1}$$

$$\mathbf{d}_m = \left(\left\langle \mathbf{S}\mathbf{S}^T \right\rangle \bullet \left(\mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A}\right)\right)^{-1} vec\left[diag\left[\mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{X}^{(m)} \langle \mathbf{S} \rangle^T\right]\right]$$

$$\boldsymbol{\Psi}^{(m)} = \frac{1}{N} diag\left[\mathbf{X}^{(m)} \mathbf{X}^{(m)T} + \mathbf{A}\mathbf{D}^{(m)} \left\langle \mathbf{S}\mathbf{S}^T \right\rangle \mathbf{D}^{(m)} \mathbf{A}^T - 2\mathbf{A}\mathbf{D}^{(m)} \langle \mathbf{S} \rangle \mathbf{X}^{(m)T}\right]$$

Consequently the EMPARAFAC algorithm can be stated as shown in Figure 1.13:

EMPARAFAC

Initialize **A, D** and $\boldsymbol{\Psi}$

iter=0, $\Delta\mathcal{L}(\boldsymbol{\theta})>\varepsilon$, $\mathcal{L}(\boldsymbol{\theta})_0=0$

while iter<criterion & $\Delta\mathcal{L}(\boldsymbol{\theta})>\varepsilon$

iter=iter+1

$$\boldsymbol{\Sigma}_{\mathbf{S}} = \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)} \right)^{-1}$$

$$\langle \mathbf{S} \rangle = \boldsymbol{\Sigma}_{\mathbf{S}} \left(\sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{X}_m \right)$$

$$\left\langle \mathbf{S}\mathbf{S}^T \right\rangle = N\boldsymbol{\Sigma}_{\mathbf{S}} + \langle \mathbf{S} \rangle \langle \mathbf{S} \rangle^T$$

$$\mathbf{a}_k = \left(\sum_{m=1}^{M} \boldsymbol{\Psi}^{(m)-1} \mathbf{X}^{(m)} \langle \mathbf{S} \rangle^T \mathbf{D}^{(m)} \right)_k \left(\sum_{m=1}^{M} \left(\boldsymbol{\Psi}^{(m)-1} \right)_{kk} \mathbf{D}^{(m)} \left\langle \mathbf{S}\mathbf{S}^T \right\rangle^T \mathbf{D}^{(m)} \right)^{-1}$$

$$\mathbf{d}_m = \left(\left\langle \mathbf{S}\mathbf{S}^T \right\rangle \bullet \left(\mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \right) \right)^{-1} vec\left[diag\left[\mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{X}^{(m)} \langle \mathbf{S} \rangle^T \right] \right]$$

$$\boldsymbol{\Psi}^{(m)} = \frac{1}{N} diag\left[\mathbf{X}^{(m)} \mathbf{X}^{(m)T} + \mathbf{A}\mathbf{D}^{(m)} \left\langle \mathbf{S}\mathbf{S}^T \right\rangle \mathbf{D}^{(m)} \mathbf{A}^T - 2\mathbf{A}\mathbf{D}^{(m)} \langle \mathbf{S} \rangle \mathbf{X}^{(m)T} \right]$$

Calculate $\mathcal{L}(\boldsymbol{\theta})_{\text{iter}}$, $\Delta\mathcal{L}(\boldsymbol{\theta}) = |\mathcal{L}(\boldsymbol{\theta})_{\text{iter}} - \mathcal{L}(\boldsymbol{\theta})_{\text{iter-1}}|$

Figure 1.13: The EMPARAFAC algorithm.

PARAFAC by VBEM

For the VBEM algorithm the following assumptions are made:

$$p(\mathbf{E}^{(m)}) = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{e}_i^{(m)} \mid \mathbf{0}, diag(\boldsymbol{\varphi}_m) \right)$$

$$p(\mathbf{S}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{s}_i \mid 0, \mathbf{I})$$

$$p(\mathbf{A} \mid \alpha) = \prod_{f=1}^{F} \mathcal{N}\left(\mathbf{a}_f \mid 0, \frac{1}{\alpha_f} \mathbf{I} \right)$$

$$p\left(\alpha \mid a^{\alpha}, b^{\alpha} \right) = \prod_{f=1}^{F} \mathcal{G}\left(\alpha_f \mid a_f^{\alpha}, b_f^{\alpha} \right)$$

$$p\left(\mathbf{d}_m \mid \boldsymbol{\mu}_m, \frac{1}{v_m} \right) = \mathcal{N}\left(\mathbf{d}_m \mid \boldsymbol{\mu}_m, \frac{1}{v_m} \mathbf{I} \right)$$

$$p\left(\boldsymbol{\varphi}_m \mid \mathbf{a}_m^{\varphi}, \mathbf{b}_m^{\varphi} \right) = \prod_{p=1}^{P} \mathcal{G}\left(\varphi_{mp} \mid a_{mp}^{\varphi}, b_{mp}^{\varphi} \right)$$

For the derivation of the VBEM algorithm consult Theorem 9 page 104. The derivation yields the algorithm shown in Figure 1.14:

VBPARAFAC

Initialize **A, D,** $\boldsymbol{\alpha}$**,** $\boldsymbol{\varphi}$**,** $\mathbf{a}^{\alpha}, \mathbf{b}^{\alpha}, \mathbf{a}^{\varphi}, \mathbf{b}^{\varphi}$

iter=0, $\Delta\mathcal{F}>\varepsilon$, $\mathcal{F}_0=0$

while iter<Criterion & $\Delta\mathcal{F}>\varepsilon$

iter=iter+1

$$\boldsymbol{\Sigma}_{\mathbf{S}} = \left(\mathbf{I} + \sum_{m=1}^{M} \left\langle \mathbf{d}_m{}^T \mathbf{d}_m \right\rangle \bullet \sum_{p=1}^{P} \left\langle \varphi_{m\,p} \right\rangle \left\langle \mathbf{a}_p{}^T \mathbf{a}_p \right\rangle \right)^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{s_i}} = \boldsymbol{\Sigma}_{\mathbf{S}} \sum_{m=1}^{M} \left\langle \mathbf{d}_m{}^T \right\rangle \left\langle \mathbf{A}^T \right\rangle \langle \varphi_m \rangle \mathbf{x}_i$$

$$\boldsymbol{\Sigma}_{\mathbf{a}}{}^{p} = \left(\langle \alpha \rangle \mathbf{I} + \sum_{m=1}^{M} \left\langle \varphi^{(m)} \right\rangle \left\langle \mathbf{d}_m \mathbf{d}_m{}^T \right\rangle \bullet \sum_{i=1}^{N} \left\langle \mathbf{s}_i \mathbf{s}_i{}^T \right\rangle \right)^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{a}}{}^{p} = \boldsymbol{\Sigma}_{\mathbf{a}}{}^{p} \left\langle \sum_{m=1}^{M} \left\langle \varphi_p{}^{(m)} \right\rangle \left\langle \mathbf{D}^{(m)} \right\rangle \sum_{i=1}^{N} \mathbf{x}_i{}^{(m)T} \langle \mathbf{s}_i \rangle \right\rangle$$

$$\boldsymbol{\Sigma}_{\mathbf{C}^{(m)}} = \left(v_m \mathbf{I} + \sum_{j=1}^{P} \left\langle \varphi_{mj} \right\rangle \left\langle \mathbf{a}_{:,j}{}^T \mathbf{a}_{:,j} \right\rangle \bullet \sum_{i=1}^{N} \left\langle \mathbf{s}_i{}^T \mathbf{s}_i \right\rangle \right)^{-1}$$

$$m_{\mathbf{C}^{(m)}} = \boldsymbol{\Sigma}_{\mathbf{C}^{(m)}} \left(v_m \boldsymbol{\mu}_m + \sum_{i=1}^{N} \sum_{j=1}^{D} \left\langle \mathbf{x}_{ij}{}^{(m)} \right\rangle \varphi_{mj} \langle \mathbf{s}_i \rangle \bullet \left\langle \mathbf{a}_{:,j} \right\rangle \right)$$

$$\hat{a}_f^{\alpha} = a_f^{\alpha} + \frac{p}{2}$$

$$\hat{b}_f^{\alpha} = b_f^{\alpha} + \frac{\left\langle \mathbf{a}_f{}^T \mathbf{a}_f \right\rangle}{2}$$

$$\left\langle \alpha_f \right\rangle = \hat{a}_f^{\alpha} \Big/ \hat{b}_f^{\alpha}$$

$$\hat{a}_{mp}^{\varphi} = a_{mp}^{\varphi} + \frac{N}{2}$$

$$\hat{b}_{mp}^{\varphi} = b_{mp}^{\varphi} + \frac{1}{2} \sum_{i=1}^{N} \left\langle \left(\mathbf{x}_{ip}{}^{(m)} - \mathbf{a}_{:,p} \mathbf{D}^{(m)} \mathbf{s}_i \right)^2 \right\rangle$$

$$\left\langle \varphi_{mp} \right\rangle = \hat{a}_{mp}^{\varphi} \Big/ \hat{b}_{mp}^{\varphi}$$

Calculate $\mathcal{F}(\boldsymbol{\theta})_{\text{iter}}, \Delta\mathcal{F}(\boldsymbol{\theta}) = |\mathcal{F}(\boldsymbol{\theta})_{\text{iter}} - \mathcal{F}(\boldsymbol{\theta})_{\text{iter-1}}|$

Figure 1.14: The VBPARAFAC algorithm.

1.5.5 PARAFAC combined with ICA

Many factor analysis models exist for the two-way array analysis of matrices. Among these models Independent Component Analysis has gotten much attention as, in many situations, it has proven efficient in finding the relevant components in the data [18].

Here I show how any two-dimensional factor analysis can be applied to create a PARAFAC decomposition using features from the two dimensional factor analyses. The method will focus on the independent component analysis as revealed in the ICAPARAFAC algorithm. Furthermore, I show that for three-way arrays and some four-way arrays the decomposition can become non-iterative.

ICAPARAFAC

First I'll define what I call Combined Independence, CI. Consider the multi-way array $\boldsymbol{\mathcal{X}} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_N}$. That $\boldsymbol{\mathcal{X}}$ is combined independent in the modalities n, n+1…, N, i.e $CI_{n,n+1,\ldots N}$ means that the matricizing of $\boldsymbol{\mathcal{X}}$ into $\mathbf{X} \in \mathcal{R}^{I_1 I_2 \ldots I_{n-1} \times I_n I_{n+1} \ldots I_N}$ can be described by the model $\mathbf{X} = \mathbf{AS}$ where the rows of **S** are mutually independent, but this is not possible including less modalities in the CI. Notice, the CI applies to any combination of modalities as the order of the multi-way array modalities is only a matter of permutation. In the two-way array case, CI_2 correspond to the normal ICA model.

In ICAPARAFAC the combined independence is first identified. **A** and **S** is then found using an ICA algorithm from matricizing the multi-way array so that the columns of **X** constitutes the modalities of combined independence. Consequently, $\mathbf{A} \in \mathcal{R}^{I_1 I_2 \ldots I_{n-1} \times F}$, $\mathbf{S} \in \mathcal{R}^{F \times I_n I_{n+1} \ldots I_N}$. To find the λ^{th} factors corresponding to each of the dimensions $I_n I_{n+1} \ldots I_N$ unmatricize the λ^{th} row of **S,** i.e. $\boldsymbol{\mathcal{S}}^{(\lambda)} \in \mathcal{R}^{I_n \times I_{n+1} \times \ldots \times I_N}$. Each dimensions λ^{th} factors can now be found by the best rank-one decomposition of $\boldsymbol{\mathcal{S}}^{(\lambda)}$ using for example the SR1PARAFAC algorithm with one factor. To find the factors of each dimension in **A** corresponding to the n-1 first modalities, two approaches can be used. The first approach is similar to finding the factors underlying **S;** unmatricize the λ^{th} column of **A** to give $\boldsymbol{\mathcal{A}}^{(\lambda)} \in \mathcal{R}^{I_1 \times I_2 \times \ldots \times I_{n-1}}$. Again find each dimensions λ^{th} factor by the best rank-one decomposition of $\boldsymbol{\mathcal{A}}^{(\lambda)}$. The second approach is to find the remaining modalities factors using ALSPARAFAC on $\boldsymbol{\mathcal{X}}$ while holding the factors found underlying **S**, i.e. the n,n+1,…,N modalities fixed. Where the first approach gives the decomposition that the best describe **A** found by the ICA, the second approach gives a better approximation to $\boldsymbol{\mathcal{X}}$. The method is described in Figure 1.15.

ICAPARAFAC

Consider the model $x_{i_1 i_2 \ldots i_N} = \sum_{\lambda=1}^{F} u_{i_1 \lambda}^{(1)} u_{i_2 \lambda}^{(2)} \cdots u_{i_N \lambda}^{(N)}$

Identify the CI

Matricize $\boldsymbol{\mathcal{X}}$ so the columns of **X** correspond to the modalities of the CI.

Solve **X=AS** using ICA

Find each $\mathbf{u}_{\lambda}^{(i)}$ for the CI dimensions by finding the best rank one decomposition of the unmatricized multi-way array corresponding to the λ^{th} row of **S.**

Find $\mathbf{u}_{\lambda}$ for the dimensions not in the CI, by either finding the best rank one decomposition of the unmatricized multi-way array corresponding to $\mathbf{a}_{\lambda}$, or by ALSPARAFAC on $\boldsymbol{\mathcal{X}}$ where $\mathbf{U}^{(i)}$ of each of the CI dimensions are kept fixed.

Figure 1.15: A PARAFAC model based on ICA.

The algorithm becomes very simple when the multi-way array $\boldsymbol{\mathcal{X}}$ is of few modalities. If **A** or **S** only holds one modality the factor of this modality is given directly by **A** or **S**. Furthermore, if **A** or **S** holds two modalities, the factors of each of these two modalities can be found from the first eigenvectors corresponding to the SVD solution as this is the same as the rank one decomposition. Finally, if $\boldsymbol{\mathcal{X}}$ is a three-way array and the number of modalities of CI is one, the two approaches to find **A** yields the same results, see also Theorem 11 page 110.

Non-iterative methods for ICA such as the Molgedey-Schuster algorithm exists [18]. Consequently, decomposing any three-way arrays and four-way arrays where the number of modalities of the CI equals 2 can be done completely non-iterative when combined with the SVD method. As a result, the calculations needed to estimate the ICAPARAFAC parameters can be considerably reduced making the algorithm much faster than the ALSPARAFAC algorithm. Especially in multi-way analysis speed is an important issue as the amount of data tend to be tremendous. Therefore, the ICAPARAFAC algorithm just described seems very promising as long as combined independence can be assumed present in the data.

1.5.6 Algorithm relations

The rank one algorithm, SR1PARAFAC, corresponds to successively running an ALSPARAFAC algorithm with one factor, subtracting the found factors from the data, and finding the next factor from this subtracted dataset. The ALSPARAFAC algorithm corresponds to an EMPARAFAC model where the prior on the hidden variable **S** is a delta function, making the E step resemble the M step of the algorithm. The weight **Φ** in the EM-PARAFAC model could also have been implemented by weighted regression in the ALSPARAFAC model. Finally, the EMPARAFAC algorithm is the special case of the VBPARAFAC algorithm where the parameters priors are assumed delta functions, see also Figure 1.16.

Whereas the ALSPARAFAC algorithm seeks to find an optimal solution in terms of explaining the most variance, i.e. reducing the sum of square error, the SR1PARAFAC algorithm seeks to consecutively find factors explaining the most of the variation. The EM method and VBEM method, however, seeks to optimize the likelihood of the observed data. This does not necessarily optimize the sum of square error as the priors affect the solution. However, the EM and VBEM method is expected to generalize well by not over fit the model to the data due to the priors restricting the variables. Although the main interest in this thesis will be a PARAFAC model that well explains the observed data in favor of the ALSPARAFAC, SR1PARAFAC and ICAPARAFAC rather than finding a model that generalizes well on new data in favor of the EM and VBEM algorithm, the latter models have only been included for completeness of PARAFAC methods. Furthermore, the statistical framework enables the evaluations of questions concerning the number of factors to include in the models by the ARD and Bayesian Information Criterion, rather than just relying on the Core Consistency Diagnostic. The Bayesian Information Criterion, BIC, has been derived for the EM-algorithm in Theorem 6 page 101, while BIC has been derived for any least square optimization algorithm in Theorem 7 page 102.

Finally, the ICAPARAFAC algorithm has been developed here to handle data that can be considered CI. Whereas the ALSPARAFAC isn't optimized to yield solutions insuring CI, this is achieved by the ICAPARAFAC algorithm.

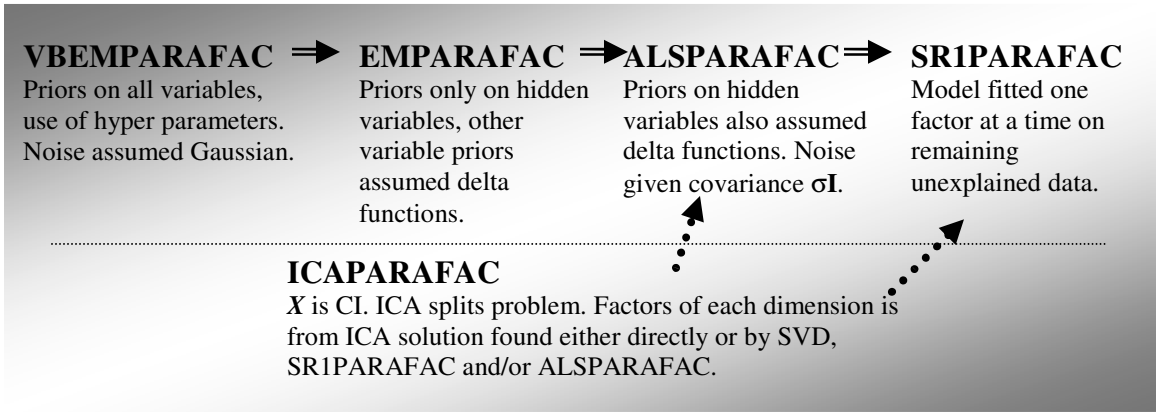


Figure 1.16: The relationship between the developed PARAFAC algorithms of this thesis.

As quoted in the beginning of this chapter, Abraham H. Maslow said: “If the only tool you have is a hammer, you tend to see every problem as a nail“. Hopefully, the tools described above will be adequate to the problems at hand.

2 EEG

"The human brain is the last, and greatest, scientific frontier. It is truly an internal cosmos that lies contained within our skulls. The more than 100 billion nerve cells and trillion supporting cells that make up your brain and mine constitute the most elaborate structure in the known universe."

Joel Davis

The term electroencephalography (EEG), as we commonly use it refers to electrical activity measured at the scalp that arises from neurons in the brain. This includes activities that arise spontaneously or in response to sensory stimuli although the latter are more commonly known as 'evoked response potentials' (EP) [29]. Finally, the EEG of sensory stimuli timed to an event is referred to as event related potentials (ERP).

As it hasn't been possible to experimentally identify the sources of the EEG signals for certain, many theories as to what constitutes the signal has arisen spanning from membrane quantum dynamical effects to K^+ fluctuations within the extracellular space. However, in this thesis only the theory which is given the most recognition in the literature will be described. This theory is primarily based on the theoretical framework explained by Paul Nunez [28], [29].

2.1 Dipoles

The electric force between two charges is defined by the well known coulombs law:

$$F = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{R^2} \qquad \textbf{eq. 2.1}$$

From coulombs law the electric field at a point $\mathbf{r}_1$ due to a point charge q located at $\mathbf{r}_2$ can be derived:

$$\mathbf{E}(\mathbf{r}_1,\mathbf{r}_2) = \frac{1}{4\pi\varepsilon_0}\frac{q}{\|\mathbf{r}-\mathbf{r}_i\|^2}\frac{(\mathbf{r}_1-\mathbf{r}_2)}{\|\mathbf{r}-\mathbf{r}_i\|} = \frac{1}{4\pi\varepsilon_0}\frac{q(\mathbf{r}_1-\mathbf{r}_2)}{\|\mathbf{r}-\mathbf{r}_i\|^3} \qquad \textbf{eq. 2.2}$$

Consequently, if there are n charges q_i located at various positions $\mathbf{r}_i$ they produce an electric field at $\mathbf{r}$ given by:

$$\vec{\mathbf{E}}(\mathbf{r}) = \sum_{i=1}^{n} \vec{\mathbf{E}}(\mathbf{r}, \mathbf{r}_i) = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{n} \frac{q(\mathbf{r} - \mathbf{r}_i)}{\|\mathbf{r} - \mathbf{r}_i\|^3} \qquad \textbf{eq. 2.3}$$

The cornerstone of the understanding of electric field behavior comes from Maxwells equations [12]:

Maxwells equations

$$\nabla \cdot \vec{\mathbf{E}} = \rho_t / \varepsilon_0 \qquad (1)$$

$$\nabla \times \vec{\mathbf{E}} = -\partial \vec{\mathbf{B}} / \partial t \qquad (2)$$

$$\nabla \cdot \vec{\mathbf{B}} = 0 \qquad (3)$$

$$\nabla \times \vec{\mathbf{B}} = \frac{\vec{\mathbf{J}}_t}{c^2 \varepsilon_0} + \frac{1}{c^2} \frac{\partial \vec{\mathbf{E}}}{\partial t} \qquad (4) \qquad \textbf{eq. 2.4}$$

$\vec{\mathbf{B}}$ is the magnetic field, ρ_t is the total charge density and $\vec{\mathbf{J}}$ is the current density.

From Maxwell's equation 2 and 4 it is seen that a change in the magnetic field results in a change in the electric field and vice versa. However, when field frequencies in the brain are less than in the order of MHz the effect of the interaction between magnetic and electric field becomes negligible [28] and the electrical potential $\mathbf{\Phi}$ can solely be determined by the electric field:

$$\vec{\mathbf{E}} = -\nabla \mathbf{\Phi} \Rightarrow$$

$$\mathbf{\Phi}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{n} \frac{q_i}{\|\mathbf{r} - \mathbf{r}_i\|} \qquad \textbf{eq. 2.5}$$

The potential of a monopole, dipole and quadrupole is shown in Figure 2.1. In general the potential due to all charges or current sources can be expressed as the following series of terms called a multipole expansion [28]:

$$\mathbf{\Phi} = (\text{monopole contribution}, r^{-1}) + (\text{dipole contribution}, r^{-2}) +$$

$$(\text{quadrupole contribution}, r^{-3}) + \ldots \qquad = \sum_{i=0}^{\infty} (2^i - \text{pole contribution}, r^{-(i+1)}) \qquad \textbf{eq. 2.6}$$

Where r is the distance to the center of the pole source

In Figure 2.1 the potential for a monopole, dipole and quadrupole is drawn.

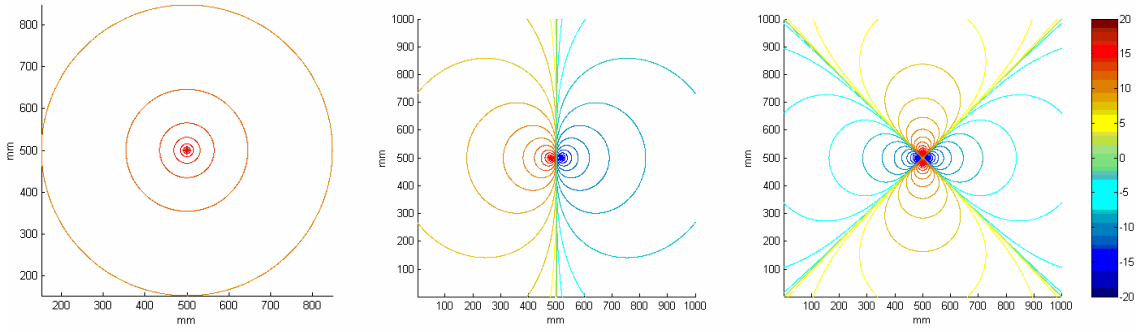


Figure 2.1: The potential for a monopole, dipole and quadrupole (units $Volt^{0.1}$)

Combining Maxwell's equation 1 and 4 yield;

$$\nabla \cdot \vec{\mathbf{J}} + \frac{\partial \rho}{\partial t} = 0 \qquad \textbf{eq. 2.7}$$

i.e. charge is neither created nor destroyed – charge is conserved. As the net charge is conserved during brain activity the monopole contribution vanishes. As all other contributions than the dipole quickly drops to zero with distance to the sources, the only contribution believed to significantly contribute to the EEG is that of the dipole. The potential of a dipole can be approximated to be, see also the derivation in Theorem 8 page 103:

$$\mathbf{\Phi}_{dipole}(\mathbf{r},\mathbf{d}) = \frac{1}{4\pi\varepsilon_0}\frac{q\|\mathbf{d}\|\cos\theta}{\|\mathbf{r}\|^2} \qquad \textbf{eq. 2.8}$$

Where **d** is the vector going from the negative charge to the positive of the dipole and θ is the angle between the vectors **r** and **d**.

As seen from the potential lines of the dipole on Figure 2.1 no or little potential is found oblique to the dipole, this is confirmed by eq. 2.8 as $\cos(90°) = 0$. Consequently, the EEG can only pick up signals from dipoles radial to the electrodes, as seen on Figure 2.2.

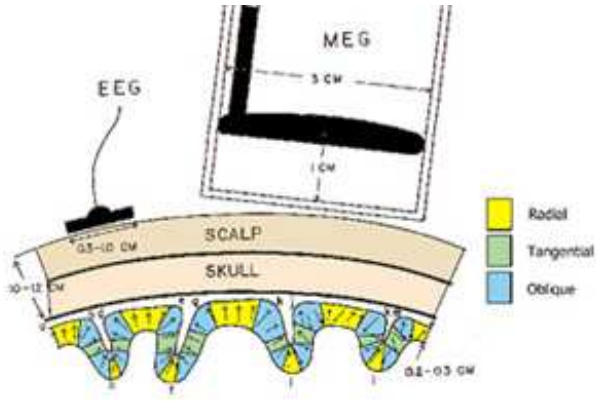


Figure 2.2: EEG can only pick up potentials from radial dipoles, whereas MEG pick up potentials from oblique dipoles (induced magnetic field is angular to current flow) (taken from [36]).

Two effects are seen with the generation of potential; capacitive current and resistive current. Nunez gives a proof that if a media is linear in both the dielectric and conductive sense i.e. the capacitor (polarization) and current is proportional to the electric field the following approximation is in general valid [28]:

$$\frac{\text{Capacitive current}}{\text{Resistive current}} \approx 0.02 \qquad \textbf{eq. 2.9}$$

Consequently, the capacitive current is in the following considered negligible. As capacitive current is ignored eq. 2.7 reduces to:

$$\nabla \cdot \vec{\mathbf{J}} = 0 \qquad \textbf{eq. 2.10}$$

Imagine two regions where all current sources are located in the first region. Expressing the current density at the boundary between the two regions as the sum of an Ohmic current (i.e. Ohms law states that current is linear to potential also makes current linear with electric field strength) and a source current $\vec{\mathbf{J}}_s$, yields:

$$\vec{\mathbf{J}} = \sigma\vec{\mathbf{E}} + \vec{\mathbf{J}}_s \qquad \textbf{eq. 2.11}$$

Where σ is the conductivity

Using first part of eq. 2.5, eq. 2.9 and eq. 2.10 this gives:

$$\nabla \cdot \sigma\nabla\mathbf{\Phi} = \nabla \cdot \mathbf{J}_s \equiv \jmath_s \qquad \textbf{eq. 2.12}$$

Where $\jmath_s$ has the dimension amperes pr. m3

Using first part of eq. 2.5, and Maxwell's first equation we also find:

$$\nabla \cdot \sigma \nabla \mathbf{\Phi} = \jmath_s \Leftrightarrow \nabla \cdot \nabla \mathbf{\Phi} = \nabla \cdot \vec{\mathbf{E}} = \jmath_s / \sigma \qquad \textbf{eq. 2.13}$$

As $\jmath_s$ is the source density, using last part of eq. 2.5 with Maxwell's first equation gives:

$$\mathbf{\Phi}(\mathbf{r},t) = \frac{1}{4\pi\sigma}\sum_{i=1}^{n}\frac{I_i(t)}{\|\mathbf{r}-\mathbf{r}_i\|} \qquad \textbf{eq. 2.14}$$

Where $I_i(t)$ is the i[th] current source at time t at position $\mathbf{r}_i$. eq. 2.14 is very important as it translates current sources into EEG measurable potentials. eq. 2.13 can also be written:

$$\nabla^2 \mathbf{\Phi} = \jmath_s / \sigma \qquad \textbf{eq. 2.15}$$

From eq. 2.15 it is seen that the current source densities can be estimated by taking the Laplacian of the measured EEG potential. Consequently, taking the Laplacian of the potential is believed to improve the spatial resolution of the EEG signal as it estimates the current sources and sinks. In practice, calculating the surface Laplacian of the EEG requires some form of interpolation of the EEG to estimate the signals between the electrodes. Often splines in the form of Legendre polynomials are used for these interpolations.

2.2 EEG and coherence

Coherence is a measure of the synchrony between sources. Let the i[th] source signal at time l be defined as E_{il}. Furthermore, let the spectral density function $q_{ii}(f)$, also called the power spectrum, and cross spectral density function $q_{ij}(f)$, also called the cross power spectrum, for two sources be given by their Fast Fourier Transform (FFT). The coherence function between the sources i and j, $\gamma_{ij}^2(f)$, is then defined by:

$$q_{ij}(f) = \frac{2\Delta t}{N}\left(\sum_{l=0}^{N-1} E_{il}\exp(-i2\pi l f \Delta t)\sum_{l=0}^{N-1} E_{jl}\exp(i2\pi l f \Delta t)\right) \Rightarrow$$
$$\gamma_{ij}^2(f) = \frac{\left|q_{ij}(f)\right|^2}{\left|q_{ii}(f)\right| \cdot \left|q_{jj}(f)\right|} \qquad \textbf{eq. 2.16}$$

If the sources are coherent at the frequency f then $\gamma_{12}^2(f) = 1$, the sources are considered incoherent if $\gamma_{12}(f) = 0$. Notice how the coherence is defined as the squared cross spectral correlation between the two sources.

Let M be the amount of coherent sources, N the amount of incoherent sources. The relative contribution of coherent to incoherent sources in the EEG is estimated to be $M/\sqrt{N}$ [28].

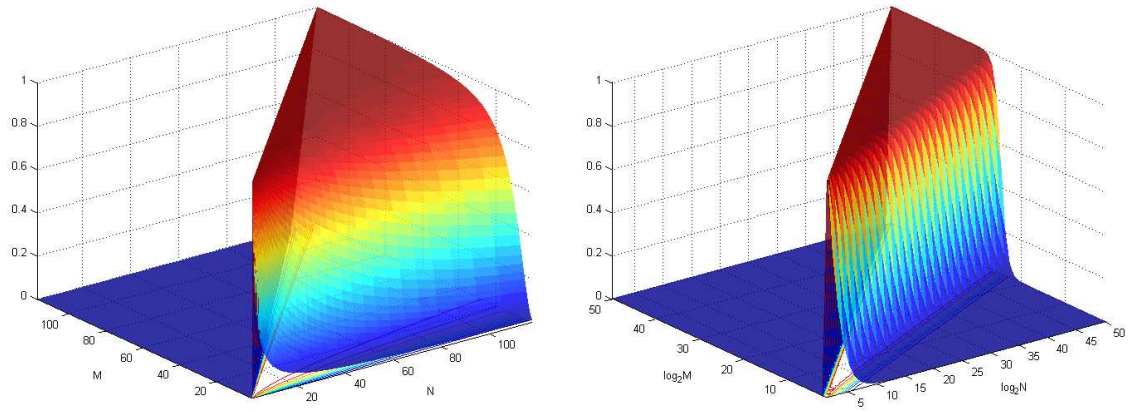


Figure 2.3: The percentage of measured signal picked up from the EEG of coherent sources m versus active sources N. Clearly, the coherent sources dominates the recorded EEG-signal even when only few of the active sources are coherent. (Notice figure not valid for M>N).

As seen on Figure 2.3 the coherent sources dramatically dominate the recorded EEG signal even though the coherent neurons are far less numerous than the incoherent neurons. The EEG signal is therefore believed to originate from the synchronous firing of parallel oriented neurons. 65-75% of the cortical neurons are oriented perpendicular to the cortical surface. These pyramidal cells have large amounts of interconnections, so it seems as if a relatively high degree of synchrony can be obtained from these neurons. Furthermore, the pyramidal neurons of the neocortex are the neurons closest to the scalp surface. According to eq. 2.6 the distance to the scalp causes pyramidal neurons to be the least reduced. Therefore, the EEG is believed to mostly originate from pyramidal cells in the neocortex [29]. Amplitude changes in the EEG with physiological state becomes, in this framework, a result of changes in the number of synchronously active neurons [28]. This is supported by current/source density studies that indicate that the pyramidal cells of layer III and V are the principal source of the EEG [35] .

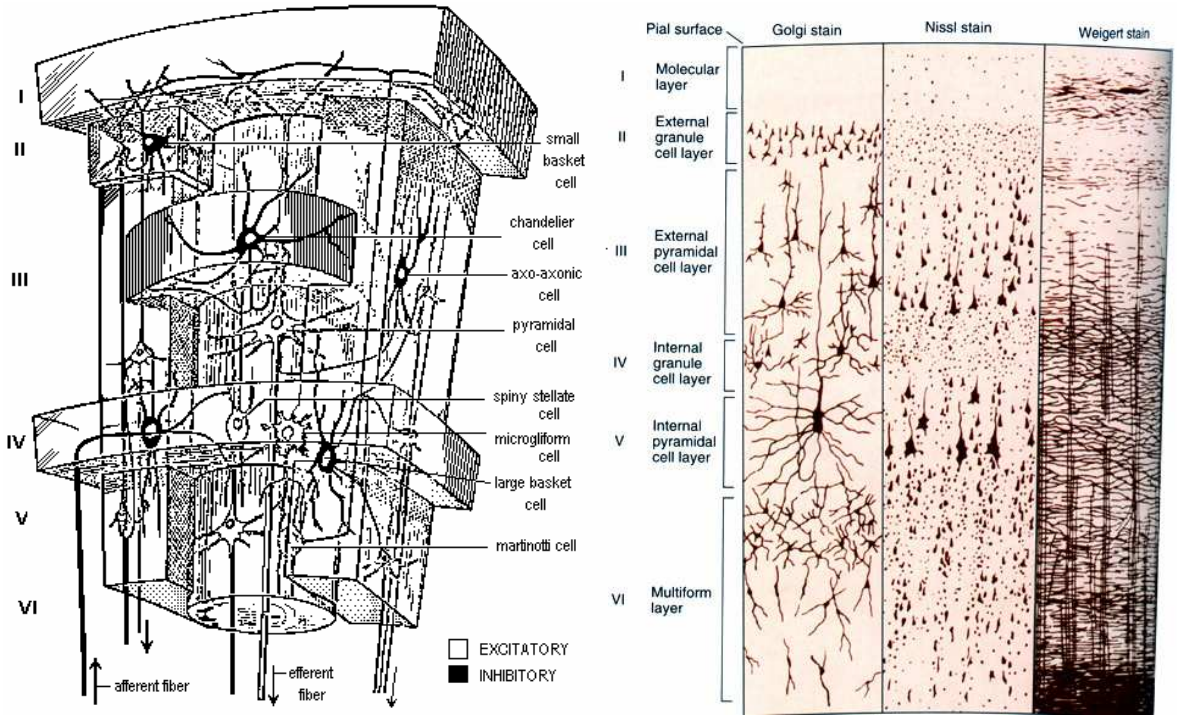


Figure 2.4: Model of the layers of neocortex and findings using Golgi, Nissl and Weigert stains. From the stains the parallel orientation of the pyramidal cell perpendicular to the cortex surface becomes evident. The pyramidal neurons constitute 70% of the neurons of cortex and posses 10^3 to 10^5 synaptic contacts. Nunez believes that each electrode pick up signal from around 30 to 40 macro-columns where each macro-column contains 10^5-10^6 neurons [29]. (Taken from [16],[37]).

The possible roles of cortical coherence are believed to be

1. **Blocking**
   Focus on a particular modality
2. **Matching**
   Transfer of data from one group of neurons to the next
3. **Binding**
   Synchronous activity between neuron groups
4. **Plasticity**
   Transfer of function from one neuron group to another

When analyzing the ERP the following measures are very useful [8]:

$$ERSP(f,t) = \frac{1}{n}\sum_{k=1}^{n}\left|F_k(f,t)\right|^2 \qquad \text{Event-related spectral perturbation}$$

$$ITPC(f,t) = \frac{1}{n}\sum_{k=1}^{n}\frac{F_k(f,t)}{\left|F_k(f,t)\right|} \qquad \text{intertrial phase coherence} \qquad \textbf{eq. 2.17}$$

$$ERPCOH^{a,b}(f,t) = \frac{1}{n}\sum_{k=1}^{n}\frac{{F_k}^a(f,t){F_k}^b(f,t)^*}{\left|{F_k}^a(f,t){F_k}^b(f,t)\right|} \qquad \text{Event related cross coherence}$$

Where n is the number of trials. Notice that the square of the ERPCOH corresponds to the coherence measure as defined in eq. 2.16.

Although coherence may appear to be an ideal measure of brain function, interpretations of experimental EEG coherence are often confounded by technical limitations. Scalp coherence between electrode sites closer than about 8 to 10 cm is typically large or moderate due only to passive current spread, volume conduction and reference electrode effects, even when the underlying cortical sources are uncorrelated [29]. This problem can to some degree be circumvented by taking the surface Laplacian as described in eq. 2.15.

2.3 Synaptic potentials and action potentials

eq. 2.8 can also be expressed in terms of current sources and sinks. Let the dipole consist of a current source at time t, $I(t)$ and a current sink $-I(t)$. Let **d** be the vector from the sink to the source and **r** the vector to the center of the dipole, then:

$$\mathbf{\Phi}_{dipole}(\mathbf{r},\mathbf{d},t) = \frac{1}{4\pi\varepsilon_0}\frac{I(t)\|\mathbf{d}\|\cos\theta}{\|\mathbf{r}\|^2} \qquad \textbf{eq. 2.18}$$

Potential differences recorded in the EEG are therefore believed to derive from coherent dipoles constituting current sources and sinks. These are thought to originate from two different processes [28]; synaptic potentials and action potentials, the latter also referred to as sodium-potassium spikes.

2.3.1 Synaptic potentials

An action potential in the presynaptic axon activates a chemical agent (transmitter) which diffuses across the synaptic cleft into the subsynaptic membrane. If the synapse is excitatory, the effect of the chemical transmitter is to increase (excitatory post synaptic potential, EPSP) or decrease (inhibitory post synaptic potential, IPSP) the permeability of the subsynaptic membrane to positive/negative ions which flow through the local surface of the membrane. The current flows across the membrane, through the intracellular fluid, back across the membrane at more distant locations, and finally back to the synapse to complete a closed loop, see Figure 2.5. This loop acts as a current source-sink dipole [28].

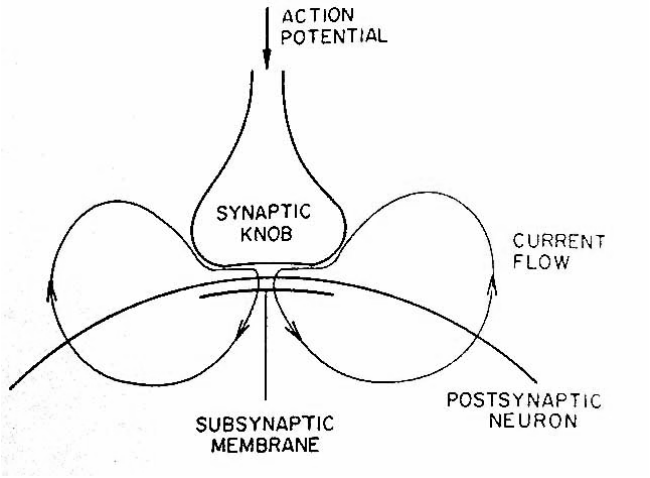


Figure 2.5: The current flow of a synaptic potential. Positive/negative ions flow through the local surface of the membrane. The current flows across the membrane, through the intracellular fluid, back across the membrane at more distant locations, and finally back to the synapse to complete a closed loop. (taken from [28]).

2.3.2 Action potentials

The action potential arises when a stimulus opens a few sodium channels. As a result, a net influx of sodium starts due to the concentration gradient of sodium and accelerates as the depolarization makes more sodium channels open. Eventually the increasing depolarization causes potassium channels to open while the sodium channels at this point closes. An outflow of potassium is caused by potassium's concentration gradient repolarizing the cell. Eventually the potassium channel closes and the resting potential is restored due to the sodium-potassium-pump, see also Figure 2.6.

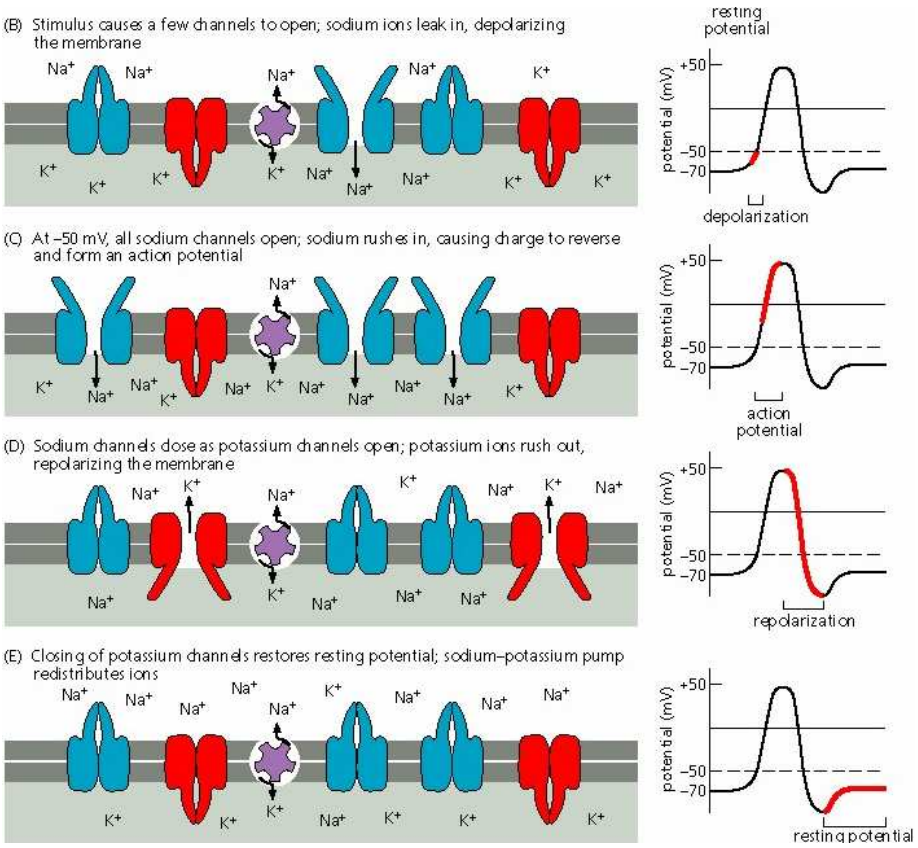


Figure 2.6: The generation of an action potential (adapted from [39])

From Figure 2.6 it is seen that the sodium currents are followed by opposite delayed potassium currents. Nonetheless, this potassium current is unable to counteract the potential generated from the sodium current as $\sigma_{Na^+} < \sigma_{K^+}$ both interior and exterior the cell as revealed on Figure 2.7.

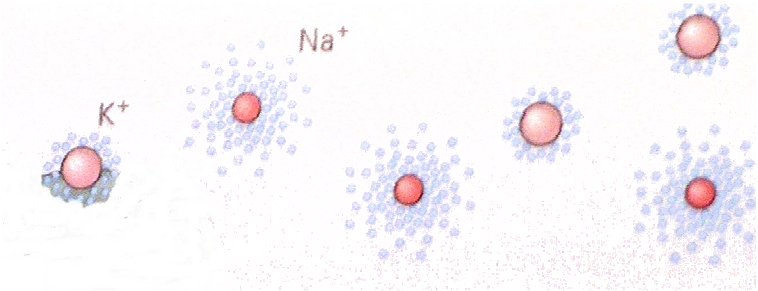


Figure 2.7: The smaller an ion is the more highly localized is its charge and the stronger its effective electric field. As a result, smaller ions attract water more strongly. Consequently, because of its larger water shell, Na^+ behaves as if it is larger than K^+ making it less mobile, i.e. giving Na^+ lower conductivity than K^+ [16]. (Figure adapted from [16]).

2.3.3 Synaptic Potentials versus Action Potentials

From eq. 2.18 it is seen that the dipole potential is linear to the distance between the current source and sink $\|\mathbf{d}\|$. As myelination of neurons dramatically increases the distance between the current sources and sinks, most EEG signals from action potentials is believed to originate from pyramidal cells of heavy myelinated axons [28]. However, as the propagation of current is much slower in unmyelinated neurons coherence is more easily achieved in the unmyelinated regions in favor of the signal originating mostly from the synaptic potentials from the dendritic parts of the pyramidal cell [11]. On the other hand, the dendrites in general aren't as well aligned as the axon's - compare in Figure 2.4 the Golgi stain emphasizing dendritic trees with the weigert stain emphasizing myelinated axonal fibers. However, the synaptic potential has duration between 5 ms. – 20 min. whereas the action potential only has duration of 1-10 ms. Consequently, coherence is more easily achieved for the synaptic potential. As a result the greatest contribution to the EEG is believed to be that of the synaptic potential [9],[16],[28].

Table 1: Processes of potential generation favoring/ disfavoring coherence.

Measured EEG	Action Potential	Synaptic Potential
Advantage	Myelinated axons →Large distance between current sources	Long duration (5 ms-20 min)→ Coherence easy Both excitatory and inhibitory effects.
Disadvantage	Short duration (1-10 ms.)→ Coherence difficult	Dendrites not so well aligned→ Coherence difficult

2.4 Features of the EEG and ERP

The EEG has been decomposed into a series of fixed broad spectral bands based, unfortunately, on history and discovery more than on a theoretical framework [24]. These bands are described in Figure 2.8. In general, the frequency of brain oscillations is negatively correlated with their amplitude [28],[30].

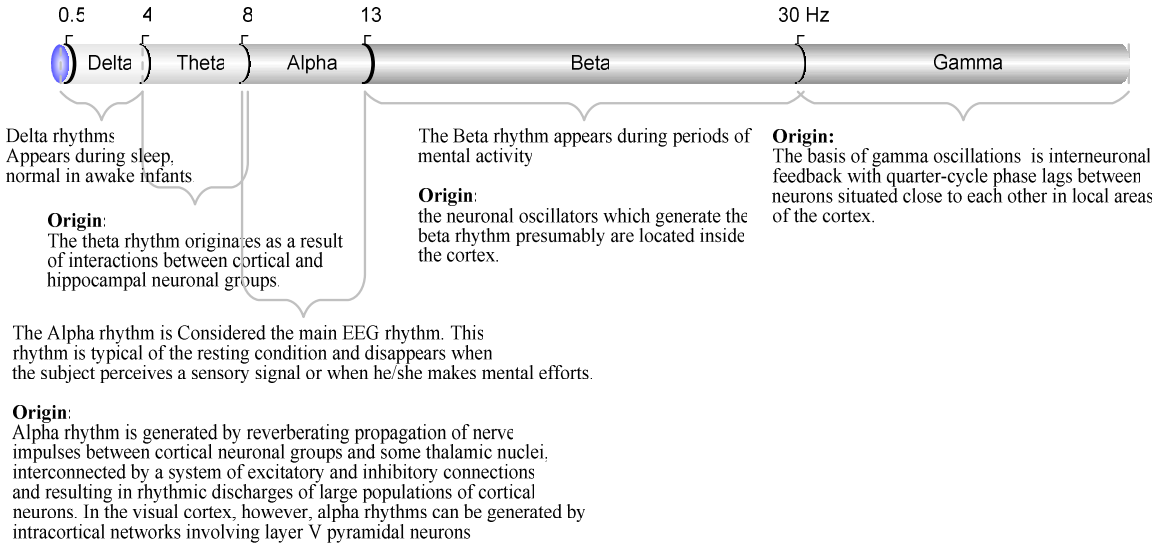


Figure 2.8: EEG rhythms and their believed origins (summary of [9],[16], [28], [34])

2.4.1 Event Related Potentials, ERP

Event Related Potentials (ERP) is the measured EEG signal timed to sensory stimuli. The EEG signal can be split into induced activity, evoked activity and noise. The induced and evoked activities are both believed to be generated from thalamic relay cells. An internally or externally paced event results not only in the 'evoked' generation of an evoked response potential (EP) but also in an 'induced' change in the ongoing EEG/MEG in form of an event-related desynchronization (ERD) or event-related synchronization (ERS) [30]. The EP represent the responses of cortical neurons due to changes in afferent activity, while ERD/ERS reflect changes in the activity of local interactions between main neurons and interneurons that control the frequency components of the ongoing EEG [30].

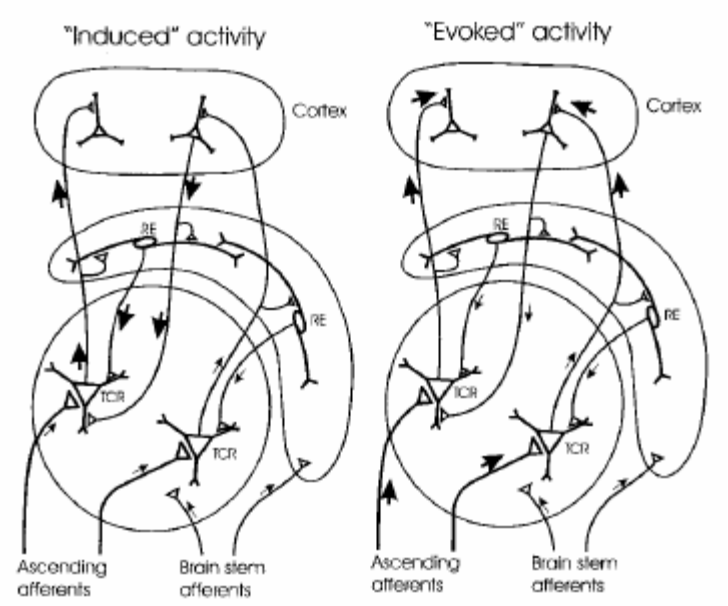


Figure 2.9: Schema for the generation of the highly frequency specific induced (ERD/ERS) and the evoked (EP) activity. TCR: thalamic relay cells, RE: reticular thalamic nucleus [30]. Where evoked activity is mainly controlled by the ascending afferents the induced is controlled by the reticular thalamic nucleus.

As revealed in Figure 2.9 of principal concern is the reticular activating system, a complex and diffuse system projecting from the brainstem to the cortex, which provides both inhibitory and excitatory inputs. This subsystem is itself under control from the cortex, as well as from collaterals of sensory pathways. The activating system is most crucially concerned with maintenance of the waking state, desynchronization of the EEG, direction of attention and governance of motivation [35] . Although EEG is believed primarily to stem from dipoles of synaptic potentials firing in synchrony mostly due to the reticular activation system, where and what exactly generates the signals remains unclear.

Evoked oscillations exhibit a strict phase-locking to the experimental event (e.g. stimulus presentation) across trials. Hence, they can be extracted from the averaged ERP, e.g. by filtering or by the ITPC of a wavelet analysis. On the other hand induced oscillations are (by definition) not at all phase-coupled to a stimulus, and show a certain degree of phase-jittering. Therefore, by averaging across trials these oscillations will cancel out completely and hence are only detectable by appropriate ways of analysis, e.g. by a single trial based wavelet analysis with subsequent averaging [6], such as the ERSP.

The components of the ERP are labeled by latency and polarity. A positive component at 100 ms is called ‘P100’ and a negative deflection at 200 ms is a ‘N200’, see Figure 2.10. Furthermore, the event related potentials that are determined by physical aspects of the stimulus are labeled, ‘exogenous’ whereas higher order processing are labeled ‘cognitive’.

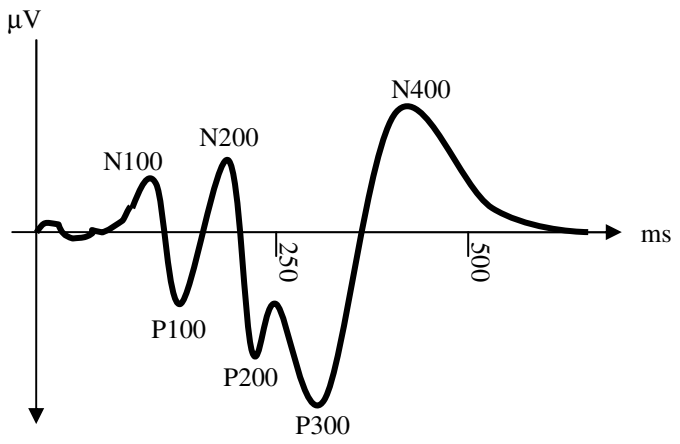


Figure 2.10: Typical significant components of the ERP

(notice: negativity is due to historically reasons directed upwards so that current flowing out of the scalp is up)

Many ERP experiments are based on the oddball paradigm: Two stimuli are presented, one of which is an infrequent target. When the target is discriminated from the other stimuli with attention, the P300 is of greater amplitude [9]. This has made especially the P300 a very interesting component of the ERP. The various most common ERP components are described in Table 2. The table is only a guideline. Experts do not agree on how to interpret all components and the components are difficult to generalize over different experimental paradigms.

Table 2: Features of ERP components

	Spike	Description
Exogenous process	P100	**Generation**
	N100	Appear to be generated in the primary receiving areas of the brain. **Identification** Both have been found to occur 100 ms after presentation of a visual stimulus. Found in the receiving areas of the brain.
	N200	**Generation** Increases in amplitude to task deviant stimuli. **Identification** Located Posterior
	P200	**Generation** Increases to novel stimuli
Cognitive process	P300b	**Generation** Attention dependent. Monotonous inverse relationship between amplitude and stimulus probability. Negatively correlated with speed of information processing as indexed by reaction times – the faster speed of processing the earlier latency. The latency increases with the time the subject needs to distinguish the rare stimulus. The amplitude increases with rarity of the stimulus and to some extent with stimulus intensity. The amplitude of the P300 has been shown to be inversely proportional to stimulus presentation probability and directly to task complexity. P300 amplitude has been shown to be inversely related to a prior probability and is influenced by the sequence of immediately preceding events, i.e., sequential event structure. Unrelated to the specific sensory areas but probably related to the parietotemporal association cortex and subcortical structures such as hippocampus and thalamus. Occurs to infrequent non-target stimuli **Identification:** In normal young adults a positive wave over the Centro-parietal scalp is seen. Can occur at any point between 280-800 ms. Posterior scalp distribution with maximum at Pz, P300 latency is age dependent, being longer in children, progressively decreasing until 18 before increasing by 1.25 ms per year In adult. Aging increases the latency, decrease the amplitude and cause forward shift in the distribution of P300
	P300a	**Generation** Mostly concerned with novel stimuli. The more the stimulus is known the more P300a approaches P300b. Its features are Similar to P300b. **Identification** Shorter latency (≈250 ms) than P300b. Situated more frontal-central than P300b and habituates rapidly. In different modalities it has also been reported as Centro-parietal.
	N400	**Generation** Found in numerous studies that have employed a task in which items were presented sequentially and subjects were asked to respond if the stimulus was unmatched (incongruent) or matched (congruent) with the preceding items. It has been proposed that it represent the associated activation of neural networks basic to stimulus integration. **Identification** Located in frontal area, and is larger in non-matched items than matched items.

2.4.2 Noise

EEG potentials are measured as the difference between two points, one on the scalp where EEG effects are strong and one (the reference electrode) hopefully isolated from these effects. Some commonly used reference sites are Cz, earlobes, mastoids (bone right behind ear), tip of nose and average reference ("reference free"). Earlobes or mastoids are generally linked either physically or mathematically in order to maintain symmetry. The average reference uses the constraint that the sum of the potentials over a spherical surface is zero and requires fairly high density recording (~128 channels). It can be improved by estimating potentials for the inferior spherical area. Spherical spline interpolation is sometimes used for these estimates. The 10-20 international system is a standardized system to place the electrodes. It relies on taking measurements between certain fixed points on the head. The name 10-20 refers to the fact that the electrodes used to be placed at points 10% or 20% of these distances [40]. Today other fractions are used, but the name 10-20 has been kept. How the recorded signal is referred can have crucial impact on the noise in the EEG as a very noise full reference can mess up the signal of all recorded channels.

Many sources of noise disturb the EEG/ERP signal. Of primary concern is muscular action due to eye movement, hearth beat etc. Furthermore, noise from electronic devices can have a great impact especially in the 50 Hz range where most electronic devices, at least in Europe, operate. Finally, volume conductance is another huge problem when dealing with the EEG/ERP signal. Both heartbeat and eye movement can be reduced by correcting the signals from recording sites by the eye (EOG) and heart (ECG). Instead of specific recordings from these sights, recordings of the influence of eye and cardiac activity can also be identified prior to the ERP experiment. As previously mentioned, volume conductance can be reduced by taking the surface Laplacian.

2.4.3 EEG/ERP and PARAFAC

Extracting the correct features of the EEG is of crucial importance. Well used methods of analyzing EEG data are; Wavelet Analysis, Neural Networks Analysis, Blind Deconvolution and Source Separation Methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA).

In this thesis, an approach described by Martinez-Montes et al. [24] and Miwakeichi et al. [25] where sources are separated using parallel factor analysis, PARAFAC, will be used. Although Harshmann in his original paper in 1970 [13] suggested the use of the PARAFAC model on EEG the use has been very limited. In 1988 Möcks [26] and in 1991 Field [10] used the model on the ERP to decompose the space-time-subject. In 1985 Cole et al.[7] used it on the ongoing EEG in a way similar to that of Miwakeichi et al.

Martinez-Montes and Miwakeichi et al. used the PARAFAC model to extract features of the ongoing EEG. They proved that PARAFAC was capable of successfully identifying the theta and alpha atoms of a cognitive task and showed furthermore the algorithms ability to identify eye blinks. In this thesis the PARAFAC decomposition will instead be applied to the ERP. The PARAFAC model seems plausible as the EEG/ERP-signals in several ways can be considered multi-way arrays as seen on Figure 2.11.

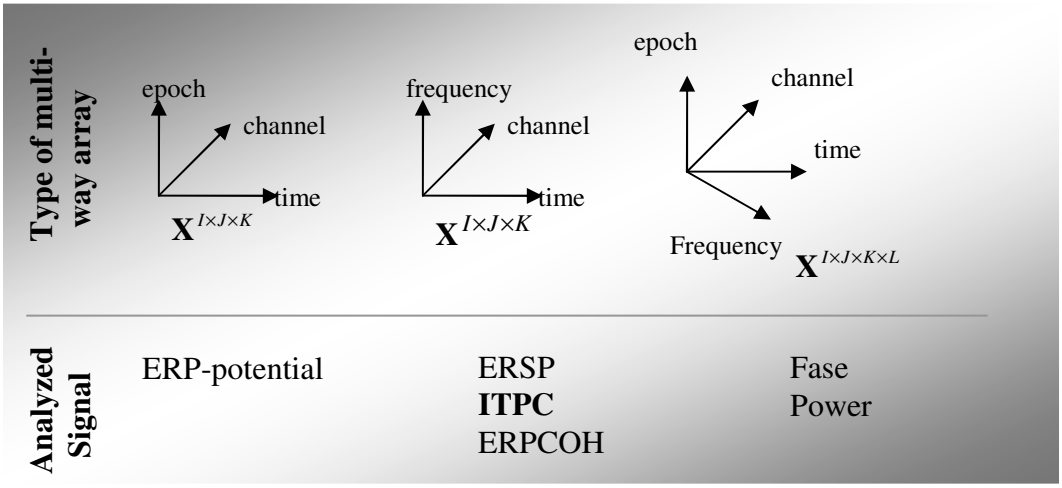


Figure 2.11: Different forms of multi-way arrays arising from the EEG.

More modalities could easily be added to Figure 2.11 denoting for example the analyzed subjects or the various conditions under which the data has been recorded. The PARAFAC analysis of the epoch averaged ERP-potential given by $channel \times time \times subject$ and the ITPC is in the following of great interest.

From eq. 1.29 we had $x_{ijk} = \sum_{\lambda=1}^{F} a_{i\lambda} b_{j\lambda} s_{k\lambda}$, in terms of the ITPC; **a** denotes the component pertaining to the topography, i.e. channels, **b** the frequency component and

s the component describing the temporal development. The model seems plausible as the ERP is believed to be a non-stationary process requiring a change in time of the factor proportions, which by the model is insured by **s**. eq. 1.29 can be restated as $x_{il} = \sum_{\lambda=1}^{F} a_{i\lambda} c_{l\lambda}$, where $c_{l\lambda} = b_{j\lambda} s_{k\lambda}$ and $L = J \cdot K$. This corresponds to the normal two-dimensional factor analysis model. Performing ICA on this model makes the assumption that the combined frequency-time components $d_{j\lambda} s_{k\lambda}$ are mutually independent, i.e. $CI_{2,3}$. The goal in the analysis of the ITPC is both to separate the multi-way array into factors that relate to different time-frequency components favoring the ICAPARAFAC algorithm but also to do data exploration favoring ALSPARAFAC as it keeps the most of the variation.

3 DATA Analysis

*"Absence of evidence is **not** evidence of absence!"*

-unknown

3.1 Simulated data

To evaluate the ability of the PARAFAC algorithms to find the components of real data, the developed methods were tested on simulated data. A 32 channel EEG sampled at 500 Hz was generated and added 50 Hz oscillations of amplitude 0.8 on all channels mimicking electronic noise. Two burst of 35 Hz sinusoidal oscillations with an amplitude of 1.0 were placed in channel 30,31 and 32 at the posterior areas resembling occipital gamma activity while one burst of 25 Hz oscillation with amplitude 1.5 were generated simultaneously at each ear at channel 11 and 15. Finally, normal distributed random noise of power 1.0 was added to all channels. The data was transformed using a complex Morlet wavelet with bandwidth parameter 2 and center frequency 1, and the power of this wavelet transformed signal analyzed. The three PARAFAC factors shown in Figure 3.3 were expected to be found from the data. On Figure 3.1 the simulated EEG data is revealed and the corresponding power of the wavelet transform is seen on Figure 3.2.

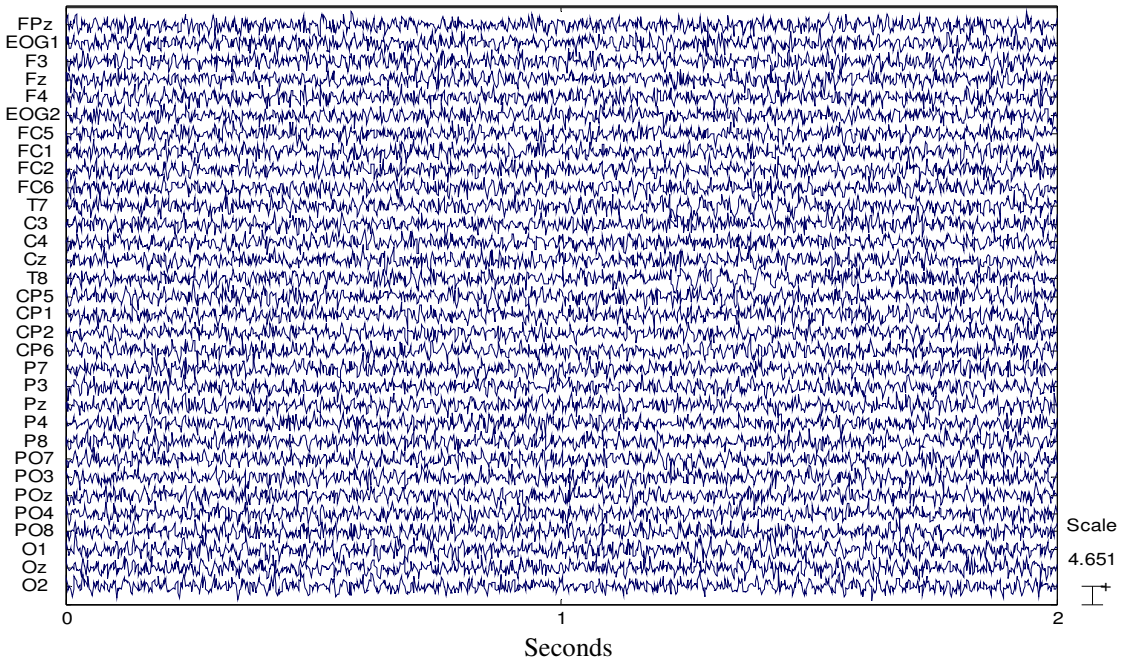


Figure 3.1: The simulated EEG-data.

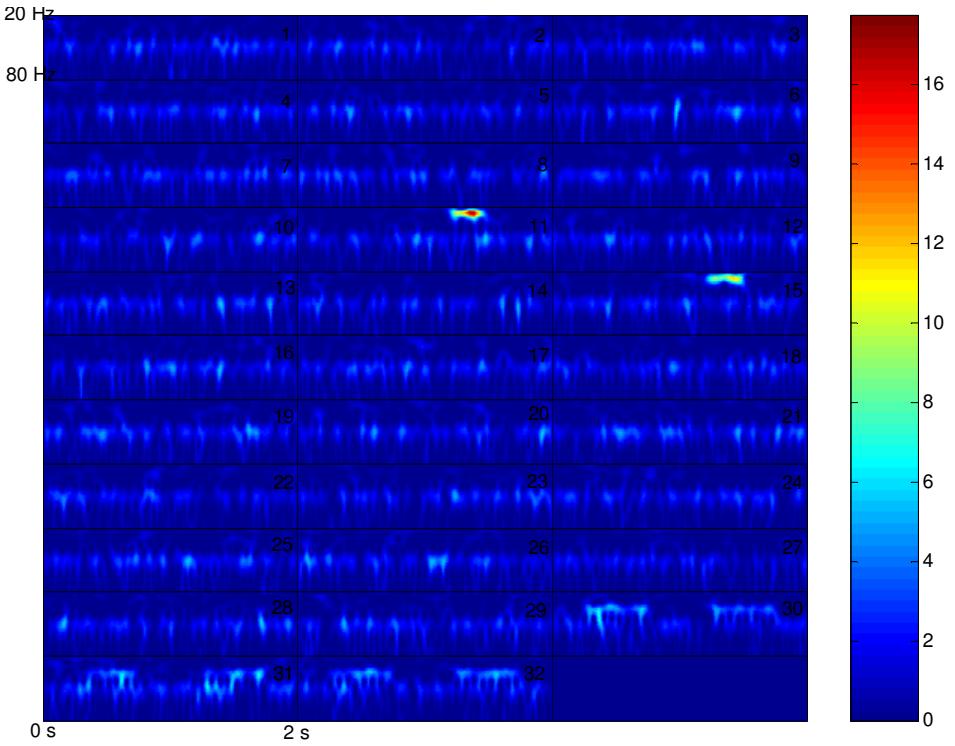


Figure 3.2: The power of the complex Morlet wavelet transform on each of the 32 channels of the simulated data.

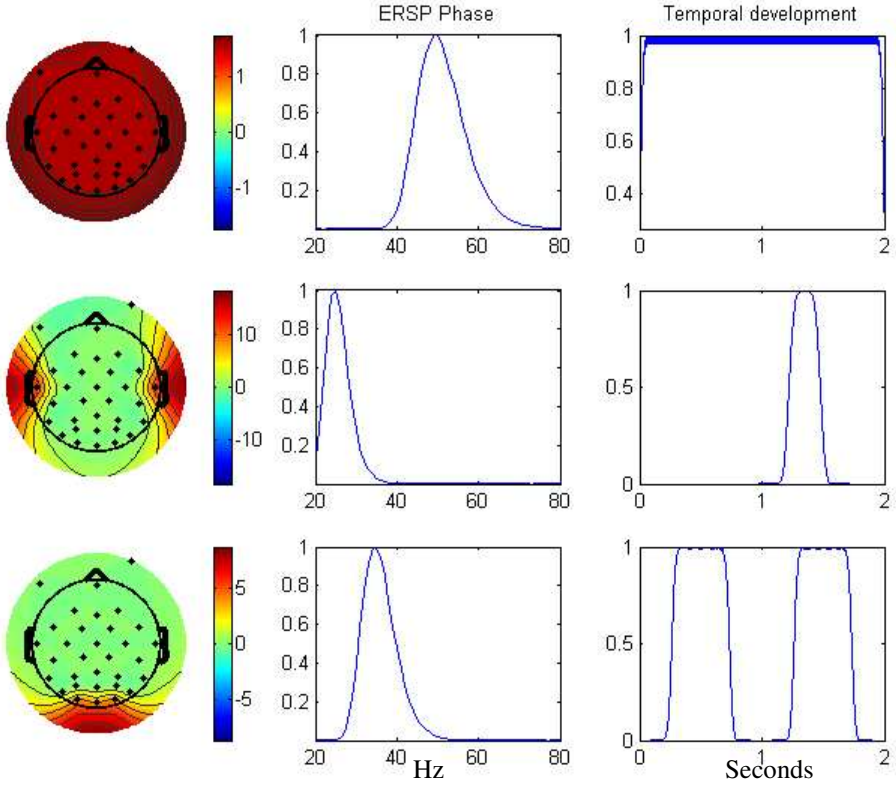


Figure 3.3: The true factors of the simulated data.

The raw simulated data was first analyzed using Independent Component analysis by the 'runica' standard method in EEGLAB [8]. As revealed in Figure 3.4 none of the independent components solely captures any of the underlying three factors. Especially the same 50 Hz oscillation present in all channels was split into individual components. Consequently, the independent component analysis did not seem efficient in accessing the various factors present in the data. Furthermore, no clear indication of the time points at which the factors were present was given by the ICA-decomposition as it is irresolvable from the EEG of the components. Thus, without any frequency information the ICA algorithm was unable to identify the factors.

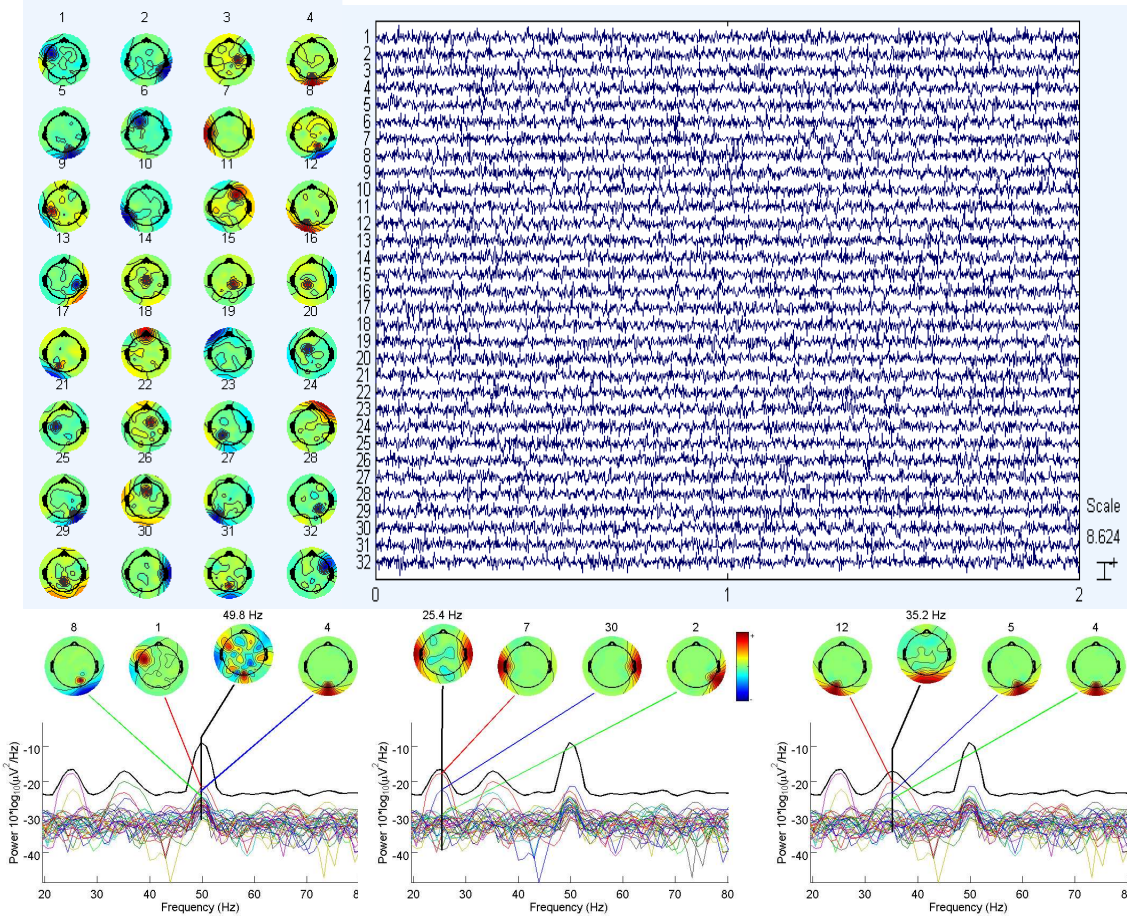


Figure 3.4: Top panel; the component map and time series of all 32 independent components. Lower panel; the maps of the three independent components contributing the most at 50 Hz, 25 Hz and 35 Hz to the specter of the EEG along with the summed map of the three components. Clearly, the ICA decomposition hasn't been able to identify the true components of the data.

The developed PARAFAC models were then tested in their ability to access the components. In the ICAPARAFAC model $CI_{2,3}$ was assumed, i.e. a combination of the time and frequency dimensions were thought independent. As non-negative solutions were desired, a non-negative matrix factorization (NMF) was compared to a decomposition based on SVD for the rank one decomposition. The NMF was optimized in a least square sense as described in [22]. Although, non-negative ICA algorithms would

be more correct to use due to the non-negative nature of the data, an ICA algorithm based on maximum likelihood described in [18] was used as it gave approximately non-negative results. For the derivation of the Bayesian Information Criterions used consult Theorem 6 and Theorem 7 page 101-102. The number of observation in the BIC measures was defined as the number of time points in the data. Furthermore, BIC was normalized by the number of observations. Although only the ALSPARAFAC corresponded to a least-square optimization the SR1PARAFAC and ICAPARAFAC algorithms also used the BIC given for a least square solution. This was done since the factors found of SR1PARAFAC and ICAPARAFAC were believed to be close to a pure least square solution.

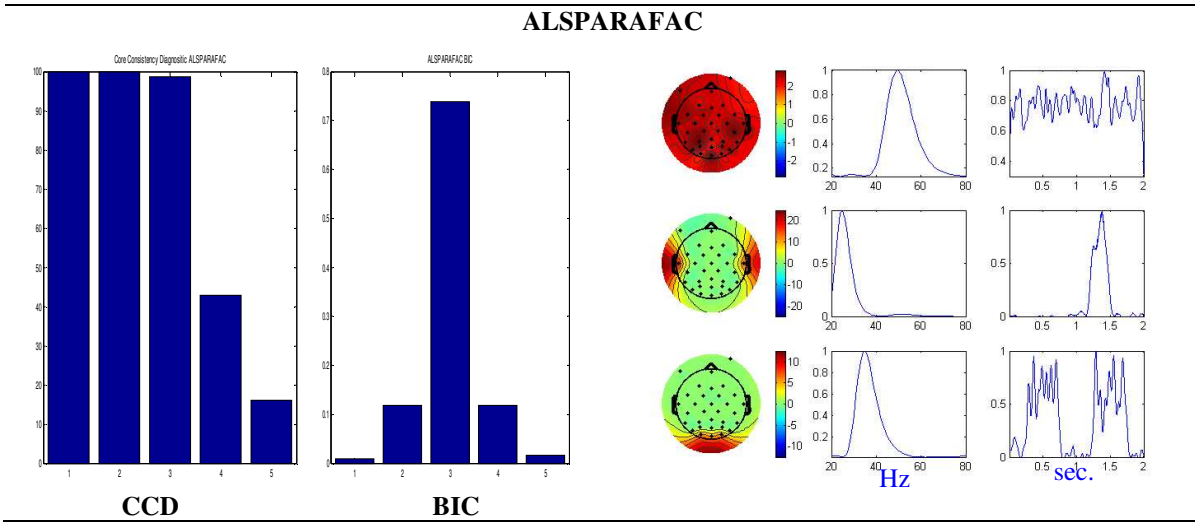


Figure 3.5: To the left; the determination of the number of factors present using ALSPARAFAC, given by the Core Consistency Diagnostic, CCD and BIC. Both the CCD and BIC clearly indicate a three component model. To the right; the factors found when fitting a three component model. Obviously, ALSPARAFAC has positively identified all three factors.

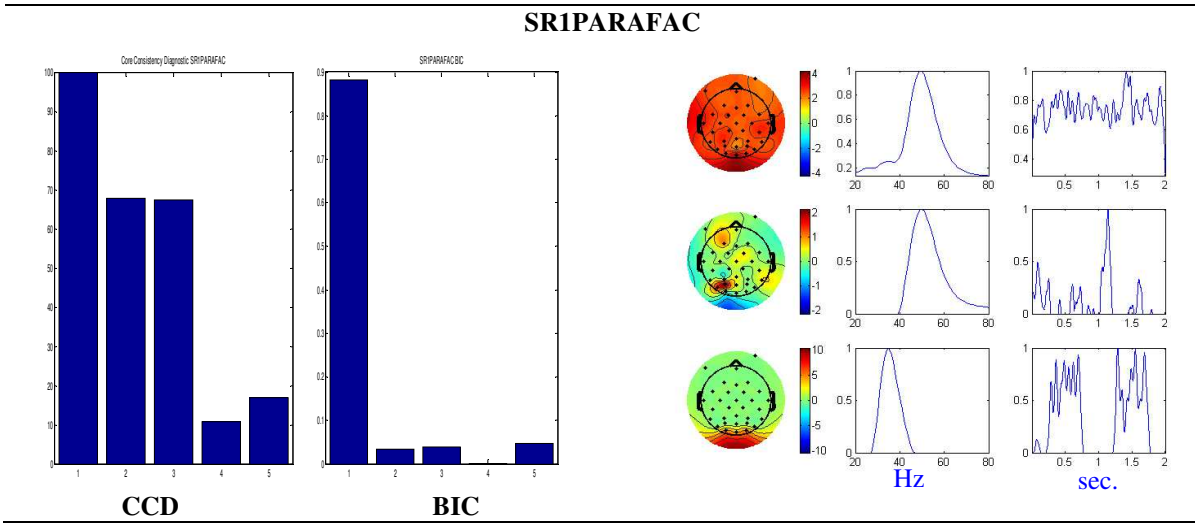


Figure 3.6: To the left; the determination of the number of factors present using the SR1PARAFAC, given by CCD and BIC. The CCD uncertainly indicates one to three components present whereas BIC give strong indication of a one component model. To the right; the factors found when fitting a three component model. The SR1PARAFAC only correctly identifies two of the three factors.

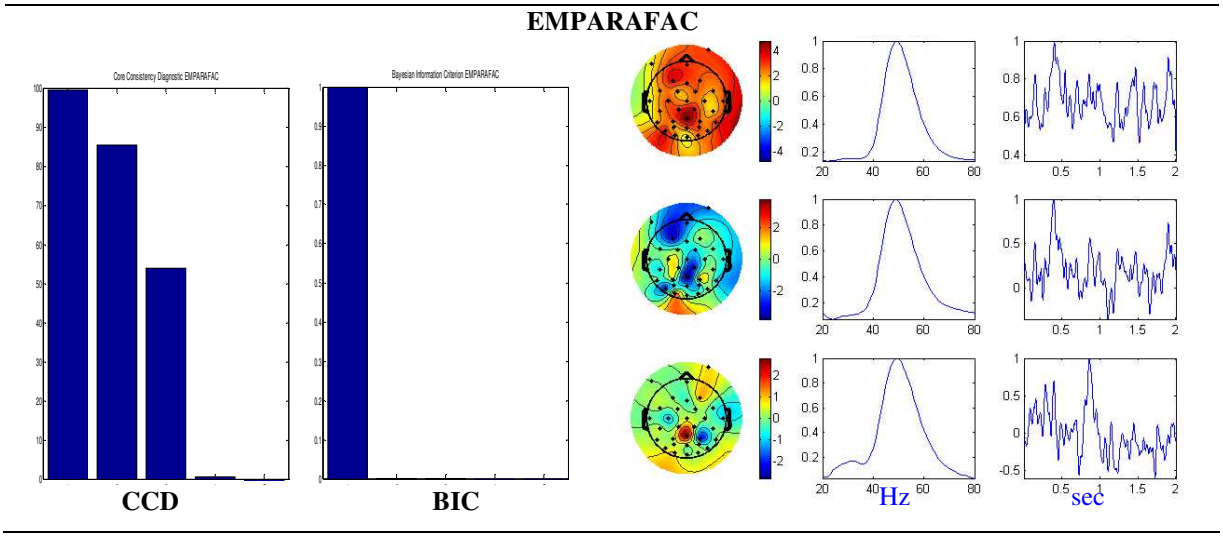


Figure 3.7: To the left; the determination of the number of factors present using EMPARAFAC, given by CCD and BIC. The CCD indicate a model having two factors whereas BIC gives sign of only one factor. To the right; the factors found when fitting a three component model. The EMPARAFAC only identify as indicated by BIC one component - the 50 Hz noise present in all channels.

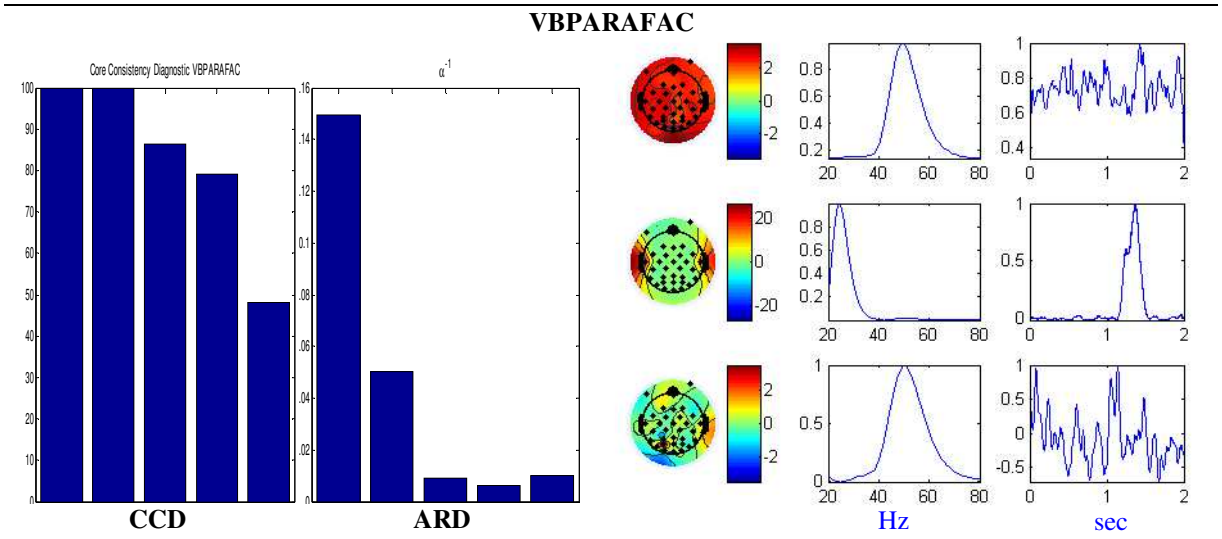


Figure 3.8: To the left; the determination of the number of factors present using VBPARAFAC, given by the CCD and ARD. The CCD is very unclear but indicate that up to four factors are present. The ARD however only reveal that one or two factors are present. To the right; the factors found when fitting a three component model. The VBPARAFAC correctly identifies as indicated by the ARD two components - the 50 Hz noise present in all channels and the 25 Hz ear activity. It is however unable to find the occipital activity[2].

[2] The VBPARAFAC ran for 10,000 iterations as suggested by [27], priors were set to be non-informative. However, for both VBPARAFAC and EMPARAFAC it was difficult to determine whether the algorithms had converged.

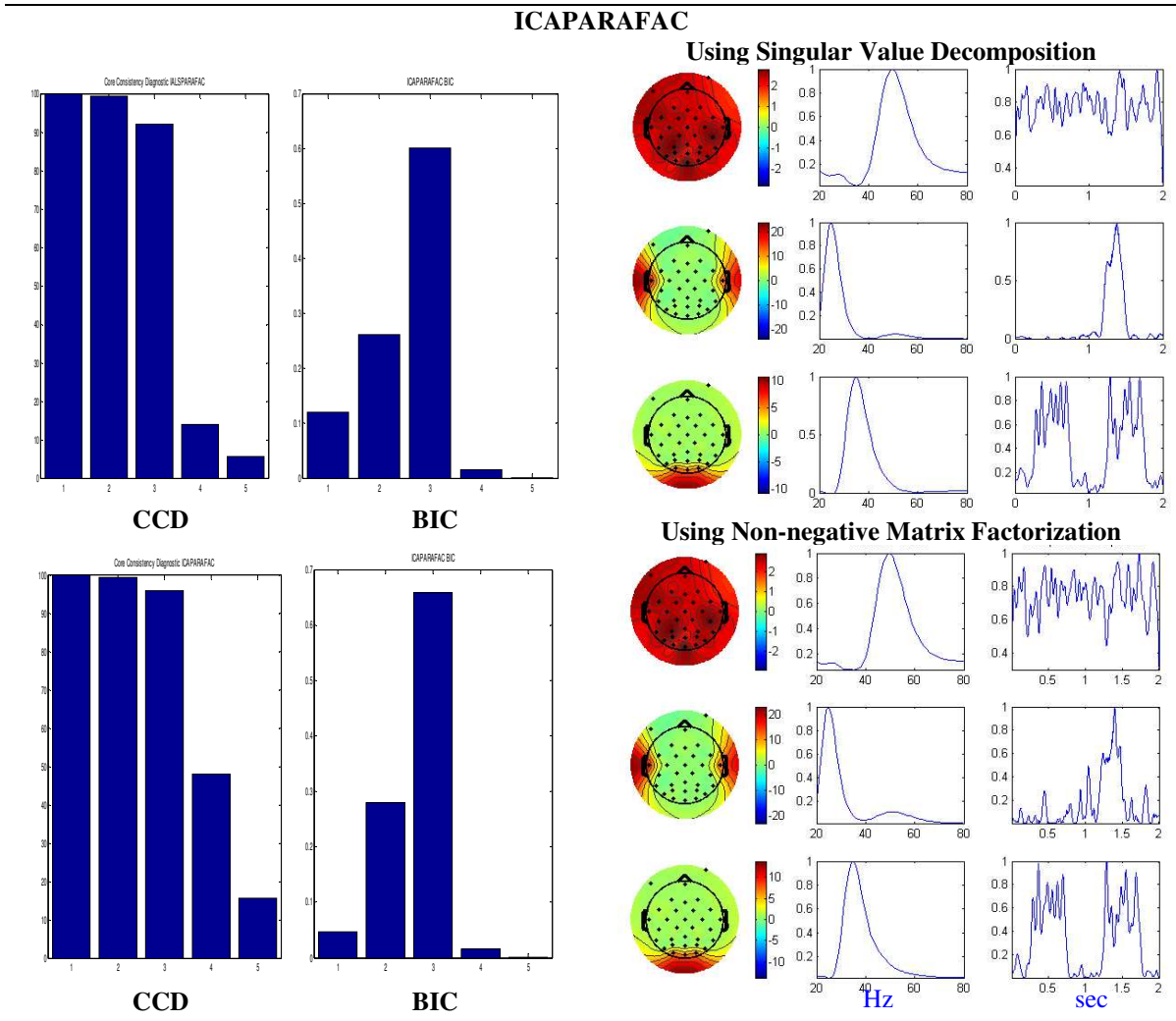


Figure 3.9: To the left; the determination of the number of factors present using ICAPARAFAC, given by CCD and BIC for an ICAPARAFAC algorithm using SVD and one implemented with NMF. The CCD and BIC of both algorithms clearly indicate a three factor model. Both methods are also able to correctly identify the three factors. The frequencies and temporal signatures are however slightly different from each other.

As seen on Figure 3.5 the Core Consistency Diagnostic clearly indicates that three factors are present in the ALSPARAFAC, this is confirmed by the Bayesian Information Criterion. The method is also able to correctly identify all three factors. A much weaker indication of a three component model is given by the CCD for the SR1PARAFAC, see Figure 3.6. The BIC for SR1PARAFAC indicate however that only one factor is present. The SR1PARAFAC is able to correctly identify two components, the 50 Hz activity in all channels and the 25 Hz ear activity, but the 35 Hz occipital activity is lost. The EMPARAFAC algorithm as seen on Figure 3.7 is only able to identify the 50 Hz activity in all channels, from BIC it is also seen that only one factor is indicated to be present in the data. From the automatic relevance determination (ARD) of the VBPARAFAC on Figure 3.8, it is seen that one to two factors are found to be present in the data whereas the CCD is very unclear but indicate that up to four factors are present. The

VBPARAFAC method correctly finds the 50 Hz and 25 Hz activity. Finally, the ICAPARAFAC based on SVD and NMF both clearly indicate from the CCD and BIC as seen on Figure 3.9 that three factors are present in the data. Both methods also correctly identify all three factors. From the results of the ICAPARAFAC algorithm using SVD or NMF didn't change the CCD or BIC. However, the temporal signatures as well as the frequency signatures were slightly altered. Notice how the SVD solution is very close to the ALSPARAFAC solution.

From the simulated data only the ALSPARAFAC algorithm and ICAPARAFAC algorithm successfully identified all the factors. For these two methods the CCD and BIC both worked well, as they strongly indicated three factors were present. The two algorithms were then compared in their ability at different noise level to identify the 25 Hz activity at the ears and 35 Hz activity at the occipital region. For each level of noise fifty ALSPARAFAC and ICAPARAFAC models were fitted to the data. The ICAPARAFAC was based on the Non-negative Matrix Factorization. Both algorithms were evaluated by how much their found factors correlated with the true underlying factors. The correlation was calculated as the average correlation taken over each of the three factor-components, i.e. as the average correlation of the topographic, frequency and temporal signatures between the real and found factors.

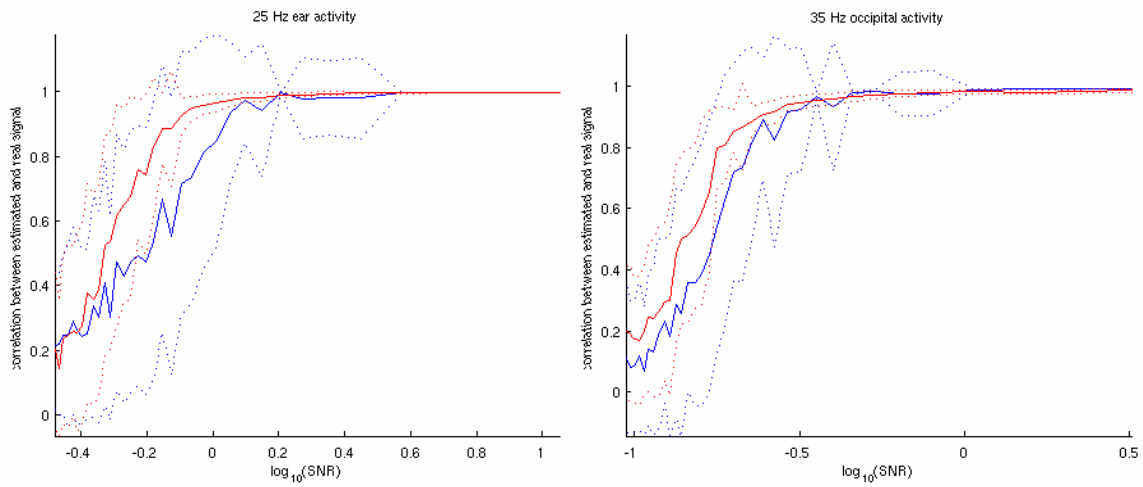


Figure 3.10: The correlation between the true and estimated factors for different signal to noise ratios (SNR). Blue corresponds to ALSPARAFAC, red to ICAPARAFAC. Dashed lines correspond to one standard deviation from the solid lines. Clearly, the ICAPARAFAC is better at finding the true components and more stable than the ALSPARAFAC method as the SNR drops.

From Figure 3.10 it is seen that the ICAPARAFAC algorithm is better at estimating both the 25 Hz ear and 35 Hz occipital activity as the signal to noise ratio drops. Both methods have more problems finding the ear activity than the occipital activity when the signal to noise ratio decreases. Whereas both methods correctly identified the occipital activity down to a SNR=$10^{-0.5}$=0.32, already around a SNR=$10^{0.1}$=1.26 the ALSPARAFAC methods have problems finding the ear activity. This stems from the fact that the occipital activity is present longer and in more channels than the ear activity making it easier to detect. Furthermore, ICAPARAFAC is more stable than ALSPARAFAC as the standard deviation of ICAPARAFAC is smaller than that of ALSPARAFAC and ALSPARAFAC

begins to be unstable earlier around a SNR=$10^{0.5}$=3.16 for the ear activity and at SNR=10^{0}=1 for the occipital activity.

Finally, the ICAPARAFAC method was compared to the ALSPARAFAC on the chemometric data set "Claus" described in [41]. The analysis is shown in Appendix D: ICA- and ALSPARAFAC on Chemometric Data. Also on this dataset ICAPARAFAC performed well.

3.2 Real data

The real data is generated from an experiment by Herrmann et al. [15] regarding gamma oscillations in the visual system. Gamma oscillations have been shown to correlate with perceptual binding, attention, arousal, object recognition and language perception. A mechanism which underlies many of the above mentioned cognitive functions is the match of sensory information with memory contents. Herrmann and colleagues argue that the so-called 'early' gamma-band activity, occurring in EEG before 150 ms after stimulus presentation, reflects a match with memory. In addition, they argue that 'late' gamma activity, which typically emerges with a latency of more than 200 ms, is a temporal signature of utilization processes such as response selection or context updating [14]. This has lead to the Match and Utilization Model, MUM, explained in Figure 3.11.

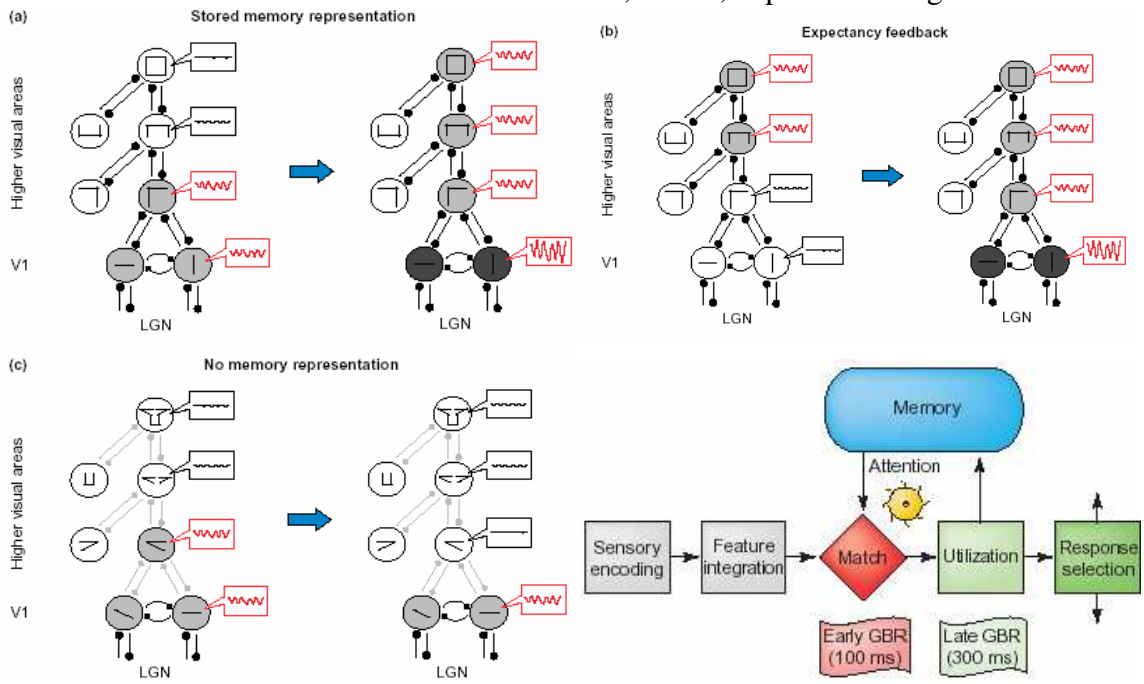


Figure 3.11: Black connections represents memory connection, gray the lack of memory connection. According to Herrmann et al. a stored memory representation will result in an enhanced gamma oscillations and synchrony as the features are matched with the memory content. Furthermore, as revealed in b the expectation of a known visual stimuli can result in enhanced gamma oscillations and synchrony as the neurons expecting the visual feature are closer to threshold. As revealed in c, when no memory representation is present, no enhancement takes place. This concept has been expanded to the Match and Utilization model, MUM, (lower right figure): Sensory coding is integrated into features, and these features are matched with memory contents around 100 ms. At around 300 ms a process denoted utilization takes place. Here updating of memory contents, selection of different behavioural responses and the reallocation of attention is believed to take place. Whereas the match at 100 ms is evoked the utilization at 300 ms is believed to be induced. Figures adapted from [14].

Herrmann et al. find that under visual stimulation a strong increase in evoked oscillations near 40 Hz over posterior areas with a latency of approximately 100 ms and a later increase in induced activity with a latency around 300 ms can be observed [6],[15]. As evoked activity is phase locked to the stimuli, the ITPC will be analyzed. Coherence in the posterior regions is expected to be found. In the following, coherence is defined as the ITPC.

Eleven healthy subjects with mean age 25.7±1.7 years participated in the experiment. All subjects had normal or corrected to normal vision. The experiment was done by Sidse Arnfred at Cognitive Research Unit, Department of Psychiatry, Hvidovre Hospital. The subjects were asked to classify objects as round or edgy by right or left clicking a computer mouse. Some of the objects had a long term memory representation (object) whereas other objects consisted of the same atoms but randomly placed not to make sense (non-objects), see also Figure 3.12. To insure the subjects were naïve to the experiment the task of classifying the objects as edgy or round was given even though no such clear interpretation of the objects was always present.

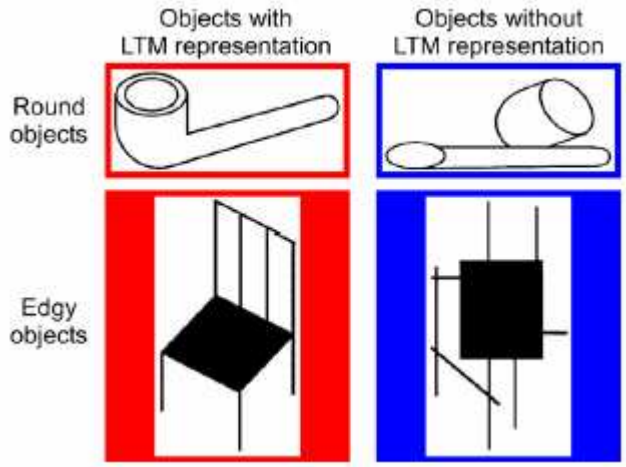


Figure 3.12: Example of stimuli with long term memory representation (object) and without (non-object) , taken from [15].

The subjects were recorded using a BIOSEMI 64 channel active electrode system, see also http://www.biosemi.com/active_electrode.htm. The EEG was referenced to the average of two channels placed at each ear, i.e. channel 65 and 66. Data was sampled at 512 Hz. The epochs were extracted from the data taking -250 to 1000 ms. from stimuli onset. Baseline activity from -250 to -100 ms. was subtracted each epoch. A total of between 102 and 105 epochs were present in both the object and non-object condition for each subject. A complex Morlet wavelet with center frequency 1 and bandwidth parameter 2 was used. Although Herrmann et al. suggest removing epochs having standard deviations more than 50 μV [15], we compared this rejection criterion with an extensive rejection analysis of the epochs in EEGLAB using independent component analysis as suggested by Makeig et al. [8]. However, we realized that since we were analyzing the ITPC, the number of epochs used was more important than the quality of

each epoch. Although some epochs were very noisy they still had the correct phase. Since more epochs reduced the noise as averages could be taken over more trials, see also Figure 3.13, we ended up accepting all epochs in the data.

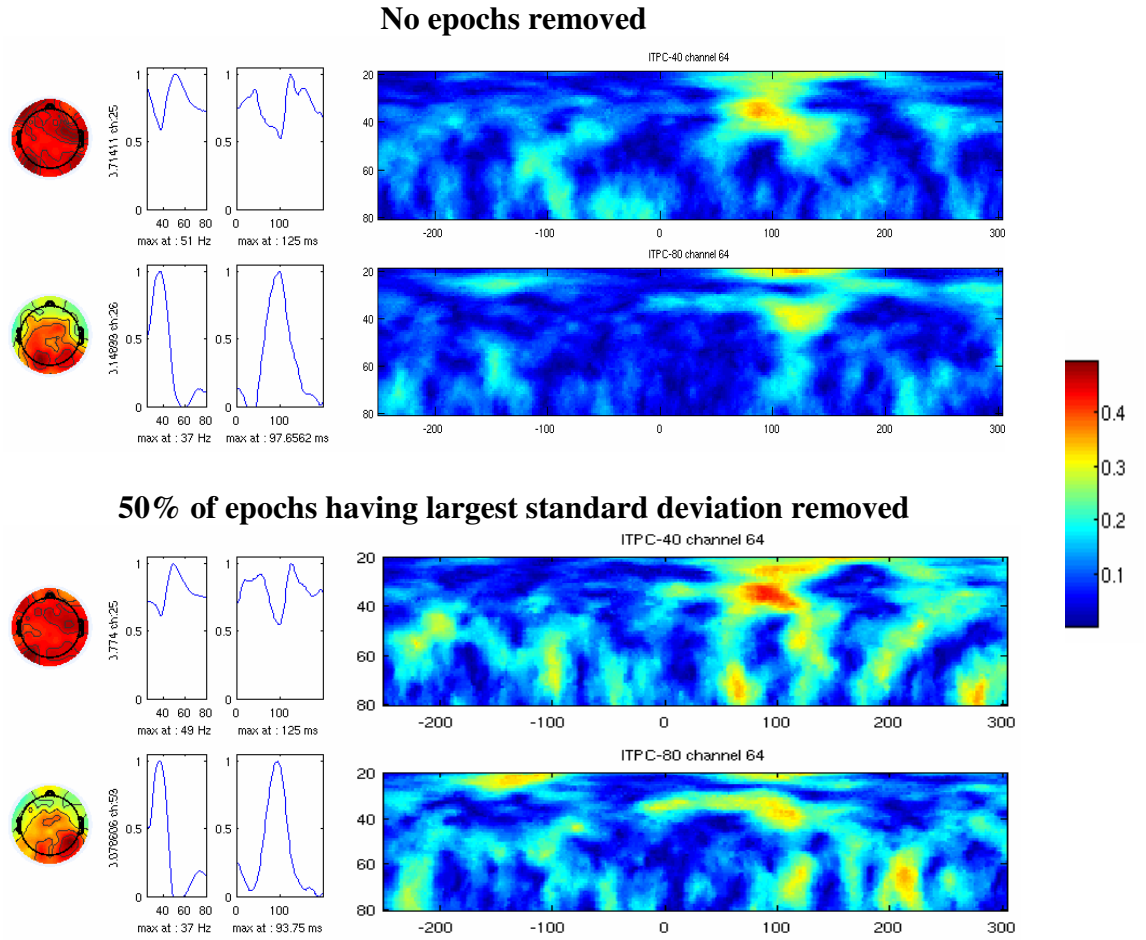


Figure 3.13: Left panel; example of a two component ALSPARAFAC analysis of the ITPC performed on all epochs of a subject and where 50% of the epochs having largest standard deviation within a 200 ms timeframe were removed. Activity at the posterior region is evident from the topographic image with all epochs whereas the removal of 50% of the epochs dramatically removes the coherence in the left occipital region. Right panel; the ITPC found in the object (40) and non-object (80) condition. Clearly the ITPC is less noise full when using all epochs; see top images, compared to removing 50%; bottom images (color scale given to the right).

Finally, the wavelet chosen also to some extent impacted the coherence found as revealed on Figure 3.14.

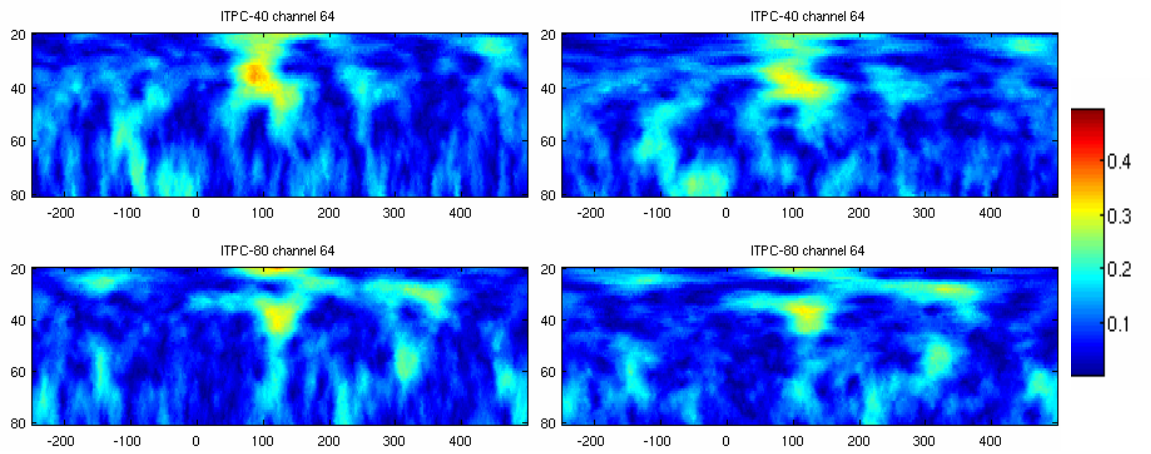


Figure 3.14: Taking the wavelet transform of the data using a complex Morlet wavelet having center frequency 1 and bandwidth parameter 2 (left figure) and bandwidth parameter 4 (right figure). The two wavelet transforms yield slightly different results (x-axis in ms, y-axis in Hz).

Prior to analyzing the data using PARAFAC, the data was analyzed similar to the way Herrmann et al. analyzed their data [15].

Analysis by Herrmann and colleagues

In their analysis, Herrmann and colleagues find the time and frequency corresponding to the coherence peak at channel 64 (equivalent to O2, placed at the center of the right hemispheres occipital lobe). However, as the whole occipital region is affected, we also analyzed the mean of the occipital region corresponding to channel 20-31 and 57-64.

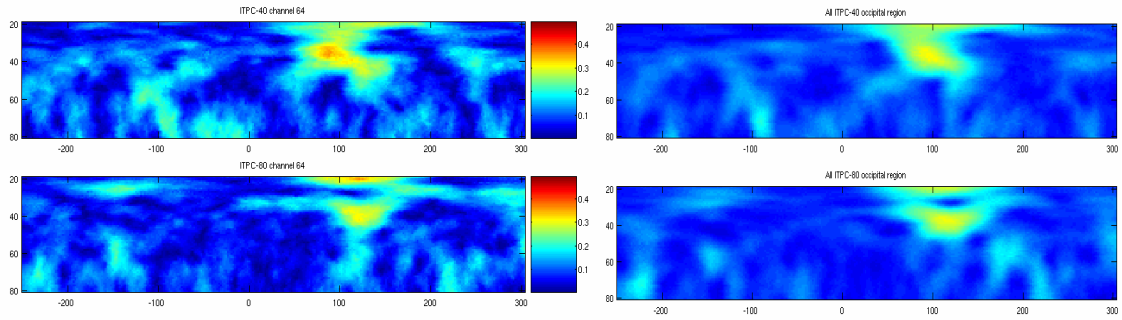


Figure 3.15: Left panel; an example of a subjects ITPC of channel 64 for the object condition top image and non-object condition bottom image. Right panel; same figure, but the average of the whole occipital region. Both panels clearly reveal gamma activity around 100 ms (x-axis in ms, y-axis in Hz).

As revealed on Figure 3.15 there is a strong coherence at around 37 Hz and 100 ms. Furthermore, the coherence seems stronger in the object situation than the non-object. Herrmann et al. find the time and frequency corresponding to each subject's peak in the

object situation and compares this coherence value with the corresponding value at same time-frequency for the non-object condition.

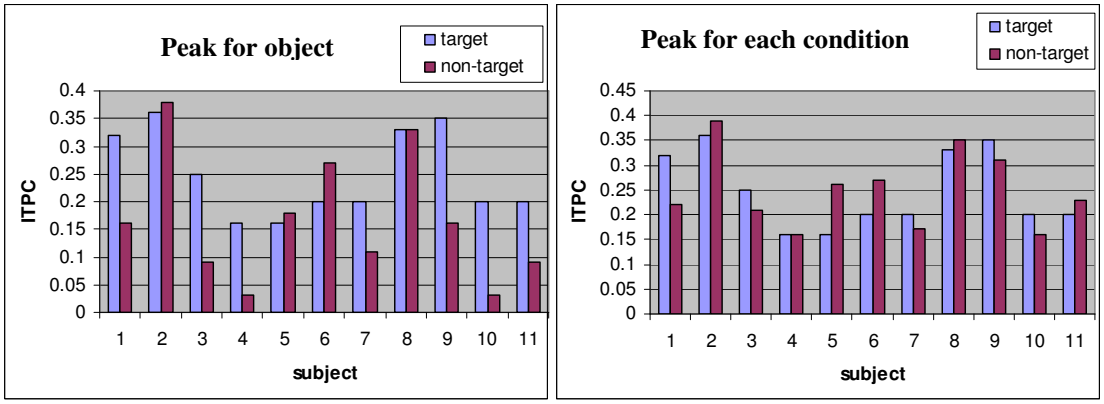


Figure 3.16: The coherence values for object and non object where object peaks and the coherence values for object and non-object where each condition has its peaks. No significant difference is in the two situations found between the conditions (target=object, non-target=non-object).

As seen on Figure 3.16 although the object condition results in higher coherence values in most situations when comparing the coherence at the peak of the object condition, it is not significant as Herrmann et al. find it to be. As using the peak of the object condition favors object we also compared the coherence at the peak of object with that of the non-object condition. Here a difference in degree of coherence seemed very random. Consequently, the finding of Herrmann et al. that the object condition is more coherent than the non-object condition seems very questionable. The same analysis performed on the whole occipital region yielded similar results.

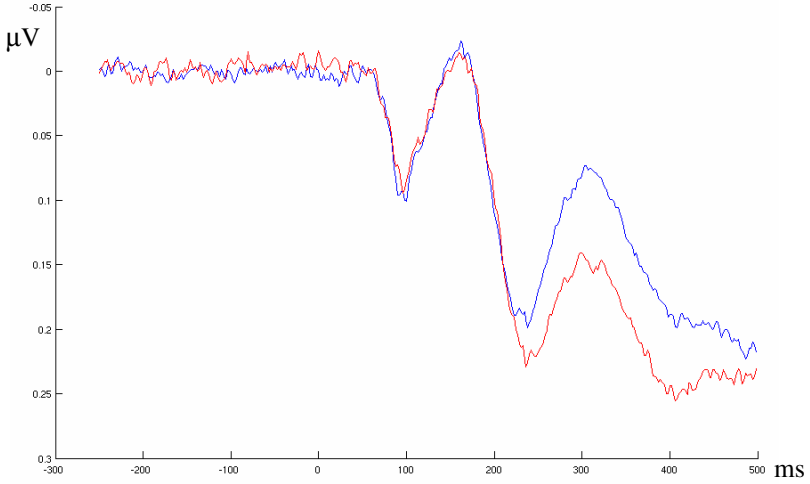


Figure 3.17: The ERP of the grand average of all subjects taken over the whole occipital region, i.e. channel 20 to 31 and 57 to 64, blue is object, red is non-object. 20% of the epoch having largest standard deviation within a 200 ms time-frame was removed. Clearly there is a difference in the ERP of object and non-object from 200-500 ms (Notice; negative is up).

From Figure 3.17 the ERP of the grand average in the occipital region reveals a P100 followed by an N100 after which a P200, N200 and a late P300 appear (around 400 ms). While no difference is present in the ERP from 0-200 ms there seems to be a difference between the ERP of object and non-object from 200-500 ms, this difference hasn't been explained by Herrmann et al. As seen on Figure 3.17 the difference between object and non-object mainly stems from the P200 and N200. In Table 2 page 54 it was explained that the P200 is known to increase with novel stimuli. As non-objects contrary to objects represent a novel stimuli every time this might explain the larger P200 for non-object. Furthermore, N200 is known to increase to task deviant stimuli. As the task was only present to keep the subjects naïve to the experiment, it is difficult to evaluate task effects as they ideally shouldn't be correlated with the condition type. However, as non-objects are probably easier to classify as edgy or round than objects (i.e. the pipe of Figure 3.12 is both edgy and round) this could explain the larger N200 for object.

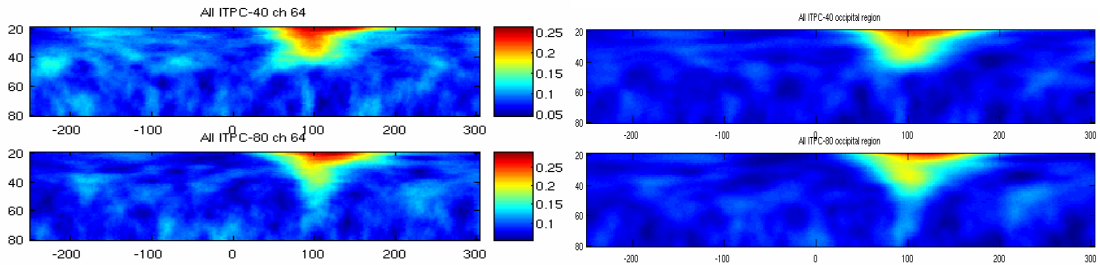


Figure 3.18: Grand average of the ITPC, object and non-object for channel 64 and the average of the whole occipital region; object in top images, non-object in bottom images. Clearly there is a strong coherence between 20-40 Hz around 100 ms. (x-axis in ms, y-axis in Hz).

Figure 3.18 shows the grand average of the ITPC for all 11 subjects at channel 64 and in the whole occipital region. It seems as if slightly more coherence is present in the object situation. Within the gamma band, the grand average at channel 64 peaks at 37 Hz and 107 ms.

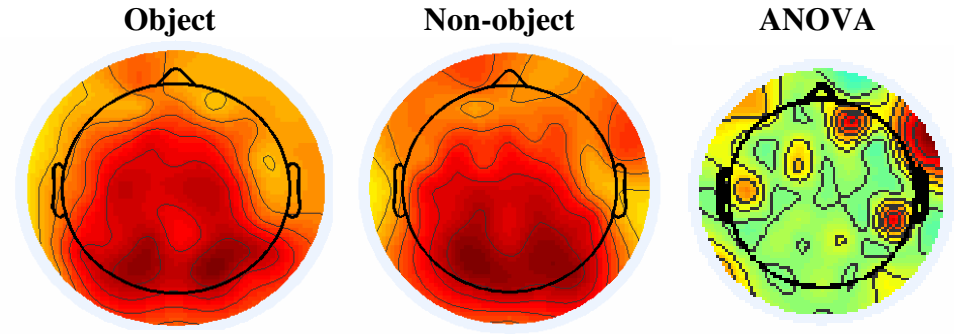


Figure 3.19: The mean coherence in all channels of the 11 subjects at 37 Hz, 107 ms for the object and non object condition (color scale is the same), and the ANOVA of the analysis of difference between object and non-object at this time-frequency point. From the ANOVA no difference between object and non-object is found in the occipital region.

From Figure 3.19 it is seen that the average coherence is more or less the same for object and non-object at the peak of the grand average for object. A test of difference between object and non-object also reveal that no significant difference is present. The largest difference is found to the frontal right where no difference is theoretically justified.

Summary of the analysis by Herrmann and colleagues

From the analysis corresponding to Herrmann and colleagues no significant difference between the object and non-object condition was found. However, coherence was clearly present in the gamma band around 50-150 ms as explained by Herrmann et al. Furthermore, a difference in the ERP between the object and non-object condition seemed to be present from 200-500 ms.

Analysis by PARAFAC

In the following, the analysis, if not otherwise stated, is performed by the ALSPARAFAC algorithm with 'row-wise' non-negativity constraints on all modes. As PARAFAC is a data exploratory tool the analysis was performed without any prior assumptions of what to expect to see from the data. First, an overall 4 way analysis was performed defined by $channel \times frequency \times time \times subject$ from 0-200 ms from stimuli onset. A Core Consistency Diagnostic was only possible to access when analyzing three-way arrays as the diagnostic was too memory consuming for MATLAB, even for a computer having 2 GB of RAM. The factors were ordered in accordance to the amount of variation they explained. Each analysis was run several times to assure stable solutions.

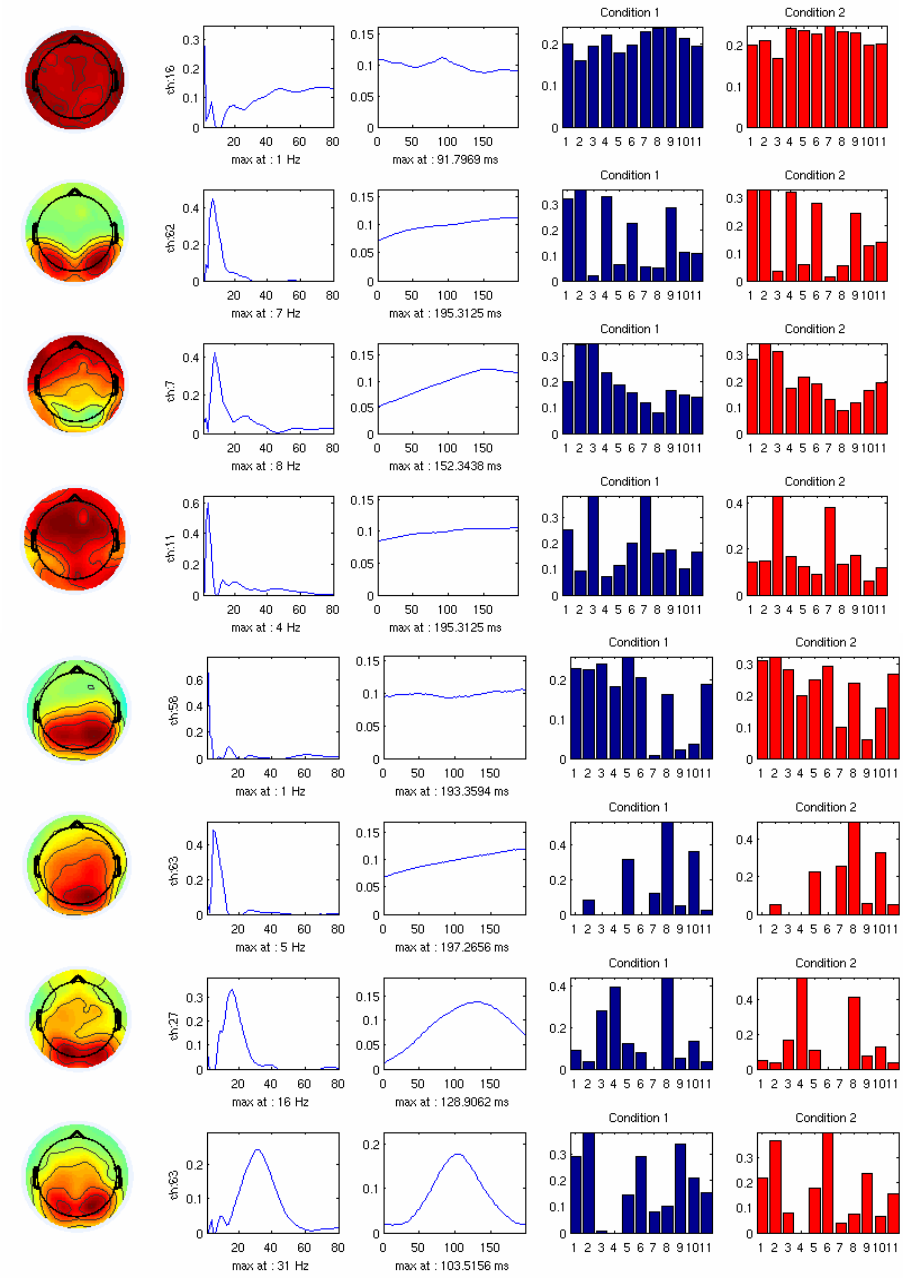


Figure 3.20: An eight component PARAFAC model fitted to the data, blue bars corresponds to object, red to non-object. Factor 2, 5, 6, 7 and 8 all indicate occipital activity. Especially factor 8 pertains to the Gamma activity around 100 ms as described by Herrmann and colleagues.

As seen on Figure 3.20 the first factor models some average activity. The second, fifth and sixth factor correspond to low frequent occipital activity relating to the ERP. For all these factors no significant difference is found between the two conditions. The eighth factor however, reveals a gamma activity in the occipital region around 100 ms corresponding to Herrmann et al.'s findings. The subjects' activities during the two conditions reveal that the last 5 subjects have more gamma activity in this factor during object than the non-object condition. Also, subject three and four seems to almost

completely lack this activity. Furthermore, factor seven reveals some beta activity around 130 ms.

An ANOVA was performed to look for differences between the two conditions for the eleven subjects. This gave an F-test value multi-way array given by $channel \times frequency \times time$ as revealed in Figure 3.21.

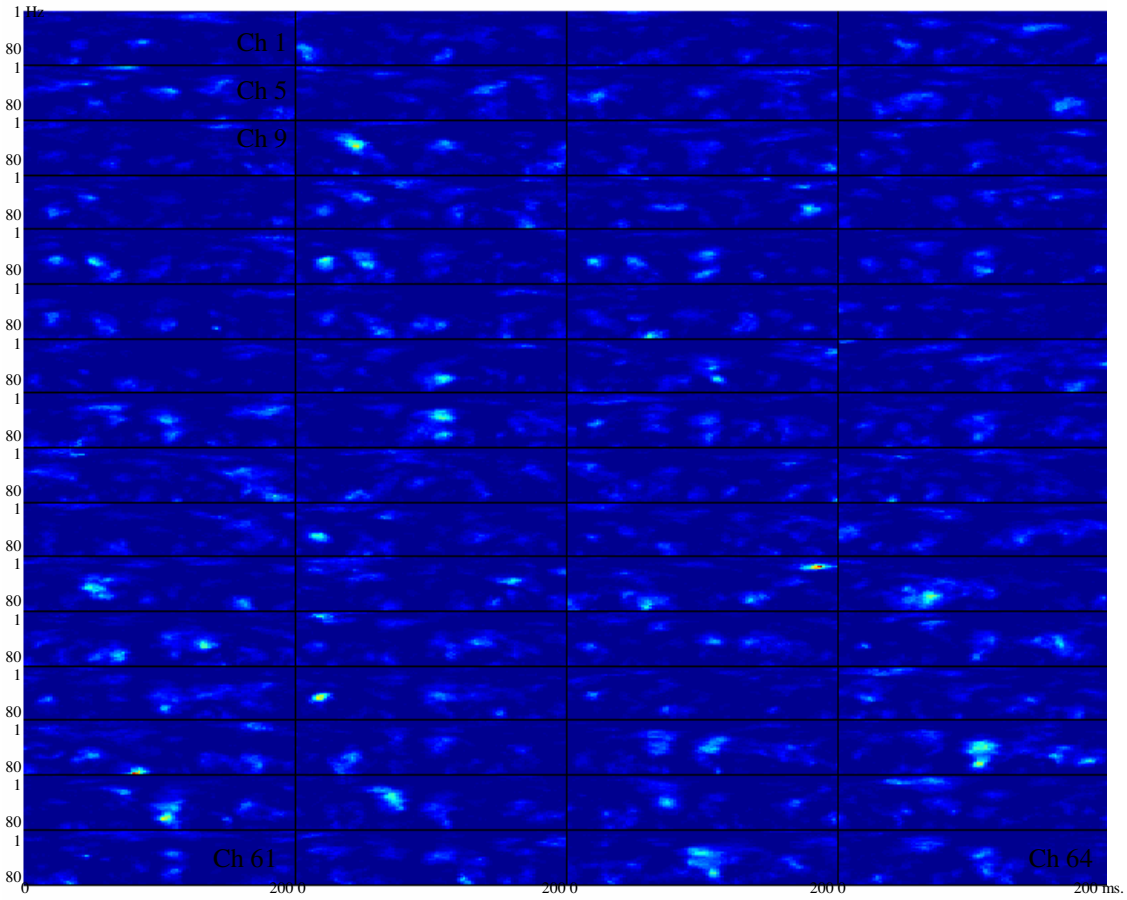


Figure 3.21: ANOVA test of difference between object and non-object in the 11 subjects, shown in a 16×4 array where each array represent a channels F-test value to given frequency-time point. From the F-values in the array it is difficult to grasp where the differences between the two conditions are present.

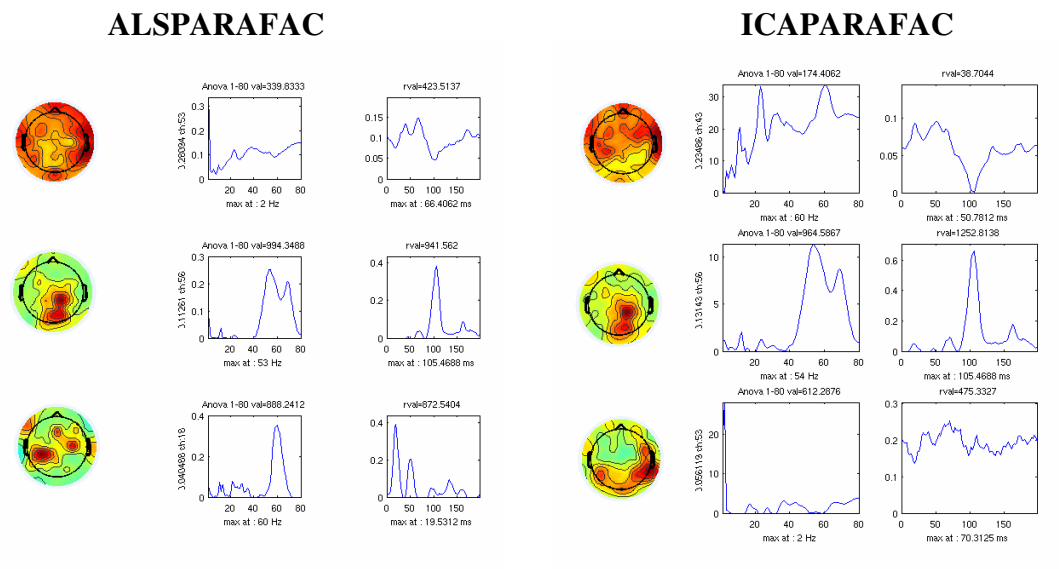


Figure 3.22: A PARAFAC model based on ALSPARAFAC and ICAPARAFAC fitted to the F-test multi-way array. Where first factor models some background activity, the second factor of both methods indicates a difference around 100 ms in the Gamma band in accordance with Herrmann and colleagues findings.

Figure 3.22 shows an ALSPARAFAC and ICAPARAFAC model fitted to the F-test multi-way array. The first factor of both algorithms models some background activity. The second factor shows that the difference between object and non-object primarily is in the occipital region in the gamma band of 30-80 Hz. It is difficult to explain what the last factor of ALSPARAFAC pertains to, but the third factor of the ICAPARAFAC model reveals a 2 Hz difference in the occipital region between the two groups corresponding to a difference in the ERP. As a result, the ANOVA clearly indicate that the difference between the two groups is as Herrmann et al. found in the Gamma band around 100 ms. As a result; the PARAFAC model is capable without any prior knowledge to identify the interesting features of the data.

Analyzing the Gamma band (30-80 Hz) by PARAFAC

To analyze the Gamma range a PARAFAC model was fitted to the data in the frequency range 30-80 Hz. Where the first and second factor of Figure 3.23 models some background activity the third factor shows the occipital gamma activity and the fourth factor reveals a central gamma activity. No systematic difference is found in the factors between the two conditions.

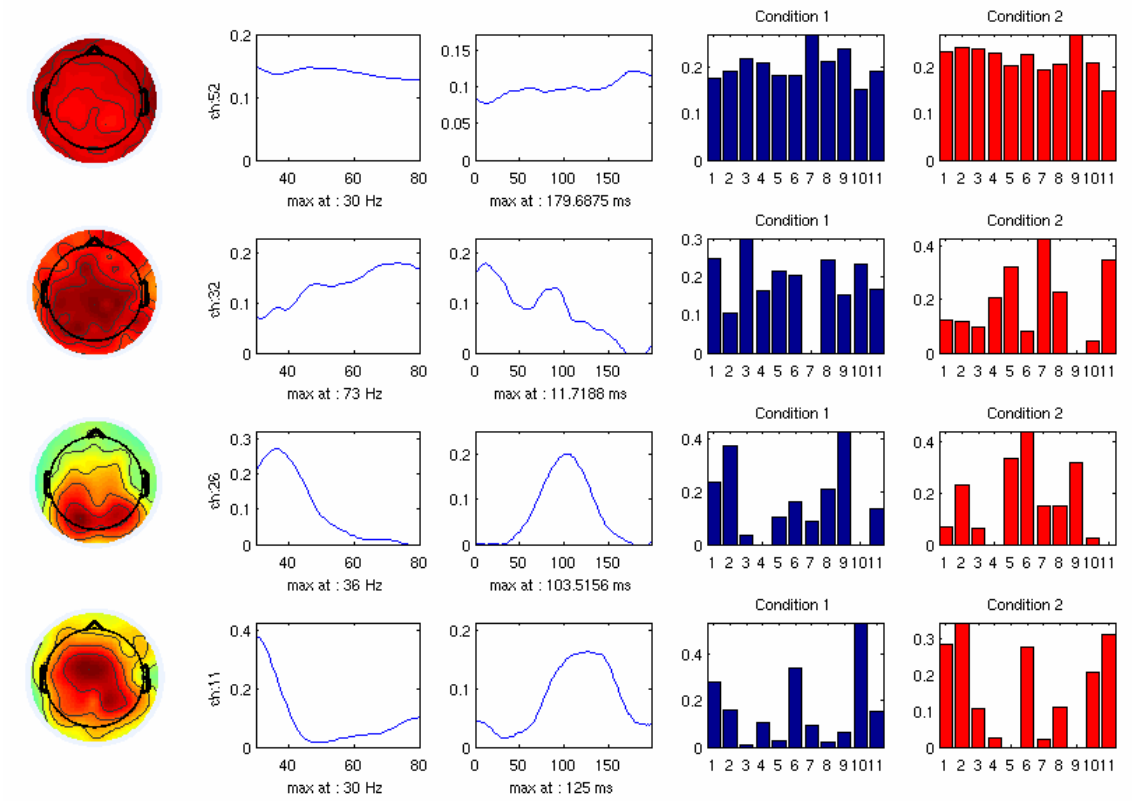


Figure 3.23: A four component PARAFAC model fitted to the data at the frequency range 30-80 Hz. Where first two factors model some average background activity, the third factor clearly reveal an occipital Gamma activity around 36 Hz at 104 ms. Finally, the last factor is more central, delayed and lower frequent.

The condition was also taken into the PARAFAC model yielding the 5-way model given in Figure 3.24. The first factor of this analysis clearly reveals some occipital gamma activity slightly more present in the object (1) than non-object (2) condition. The second factor pertains only to the non-object condition. It models a slightly more frontal, higher frequency activity around 100 ms.

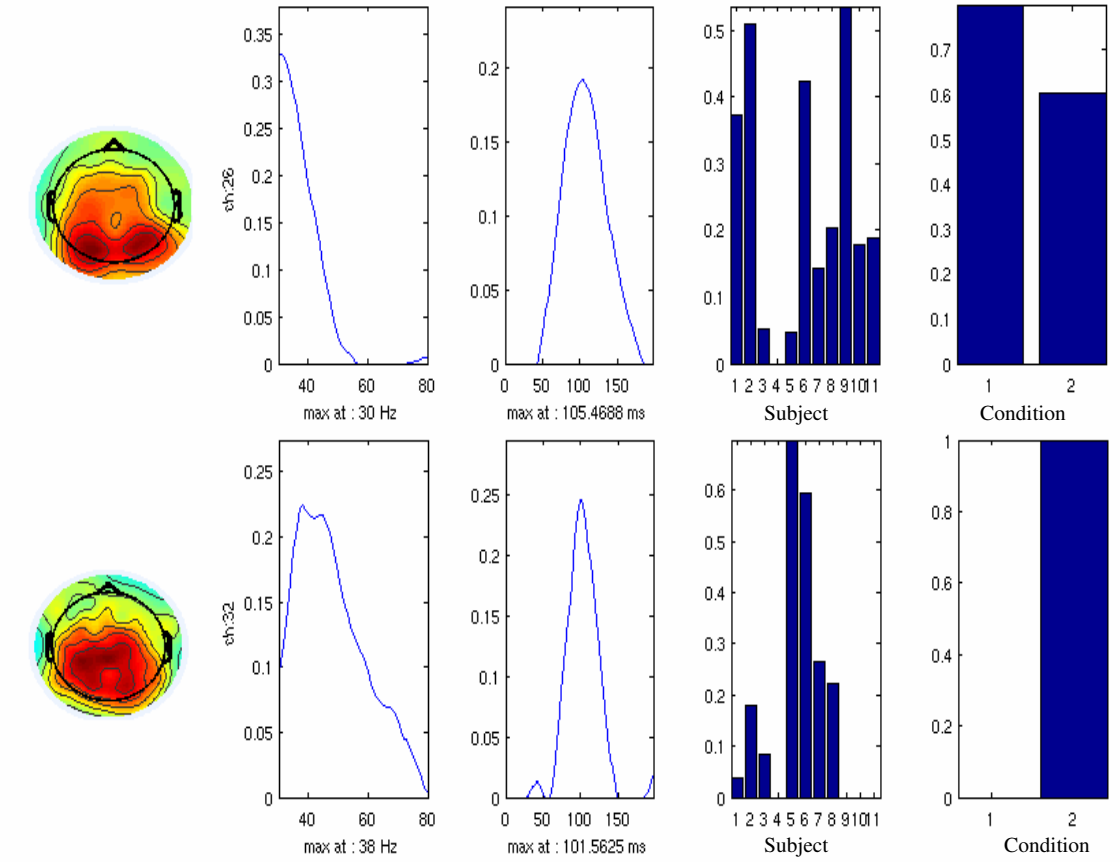


Figure 3.24: A PARAFAC model fitted to the data where condition was taken as an extra modality, 1 is object, 2 is non-object. As only two components could be found due to the limitation of only two conditions baseline activity was subtracted before fitting the PARAFAC. The first factor clearly represents the occipital gamma activity around 100 ms. This factor is mostly present in the object condition but weak in subject 3, 4 and 5. The second factor is higher frequent, slightly more central and pertains only to the non-object condition. The two factors indicate that the object condition is lower frequent whereas the non-object is slightly higher frequent and more central.

Furthermore, a PARAFAC model was fitted to each condition.

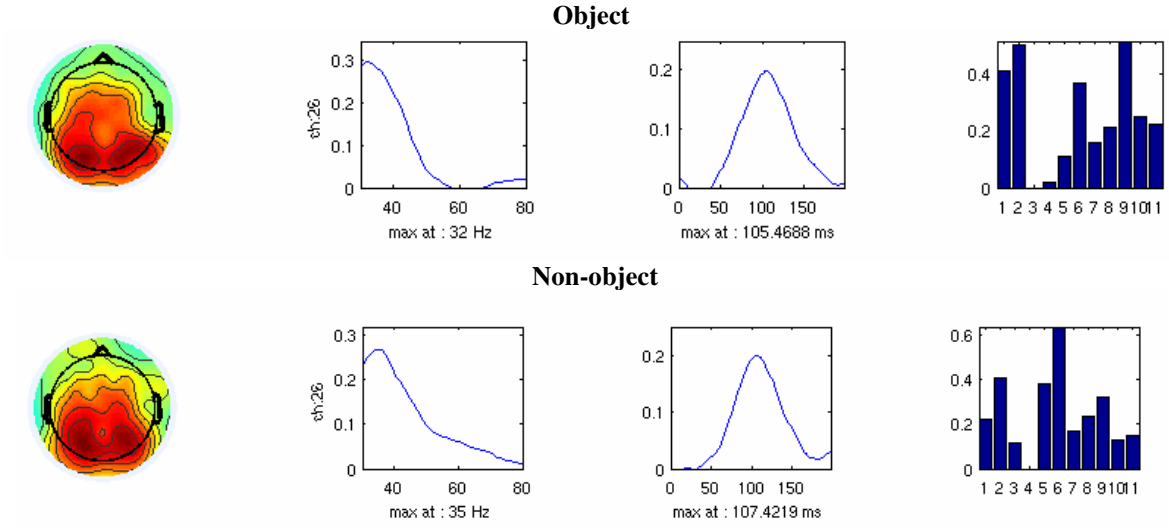


Figure 3.25: Top panel; a PARAFAC model fitted to the object condition. Bottom panel; a PARAFAC model fitted to the non-object condition. Baseline activity subtracted. Again it is revealed that both conditions have clear gamma activity around 100 ms. However, subject 3 and 4 seem to lack the activity in both conditions. Comparing the object with the non-object condition it is seen that the non-object is slightly higher frequent and delayed.

As seen on Figure 3.25 both object and non-object have clear gamma activity around 100 ms in the occipital region. However, the object condition peaks at 32 Hz, 105 ms whereas the non-object peaks at 35 Hz 107 ms. In both conditions subject 3 and 4 have practically no gamma activity in the occipital region.

In addition, a PARAFAC model was fitted to the ANOVA of the gamma band.

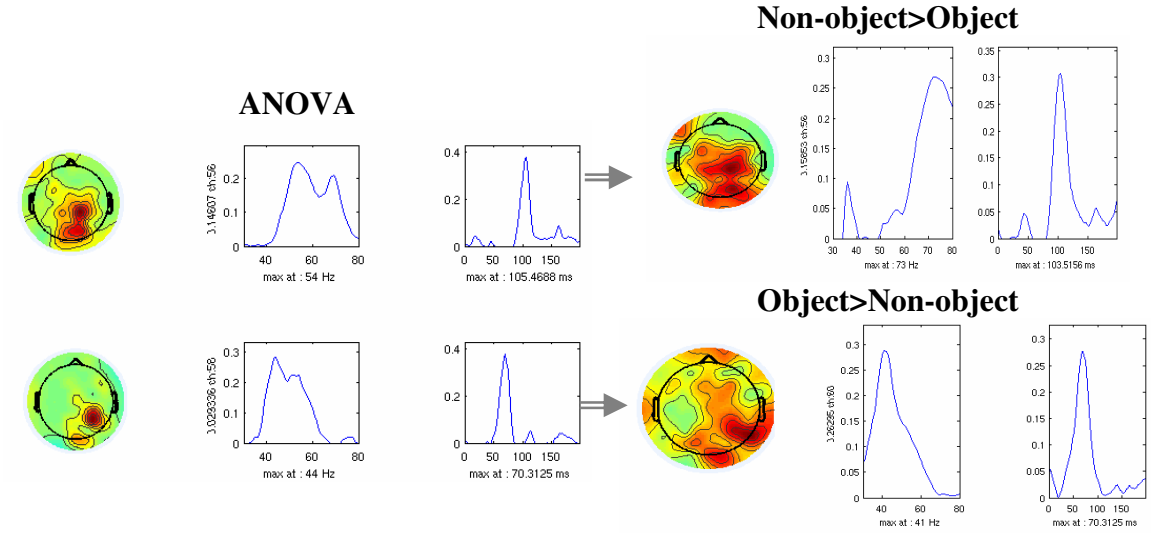


Figure 3.26: Left figure; a PARAFAC based on the F-test value of the gamma band. Top, right figure; a PARAFAC model fitted to regions where non-object is more coherent than object. Bottom, right; a PARAFAC model fitted to regions where object is more coherent than non-object. As seen from the first factor of the ANOVA this factor pertains to the situation where non-target on average is more coherent than target whereas the second factor of the ANOVA corresponds to a situation where object on average is more coherent than non-object. Consequently, object is more coherent early and at lower frequencies whereas non-object is more coherent later and at higher frequencies (baseline activity removed from the data).

From the ANOVA of Figure 3.26 the second factor corresponds to the second factor found in Figure 3.22. Furthermore, the third factor found in the ANOVA of Figure 3.26 also reveals the presence of an earlier and less high frequent difference between the two groups. Analyzing when object is larger than non-object and when non-object is larger than the object condition, it is seen that the first factor of the ANOVA corresponds to the factor where non-object is more coherent than the object condition, whereas the second factor of the ANOVA matches the situation where object is more coherent than the non-object condition. Consequently, the difference between the object and non-object condition is mainly due to the fact that object is coherent earlier and at lower frequencies than non-object.

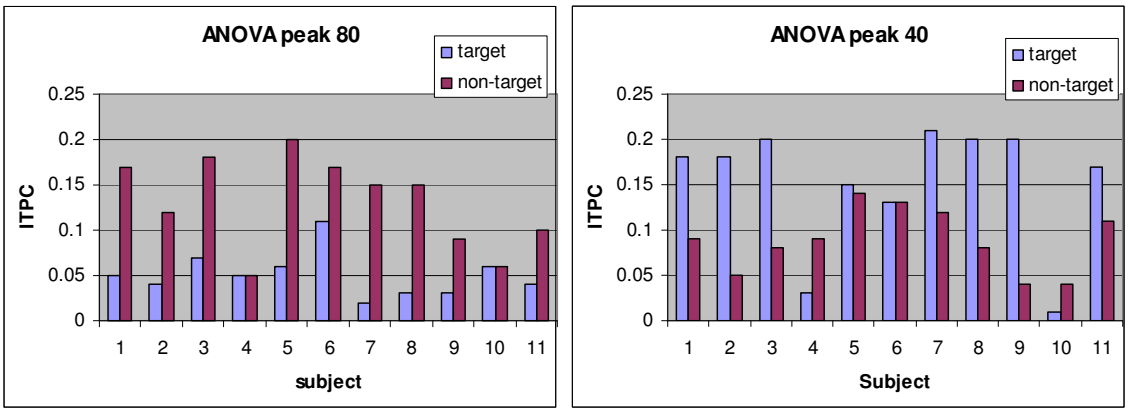


Figure 3.27: The coherence value at the peak of the ANOVA of factor 1 denoted peak 80 and of factor 2 denoted peak 40 found in Figure 3.26. As seen on Figure 3.26 the first factor of the ANOVA corresponds to a situation where non-object is more coherent than object (target) whereas the second factor of the ANOVA pertains to a situation where object is more coherent than non-object.

In Figure 3.27 the same pattern reveals itself. The first factor of the ANOVA in Figure 3.26 corresponds to the situation where non-object is more coherent than object whereas the second factor corresponds to the situation where object is more coherent than the non-object condition.

The PARAFAC model was also fitted to the ITPC of each subject given by the multi-way array $channel \times frequency \times time$.

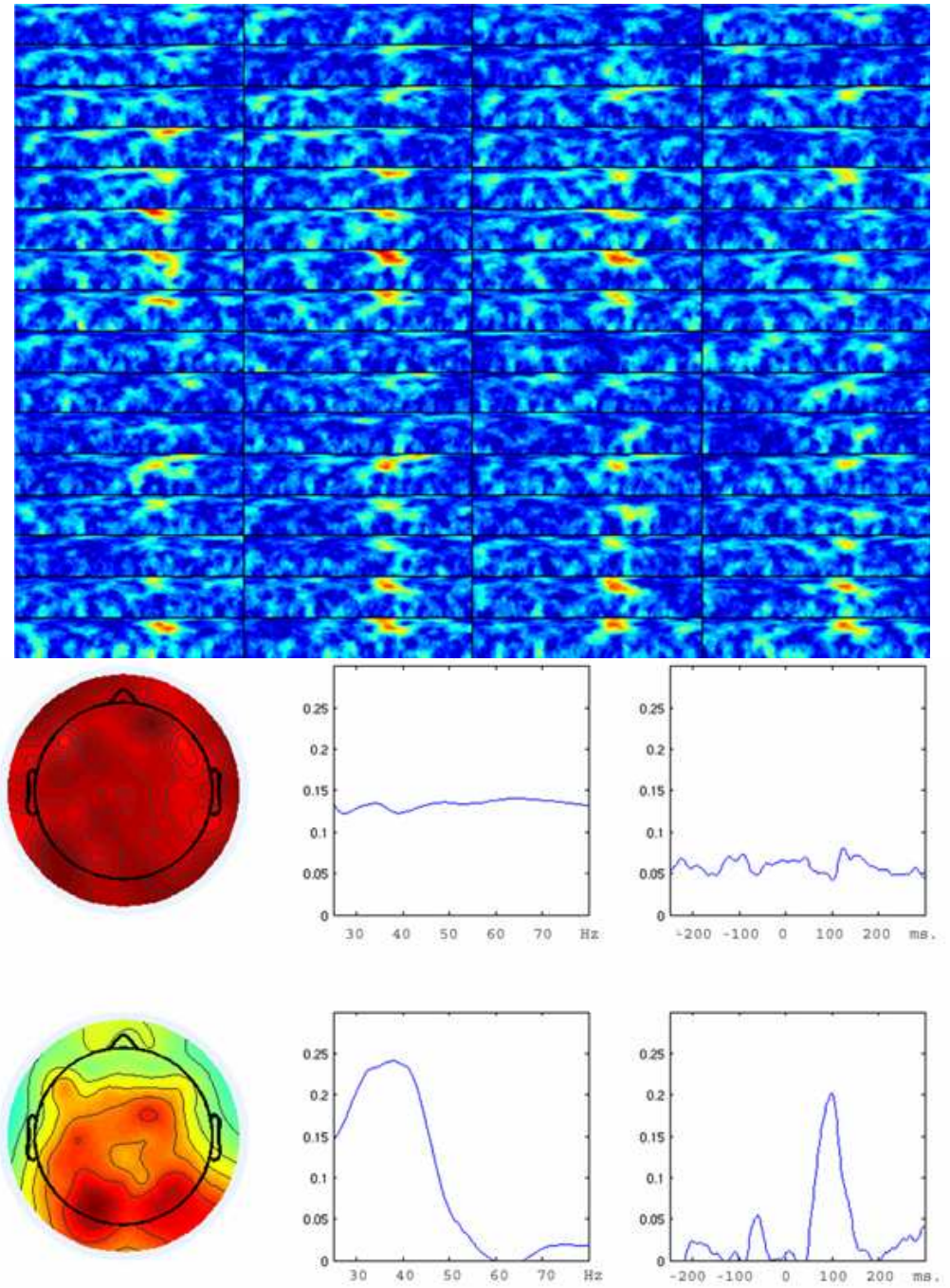


Figure 3.28:Top panel; the ITPC multi-way array given by channel×frequency×time of a subject shown in a 16×4 array of channels where x-axis corresponds to time from -250-300 ms and y-axis frequency from 20-80 Hz. Bottom panel; a PARAFAC model fitted to this ITPC. Where the first factor shows some background activity the second factor clearly reveals the occipital Gamma activity around 100 ms.

For each subject, the gamma peak in the occipital region at 50-150 ms was identified in time and frequency by the factor corresponding to the second factor in the PARAFAC

decomposition revealed on Figure 3.28. Notice that the first factor corresponds to some baseline activity.

Having identified the peak of each subject, the coherence at each subject's frequency-time point was found for all channels. Finally, the mean of these topographic maps were calculated as revealed on Figure 3.29.

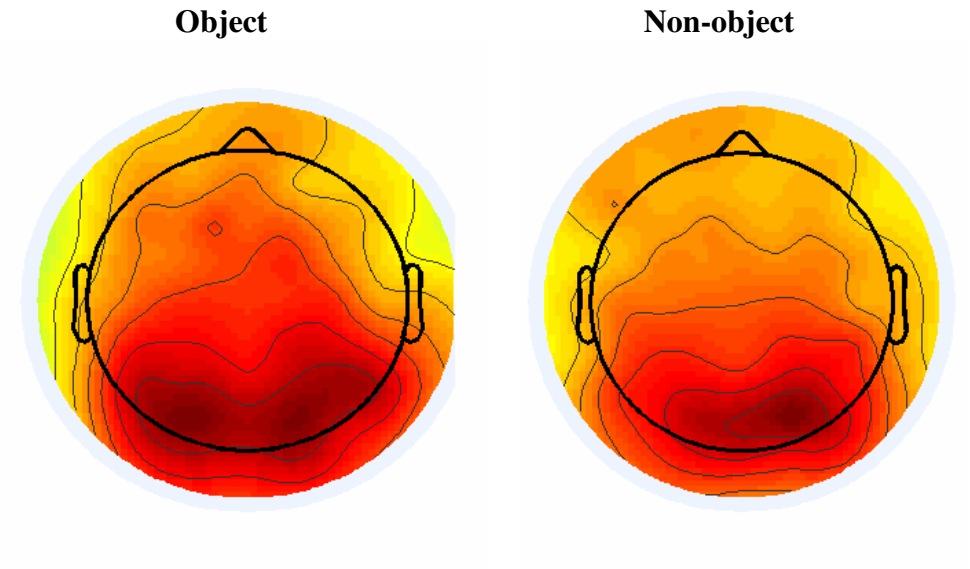


Figure 3.29: The mean coherence for all subjects at their gamma-peak for object and non-object. Coherence seems to be present in a larger region for the object condition than the non-object condition.

Figure 3.29 indicate that the coherence at the peak is high at a much larger region for object than for non-object. As channel 64 which was the basis of Herrmann et al.'s analysis lies right at the peak of both object and the non-object in Figure 3.29 this might be why the findings of difference between the two conditions in Herrmann and colleagues' analysis was poor. Had Herrmann and colleagues' analysis been based on a channel in the left hemisphere, the difference in coherence between object and non-object might have been stronger.

Finally, the PARAFAC model was used to analyze the ERP. This has been done previously by Field et al [10]. Field and colleagues found that a problem of degeneracy arose when fitting the PARAFAC model to the ERP. As a solution they proposed introducing an orthogonality constraint on the dimension representing the temporal development of the ERP. The orthogonality constraint will here be compared to imposing non-negativity as we suggest. The non-negativity can simply be assured by adding a positive constant to the ERP. Prior to analyzing the ERP, 20 % of the epochs having largest standard deviation within a 200 ms time window were removed to get rid of eye and muscle artifacts.

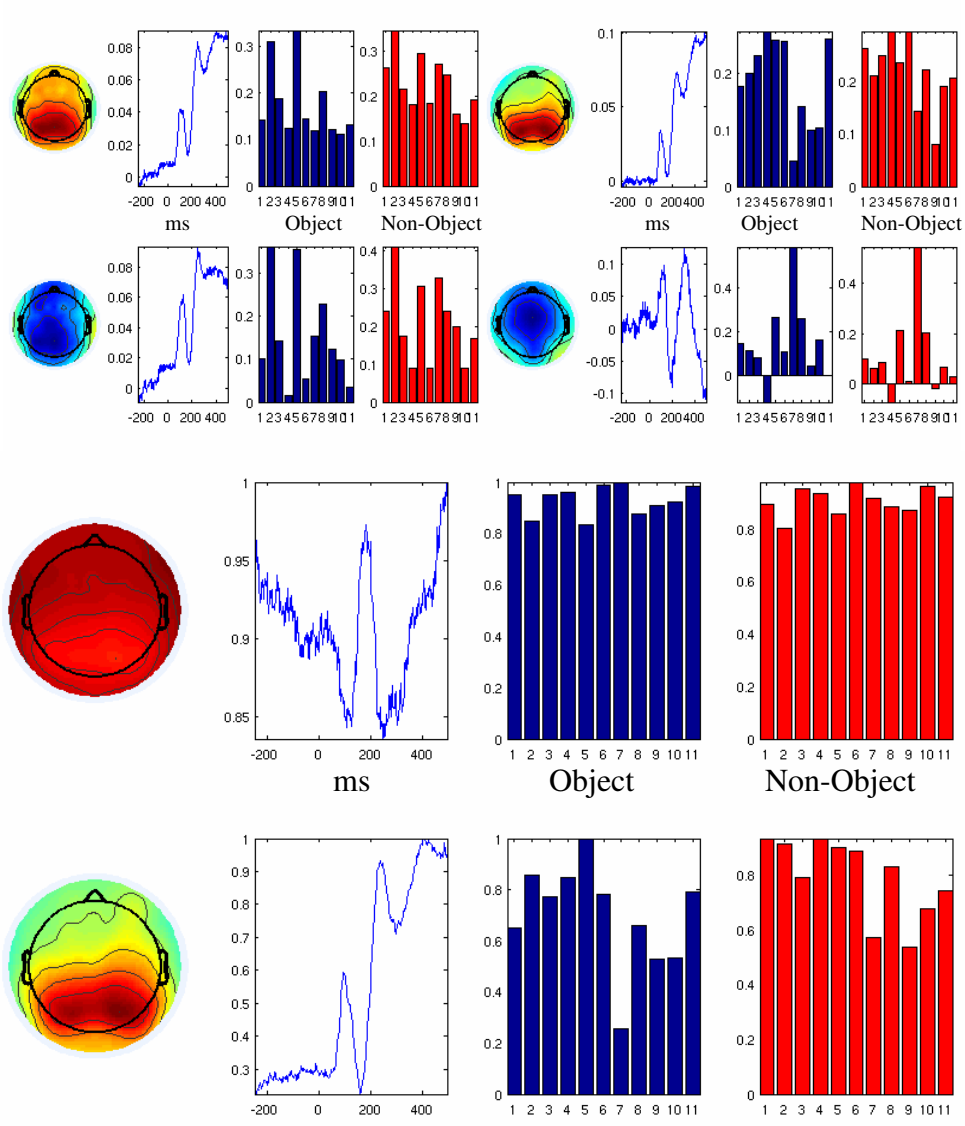


Figure 3.30: Top left panel; analyzing the ERP unconstrained. Top right panel; imposing an orthogonality constraint on the ERP. Bottom panel; imposing non-negativity by addition of a constant. Blue bars correspond to the object condition, red bars to non-object, notice; positive is here upward on the ERP. Neither the unconstrained nor the orthogonality constrained PARAFAC models are able to find the true ERP. This is however found for the non-negativity constrained model where the ERP correctly is split into a frontal and an occipital part.

As seen on Figure 3.30 the unconstrained solution yields highly degenerate factors. The ERP of the second factor is almost identical to the ERP of the first but with opposite sign as revealed in the topographic maps. Imposing the orthogonality constraint insures no degeneracy in the ERP. However, a few subjects have negative coefficients and the justification for the two ERP's to be orthogonal in reality is very questionable. Imposing non-negativity however yields excellent results. The non-negative PARAFAC algorithm has split the ERP into two easy interpretable components. The first component models a mostly frontal ERP whereas the second component beautifully models the ERP of the occipital region. This occipital ERP seems to be more present in the non-object than the

object condition. This corresponds to the findings of Figure 3.17 where the P200 was stronger for the non-object condition.

A PARAFAC model was then fitted to the ERP of each condition.

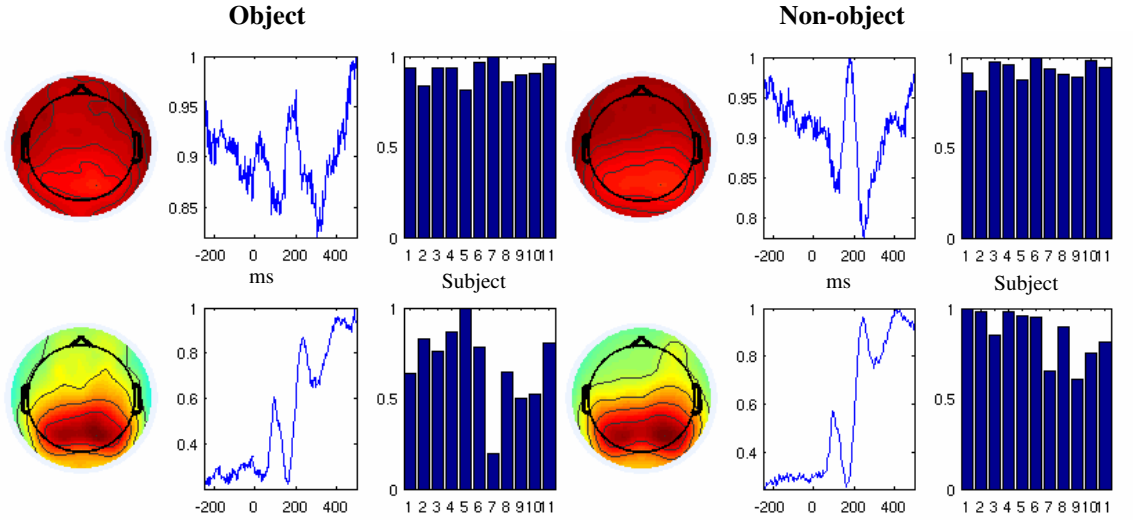


Figure 3.31: The ERP for object, left figure, and non-object, right figure. Where the occipital ERP is very similar in the two conditions the frontal ERP has a stronger P200 in the non-object condition than the object condition.

From the first factor of Figure 3.31 it is seen that the P200 of non-object is much stronger than for object frontally as also found to be the case in the occipital region as revealed in Figure 3.17. Again, this is due to the fact that the non-object represented novel stimuli.

Summary of the PARAFAC analysis

From the PARAFAC analysis of the 11 subjects it was seen that the main activity was in the occipital region corresponding to an experiment having to do with visual stimuli. Furthermore, the difference between the two conditions was mostly present in the gamma band around 100 ms. In addition, this difference was primarily due to a delayed coherent signal at higher frequencies for non-object than the object condition. At the peak of each of the two conditions it seemed as if coherence was present in a larger region in the object situation than the non-object situation. Finally, imposing non-negativity to the ERP by addition of a constant made the PARAFAC model able to correctly split the ERP into an occipital and a frontal factor. This was not possible for an unconstrained model while imposing orthogonality as previously done didn't yield results that were as satisfying. From the ERP's it was found that non-object also had a stronger P200 frontal.

4 Discussion

"It is a good morning exercise for a research scientist to discard a pet hypothesis every day before breakfast. It keeps him young."

Konrad Lorenz

PARAFAC and simulated data

The test of the PARAFAC algorithms on simulated EEG/ERP data revealed that both ALSPARAFAC and the ICAPARAFAC effectively found the right factors in the data. This was however not possible from the simple 2-D analysis using an ICA-algorithm on the raw data. Consequently, PARAFAC seemed effective in the analysis of the frequency changes in the EEG/ERP when using these two algorithms.

The Core Consistency Diagnostic and Bayesian Information Criterion both proved effective in accessing the correct number of factors in the data for the models that performed the best, i.e. ICAPARAFAC and ALSPARAFAC.

ICAPARAFAC performed better than ALSPARAFAC at estimating the true factors as the signal to noise ratio dropped. As the ICAPARAFAC algorithm also was faster than the ALSPARAFAC algorithm it has great potentials. In Appendix D: ICA- and ALSPARAFAC on Chemometric Data it was revealed that on chemometric data the ICAPARAFAC algorithm also performs well. Therefore, the new ICAPARAFAC algorithm seems promising in a wide range of fields where Combined Independence can be assumed in the data.

Herrmann and colleagues analysis vs. the PARAFAC analysis

Unfortunately, only 11 subjects were analyzed giving quite an uncertain picture to base a conclusive comparison of Herrmann et al.'s findings with the findings of PARAFAC. Furthermore, 2 of the 11 subjects lacked completely clear coherent gamma activity in the occipital region. However, the fact that Herrmann et al. by the basis of the peak of the object condition conclude that object is more coherent than the non-object seems wrong as it favors the object condition. Furthermore, they only examined channel 64 instead of taking the whole occipital region into considerations which is questionable – it might be that channel 64 was chosen for the only reason that it was the most significant. The PARAFAC analysis indicates that the difference found is mostly due to the fact that the non-object peak later and at higher frequencies than the object condition, rather than having to do with non-object in general being less coherent. Furthermore, the analysis based on the individually found peaks in the occipital region for object and non-object revealed that object seemed coherent in a larger region than the non-object.

The PARAFAC analysis was capable of integrating the information present in all channels into simple interpretable components. This made the analysis of Herrmann and colleagues' paradigm much more complete by PARAFAC than by their own proposed analysis. In addition, it was seen that PARAFAC was able to work in several ways

analyzing both the averaged ERP and the wavelet transformed data over several different modalities taking into account subject and condition variability. Finally, when analyzing the ERP it was revealed that imposing non-negativity worked much better than forcing orthogonality.

The paradigm Herrmann et al. use isn't strong. Although the examples shown from the paradigm in Figure 3.12 are definitely recognizable as a pipe and chair whereas no such interpretation is present in the non-object situation, the non-object almost looks like a broken pipe. Furthermore, there are recognizable objects in the non-object condition such as a square, a cylinder and a bowl which in itself might have long term memory representations. This weakens the difference between object and non-object - weakening the whole paradigm.

5 Conclusion

"The important thing is not to stop questioning. "

Albert Einstein

It was shown that PARAFAC is an effective data exploratory tool in the analysis of the Inter Trial Phase Coherence of the ERP. Furthermore, it was revealed that imposing non-negativity when decomposing the ERP by PARAFAC seemed to solve the problem of degeneracy completely. Here PARAFAC was only used to analyze the ERP and ITPC. However, there is no reason why PARAFAC shouldn't also be an effective tool in the analysis of the ERSP and ERPCOH. These measures were not analyzed as the main interest was to access the Inter Trial Phase Coherence of the data. Furthermore, both measures are very susceptible to noise requiring an extensive preprocessing of the data. However, future work will focus on PARAFAC's ability to analyze these measures.

In the analysis of Herrmann et al.'s gamma band coherence it was found that the non-object was slightly more delayed and higher frequent than the object condition. New experiments are presently conducted to confirm these findings. Future work will also be done to use PARAFAC in analyzing ERP data from other experiments having multiple conditions including different forms of sensory stimuli. As PARAFAC is capable of analyzing complicated multi-modal data it might likely shed new light on these data.

The ICAPARAFAC algorithm also showed promising result, being a great alternative to the popular ALSPARAFAC algorithm when combined independence can be assumed. Work lies ahead in improving the underlying Independent Component Analysis algorithm both to deal with non-negativity as well as being optimized to find the correct components in the ERP.

The PARAFAC model analyzed the wavelet transformed data. As the wavelet transform corresponds to a convolution in which random noise ideally becomes a constant factor at all frequencies analyzing the data in the frequency domain is in itself an efficient way to handle noise full data. Furthermore, the PARAFAC algorithms were able to separate systematic oscillatory noise such as 50 Hz noise from electric devices into a designated component. As it is possible to reconstruct the signal from the wavelet coefficients [32] , the EEG/ERP corresponding to each factor of the PARAFAC can also be reconstructed. Thus, the PARAFAC analysis might also work well in reducing systematic noise. Furthermore, knowing the signatures of the noise from a training set can be used to find the noise-signature in one of the modalities from a test set by keeping the components of all other modalities found from the training set fixed on the test set.

The PARAFAC analysis of the coherence relied on the wavelet analysis accessing the correct temporal frequency information of the data. In this thesis the wavelet suggested by Herrmann et al. [14],[15] was used. However, Figure 3.14 showed that the choice of wavelet had an impact on the coherence found. Consequently, the influence of the

wavelet's form used in the analysis of the ERP should also in future work be explored. Maybe even waveforms corresponding to more physiologic burst of neuronal activity could be developed and analyzed.

Wavelet analyzed data is known to be overcomplete, i.e. having more data than information present in the data. Methods such as various matching pursuit algorithms have tried to resolve this problem, reducing the wavelet data to an information level corresponding to the original signal [17]. Within this field, PARAFAC seems applicable as it is an effective tool for data reduction as proven for multiple image compression [33]. Consequently, work analyzing PARAFAC's ability to solve the overcomplete representation in the wavelet analysis could also prove an important application of the model.

Although the functional magnetic resonance imaging, fMRI, has taken much focus away from the EEG and ERP, EEG/ERP still has great potentials. First of all the EEG is much easier to use for experimentation as the subjects can stay in a natural environment rather than having to be put inside a scanner. Furthermore, the EEG offers a much higher temporal resolution below the micro second range whereas the fMRI still works in the ms range and is most probably limited to this range. Consequently, ways of integrating EEG and fMRI has lately gotten much attention. Hopefully, the PARAFAC analysis of the EEG can help in this work decomposing the EEG signal into atoms that can be related to the fMRI signal. This application of the PARAFAC model has already gotten some attention [24]. As PARAFAC shows promising results in the field of EEG its application to MEG will very likely also be good. Consequently, multi-way array analysis will probably in the future become an important tool in the analysis of brain-recordings from a variety of scanning techniques.

Although Harshman proposed the use of PARAFAC on EEG in1970, the use has been very limited. Why this is so is hard to understand given its wide usability to explore the EEG/ERP. However, the limited use might have been the consequence of the fact that previous works didn't resolve the problem of degeneracy by imposing non-negativity. Furthermore, the PARAFAC analysis is very memory consuming and slow, putting great demands on computer power. In this analysis 2 GB of RAM was required in order to simultaneously analyze the ITPC of the 11 subjects in both conditions having 64 channels of 200 ms data sampled at 512 Hz at 60 different frequencies. These computer requirements might be the reason for the limited use of PARAFAC so far in the field of EEG/ERP research. Hopefully, however, PARAFAC will turn out to be an important tool that will be widely used in the future when analyzing brain-data.

References

[1] Andersson, Claus A.; Bro, Rasmus "*The N-way Toolbox for MATLAB"* Chemometrics and Intelligent Laboratory Systems 52 2000 1–4, Toolbox can be downloaded from: http://www.models.kvl.d//source

[2] Arnfred, Sidse *"Proprioceptive Event Related Potentials: Gating and Task effects" Article in press. Clinical Neurophysiology*

[3] Beal, J. Matthew *"VARIATIONAL ALGORITHMS FOR APPROXIMATE BAYESIAN INFERENCE"* Thesis submitted for the degree of Doctor of Philosophy of the University of London May 2003

[4] Bishop, Christopher M. *"Neural Networks for Pattern Recognition"* Oxford University Press 1995

[5] Bro, Rasmus *"Multi-way Analysis in the Food Industry Models, Algorithms, and Applications"* Ph.D. thesis

[6] Busch, Niko A; Debener, Stefan; Kranczioch, Cornelia; Engel,Andreas K.; Herrmann et al., Christoph S. *"Size matters: effects of stimulus size, duration and eccentricity on the visual gamma-band response"* Clinical Neurophysiology 115 (2004) 1810–1820

[7] Cole, H.W.; Ray, W.J. *"EEG correlates of emotional tasks related to attentional demands"* Int. J. Psychophysiol 3(1): 33-41, 1985

[8] Delorme, Arnaud; Makeig, Scott *"EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis"* Journal of Neuroscience Methods 134 (2004), 9-21

[9] EL-Bab, Mohamed Fath *"COGNITIVE EVENT RELATED POTENTIALS DURING A LEARNING TASK"*, A thesis presented for the degree of Doctor of Philosophy, Department of Clinical Neurological Sciences Faculty of Medicine, University of Southampton, United Kingdom, April 2001

[10] Field, Aaron S.; Graupe, *Daniel "Topographich Component (Parallel Factor) analysis of Multichannel Evoked Potentials: Practical Issues in Trilinear Spatiotemporal Decomposition"* Brain Topographa, Vol. 3, Nr. 4, 1991

[11] Grasman, R. P. *"Sensor array signal processing and the neuro-electromagnetic inverse problem in functional connectivity analysis of the brain"* PhD-thesis, University of Amsterdam, Amsterdam (2004).

[12] Griffiths, David J. *"Introduction to Electrodynamics"* Third edition Prentice Hall, 1999

[13] Harshman , Richard A. *"FOUNDATIONS OF THE PARAFAC PROCEDURE: MODELS AND CONDITIONS FOR AN "EXPLANATORY" MULTIMODAL FACTOR ANALYSIS"* UCLA, December, 1970

[14] Herrmann, Christoph S.; Munk , Matthias H.J.; Engel , Andreas K. "*Cognitive functions of gamma-band activity: memory match and utilization"* TRENDS in Cognitive Sciences Vol.8 No.8 August 2004

[15] Herrmann, Christoph S; Lenz, Daniel; Junge, Stefanie ; Busch, Niko A; Maess, Burkhard *"Memory-matches evoke human gamma-responses"* BMC Neuroscience 2004, 5:13

[16] Kandel, Eric R., Schwartz, James H., Jessell, Thomas M. *"Principles of Neural Science"* McGrawHill, fourth edition 2000, p. 106-108, 326

[17] Koenig, T.; Marti-Lopez, F.; Valdes-Soas, P. *"Topographic Time-Frequency Decomposition of the EEG"* NeuroImage 14, p. 383-390, 2001

[18] Kolendra, Thomas *"Adaptive tools in virtual environments",* IMM-2002

[19] Lathauwer, Lieven De.; Moor, Bart De; Vandewalle, Joos *"A MULTILINEAR SINGULAR VALUE DECOMPOSITION"* SIAM J. MATRIX ANAL. APPL. Vol. 21, No. 4, pp. 1253–1278

[20] Lathauwer ,Lieven De; Moor, Bart De; Vandewalle, Joos *"ON THE BEST RANK-1 AND RANK-(R1,R2, . . . , RN) APPROXIMATION OF HIGHER-ORDER TENSORS."* SIAM J. MATRIX ANAL. APPL. c Vol. 21, No. 4, pp. 1324–1342

[21] Lathauwer, L. De ; Moor, B. De ; Vandewalle, J. "*Independent component analysis based on higher-order statistics only"*, Proceedings of the IEEE Signal Processing Workshop on Statistical Signal and Array Processing, Corfu, Greece, 1996, p. 356–359.

[22] Lee, Daniel D.; Seung, H. Sebastian *"Algorithms for Non-negative Matrix Factorization"* Adv. Neural Info. Proc. Syst. 13, 556-562 (2001)

[23] Martin, Carla D. Moravitz *"Tensor Decompositions Workshop Discussion Notes American Institute of Mathematics (AIM) Palo Alto, CA"* July 19-23, 2004

[24] Martínez-Montes, E.; Valdés-Sosa, P.A.; Miwakeichi, F.; Goldman, R. I.; Cohen, MS. *"Concurrent EEG/fMRI analysis by multiway Partial Least Squares"* NeuroImage 22, p. 1023-1034, 2004.

[25] Miwakeichi, Fumikazu; Martínez-Montes, Eduardo; Valdés-Sosa, Pedro A.; Nishiyama, Nobuaki; Mizuhara, Hiroaki; Yamaguchi, Yoko *"Decomposing EEG data into space-time-frequency components using Parallel Factor Analysis"* NeuroImage 22 (2004) p. 1035-1045.

[26] Möcks, Joachim *"Decomposing Event-Related Potentials: A New Topographic Component Model"*, Biological Psychology 26 (1988), 199-215

[27] Nielsen, Frederik Brink *"Variational Approach to Factor Analyssi and Related Models"* Thesis, Technical University of Denmark, 2004.

[28] Nunez, Paul L. *"Electric Fields Of the Brain"*, Oxford University Press 1981, p.50-300, p. 445-450

[29] Nunez, Paul L. *"Neocortical Dynamics and Human EEG Rhythms"* Oxford university press 1995 chapter 1,4

[30] Pfurtscheller, G., Lopes da Silva, F.H. *"Event-related EEG/MEG synchronization and desynchronization: basic principles"*, Clinical Neurophysiology 110 (1999) p. 1842-1857

[31] Sidiropoulos, Nicholas D., Bro, Rasmus *"On the uniqueness of multilinear decomposition of N-way arrays"* J. Chemometrics 2000; 14 p. 229–239

[32] Todorovska, M.I., *"Estimation of Instantaneous Signals Using the Continuous Wavelet Transform"*, University of Southern California, Department of Civil Engineering, Report CE-01-07, 2001.

[33] Wang, Hongcheng; Ahuja , Narendra *"Compact Representation of Multidimensional Data Using Tensor Rank-One Decomposition"* Beckman Institute, University of Illinois at Urbana-Champaign, USA

[34] Windhorst, U. and Johansson, H "*Modern Techniques in Neuroscience Research"*, Springer Verlag 1999

[35] Wrightt , J.J., Kydd, R.R. *"The electroencephalogram and cortical neural networks"*, Network 3 1992 page 341-362

[36] http://www.univ.trieste.it/~brain/Segnali/Segnali2.html

[37] http://www.departments.bucknell.edu/linguistics/lectures/05lect14.html

[38] http://www.benbest.com/science/anatmind/FigV5.gif

[39] http://bio.winona.msus.edu/bates/Human%20Bio/Images/hb10_05.jpg

[40] http://psyphz.psych.wisc.edu/. ../EEGintro.htm

[41] http://www.models.kvl.dk/research/data/Amino_Acid_fluo/index.asp

Appendix A: Theorems with proofs

Theorem 1 (Jensens Inequality)

Let $f : R \to R$ be a convex function, $\lambda_i > 0 \,\forall i$ and $\sum_i \lambda_i = 1$ then

$$f\left(\sum_{j=1}^{M} \lambda_j x_j\right) \geq \sum_{j=1}^{M} \lambda_j f\left(x_j\right)$$

Proof:

Let $x_i \in [a;b]$ where a,b$\in \mathcal{R}$ and let $t \in [0;1]$. As f is convex the following holds:

$f((1-t)a + tb) \geq (1-t)f(a) + tf(b)$

as $x_i \in [a;b]$ it follows that $x_i = (1-t_i)a + t_i b$. This gives:

$$\sum_i \lambda_i x_i = \sum_i \lambda_i ((1-t_i)a + t_i b) = a\sum_i \lambda_i (1-t_i) + b\sum_i \lambda_i t_i = (1-t)a + tb \in [a;b] \;\; as \;\; t = \sum_i \lambda_i t_i$$

Proof now done by induction;

M=2:

$$f(\lambda_1 x_1 + \lambda_2 x_2) = f((1-\lambda_2)x_1 + \lambda_2 x_2) \geq (1-\lambda_2) f(x_1) + \lambda_2 f(x_2) = \lambda_1 f(x_1) + \lambda_2 f(x_2)$$

Assume $f\left(\sum_{j=1}^{M} \lambda_j x_j\right) \geq \sum_{j=1}^{M} \lambda_j f(x_j)$ valid for M-1. We now have:

$$f\left(\sum_{j=1}^{M} \lambda_j x_j\right) = f\left(\sum_{j=1}^{M-1} \lambda_j x_j + \lambda_M x_M\right) = f\left((1-\lambda_M)\sum_{j=1}^{M-1} \frac{\lambda_j}{(1-\lambda_M)} x_j + \lambda_M x_M\right) \geq$$

$$(1-\lambda_M) f\left(\sum_{j=1}^{M-1} \frac{\lambda_j}{(1-\lambda_M)} x_j\right) + \lambda_M f(x_M) \geq (1-\lambda_M)\sum_{j=1}^{M-1} \frac{\lambda_j}{(1-\lambda_M)} f(x_j) + \lambda_M f(x_M) = \sum_{j=1}^{M} \lambda_j f(x_j)$$

Theorem 2

Given the following model and assumptions:

$$\begin{aligned}
&\textbf{Model :} && \mathbf{X}^{(i)} = \mathbf{A}\mathbf{D}^{(i)}\mathbf{S} + \mathbf{E}^{(i)} \\
&\textbf{Assumptions :} && \mathcal{N}\left(\mathbf{e}^{(i)} \mid \mathbf{0}, \boldsymbol{\Psi}^{(i)}\right) \quad (1) \\
& && \mathcal{N}\left(\mathbf{s}_j \mid \mathbf{0}, \mathbf{I}\right) \quad (2) \\
& && \boldsymbol{\Psi}^{(i)}, \mathbf{D}_i \text{ diagonal} \quad (3)
\end{aligned}$$

The following holds:

$$p\left(\mathbf{s}_i \middle| \mathbf{A}, \left\{\mathbf{x}_i^{(m)}, \mathbf{D}^{(m)}, \boldsymbol{\Psi}^{(m)}\right\}_{m=1}^{M}\right) \propto \exp\left(\mathbf{s}_i{}^T \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} - \frac{1}{2}\mathbf{s}_t^T \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A}\mathbf{D}^{(m)}\right)\mathbf{s}_i\right)$$

Proof:

$$p\left(\mathbf{s}_i \middle| \mathbf{A}, \left\{\mathbf{x}_i^{(m)}, \mathbf{D}^{(m)}, \boldsymbol{\Psi}^{(m)}\right\}_{m=1}^{M}\right) \underset{1}{=}$$

$$\frac{p\left(\mathbf{x}_i^{(m)} \middle| \mathbf{s}_i, \mathbf{A}, \left\{\mathbf{D}^{(m)}, \boldsymbol{\Psi}^{(m)}\right\}_{m=1}^{M}\right) p\left(\mathbf{s}_i \middle| \mathbf{A}, \left\{\mathbf{D}^{(m)}, \boldsymbol{\Psi}_j^{(m)}\right\}_{m=1}^{M}\right)}{p\left(\mathbf{x}_i^{(m)} \middle| \mathbf{A}, \left\{\mathbf{D}^{(m)}, \boldsymbol{\Psi}_j^{(m)}\right\}_{m=1}^{M}\right)} \underset{2}{\propto}$$

$$p\left(\mathbf{x}_i^{(m)} \middle| \mathbf{s}_i, \mathbf{A}, \left\{\mathbf{D}^{(m)}, \boldsymbol{\Psi}^{(m)}\right\}_{m=1}^{M}\right) \cdot p\left(\mathbf{s}_i \middle| \mathbf{A}, \left\{\mathbf{D}^{(m)}, \boldsymbol{\Psi}^{(m)}\right\}_{m=1}^{M}\right) \underset{3}{\propto}$$

$$\exp\left(-\frac{1}{2}\sum_{m=1}^{M}\left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T \boldsymbol{\Psi}^{(m)-1}\left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right) \exp\left(-\frac{1}{2}\left(\mathbf{s}_i - \mathbf{0}\right)^T \mathbf{I}^{-1}\left(\mathbf{s}_i - \mathbf{0}\right)\right) =$$

$$\exp\left(-\frac{1}{2}\sum_{m=1}^{M}\left(\mathbf{x}_i^{(m)T} \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} + \left(\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i - \mathbf{x}_i^{(m)T} \boldsymbol{\Psi}^{(m)-1} \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i - \left(\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)}\right) - \frac{1}{2}\mathbf{s}_i^T \mathbf{I}^{-1} \mathbf{s}_i\right) \underset{4}{\propto}$$

$$\exp\left(\mathbf{s}_i{}^T \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} - \frac{1}{2}\mathbf{s}_t^T \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A}\mathbf{D}^{(m)}\right)\mathbf{s}_i\right)$$

1. Follows from Bayes theorem
2. Denominator is a normalization constant
3. Follows from assumption 1 and 2
4. Result of the fact that $\mathbf{x}_i^{(m)T} \boldsymbol{\Psi}^{(m)-1} \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i = \left(\mathbf{x}_i^{(m)T} \boldsymbol{\Psi}^{(m)-1} \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T = \mathbf{s}_i^T \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)}$ and $\sum_{m=1}^{M} \mathbf{x}_i^{(m)T} \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} = \text{Constant}$

Theorem 3 (The E step of PARAFAC)

Given the model, assumptions and result of Theorem 2 the following holds:

$$\boldsymbol{\Sigma}_{\mathbf{S}} = \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)} \right)^{-1}$$

$$\langle \mathbf{S} \rangle = \boldsymbol{\Sigma}_{\mathbf{S}} \left(\sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{X}_m \right)$$

$$\left\langle \mathbf{S}\mathbf{S}^T \right\rangle = N\boldsymbol{\Sigma}_{\mathbf{S}} + \langle \mathbf{S} \rangle \langle \mathbf{S} \rangle^T$$

Proof:

As the posterior distribution of **S** i.e. the distribution of **S** conditioned on the model parameter is Gaussian distributed we have:

$$p(\mathbf{S}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p\left(\mathbf{s}_i \middle| \boldsymbol{\theta}\right) \propto \prod_{i=1}^{N} \exp\left(-\frac{1}{2} \left(\mathbf{s}_i - \langle \mathbf{s}_i \rangle \right)^T \Sigma_{\mathbf{S}}^{-1} \left(\mathbf{s}_i - \langle \mathbf{s}_i \rangle \right) \right)$$

$$= \exp\left(\sum_{i=1}^{N} \left(-\frac{1}{2} \mathbf{s}_i^T \Sigma_{\mathbf{S}}^{-1} \mathbf{s}_i + \mathbf{s}_i^T \Sigma_{\mathbf{S}}^{-1} \langle \mathbf{s}_i \rangle - \frac{1}{2} \langle \mathbf{s}_i \rangle^T \Sigma_{\mathbf{S}}^{-1} \langle \mathbf{s}_i \rangle \right) \right)$$

$$\underbrace{\propto_1 \exp\left(\sum_{i=1}^{N} \left(-\frac{1}{2} \mathbf{s}_i^T \Sigma_{\mathbf{S}}^{-1} \mathbf{s}_i + \mathbf{s}_i^T \Sigma_{\mathbf{S}}^{-1} \langle \mathbf{s}_i \rangle \right) \right)}_{result\,1}$$

1. Follows as $\frac{1}{2} \langle \mathbf{S} \rangle^T \Sigma_{\mathbf{S}}^{-1} \langle \mathbf{S} \rangle$ is a constant

As:

$$p\left(\mathbf{s}_i \middle| \mathbf{A}, \left\{ \mathbf{x}_i^{(m)}, \mathbf{D}^{(m)}, \boldsymbol{\Psi}^{(m)} \right\}_{m=1}^{M} \right) \propto \exp\left(\mathbf{s}_i^T \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} - \frac{1}{2} \mathbf{s}_i^T \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)} \right) \mathbf{s}_i \right)$$

It follows

$$p(\mathbf{S} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p\left(\mathbf{s}_i \middle| \mathbf{A}, \left\{ \mathbf{x}_i^{(m)}, \mathbf{D}^{(m)}, \boldsymbol{\Psi}^{(m)} \right\}_{m=1}^{M} \right) \propto$$

$$\prod_{i=1}^{N} \exp\left(\mathbf{s}_i^T \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} - \frac{1}{2} \mathbf{s}_i^T \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)} \right) \mathbf{s}_i \right) =$$

$$\underbrace{\exp\left(\sum_{i=1}^{N} \left(\mathbf{s}_i^T \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{x}_i^{(m)} - \frac{1}{2} \mathbf{s}_i^T \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}^{(m)} \mathbf{A}^T \boldsymbol{\Psi}^{(m)-1} \mathbf{A} \mathbf{D}^{(m)} \right) \mathbf{s}_i \right) \right)}_{result\,2}$$

When comparing result 1 with result 2, it is immediately seen that:

$$\boldsymbol{\Sigma}_{\mathbf{S}} = \left(\mathbf{I} + \sum_{m=1}^{M} \mathbf{D}_m \mathbf{A}^T \boldsymbol{\Psi}_m^{-1} \mathbf{A} \mathbf{D}_m \right)^{-1}$$

$$\langle \mathbf{S} \rangle = \boldsymbol{\Sigma}_{\mathbf{S}} \left(\sum_{m=1}^{M} \mathbf{D}_m \mathbf{A}^T \boldsymbol{\Psi}_m^{-1} \mathbf{X}_m \right)$$

The last part of the theorem follows by noting that:

$$Cov(\mathbf{X}) = \int (\mathbf{X} - \langle \mathbf{X} \rangle)(\mathbf{X} - \langle \mathbf{X} \rangle)^T \, p(\mathbf{X}) d\mathbf{X} =$$

$$\int \mathbf{X}\mathbf{X}^T p(\mathbf{X}) d\mathbf{X} - \int \mathbf{X} \langle \mathbf{X} \rangle^T p(\mathbf{X}) d\mathbf{X} - \int \langle \mathbf{X} \rangle \mathbf{X}^T p(\mathbf{X}) d\mathbf{X} + \int \langle \mathbf{X} \rangle \langle \mathbf{X} \rangle^T p(\mathbf{X}) d\mathbf{X} =$$

$$\langle \mathbf{X}\mathbf{X}^T \rangle - \langle \mathbf{X} \rangle \langle \mathbf{X} \rangle^T - \langle \mathbf{X} \rangle \langle \mathbf{X} \rangle^T + \langle \mathbf{X} \rangle \langle \mathbf{X} \rangle^T = \langle \mathbf{X}\mathbf{X}^T \rangle - \langle \mathbf{X} \rangle \langle \mathbf{X} \rangle^T$$

As each $\mathbf{s}_i$ has the covariance $\sum_{\mathbf{S}}$ it follows $\mathrm{cov}(\mathbf{S}) = N \sum_{\mathbf{S}}$. Therefore:

$$Cov(\mathbf{S}) = \langle \mathbf{S}\mathbf{S}^T \rangle - \langle \mathbf{S} \rangle \langle \mathbf{S} \rangle^T \Leftrightarrow \langle \mathbf{S}\mathbf{S}^T \rangle = Cov(\mathbf{S}) + \langle \mathbf{S} \rangle \langle \mathbf{S} \rangle^T = N \sum_{\mathbf{S}} + \langle \mathbf{S} \rangle \langle \mathbf{S} \rangle^T$$

Theorem 4 : (The likelihood of EMPARAFAC)

Given the model with assumptions of Theorem 2 the following holds:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{N}{2}\sum_{m=1}^{M}\ln\left|\boldsymbol{\Psi}^{(m)}\right| - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\mathbf{x}_i^{m\,T}\boldsymbol{\Psi}^{(m)-1}\mathbf{x}_i^{(m)} + tr\left(\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\left\langle \mathbf{s}\mathbf{s}_i\right\rangle\right) - 2\mathbf{x}_i^{(m)T}\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\left\langle \mathbf{s}_i\right\rangle + const$$

Proof:

To prove the statement we make use the following result:

$$\left\langle \mathbf{s}_i^{\,T}\mathbf{W}\mathbf{s}_i\right\rangle = tr\left(\mathbf{W}\,\mathrm{cov}(\mathbf{s}_i,\mathbf{s}_i)\right) + \left\langle \mathbf{s}_i^{\,T}\right\rangle\mathbf{W}\left\langle \mathbf{s}_i\right\rangle \qquad (*)$$

Writing out the likelihood we get:

$$\mathcal{L}(\boldsymbol{\theta}) \underset{1}{=} \left\langle \ln\prod_{i=1}^{N}\prod_{m=1}^{M}(2\pi)^{-d/2}\left|\boldsymbol{\Psi}^{(m)}\right|^{-\frac{1}{2}}\exp\left[\left(\mathbf{x}_i^{(m)}-\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T\boldsymbol{\Psi}^{(m)-1}\left(\mathbf{x}_i^{(m)}-\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right]\right\rangle =$$

$$-\frac{dMN}{2}\ln(2\pi) - \frac{N}{2}\sum_{m=1}^{M}\ln\left|\boldsymbol{\Psi}^{(m)}\right| - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\left(\mathbf{x}_i^{m\,T}\boldsymbol{\Psi}^{(m)-1}\mathbf{x}_i^{(m)} + \right.$$

$$\left.\left\langle \mathbf{s}_i^{\,T}\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right\rangle - \left\langle 2\mathbf{x}_i^{(m)T}\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right\rangle\right) \underset{2}{=}$$

$$-\frac{dMN}{2}\ln(2\pi) - \frac{N}{2}\sum_{m=1}^{M}\ln\left|\boldsymbol{\Psi}^{(m)}\right| - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\left(\mathbf{x}_i^{m\,T}\boldsymbol{\Psi}^{(m)-1}\mathbf{x}_i^{(m)} + \right.$$

$$\left.\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\,\mathrm{cov}(\mathbf{s}_i,\mathbf{s}_i) + \left\langle \mathbf{s}_i^{\,T}\right\rangle\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\left\langle \mathbf{s}_i\right\rangle - 2\mathbf{x}_i^{(m)T}\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\left\langle \mathbf{s}_i\right\rangle\right) \underset{3}{=}$$

$$-\frac{N}{2}\sum_{m=1}^{M}\ln\left|\boldsymbol{\Psi}^{(m)}\right| - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\left(\mathbf{x}_i^{m\,T}\boldsymbol{\Psi}^{(m)-1}\mathbf{x}_i^{(m)} + tr\left(\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\left\langle \mathbf{s}_i\mathbf{s}_i^{\,T}\right\rangle\right) - 2\mathbf{x}_i^{(m)T}\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\mathbf{D}^{(m)}\left\langle \mathbf{s}_i\right\rangle\right) + const$$

1. Follows from assumption 1, the expectation is taken as the expectation of **S**, making the likelihood only depend on **θ**.
2. Follow from assumption 3.
3. Follows from (*) and assumption 2.

Theorem 5 (The M step of PARAFAC)

Given the likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{N}{2}\sum_{m=1}^{M}\ln\left|\boldsymbol{\Psi}^{(m)}\right| - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\mathbf{x}_i^{m\,T}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{x}_i^{(m)} + tr\left(\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}\mathbf{s}_i\right\rangle\right) - 2\mathbf{x}_i^{(m)^T}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\right\rangle + const$$

The following holds:

$$\mathbf{a}_k = \left(\sum_{m=1}^{M}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{X}^{(m)}\left\langle\mathbf{S}\right\rangle^T\mathbf{D}^{(m)}\right)_k\left(\sum_{m=1}^{M}\left(\boldsymbol{\Psi}^{(m)^{-1}}\right)_{kk}\mathbf{D}^{(m)}\left\langle\mathbf{S}\mathbf{S}^T\right\rangle^T\mathbf{D}^{(m)}\right)^{-1}$$

$$\mathbf{d}_m = \left(\left\langle\mathbf{S}\mathbf{S}^T\right\rangle \bullet \left(\mathbf{A}^T\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\right)\right)^{-1} vec\left[diag\left[\mathbf{A}^T\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{X}^{(m)}\left\langle\mathbf{S}\right\rangle^T\right]\right]$$

$$\boldsymbol{\Psi}^{(m)} = \frac{1}{N}diag\left[\mathbf{X}^{(m)}\mathbf{X}^{(m)T} + \mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{S}\mathbf{S}^T\right\rangle\mathbf{D}^{(m)}\mathbf{A}^T - 2\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{S}\right\rangle\mathbf{X}^{(m)T}\right]$$

proof:

$$\mathbf{a}_k = \left(\sum_{m=1}^{M}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{X}^{(m)}\left\langle\mathbf{S}\right\rangle^T\mathbf{D}^{(m)}\right)_k\left(\sum_{m=1}^{M}\left(\boldsymbol{\Psi}^{(m)^{-1}}\right)_{kk}\mathbf{D}^{(m)}\left\langle\mathbf{S}\mathbf{S}^T\right\rangle^T\mathbf{D}^{(m)}\right)^{-1}$$

$$\frac{\partial\mathcal{L}(\boldsymbol{\theta})}{\partial\mathbf{A}} = \frac{\partial\left(\sum_{i=1}^{N}\sum_{m=1}^{M}tr\left(\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\mathbf{s}_i^T\right\rangle\right) - 2\mathbf{x}_i^{(m)^T}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\right\rangle\right)}{\partial\mathbf{A}} =_1$$

$$\frac{\partial\left(\sum_{i=1}^{N}\sum_{m=1}^{M}tr\left(\mathbf{A}^T\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\mathbf{s}_i^T\right\rangle\mathbf{D}^{(m)}\right) - 2\mathbf{x}_i^{(m)^T}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\right\rangle\right)}{\partial\mathbf{A}} =_2$$

$$\sum_{i=1}^{N}\sum_{m=1}^{M}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\mathbf{s}_i^T\right\rangle\mathbf{D}^{(m)} + \boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\left(\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\mathbf{s}_i^T\right\rangle\mathbf{D}^{(m)}\right)^T - 2\boldsymbol{\Psi}^{(m)^{-1}}2\mathbf{x}_i^{(m)}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\right\rangle^T =_3$$

$$\sum_{i=1}^{N}\sum_{m=1}^{M}2\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\mathbf{s}_i^T\right\rangle\mathbf{D}^{(m)} - 2\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{x}_i^{(m)}\left\langle\mathbf{s}_i\right\rangle^T\mathbf{D}^{(m)} \Rightarrow$$

$$\sum_{i=1}^{N}\sum_{m=1}^{M}2\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{s}_i\mathbf{s}_i^T\right\rangle\mathbf{D}^{(m)} - 2\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{x}_i^{(m)}\left\langle\mathbf{s}_i\right\rangle^T\mathbf{D}^{(m)} = 0 \Leftrightarrow_4$$

$$\sum_{m=1}^{M}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{A}\mathbf{D}^{(m)}\left\langle\mathbf{S}\mathbf{S}^T\right\rangle\mathbf{D}^{(m)} = \sum_{m=1}^{M}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{X}^{(m)}\left\langle\mathbf{S}\right\rangle^T\mathbf{D}^{(m)} \Rightarrow$$

$$\sum_{m=1}^{M}\boldsymbol{\Psi}_{kk}^{(m)^{-1}}\mathbf{a}_k\mathbf{D}^{(m)}\left\langle\mathbf{S}\mathbf{S}^T\right\rangle\mathbf{D}^{(m)} = \left(\sum_{m=1}^{M}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{X}^{(m)}\left\langle\mathbf{S}\right\rangle^T\mathbf{D}^{(m)}\right)_k \Leftrightarrow$$

$$\mathbf{a}_k = \left(\sum_{m=1}^{M}\boldsymbol{\Psi}^{(m)^{-1}}\mathbf{X}^{(m)}\left\langle\mathbf{S}\right\rangle^T\mathbf{D}^{(m)}\right)_k\left(\sum_{m=1}^{M}\boldsymbol{\Psi}_{kk}^{(m)^{-1}}\mathbf{D}^{(m)}\left\langle\mathbf{S}\mathbf{S}^T\right\rangle\mathbf{D}^{(m)}\right)^{-1}$$

1. follows as tr(**ABC**)=tr(**BCA**)=tr(**CAB**)

2. result of $\frac{\partial tr\left(\mathbf{A}^T\mathbf{BAC}\right)}{\partial \mathbf{A}} = \mathbf{BAC} + \mathbf{B}^T\mathbf{AC}^T$ and $\frac{\partial \mathbf{b}^T\mathbf{Ac}}{\partial \mathbf{A}} = \mathbf{bc}^T$
3. comes by $\left(\mathbf{D}^{(m)}\left\langle \mathbf{s}_i\mathbf{s}_i{}^T\right\rangle \mathbf{D}^{(m)}\right)^T = \left(\mathbf{D}^{(m)}\left\langle \mathbf{s}_i\mathbf{s}_i{}^T\right\rangle \mathbf{D}^{(m)}\right)$
4. consequence of $\sum_{i=1}^{N}\left\langle \mathbf{s}_i\mathbf{s}_i{}^T\right\rangle = N\left\langle \mathbf{SS}_i{}^T\right\rangle$ and $\sum_{i=1}^{N}\mathbf{x}_i^{(m)}\left\langle \mathbf{s}_i\right\rangle^T = N\mathbf{X}^{(m)}\left\langle \mathbf{S}\right\rangle^T$

$$\mathbf{d}_m = \left(\left\langle \mathbf{SS}^T\right\rangle \bullet \left(\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\right)\right)^{-1} vec\left[diag\left[\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{X}^{(m)}\left\langle \mathbf{S}\right\rangle^T\right]\right]$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{D}^{(m)}} = \frac{\partial\left(\sum_{i=1}^{N}\sum_{m=1}^{M} tr\left(\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{AD}^{(m)}\left\langle \mathbf{s}_i\mathbf{s}_i{}^T\right\rangle\right) - 2\mathbf{x}_i^{(m)T}\boldsymbol{\Psi}^{(m)-1}\mathbf{AD}^{(m)}\left\langle \mathbf{s}_i\right\rangle\right)}{\partial \mathbf{D}^{(m)}} =$$

$$\sum_{i=1}^{N}\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{AD}^{(m)}\left\langle \mathbf{s}_i\mathbf{s}_i{}^T\right\rangle + \left(\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\right)^T\mathbf{D}^{(m)}\left\langle \mathbf{s}_i\mathbf{s}_i{}^T\right\rangle^T - 2\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{x}_i^{(m)}\left\langle \mathbf{s}_i\right\rangle^T =$$

$$2\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{AD}^{(m)}\left\langle \mathbf{SS}^T\right\rangle - 2\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{X}^{(m)}\left\langle \mathbf{S}\right\rangle^T = 0 \Rightarrow$$

$$\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{AD}^{(m)}\left\langle \mathbf{SS}^T\right\rangle = \mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{X}^{(m)}\left\langle \mathbf{S}\right\rangle^T \Rightarrow$$

$$\mathbf{d}_m = \left(\left\langle \mathbf{SS}^T\right\rangle \bullet \left(\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{A}\right)\right)^{-1} vec\left(diag\left[\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{X}^{(m)}\left\langle \mathbf{S}\right\rangle^T\right]\right)$$

where $\mathbf{D}^{(m)} = diag(\mathbf{d}_m)$

$$\boldsymbol{\Psi}^{(m)} = \frac{1}{N}diag\left[\mathbf{X}^{(m)}\mathbf{X}^{(m)T} + \mathbf{AD}^{(m)}\left\langle \mathbf{SS}^T\right\rangle \mathbf{D}^{(m)}\mathbf{A}^T - 2\mathbf{AD}^{(m)}\left\langle \mathbf{S}\right\rangle \mathbf{X}^{(m)T}\right]$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\Psi}^{(m)-1}} = \frac{\partial -\frac{N}{2}\sum_{m=1}^{M}\ln\left|\boldsymbol{\Psi}^{(m)}\right| - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\left(\mathbf{x}_i^{mT}\boldsymbol{\Psi}^{(m)-1}\mathbf{x}_i^{(m)} + tr\left(\mathbf{D}^{(m)}\mathbf{A}^T\boldsymbol{\Psi}^{(m)-1}\mathbf{AD}^{(m)}\left\langle \mathbf{ss}_i\right\rangle\right) - 2\mathbf{x}_i^{(m)T}\boldsymbol{\Psi}^{(m)-1}\mathbf{AD}^{(m)}\left\langle \mathbf{s}_i\right\rangle\right)}{\partial \boldsymbol{\Psi}^{(m)-1}} =$$

$$-\frac{N}{2}\boldsymbol{\Psi}^{(m)} - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\mathbf{x}_i^{(m)}\mathbf{x}_i^{(m)T} + \mathbf{AD}^{(m)}\left\langle \mathbf{s}_i\mathbf{s}_i{}^T\right\rangle \mathbf{D}^{(m)}\mathbf{A}^T - 2\mathbf{AD}^{(m)}\left\langle \mathbf{s}_i\right\rangle \mathbf{x}_i^{(m)T} = 0 \Rightarrow$$

$$\boldsymbol{\Psi}^{(m)} = \frac{1}{N}diag\left(\sum_{m=1}^{M} X^{(m)}\mathbf{X}^{(m)T} + \mathbf{AD}^{(m)}\left\langle \mathbf{SS}^T\right\rangle \mathbf{D}^{(m)}\mathbf{A}^T - 2\mathbf{AD}^{(m)}\left\langle \mathbf{S}\right\rangle \mathbf{X}^{(m)T}\right)$$

1. follows by: $\frac{\partial \ln\left|\mathbf{A}\right|}{\partial \mathbf{A}} = \left(\mathbf{A}^T\right)^{-1}$, $\frac{\partial \mathbf{b}^T\mathbf{A}^{-1}\mathbf{c}}{\partial \mathbf{A}} = -\mathbf{A}^{-\mathrm{T}}\mathbf{bc}^T\mathbf{A}^{-T}$ and
   $\frac{\partial tr\left(\mathbf{BA}^{-1}\mathbf{C}\right)}{\partial \mathbf{A}} = -\left(\mathbf{A}^{-1}\mathbf{CBA}^{-1}\right)^T$

Theorem 6 (BIC for EMPARAFAC)

The Bayesian Information Criterion (BIC) for the EMPARAFAC model is given by:

$$p(\mathbf{D} \mid \mathrm{F}\ ) \approx \left(\prod_{m=1}^{M} \left((2\pi)^{-d/2} \left| \mathbf{\Psi}^{(m)} \right|^{-\frac{1}{2}} \right)^{N} \right) \exp\left(-0.5 \cdot \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right)^T \mathbf{\Psi}^{(m)-1} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right) \right) N^{-\frac{D}{2}}$$

F is the number of factors in the PARAFAC model, and $D = F(P+M)$ is the number of 'effectively' free parameters, where P is the number of rows of $\mathbf{X}^{(m)}$.

Proof:

From eq. 1.24 we have $p(\mathbf{D} \mid F) \approx p(\mathbf{D} \mid \boldsymbol{\theta}_{ML}, F) N^{-\frac{D}{2}}$. Combining this with eq. 1.53 we get:

$$p(\mathbf{D} \mid \boldsymbol{\theta}_{ML}, F) = \int p\left(\left\{ \mathbf{X}^{(m)} \right\}_{m=1}^{M} \mid \mathbf{S}, \mathbf{A}, \left\{ \mathbf{D}^{(m)}, \mathbf{\Psi}^{(m)} \right\}_{m=1}^{M} \right) p(\mathbf{S}) d\mathbf{S} =$$

$$\int \prod_{i=1}^{N} p\left(\left\{ \mathbf{x}_i^{(m)} \right\}_{m=1}^{M} \mid \mathbf{s_i}, \mathbf{A}, \left\{ \mathbf{D}^{(m)}, \mathbf{\Psi}^{(m)} \right\}_{m=1}^{M} \right) p(\mathbf{s_i}) d\mathbf{s_i} =$$

$$\left(\prod_{m=1}^{M} \left((2\pi)^{-d/2} \left| \mathbf{\Psi}^{(m)} \right|^{-\frac{1}{2}} \right)^{N} \right) \exp\left(-0.5 \cdot \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \langle \mathbf{s}_i \rangle \right)^T \mathbf{\Psi}^{(m)-1} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \langle \mathbf{s}_i \rangle \right) \right) \Rightarrow$$

$$p(\mathbf{D} \mid \boldsymbol{\theta}_{ML}, M) N^{-\frac{D}{2}} =$$

$$\left(\prod_{m=1}^{M} \left((2\pi)^{-d/2} \left| \mathbf{\Psi}^{(m)} \right|^{-\frac{1}{2}} \right)^{N} \right) \exp\left(-0.5 \cdot \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \langle \mathbf{s}_i \rangle \right)^T \mathbf{\Psi}^{(m)-1} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \langle \mathbf{s}_i \rangle \right) \right) N^{-\frac{D}{2}}$$

The true number of free parameters is $D = F(P+M) + PM$, however the free parameters of $\mathbf{\Psi}$ is not directly part of the factor model, and does consequently not help improving the fit corresponding to having PM parameters. Therefore, the free parameters are only considered to be $D = F(P+M)$. As N and D depends on which modalities M, N and P pertains to, the BIC measure is greatly dependent on how each of the 3-way-array's ways are defined in terms of the model $\mathbf{X}^{(i)} = \mathbf{A}\mathbf{D}^{(i)}\mathbf{S}$.

Theorem 7: (BIC for Least Square optimization of PARAFAC)

The Bayesian Information Criterion (BIC) for a least square optimization of PARAFAC as defined in eq. 1.30 is when assuming $\mathbf{X^{(m)}}$ and $\mathbf{S}$ to be i.i.d., given by:

$$\log(p(\mathbf{D} \mid F) \approx \log(p(\mathbf{D} \mid \boldsymbol{\theta}, F) N^{-\frac{D}{2}}) = -\tfrac{1}{2} MN \log(\sigma) - \tfrac{D}{2} \log N \quad + const$$

F is the number of factors in the PARAFAC model, and $D = F(P + M + N)$ is the number of free parameters, where P is the number of rows of $\mathbf{X}^{(m)}$.

Proof:

$$p(\mathbf{D} \mid \boldsymbol{\theta}, F) = p\left(\left\{ \mathbf{X}^{(m)} \right\}_{m=1}^{M} \mid \mathbf{S}, \mathbf{A}, \sigma, \left\{ \mathbf{D}^{(m)} \right\}_{m=1}^{M} \right) =$$

$$\prod_{i=1}^{N} p\left(\left\{ \mathbf{x}_i^{(m)} \right\}_{m=1}^{M} \mid \mathbf{s}_i, \mathbf{A}, \sigma, \left\{ \mathbf{D}^{(m)} \right\}_{m=1}^{M} \right) =$$

$$\left(\prod_{m=1}^{M} \left((2\pi)^{-d/2} \left| \sigma \cdot \mathbf{I} \right|^{-\frac{1}{2}} \right)^{N} \right) \exp\left(-0.5 \cdot \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right)^T (\sigma \cdot \mathbf{I})^{-1} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right) \right) \Rightarrow$$

$$p(\mathbf{D} \mid \boldsymbol{\theta}, M) N^{-\frac{D}{2}} =$$

$$\left(\prod_{m=1}^{M} \left((2\pi)^{-d/2} \sigma^{-\frac{1}{2}} \right)^{N} \right) \exp\left(-0.5 \cdot \frac{1}{\sigma} \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right)^T \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right) \right) N_0^{-\frac{D}{2}} =$$

$$\left(\prod_{m=1}^{M} \left((2\pi)^{-d/2} \sigma^{-\frac{1}{2}} \right)^{N} \right) \exp(-0.5NM) N_0^{-\frac{D}{2}} \Rightarrow$$

$$\log(p(\mathbf{D} \mid \boldsymbol{\theta}, M) N^{-\frac{D}{2}}) = -\tfrac{1}{2} MN \log(\sigma) - \tfrac{D}{2} \log N \quad + const$$

1. Consequence of $\sigma = \dfrac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right)^T \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right)$

As N depends on which modality is considered observations, the BIC measure is greatly dependent on how each of the 3-way-array's ways are defined in terms of the model $\mathbf{X}^{(i)} = \mathbf{A}\mathbf{D}^{(i)}\mathbf{S}$.

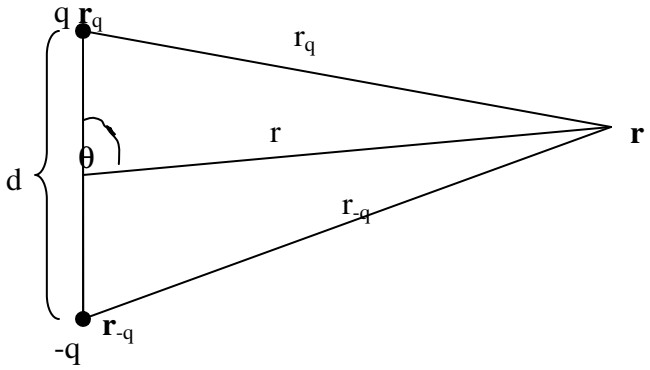
Theorem 8: (Potential of a dipole field)

The potential in a distance r of the center of a dipole, where the charges (q and –q) are separated by the distance d is given by:

$$\mathbf{\Phi}_{dipole}(\mathbf{r},\mathbf{d}) = \frac{1}{4\pi\varepsilon_0}\frac{qd\cos\theta}{\|\mathbf{r}\|^2}$$

where θ is the angle between $\mathbf{r}$ and the midpoint of the line connecting the two charges.

Proof:



$$r_q = \sqrt{r^2 + \tfrac{d}{2}^2 - 2r\tfrac{d}{2}\cos\theta} = \sqrt{r^2 + \tfrac{d}{2}^2 - rd\cos\theta}$$

$$r_{-q} = \sqrt{r^2 + \tfrac{d}{2}^2 - 2r\tfrac{d}{2}\cos(\pi-\theta)} = \sqrt{r^2 + \tfrac{d}{2}^2 + rd\cos(\theta)}$$

$$r = \|\mathbf{r}\|$$

$$d = \|\mathbf{d}\|$$

$$\left\|\vec{\mathbf{E}}(\mathbf{r},\mathbf{d})\right\| = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{r^2 + \tfrac{d}{2}^2 - rd\cos\theta} - \frac{q}{r^2 + \tfrac{d}{2}^2 + rd\cos(\theta)}\right) =$$

$$\frac{q}{4\pi\varepsilon_0}\left(\frac{2rd\cos\theta}{\left(r^2 + \tfrac{d}{2}^2 - rd\cos\theta\right)\left(r^2 + \tfrac{d}{2}^2 + rd\cos(\theta)\right)}\right) \overset{1}{\approx}$$

$$\frac{q}{4\pi\varepsilon_0}\left(\frac{2rd\cos\theta}{r^4}\right) = \frac{q}{4\pi\varepsilon_0}\left(\frac{2d\cos\theta}{r^3}\right)$$

$$-\nabla\mathbf{\Phi}(\mathbf{r},\mathbf{d}) = \vec{\mathbf{E}}(\mathbf{r},\mathbf{d}) \Rightarrow$$

$$\mathbf{\Phi}(\mathbf{r},\mathbf{d}) = \frac{1}{4\pi\varepsilon_0}\frac{qd\cos\theta}{r^2} = \frac{1}{4\pi\varepsilon_0}\frac{qd\cos\theta}{r^2}$$

1. Follows from the fact that $r >> d$

Theorem 9 (Derivation of the VBEM algorithm for PARAFAC)

For the PARAFAC model of eq. 1.30, the following assumptions are made:

$$p(\mathbf{E}^{(m)}) = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{e}_i{}^{(m)} \mid 0, diag(\varphi_m)\right)$$

$$p(\mathbf{S}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{s}_i \mid 0, \mathbf{I})$$

$$p(\mathbf{A} \mid \alpha) = \prod_{f=1}^{F} \mathcal{N}\left(\mathbf{a}_f \mid 0, \tfrac{1}{\alpha_f}\mathbf{I}\right)$$

$$p\left(\alpha \mid a^{\alpha}, b^{\alpha}\right) = \prod_{f=1}^{F} \mathcal{G}\left(\alpha_f \mid a_f^{\alpha}, b_f^{\alpha}\right)$$

$$p\left(\mathbf{d}_m \mid \boldsymbol{\mu}_m, \tfrac{1}{v_m}\right) = \mathcal{N}\left(\mathbf{d}_m \mid \boldsymbol{\mu}_m, \tfrac{1}{v_m}\mathbf{I}\right)$$

$$p\left(\varphi_m \mid \mathbf{a}_m^{\varphi}, \mathbf{b}_m^{\varphi}\right) = \prod_{p=1}^{P} \mathcal{G}\left(\varphi_{mp} \mid a_{mp}^{\varphi}, b_{mp}^{\varphi}\right)$$

where $\mathbf{d}_m$ corresponds to the diagonal of $\mathbf{D}^{(m)}$

These yields the following update rules:

$$\boldsymbol{\Sigma}_{\mathbf{S}} = \left(\mathbf{I} + \sum_{m=1}^{M} \left\langle \mathbf{d}_m{}^T \mathbf{d}_m \right\rangle \bullet \sum_{p=1}^{P} \left\langle \varphi_{m\,p} \right\rangle \left\langle \mathbf{a}_p{}^T \mathbf{a}_p \right\rangle\right)^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{s_i}} = \boldsymbol{\Sigma}_{\mathbf{S}} \sum_{m=1}^{M} \left\langle \mathbf{d}_m{}^T \right\rangle \left\langle \mathbf{A}^T \right\rangle \langle \varphi_m \rangle \mathbf{d}_i$$

$$\boldsymbol{\Sigma}_{\mathbf{a}}{}^{p} = \left\langle \left(\alpha \mathbf{I} + \sum_{i=1}^{N}\sum_{m=1}^{M} \mathbf{s}_i{}^T \mathbf{D}^{(m)} \varphi_j{}^{(m)} \mathbf{D}^{(m)} \mathbf{s}_i\right)\right\rangle^{-1} = \left(\langle \alpha \rangle \mathbf{I} + \sum_{m=1}^{M} \left\langle \varphi^{(m)} \right\rangle \left\langle \mathbf{d}^{(m)} \mathbf{d}^{(m)T} \right\rangle \bullet \sum_{i=1}^{N} \left\langle \mathbf{s}_i \mathbf{s}_i{}^T \right\rangle\right)^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{a}}{}^{p} = \boldsymbol{\Sigma}_{\mathbf{a}}{}^{p} \left\langle \sum_{i=1}^{N}\sum_{m=1}^{M} \varphi_p{}^{(m)} \mathbf{x}_i{}^{(m)T} \mathbf{D}^{(m)} \mathbf{s}_i \right\rangle = \boldsymbol{\Sigma}_{\mathbf{a}}{}^{p} \left\langle \sum_{m=1}^{M} \left\langle \varphi_p{}^{(m)} \right\rangle \left\langle \mathbf{D}^{(m)} \right\rangle \sum_{i=1}^{N} \mathbf{x}_i{}^{(m)T} \langle \mathbf{s}_i \rangle \right\rangle$$

$$\boldsymbol{\Sigma}_{\mathbf{D}^{(m)}} = \left(v_m \mathbf{I} + \sum_{j=1}^{P} \left\langle \varphi_{mj} \right\rangle \left\langle \mathbf{a}_{:,j}{}^T \mathbf{a}_{:,j} \right\rangle \bullet \sum_{i=1}^{N} \left\langle \mathbf{s}_i{}^T \mathbf{s}_i \right\rangle\right)^{-1}$$

$$m_{\mathbf{D}^{(m)}} = \boldsymbol{\Sigma}_{\mathbf{D}^{(m)}} \left(v_m \boldsymbol{\mu}_m + \sum_{i=1}^{N}\sum_{j=1}^{D} \left\langle x_{ij}{}^{(m)} \right\rangle \varphi_{mj} \langle \mathbf{s}_i \rangle \bullet \langle \mathbf{a}_{:,j} \rangle \right)$$

$$\hat{a}_f^{\alpha} = a_f^{\alpha} + \frac{p}{2}$$

$$\hat{b}_f^{\alpha} = b_f^{\alpha} + \frac{\left\langle \mathbf{a}_f{}^T \mathbf{a}_f \right\rangle}{2}$$

$$\hat{a}_{mp}^{\varphi} = a_{mp}^{\varphi} + \frac{N}{2}$$

$$\hat{b}_{mp}^{\varphi} = b_{mp}^{\varphi} + \frac{1}{2}\sum_{i=1}^{N} \left\langle \left(x_{ip}{}^{(m)} - \mathbf{a}_{:,p} \mathbf{D}^{(m)} \mathbf{s}_i\right)^2 \right\rangle$$

Proof:

$$p(\mathbf{x}_i^{(m)}, \mathbf{s}_i, \mathbf{D}^{(m)}, \varphi_m, \mathbf{A}, \alpha) = \underbrace{k_1 |\varphi_m|^{1/2} \exp\left(-\tfrac{1}{2} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right)^T \varphi_m \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right) \right)}_{p\left(\mathbf{d}_i | \mathbf{s}_i, \mathbf{C}^{(m)}, \varphi_m, \alpha\right)} \cdot$$

$$\underbrace{k_2 \exp\left(-\tfrac{1}{2} (\mathbf{s}_i)^T \mathbf{I} (\mathbf{s}_i) \right)}_{p(\mathbf{s}_i)} \cdot$$

$$\underbrace{k_3 \left| \tfrac{1}{v_m} \mathbf{I} \right|^{-1/2} \exp\left(-\tfrac{1}{2} (\mathbf{d}_m - \boldsymbol{\mu}_m)^T (v_m \mathbf{I}) (\mathbf{d}_m - \boldsymbol{\mu}_m) \right)}_{p(\mathbf{D}_m)} \cdot$$

$$\underbrace{\prod_{p=1}^{P} \frac{1}{\Gamma\left(a_{mp}^{\varphi}\right) \frac{1}{b_{mp}^{\varphi}}^{a_{mp}^{\varphi}}} \varphi_{mp}^{\,a_{mp}^{\varphi} - 1} \exp\left(-\varphi_{mp} b_{mp}^{\varphi} \right)}_{p(\varphi_m)} \cdot$$

$$\underbrace{\prod_{f=1}^{F} k_4 \left| \tfrac{1}{\alpha_f} \mathbf{I} \right|^{-1/2} \exp\left(-\tfrac{1}{2} (\mathbf{a}_f)^T \alpha \mathbf{I} (\mathbf{a}_f) \right)}_{p(\mathbf{a}_f | \alpha)} \cdot$$

$$\underbrace{\prod_{f=1}^{F} \frac{1}{\Gamma\left(a_f^{\alpha}\right) \frac{1}{b_f^{\alpha}}^{a_f^{\alpha}}} \alpha_f^{\,a_f^{\alpha} - 1} \exp\left(-\alpha_f b_f^{\alpha} \right)}_{p(\alpha)}$$

$$\ln p(\mathbf{X}, \mathbf{S}, \left\{ \mathbf{D}^{(m)}, \varphi_m \right\}_{m=1}^{M}, \mathbf{A}, \alpha) = \sum_{i=1}^{N} \left(\underbrace{\sum_{m=1}^{M} \tfrac{1}{2} \ln |\varphi_m| - \tfrac{1}{2} \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right)^T \varphi_m \left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)} \mathbf{s}_i \right)}_{p\left(\mathbf{x}_i^{(m)} | \mathbf{s}_i, \mathbf{D}_m, \varphi_m, \alpha\right)} \underbrace{- \tfrac{1}{2} (\mathbf{s}_i)^T \mathbf{I} (\mathbf{s}_i)}_{p(\mathbf{s}_i)} \right) -$$

$$\underbrace{\tfrac{1}{2} \ln \left| \tfrac{1}{v_m} \mathbf{I} \right| - \tfrac{1}{2} (\mathbf{d}_m - \boldsymbol{\mu}_m)^T v_m \mathbf{I} (\mathbf{d}_m - \boldsymbol{\mu}_m)}_{p\left(\mathbf{D}^{(m)}\right)} + \underbrace{\sum_{p=1}^{P} -\ln \Gamma\left(a_{mp}^{\varphi}\right) - a_{mp}^{\varphi} \ln \left(\frac{1}{b_{mp}^{\varphi}} \right) + \left(a_{mp}^{\varphi} - 1 \right) \ln \varphi_{mp} - \varphi_{mp} b_{mp}^{\varphi}}_{p(\varphi_m)} +$$

$$\underbrace{\sum_{f=1}^{F} -\tfrac{1}{2} \ln \left| \tfrac{1}{\alpha_f} \mathbf{I} \right| - \tfrac{1}{2} (\mathbf{a}_f)^T \alpha_f \mathbf{I} (\mathbf{a}_f)}_{p(\mathbf{a}_k | \alpha)} + \underbrace{\sum_{f=1}^{F} -\ln \Gamma\left(a_f^{\alpha}\right) - a_f^{\alpha} \ln \left(\frac{1}{b_f^{\alpha}} \right) + \left(a_f^{\alpha} - 1 \right) \ln \alpha_f - \alpha_f b_f^{\alpha}}_{p(\alpha)}$$

$$q(\mathbf{s}_i): \sum_{m=1}^{M}\left(-\tfrac{1}{2}\left(\mathbf{x}_i^{(m)}-\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T\varphi_m\left(\mathbf{x}_i^{(m)}-\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right)-\tfrac{1}{2}(\mathbf{s}_i)^T\mathbf{I}(\mathbf{s}_i)\propto$$

$$\sum_{m=1}^{M}\left(-\left(\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T\varphi_m\left(\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)-\mathbf{x}_i^{(m)}\varphi_m{}^{-1}\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)-(\mathbf{s}_i)^T\mathbf{I}(\mathbf{s}_i)=$$

$$\sum_{m=1}^{M}\left(-\mathbf{s}_i{}^T\mathbf{D}^{(m)T}\mathbf{A}^T\varphi_m\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i+\mathbf{x}_i^{(m)}\varphi_m{}^{-1}\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)-(\mathbf{s}_i)^T\mathbf{I}(\mathbf{s}_i)=$$

$$-\mathbf{s}_i{}^T\left(\mathbf{I}+\sum_{m=1}^{M}\mathbf{D}^{(m)T}\mathbf{A}^T\varphi_m\mathbf{A}\mathbf{D}^{(m)}\right)\mathbf{s}_i+\sum_{m=1}^{M}\mathbf{s}_i{}^T\mathbf{D}^{(m)T}\mathbf{A}^T\varphi_m\mathbf{x}_i^{(m)}\Rightarrow$$

$$\boldsymbol{\Sigma}_{\mathbf{S}}=\left(\mathbf{I}+\sum_{m=1}^{M}\left\langle\mathbf{D}^{(m)T}\mathbf{A}^T\varphi_m\mathbf{A}\mathbf{D}^{(m)}\right\rangle\right)^{-1}=\left(\mathbf{I}+\sum_{m=1}^{M}\left\langle\mathbf{d}_m{}^T\mathbf{d}_m\right\rangle\bullet\sum_{p=1}^{P}\left\langle\varphi_{mp}\right\rangle\left\langle\mathbf{a}_p{}^T\mathbf{a}_p\right\rangle\right)^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{s_i}}=\boldsymbol{\Sigma}_{\mathbf{S}}\sum_{m=1}^{M}\left\langle\mathbf{D}^{(m)T}\mathbf{A}^T\varphi_m\mathbf{x}_i^{(m)}\right\rangle=\boldsymbol{\Sigma}_{\mathbf{S}}\sum_{m=1}^{M}\left\langle\mathbf{D}^{(m)T}\right\rangle\left\langle\mathbf{A}^T\right\rangle\left\langle\varphi_m\right\rangle\mathbf{x}_i^{(m)}$$

1. $\left\langle\mathbf{D}^{(m)T}\mathbf{A}^T\varphi_m\mathbf{A}\mathbf{D}^{(m)}\right\rangle=\left\langle\mathbf{D}^{(m)T}\sum_{p=1}^{P}\mathbf{a}_p{}^T\varphi_{mp}\mathbf{a}_p\mathbf{D}^{(m)}\right\rangle=\left\langle\mathbf{d}_m{}^T\mathbf{d}_m\right\rangle\bullet\sum_{p=1}^{P}\left\langle\varphi_{mp}\right\rangle\left\langle\mathbf{a}_p{}^T\mathbf{a}_p\right\rangle$

$$q(\mathbf{A}):\sum_{i=1}^{N}\sum_{m=1}^{M}\left(-\left(\mathbf{x}_i{}^{(m)}-\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T\varphi_m\left(\mathbf{x}_i{}^{(m)}-\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right)-\sum_{f=1}^{F}\left(\mathbf{a}_f\right)^T\left(\alpha_f\mathbf{I}\right)\left(\mathbf{a}_f\right)=$$

$$\sum_{i=1}^{N}\sum_{m=1}^{M}-\mathbf{s}_i{}^T\mathbf{D}^{(m)T}\mathbf{A}^T\varphi_m\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i+2\mathbf{x}_i{}^{(m)T}\varphi_m\mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i-\sum_{f=1}^{F}\left(\mathbf{a}_f\right)^T\alpha\mathbf{I}\left(\mathbf{a}_f\right)=$$

$$-\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{p=1}^{P}\left(\mathbf{a}_{p,:}\left(\mathbf{s}_i{}^T\mathbf{D}^{(m)T}\varphi_{mp}\mathbf{D}^{(m)}\mathbf{s}_i\right)\mathbf{a}_{p,:}{}^T-2\mathbf{a}_{p,:}\varphi_{mp}\mathbf{x}_i{}^{(m)T}\mathbf{D}^{(m)}\mathbf{s}_i\right)-\sum_{f=1}^{F}\left(\mathbf{a}_f\right)^T\alpha\mathbf{I}\left(\mathbf{a}_f\right)=$$

$$\Rightarrow$$

$$\boldsymbol{\Sigma}_{\mathbf{a}}^{p}=\left\langle\left(\alpha\mathbf{I}+\sum_{i=1}^{N}\sum_{m=1}^{M}\mathbf{s}_i{}^T\mathbf{D}^{(m)}\varphi_{mp}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right\rangle^{-1}=\left(\langle\alpha\rangle\mathbf{I}+\sum_{m=1}^{M}\langle\varphi_m\rangle\left\langle\mathbf{d}_m\mathbf{d}_m{}^T\right\rangle\bullet\sum_{i=1}^{N}\left\langle\mathbf{s}_i\mathbf{s}_i{}^T\right\rangle\right)^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{a}}^{p}=\boldsymbol{\Sigma}_{\mathbf{a}}^{p}\left\langle\sum_{i=1}^{N}\sum_{m=1}^{M}\varphi_{mj}\mathbf{x}_i{}^{(m)T}\mathbf{D}^{(m)}\mathbf{s}_i\right\rangle=\boldsymbol{\Sigma}_{\mathbf{a}}^{p}\left\langle\sum_{m=1}^{M}\langle\varphi_{mj}\rangle\left\langle\mathbf{D}^{(m)}\right\rangle\sum_{i=1}^{N}\mathbf{x}_i{}^{(m)T}\langle\mathbf{s}_i\rangle\right\rangle$$

$$q(\boldsymbol{\alpha}_f):\underbrace{-\tfrac{1}{2}\ln\left|\tfrac{1}{\alpha_f}\mathbf{I}\right|-\tfrac{1}{2}\left(\mathbf{a}_f\right)^T\left(\alpha_f\mathbf{I}\right)\left(\mathbf{a}_f\right)}_{p(\mathbf{a}_f|\alpha_f)}\underbrace{-\ln\Gamma\left(a_f^{\alpha}\right)-a_f^{\alpha}\ln\left(\frac{1}{b_f^{\alpha}}\right)+\left(a_f^{\alpha}-1\right)\ln\alpha_f-\alpha_f b_f^{\alpha}}_{p(\alpha_f)}=$$

$$\frac{p}{2}\ln\alpha_f-\tfrac{1}{2}\alpha_f\left(\mathbf{a}_f\right)^T\left(\mathbf{a}_f\right)-\ln\Gamma\left(a_f^{\alpha}\right)-a_f^{\alpha}\ln\left(\frac{1}{b_f^{\alpha}}\right)+\left(a_f^{\alpha}-1\right)\ln\alpha_f-\alpha_f b_f^{\alpha}$$

$$\hat{a}_f^{\alpha}=a_f^{\alpha}+\frac{p}{2}$$

$$\hat{b}_f^{\alpha}=b_f^{\alpha}+\left\langle\frac{\left(\mathbf{a}_f\right)^T\left(\mathbf{a}_f\right)}{2}\right\rangle$$

$$\left\langle\boldsymbol{\alpha}_f\right\rangle=\frac{\hat{a}_f^{\alpha}}{\hat{b}_f^{\alpha}}$$

$$q(\varphi_{mp}): \sum_{i=1}^{N}\left(\tfrac{1}{2}\ln\varphi_{mj} - \tfrac{1}{2}\left(x_{ij}^{(m)} - \mathbf{a}_{:,j}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T \varphi_{mj}\left(x_{ij}^{(m)} - \mathbf{a}_{:,j}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right) + \left(a_{mj}^{\varphi} - 1\right)\ln\varphi_{mj} - \varphi_{mj} b_{mj}^{\varphi} =$$

$$\left(a_{mp}^{\varphi} - 1 + \frac{N}{2}\right)\ln\varphi_{mj} - \varphi_{mj}\left(b_{mj}^{\varphi} + \sum_{i=1}^{N}\frac{1}{2}\left(x_{ij}^{(m)} - \mathbf{a}_{:,j}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T\left(x_{ij}^{(m)} - \mathbf{a}_{:,j}B^{(m)}\mathbf{s}_i\right)\right)$$

$$\hat{a}_{mp}^{\varphi} = a_{mp}^{\varphi} + \frac{N}{2}$$

$$\hat{b}_{mp}^{\varphi} = b_{mp}^{\varphi} + \frac{1}{2}\sum_{i=1}^{N}\left\langle\left(x_{ip}^{(m)} - \mathbf{a}_{:,p}\mathbf{D}^{(m)}\mathbf{s}_i\right)^2\right\rangle$$

$$\left\langle\varphi_{mp}\right\rangle = \frac{a_{mp}^{\varphi}}{b_{mp}^{\varphi}}$$

$$q(\mathbf{D}^{(m)}): \sum_{i=1}^{N}\left(-\tfrac{1}{2}\left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)^T \varphi_{mj}\left(\mathbf{x}_i^{(m)} - \mathbf{A}\mathbf{D}^{(m)}\mathbf{s}_i\right)\right) - \tfrac{1}{2}\left(\mathbf{d}_m - \boldsymbol{\mu}_m\right)^T v_m\mathbf{I}\left(\mathbf{d}_m - \boldsymbol{\mu}_m\right) =$$

$$\sum_{i=1}^{N}\sum_{p=1}^{P}\left(-\tfrac{1}{2}\left(x_{ij}^{(m)} - \mathbf{a}_{:,j}\mathbf{d}_m\mathbf{s}_i\right)^T \varphi_{mj}\left(x_{ij}^{(m)} - \mathbf{s}_i\mathbf{a}_{:,j}\mathbf{d}_m\right)\right) - \tfrac{1}{2}\left(\mathbf{d}_m - \boldsymbol{\mu}_m\right)^T v_m\mathbf{I}\left(\mathbf{d}_m - \boldsymbol{\mu}_m\right) =$$

$$\sum_{i=1}^{N}\left(-\sum_{p=1}^{P}\left(\boldsymbol{\mu}_m{}^T v_m\mathbf{I} + x_{ip}^{(m)}\varphi_{mp}\mathbf{s}_i\mathbf{a}_{:,p}\right)\mathbf{d}_m - \tfrac{1}{2}\mathbf{d}_m{}^T\left(v_m\mathbf{I} + \varphi_{mp}{}^{-1}\mathbf{s}_i{}^T\mathbf{s}_i\mathbf{a}_{:,p}{}^T\mathbf{a}_{:,p}\right)\mathbf{d}_m\right)$$

$$\boldsymbol{\Sigma}_{\mathbf{D}^{(m)}} = \left(v_m\mathbf{I} + \sum_{p=1}^{P}\left\langle\varphi_{mp}\right\rangle\left\langle\mathbf{a}_{:,p}{}^T\mathbf{a}_{:,p}\right\rangle \bullet \sum_{i=1}^{N}\left\langle\mathbf{s}_i{}^T\mathbf{s}_i\right\rangle\right)^{-1}$$

$$m_{\mathbf{D}^{(m)}} = \boldsymbol{\Sigma}_{\mathbf{D}^{(m)}}\left(v_m\boldsymbol{\mu}_m + \sum_{i=1}^{N}\sum_{p=1}^{P}\left\langle x_{ip}^{(m)}\right\rangle\varphi_{mp}\left\langle\mathbf{s}_i\right\rangle \bullet \left\langle\mathbf{a}_{:,p}\right\rangle\right)$$

Theorem 10: (Column-wise non-negativity constraints)

Consider the PARAFAC model as formulated by eq. 1.31, sought optimized in an unweighted least square sense. Let T be defied as $\mathbf{T} = \mathbf{X}^{JK \times I} - \left(\mathbf{B}_{-F} \middle| \otimes \middle| \mathbf{C}_{-F}\right)\mathbf{A}_{-F}^{T}$ where $\mathbf{M}_{-F}$ denotes that the F[th] column has been removed from $\mathbf{M}$. Then finding the F[th] column of A subject to the constraint that the F[th] column of A has to be non-negative corresponds to finding the optimal value of the unconstrained problem and setting all negative values equal zero.

Proof (adapted from [3])

The optimal choice of **A** unconstrained is the value that minimizes:

$\min_{\mathbf{a}_F} \left\| \mathbf{T} - \left(\mathbf{b}_F \middle| \otimes \middle| \mathbf{c}_F\right)\mathbf{a}_F^T \right\|^2$, let this value be denoted $\boldsymbol{\alpha}$, i.e. $\boldsymbol{\alpha} = \frac{\mathbf{T}^T\mathbf{z}}{\mathbf{z}^T\mathbf{z}}$, where $\mathbf{z} = \left(\mathbf{b}_F \middle| \otimes \middle| \mathbf{c}_F\right)$.

Furthermore, let $\mathbf{E} = \mathbf{T} - \mathbf{z}\boldsymbol{\alpha}^T$. It then follows that:

$$\min_{\mathbf{a}_F} \left\| \mathbf{E} + \mathbf{z}\boldsymbol{\alpha}^T - \mathbf{z}\mathbf{a}_F^T \right\|^2 = \min_{\mathbf{a}_F} \left\| \mathbf{E} + \mathbf{z}(\boldsymbol{\alpha} - \mathbf{a}_F)^T \right\|^2 =$$

$$\min_{\mathbf{a}_F} \left(tr\left(\mathbf{E}^T\mathbf{E}\right) + 2tr\mathbf{E}^T\left(\mathbf{z}(\boldsymbol{\alpha} - \mathbf{a}_F)^T\right) + tr\left[(\boldsymbol{\alpha} - \mathbf{a}_F)\mathbf{z}^T\mathbf{z}(\boldsymbol{\alpha} - \mathbf{a}_F)^T\right] \right) =$$

$$\min_{\mathbf{a}_F} \left(tr\left[(\boldsymbol{\alpha} - \mathbf{a}_F)\mathbf{z}^T\mathbf{z}(\boldsymbol{\alpha} - \mathbf{a}_F)^T\right] \right) = \min_{\mathbf{a}_F} \left(\sum_{i=1}^{p} (a_{pf} - \alpha_p)^2 \right)$$

1. Equality holds as $\mathbf{E}^T\mathbf{E}$ is a constant, furthermore,

   $$\mathbf{E}^T\left(\mathbf{z}(\boldsymbol{\alpha} - \mathbf{a}_F)^T\right) = (\mathbf{T}^T - \frac{\mathbf{T}^T\mathbf{z}}{\mathbf{z}^T\mathbf{z}}\mathbf{z}^T)\mathbf{z}(\boldsymbol{\alpha} - \mathbf{a}_F)^T = (\mathbf{T}^T\mathbf{z} - \frac{\mathbf{T}^T\mathbf{z}}{\mathbf{z}^T\mathbf{z}}\mathbf{z}^T\mathbf{z})(\boldsymbol{\alpha} - \mathbf{a}_F)^T =$$

   $$(\mathbf{T}^T\mathbf{z} - \mathbf{T}^T\mathbf{z})(\boldsymbol{\alpha} - \mathbf{a}_F)^T = 0$$

2. Follows as $\mathbf{z}^T\mathbf{z}$ is a constant.

Notice: Proof also holds for weighted regression as $\boldsymbol{\alpha} = \frac{\mathbf{T}^T\boldsymbol{\Phi}^{-1}\mathbf{z}}{\mathbf{z}^T\boldsymbol{\Phi}^{-1}\mathbf{z}}$ and

$$\mathbf{E}^T\boldsymbol{\Phi}^{-1}\left(\mathbf{z}(\boldsymbol{\alpha} - \mathbf{a}_F)^T\right) = (\mathbf{T}^T - \frac{\mathbf{T}^T\boldsymbol{\Phi}^{-1}\mathbf{z}}{\mathbf{z}^T\boldsymbol{\Phi}^{-1}\mathbf{z}}\mathbf{z}^T)\boldsymbol{\Phi}^{-1}\mathbf{z}(\boldsymbol{\alpha} - \mathbf{a}_F)^T = 0$$

Theorem 11: (Regarding ICAPARAFAC)

Consider the 3-way array $\boldsymbol{\mathcal{X}} \in R^{I\times J\times K}$ having the PARAFAC decomposition $x_{ijk} = \sum_{\lambda}^{F} b_{i\lambda} c_{j\lambda} s_{k\lambda}$. Furthermore, let $\boldsymbol{\mathcal{X}}$ be CI_3. Finding the factors corresponding to the first and second dimension by SVD is then the same as finding the factors by ALSPARAFAC.

Proof:

As $\boldsymbol{\mathcal{X}}$ is CI_3 it can be written as $\mathbf{X}^{IJ\times K} = \mathbf{A}^{IJ\times F}\mathbf{S}^{F\times K}$ where **A** and **S** is found by ICA. Finding **B** and **C** using ALSPARAFAC becomes the problem of minimizing:

$$\sum_{ijk}\left(\boldsymbol{\mathcal{X}}_{ijk} - \sum_{\lambda=1}^{F} s_{i\lambda} b_{j\lambda} c_{k\lambda}\right)^2 = \sum_{ijk}\left(\sum_{q=1}^{F} s_{kq} a_{(ij)q} - \sum_{\lambda=1}^{F} s_{k\lambda} b_{i\lambda} c_{j\lambda}\right)^2 =$$

$$\sum_{k}\left(\sum_{q,\lambda=1}^{F} s_{kq} s_{k\lambda} \sum_{ij}\left(a_{(ij)q} - b_{iq} c_{jq}\right)\left(a_{(ij)\lambda} - b_{i\lambda} c_{j\lambda}\right)\right) = tr\left((\mathbf{A}-\mathbf{M})\boldsymbol{\Psi}(\mathbf{A}-\mathbf{M})^T\right) \qquad \text{eq. 0.1}$$

1. Consequence of $m_{(ij)\lambda} = b_{i\lambda} c_{j\lambda}$, $\boldsymbol{\Psi} = \mathbf{S}\mathbf{S}^T$.

As the rows of **S** are independent $\boldsymbol{\Psi} = \mathbf{S}\mathbf{S}^T$ is a diagonal matrix. Consequently, there is no interaction between the columns of (**A-M**). Therefore, the minimization problem can be split into minimizing the squared error of each column of (**A-M**). Let $q_{ij}^{(\lambda)} = a_{(ij)\lambda}$, i.e. $\mathbf{Q} \in R^{I\times J}$ is the unmatriziced version of the λ^{th} column of **A**. The goal is now to find two vectors $\mathbf{b}_\lambda, \mathbf{c}_\lambda$ so $\left\|\mathbf{Q}^{(\lambda)} - \mathbf{b}_\lambda \mathbf{c}_\lambda^T\right\|_F$ is minimized. However, this minimization problem is solved by the singular valued decompositions, as the first vectors of the SVD decomposition explains the most of the variation of $\mathbf{Q}^{(\lambda)}$, i.e. $\left[\mathbf{U},\mathbf{T},\mathbf{V}^T\right] = SVD(\mathbf{Q}^{(\lambda)}) \Rightarrow \mathbf{b}_\lambda = \mathbf{u}_1, \mathbf{c}_\lambda = t_{11}\mathbf{v}_1$.

Appendix B: Multi-way array algebra

This compilation of definitions and manipulations is if not otherwise stated taken from the work of Lathauwer, Moor and Vandewalle [19] .

Definition 1: The n-mode Multiplication $(\times_n)$

The n-mode Multiplication of the multi-way array $\boldsymbol{\mathcal{X}} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_N}$ by a matrix $\mathbf{U} \in \mathcal{R}^{J_n \times I_n}$, denoted by $\boldsymbol{\mathcal{X}} \times_n \mathbf{U}$, is an $(I_1 \times I_2 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N)$-multi-way array of which the entries are given by:

$$(\boldsymbol{\mathcal{X}} \times_n \mathbf{U})_{i_1 i_2 \cdots i_{n-1} j_n i_{n+1} \cdots i_N} \equiv \sum_{i_n} x_{i_1 i_2 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} u_{j_n i_n}$$

Notice: let $\mathbf{V} \in \mathcal{R}^{J_m \times I_m}$ then:

$$\boldsymbol{\mathcal{X}} \times_n \mathbf{U} \times_m \mathbf{V} = \boldsymbol{\mathcal{X}} \times_m \mathbf{V} \times_n \mathbf{U}$$

Definition 2: The scalar product of two multi-way arrays

The scalar product $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle$ of two multi-way arrays $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathcal{R}^{I_1 \times I_2 \ldots \times I_N}$ is defined as:

$$\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle \equiv \sum_{i_1} \sum_{i_2} \cdots \sum_{i_n} b_{i_1 i_2 \ldots i_n} a_{i_1 i_2 \ldots i_n}$$

Definition 3: The norm of a multi-way array (Frobenius-norm)

The norm of a multi-way array is given by:

$$\|\boldsymbol{\mathcal{A}}\| = \sqrt{\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{A}} \rangle}$$

Definition 4: The rank of multi-way arrays

An N-way array $\boldsymbol{\mathcal{A}}$ has rank-1 when it equals the outer product of N vectors, i.e. $\boldsymbol{\mathcal{A}} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \ldots \circ \mathbf{u}^{(N)}$

Furthermore, the rank of an arbitrary multi-way array $\boldsymbol{\mathcal{A}}$ denoted by $R = rank(\boldsymbol{\mathcal{A}})$, is the minimal number of rank-1 multi-way arrays that yield $\boldsymbol{\mathcal{A}}$ in a linear combination.

Notice: The formulation for rank-1 can be restated as $a_{i_1 i_2 \ldots i_N} = u^{(1)}_{i_1} u^{(2)}_{i_2} \cdots u^{(N)}_{i_N}$.

Definition 5: The rank of a matrix

The normal rank of a matrix $r_{\mathbf{A}} := rank(\mathbf{A}) = r \Leftrightarrow \mathbf{A}$ contains at least a collection of r linearly independent columns, and this fails for r+1 columns. (Taken from [31])

Definition 6: The k-rank of a matrix (defined by J.B. Kruskal 1977)

The k-rank $k_{\mathbf{A}} = \mathrm{r} \Leftrightarrow$ r columns of **A** are linearly independent, but this fails for at least one set of r+1 columns (Taken from [31]). Mathematically this can be expressed as:

$k_{\mathbf{A}} = \arg\min_{N}(\mathbf{A} \mid \exists i; \mathbf{a}_i = \sum_{j \neq i}^{N} c_j \mathbf{a}_j)$, where N denotes the amount of columns of **A** used to generate $\mathbf{a}_i$.

Definition 7: Diagonal multi-way arrays

A multi-way array is called diagonalizable if the core multi-way array $\boldsymbol{S}$ of the HOSVD fulfills $s_{i_1 i_2 \ldots i_N} = 0 \,\text{unless}\, i_1 = i_2 = \ldots = i_N$.

Appendix C: MATLAB implementation of multi-way array manipulations

In the literature, I didn't find any fast implementations of the various multi-way array manipulations except for the first matricizing function given here. Therefore, I have suggested some very fast MATLAB implementations for these manipulations:

```
function X=matricizing(Y,n)

% Matricizes the multiway array Y around dimension n to give the matrix X
% Input:
% Y     Multiway array
% n     Dimension to use for matrizicing
% Output:
% X     Matrix of matricized multiway array

N=ndims(Y);
X=reshape(permute(Y, [n 1:n-1 n+1:N]),size(Y,n),prod(size(Y))/size(Y,n));
```

```
function Y=unmatricizing(X,n,D)

% The inverse operation of matricizing, i.e. recreates the multiway
% array Y that was matricizes around n to give the matrix X.
%
% input:
% D     a vector containing the original dimensions of Y
% n     the dimension along which the matricizing was originally performed
% X     a matrix corresponding to the matricizing of Y around n
% output:
% Y     multi-way array

if n==1
    perm=[1:length(D)];
else
    perm=[2:n 1 n+1:length(D)];
end

Y=permute(reshape(X,D([n 1:n-1 n+1:length(D)])),perm);
```

```
function A=tmult(T,M,n)
% the n-mode multiplication or tensor multiplication:
% Let T be an IlxI2x...xIn-lxInxIn+lx...xIN multi-way array
% and M be an JxIn then
% T x_n M gives an IlxI2x...xIn-lxJxIn+lx...xIN multi-way array
% Input:
% T     Multi-way array of dimensions IlxI2x...xIn-lxInxIn+lx...xIN
% M     JxIn-matrix
% n     the dimension to do the multiplication;
% Output:
% A     Multi-way array of dimensions IlxI2x...xIn-lxJxIn+lx...xIN

Dt=size(T);
Dm=size(M);

Tn=matricizing(T,n);
Tnew=M*Tn;
Dt(n)=Dm(1);
A=unmatricizing(Tnew,n,Dt);
```

```
function T=outerprod(FACT)

% The outer products:
% T=sum_f (U1f o U2f o U3f o ...o UNf) where Uif is a vector corresponding
% to the f'th factor of the i'th dimension.
%
% Input:
% FACT      Cell array containing the factor-vectors corresponding to
%           each of the three dimensions, i.e. Ui=FACT{i}
% Output:
% T         The multi-way array created from the outer product of each
%           dimensions vector.

T=0;
for i=1:size(FACT{1},2)
    Y=1;
    for j=1:length(FACT)
        U=FACT{j};
        Y=tmult(Y,U(:,i),j);
    end
    T=T+Y;
end
```

```
function [U,S]=HOSVD(X)
% The Higher Order Singular Value Decomposition as defined by
% Lathauwer, Lieven De., Moor, Bart De, Vandewalle, Joos
% "A MULTILINEAR SINGULAR VALUE DECOMPOSITION"
% SIAM J. MATRIX ANAL. APPL. Vol. 21, No. 4, pp. 1253 1278:
%
% Model:
% X=S x1 U1 x2 U2 ... xN UN, i.e. a TUCKER model restricted so
% Ui orthonormal being the orderen eigenvectors of the matrix
% found by matricizing(X,i). S the resulting core multi-way array.
%
% Input:
% X     Multi-way array
% Output:
% U     Cell of eigenvectors, i.e. U{i} corresponds to the eigenvectors of
%       the matrix M=matricizing(X,i).
% S     Multi-way array of same size as X

S=X;
for k=1:ndims(X)
    Xk=matricizing(X,k);
    [u,s]=eig(Xk*Xk'); %cheaper than the SVD since we only want U
    u=fliplr(u);
    U{k}=u;
    S=tmult(S,U{k}',k)
end
```

Appendix D: ICA- and ALSPARAFAC on Chemometric Data

The field in which the PARAFAC model has been the most applied is probably analyzing chemometric data. Here the dataset 'Claus' consisting of five samples containing different amounts of tyrosine, tryptophane and phenylalanine measured by fluorescence in the excitation spectra 240 to 300 nm and emission spectra 250 to 450 nm were analyzed. For a description of the dataset see [41]. The data set has the dimensions $sample \times emission \times excitation$. The ICAPARAFAC algorithm seems well justified as the goal is from the samples to find factors that are independent when emission and excitation is combined, i.e. $CI_{2,3}$.

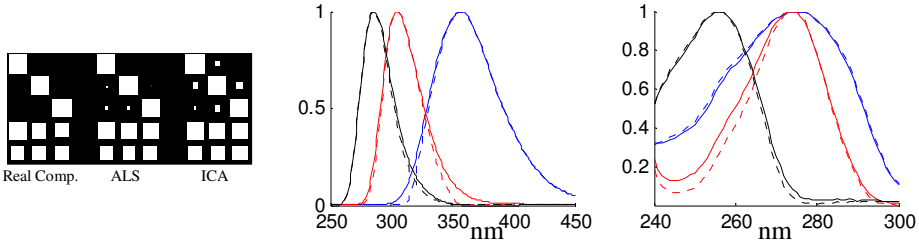


Figure 0.1: Left most panel; the true mixing of the five samples, the mixing found by ALSPARAFAC and mixing found by ICAPARAFAC. Center panel; the emission specters. Right panel; the excitation specters. ALSPARAFAC (solid) ICAPARAFAC (dashed).

From Figure 0.1 it is seen that the ICAPARAFAC algorithm is slightly worse at estimating the true mixing of the factors in the samples than the ALSPARAFAC algorithm. The ICAPARAFAC model seems however better at estimating the emission and excitation specters as they are slightly better defined. Consequently, the ICAPARAFAC method seems effective in analyzing this chemometric dataset.