M.Sc. Thesis
Master of Science in Engineering

DTU Compute
Department of Applied Mathematics and Computer Science

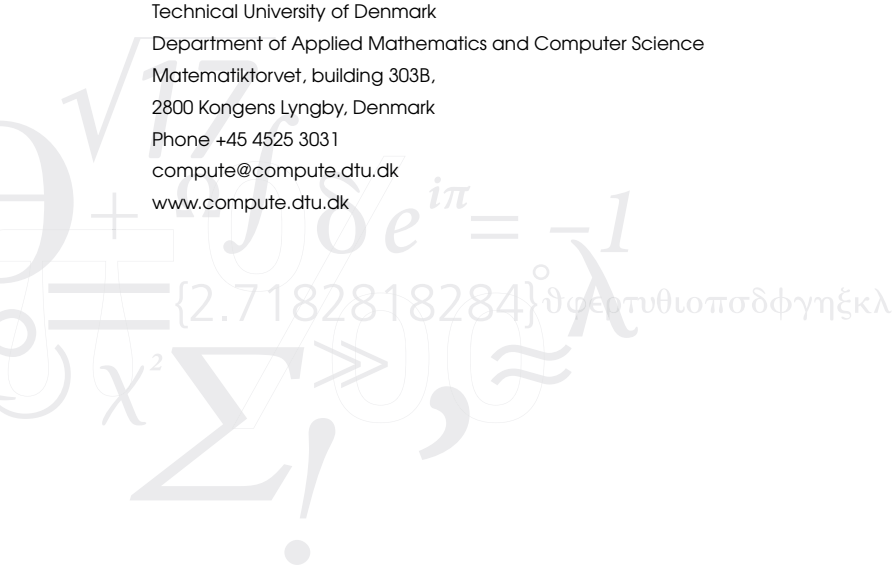
Machine Learning for Social EEG

A Bayesian approach to correlated component analysis and recording simultaneous multiple subject EEG

Simon Kamronn (s082825)
Andreas Trier Poulsen (s083500)

Kongens Lyngby 2013

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Matematiktorvet, building 303B,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk



Short Contents

Summary

This thesis describes the derivation and implementation of BCoCA, a Bayesian approach to Correlated component analysis, which was introduced in Dmochowski et al. [2012]. BCoCA enables the comparisons between more than two subjects at the same time, and relaxes the constraint of equal weights with an adaptable parameter controlling the similarity between the weights for each dataset, with the purpose of locating neural activations that are synchronised within and between brains.

The thesis outlines the principles of variational inference, the method of approximation used to derive the updates for BCoCA as well as a cost effective way to calculate its corresponding lower bound, which can be used as a measure of performance and to estimate the time of convergence. To show its capabilities BCoCA will be tested on simulated data under varying conditions, on real EEG datasets from two other experiments and will finally be used to analyse the results of an experiment conducted for this thesis.

The presented study will investigate whether neural correlations are detectable using consumer-grade hardware, with the specific goal to examine the difference between neural correlation originating from emotionally arousing and neutral films as done in Dmochowski et al. [2012]. To expand on their experimental setup and investigate the effect of experiencing an emotionally laden stimulus in a group as compared to experiencing it alone, simultaneous EEG of nine subjects were recorded. In total were 42 subjects used for the experiments.

It was shown that neural correlation is detectable using consumer-grade hardware and that it was possible to reproduce some of the results in Dmochowski et al. [2012], showing that there is a significant difference between neural correlation originating from emotionally arousing and neutral films, respectively. The results were further established by comparing scenes with periods of significant correlation and scalp projections of the neural activity. The latter showed higher activation in areas related to emotion for the emotionally intense *Sophie's Choice* compared to the suspenseful but otherwise emotionally indifferent *Bang! You're Dead*. It was unfortunately not possible to determine, whether the effect of experiencing an emotionally laden stimulus in a group is significantly different to experiencing it alone. We maintain the belief that there is a difference, but further processing is needed to reveal it.

Resumé

Denne afhandling beskriver udledningen og implementeringen af BCoCA, en Bayesisk tilgang til *correlated component analysis*, som blev introduceret i Dmochowski et al. [2012]. BCoCA muliggør sammenhold mellem mere end to dataset på samme tid, og lemper betingelsen af en ens vægte med en adaptiv parameter der kontrollerer ligheden mellem vægtene for det enkelte dataset, med det mål at lokalisere neurale aktiveringer der er synkroniseret inden i og mellem hjerner.

Denne afhandling klarlægger principperne bag variationel inferens, approksimationsmetoden der bruges til at udlede opdateringerne for BCoCA samt en beregningseffektiv måde at udregne *lower bound*, som kan bruges som mål for præstation og til at afgøre tidspunktet, hvor konvergering er opnået. For at vise BCoCA's evner vil den blive testet på simuleret data under varierende omstændigheder, på rigtige EEG datasæt fra to andre experimenter og vil slutteligt blive brugt til at analysere resultaterne fra et eksperiment udført til denne afhandling.

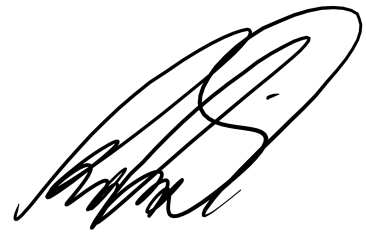
Dette eksperiment vil undersøge hvorvidt neurale korrelationer er målbare ved at bruge hardware af forbruger kvalitet, med det specifikke mål at undersøge forskellen mellem neural korrelation med oprindelse fra følelsesladede og neutrale film som gjort i Dmochowski et al. [2012]. For at udvide på deres forsøgsopstilling og undersøge effekten af at opleve et emotionelt ladet stimulus i en gruppe i forhold til at opleve det alene, blev der foretaget simultan optagelse af EEG på ni forsøgspersoner. I alt blev 42 forsøgspersoner brugt til eksperimenterne.

Det blev vist at neural korrelation er målbart ved brug af hardware af forbruger kvalitet, og at det var muligt at reproducere nogle af resultaterne fra Dmochowski et al. [2012], og herved vise at der er signifikant forskel mellem neurale korrelationer med oprindelse fra følelsesladede og neutrale film. Resultaterne blev yderligere fastslået ved at sammenligne scener med perioder med signifikante korrelationskoefficienter og skalp projektioner over den neurale aktivitet. Sidstnævnte viste større aktivering i områder relateret til følelser for den følelsesmæssigt intense *Sophie's Choice* sammenlignet med den spændingsfyldte men ellers følelsesmæssigt indifferente *Bang! You're Dead*. Det var desværre ikke muligt at afgøre, om effekten af at opleve et følelsesmæssigt ladet stimulus i en gruppe er signifikant forskellig fra at opleve det alene. Vi fastholder den tro, at der er en forskel, men at yderligere undersøgelser er nødvendige for at afsløre den.

Preface

This thesis was prepared at the department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfilment of the requirements for acquiring a M.Sc. in Medicine and Technology.

Kongens Lyngby, December 17, 2013



Simon Kamronn (s082825)

Andreas Trier Poulsen (s083500)

Acknowledgements

First and foremost we would like to thank our supervisor Professor Lars Kai Hansen at DTU Compute for his guidance and the many hours he has invested in us and this project. For our many interesting conversations, for helping us when we got stuck and for pushing us towards new possibilities when we reached our initial goals. A special thanks goes to Arkadiusz Stopczynski, Ivana Konvalinka and Camilla Falk Jensen for their invaluable inputs to the experimental setup and to Michal Radwańnski for helping prepare nine subjects at the same time. We would also like to thank our many subjects, too many to name, but important nonetheless. Thomas Ølund also deserves thanks for proofreading this thesis and lending us his camera and finally a thank you to Sofie for being our first test subject, for proofreading, and for just being there.

Contents

CHAPTER 1

Introduction

The last two decades has shown great advancements in digital signal processing, machine learning and measurement of biopotentials, which has given us a greater insight in the workings of the human body, as well as the ability to measure and compare its functions and states. Measurement of the brain is no exception to these advances, but due to its high level of complexity we still know very little about the internal processes of the brain and their effects on the mind. Emotions, behaviour and affective states in the context of social cognition has with technological advancements lately become a very active research area. An area in which we, with this thesis, will try to make a contribution.

When estimating neural activity the main approach has so far been through discrete event related designs, e.g. Brain Computer Interface (BCI; Blankertz et al. 2007). In BCI the trial consists of a training phase, where the subject is repeatedly exposed to the same stimulus be it visual, auditory or somatosensoric, and common event markers are estimated.

The concept of investigating neural responses in "natural" conditions was first proposed in a fMRI study by Hasson, Nir, et al. [2004] that found remarkable inter-subject synchronisation between subjects having viewed the same film. This experimental concept has later been adopted by Dmochowski et al. [2012] but by using EEG instead, because voxel-wise correlations in blood oxygenation level dependent (BOLD) signals are unable to capture weak activity over distant regions, and the poor temporal resolution of fMRI inhibits precise estimation of the times of synchronisation. Apart from high temporal resolution, EEG also has the advantage of being unobtrusive, making social experiments possible which could not be done in a MRI scanner. Unlike the BOLD signal, which contains a delay between the instant of cortical activity and the measured signal, EEG recordings capture the activity instantly meaning that the mixing of underlying sources are nearly linear [Nunez 1974], though the cortical interactions might not be.

As the brain is a complex and continuously working organ with transient states, Dmochowski et al. [2012] has proposed a signal decomposition method, which works continuously. Since the results are continuous and transient, the cognitive changes cannot be tracked using event markers. Instead the correlation with other subjects which have been exposed to the same stimulus is used. Dmochowski et al. [2012] employs this method in the search of a neural measure of cognitive engagement. A measure, which have yet to receive a general definition but can have many applications

in areas such as neuromarketing, quantitative measurements of entertainment and attention deficit disorders.

The method to extract this measure of attention consists of the signal decomposition of two EEG datasets from subjects experiencing the same continuous stimulus, named Correlated Component Analysis (here abbreviated as CoCA). In their experiments Dmochowski et al. used the viewing of videoclips from three films, with varying levels of suspense, as stimulation. The underlying idea of CoCA is similar to that of Canonical Correlation Analysis (CCA), in that the goal is to find weights, $\mathbf{W}$, that maximises the correlation between two datasets after filtering [Hotelling 1936]. However, CoCA differentiates itself by finding *one* set of weights that works for filtering both datasets.

In this project we will present a Bayesian version of CoCA (BCoCA) and expand it to accommodate multiple datasets. Instead of requiring that all datasets have the same $\mathbf{W}$, BCoCA will introduce a parameter, λ, which regularises how similar the weights belonging to each dataset are. This parameter can be fixed as a predefined constant or be estimated based on data through automatic relevance determination (ARD; L. K. Hansen et al. 1994; MacKay 1996).

Apart from testing the proficiency of BCoCA on synthetic data, we will test it on data from experiments conducted in collaboration with Arkadiusz Stopczynski, Michal Radwańnski, Ivana Konvalinka and Lars Kai Hansen. The experimental paradigm is inspired by Dmochowski et al. though with increased focus on the social aspect of experiencing a film. Exploiting that BCoCA can compare multiple datasets, experiments were conducted with people either watching the film alone or in a group.

This thesis will cover both the assumptions, mathematical derivation and testing of BCoCA, and the social experiments. The chapters are organized as follows,

- Chapter 1 contains this introduction followed by a general review of relevant neural physiology, psychological theories on perception and social cognition, and a short introduction to latent variable models.
- Chapter 2 presents the thoughts behind BCoCA. First the necessary theoretical background of variational inference will be discussed. After a mathematical derivation, the final algorithm will be shown and tested on synthetic data, as well as EEG from two other experiments, and compared against relevant algorithms. To improve comparisons we will use the same data for testing as in [Klami 2013] as well as construct our own.
- Chapter 3 describes the experiments, the hardware and software used, and how the experiments were conducted.
- Chapter 4 contains an analysis of the data from the experiments using CoCA and BCoCA.
- Chapter 5 is a discussion of the main results of the thesis and areas of improvement for future work.

1.1 Electroencephalography

Sitting in the park with a cup of coffee while listening to music and wondering what the girl across from you is thinking about seems like a trivial, even relaxing, way to spend a Sunday afternoon. But the heat from the sun, the taste of the coffee, the sound from the music, the view of the girl, the subconscious processing of her posture and facial expression and the active reasoning of her mental state is actually using a number of different brain networks. Complex and intersecting networks that can be divided into subregions each consisting of billions of neurons.

Neurons are highly specialised cells with the fundamental task of receiving, conducting and transmitting signals. To receive signals each neuron has up to 100,000 tentacle-like extrusions called dendrites, that all end in the cell body, which continues into the axon (figure 1.1) [Alberts et al. 2010]. The axon transmits signals from the cell body to the terminal, which is part of the synaptic contacts with the dendrites of other neurons. Due to an imbalance of positively and negatively charged ions across the membrane, the neurons have negative resting membrane potential. An activation of the synapses elicits a post-synaptic potential which is a rapid increase of membrane

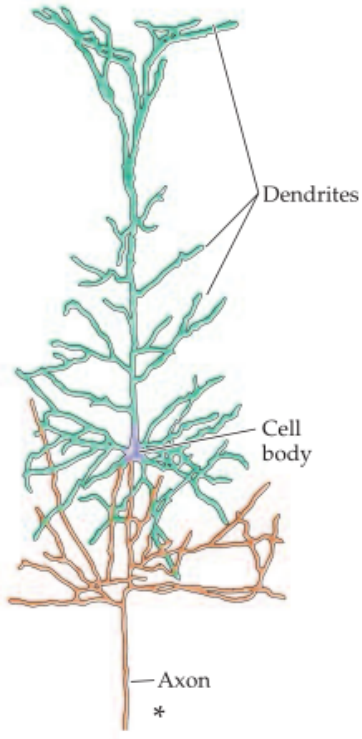


Figure 1.1: Pyramidal neuron in cerebral cortex [Purves 2004]

potential at the apex of the dendrites. Due to the structure of the membrane the potential propagates along the dendrite to the cell body and if the cumulated signal in the cell body is sufficiently strong, an explosion of electrical activity in the plasma membrane will stimulate an action potential to travel along the axon to the nerve terminal [Purves 2004].

The potential is propagated along dendrites and axons by increased influx of sodium ions to increase the potential and a later efflux of potassium to repolarise the membrane to its resting state. These currents of charged ions is a kind of secondary current in the extracellular fluid and creates a dipole between the post-synaptic terminal of the apical dendrite and the cell body, the soma, of the pyramidal cells in the cortex. The ions circles in order to reach an equilibrium of ionic concentration. One neuron produces a very small electric field but vertically aligned and synchronously activated neurons creates a dipole strong enough to measure at the scalp through cerebrospinal fluid, bone and skin [Nunez 1974; Saab 2008].

1.2 Perception and social cognition

Psychological studies has up to about a decade ago been mostly focused on *one-brain* experiments. Experiments in which subjects merely observes the environment or other persons. Though these studies are too simplistic to model many aspects of social cognition, they have yielded many interesting results in low-level cognition. One of the theories that emerged from these experiments is the mirror neuron system (MNS) which is a network of brain regions that imitates the movement of e.g. an arm grasping a cup of coffee exactly as if the subject itself performed the action. The mirror system plays a key role in the ability to not only interpret the goal of a movement but also the intention, the *why* [Rizzolatti et al. 2008]. This natural urge to imitate a response actively needs to be suppressed in most situations and might actually be a disadvantage in a situation when a non-identical complementary action is needed [Sartori et al. 2012; Kourtis et al. 2013].

The recording of EEG on multiple subjects simultaneously (hyperscanning; Babiloni, Cincotti, et al. 2006) is emerging and studies on a more complex level are becoming a possibility. The two-body problem has been neglected, but has recently received greater interest across multiple modalities in fields spanning from game theory to transmitting emotions through facial expressions [Babiloni and Astolfi 2012]. Though hyperscanning and social interaction experiments are still relatively new, critics are already pointing out issues with moving from individual to social cognitive theories [Dumas 2011]. A typical experiment in the *one-brain* theory is often formulated as a turn-taking paradigm in which participants take turns to act while the other perceive, creating a perception-action loop. A natural way to emulate a real world scenario but without the dynamics of true social interaction. The isolated paradigms of standard cognitive science only incorporates information-flow from the environment to the observer, but this approach is inadequate in the paradigm of *embodied cognition*. Interaction and emotional engagement between people are dynamic processes that

couples them in a unit that is not readily separable [Schilbach et al. 2013; Hasson, Ghazanfar, et al. 2012]. These inter-personal interactions can be crucial to understanding the mechanisms of social cognition and so far hyperscanning is the only method to tap into inter-brain processes [Konvalinka et al. 2012]. To this purpose EEG is becoming an increasingly popular modality due to its high temporal resolution and recent advances in mobile equipment [Stopczynski et al. 2013].

Dumas et al. [2010] tried to follow this dynamic interaction theory with an experiment that incorporates the spontaneous interaction in a semi-natural actor-imitator relationship using a two-way video setup and dual-EEG monitoring. The alpha-mu activity in the right centro-parietal region was found to synchronise between two brains of which one was engaged in acting and the other imitating. This area has previously been shown to be included in the mirror neuron system when observing others while retaining ones perspective of self [Tognoli et al. 2007] and mu rhythms in translating the observed into action [Pineda 2005].

Joint attention is popularly perceived as a shared gaze on an object or focus on a task. Using a dual-EEG configuration, the online interaction between face-to-face subjects engaged in either spontaneous or directed attention showed to equally decrease the oscillatory activity between 11 and 13 Hz in the left centro-parieto-occipital region when gazing on the same object as compared to different objects [Lachat et al. 2012]. The decreasing signal power was thus not due to dynamic interaction between the subjects (undirected attention) but perhaps an *interpersonal coordination component* of social interaction or just the mere knowledge that someone else gaze on the same object. The latter is supported by a recent study showing that the attention relations of others to the environment affects the attention of oneself [Böckler et al. 2012].

The mirror mechanism is not only involved in bodily emulation, it is also thought to help mediate the understanding of emotions through empathy, the capacity to understand affective experiences in other persons [Rizzolatti et al. 2008; Enticott et al. 2008]. The definition of empathy is arguably divided into three components; perspective-taking, emotion regulation and affective response, and emotional contagion [Decety et al. 2006]. The perception of emotion is on a neural level closely related to the MNS while the regulation of emotion is related to the ability to distinguish self from others using higher cognitive processes [Preston et al. 2002]. The research of empathy has mostly been through the observation of pain and the common finding is that the vicarious experience of pain activates the same affective brain areas as if the experience was first hand [Bufalari et al. 2007; Singer 2012]. Likewise does viewing a film depicting a face expressing disgust activate the same neural areas in the anterior insula and anterior cingulate cortex in the observer as the model [Wicker et al. 2003]. Even though these studies present clear indication of neural areas correlating with emotions it has not been possible to map individual emotions to specific areas. The idea of an emotion being localised to a single area is unreasonable but not the idea of a neural signature across multiple areas. Using fMRI recordings of method actors imitating emotions Kassam et al. [2013] trained a classifier that was able to identify emotions across subjects with significant accuracy.

Intuitively, many factors modulate the neural response of empathy, including the intensity of the emotion, attention to the stimulus and characteristics of the empathizer [Hein et al. 2008]. By showing a clip from the film *Sophie's Choice*, depicting an intense personal loss, Raz, Jacob, et al. [2013] showed high correlation between fMRI recordings and sadness rating of the film. They believe that because of a strong personal grievance in a real-time setting the 'first-person' perspective of the actor is adopted through *embodied simulations* (ES), e.g. facial expressions, and to a lesser degree through *theory of mind* (ToM), a cognitive representation of another's state. Interestingly, emphatic reaction correlated significantly with the sadness rating except during the most distressing part of the film. They theorise that the distinction between self and other is undermined in this moment and to protect itself, the mind distances itself from the emotional object.

Shteynberg et al. [2013] showed that the mood of the participants in a study has a significant effect on the how strongly the person reacts to a stimuli and that shared attention increased collaborative processing which in turn increased the influence of mood on evaluation. The latter finding is particular interesting because it directly supports the hypothesis that there is a difference in the neural activation when watching a film in a group as opposed to alone. This is supported by a study by Nummenmaa et al. [2012] where it was found that emotions of negative valence induce inter-subject synchronization of brain activity leading to similar perceptions and thus enabling understanding and prediction of the actions of others. Positive emotions showed a lower level of synchronization which could be explained by the different neural stimulations between negative and positive valence. Where negative emotions activates possible survival-neurons in the default-mode network, positive emotion encourage 'exploration of the environment' leading to more individual brain activation patterns.

The concept *social sharing of emotions* [Rimé et al. 1998] argues that sharing negative emotional events facilitate the cognitive processing of the emotions. Though negative emotions elicits sharing, the sharing itself may not decrease the psychological stress induced. Instead it was found to strengthen the ties to ones social network and to alter the emotional climate of the network. Studies have in line with this theory shown that viewing films of negative valence significantly increases subjects inclination to sharing the content of the film and emotions experienced during watching. The intensity of the emotional impact affects the level of sharing but the relationship was not found to be linear [Luminet et al. 2000]. In a recent review Rimé [2009] states that emotions are a reaction to a discrepancy between assumptions of what is expected and what is reality. When expectations are disconfirmed the cognitive work of a search for meaning is initiated and assumptions about the world is altered. One way to seek meaning is by comparing views with the social network. By sharing emotions the group can reach agreement of assumptions and views of the world and in turn create a social reality within the network. Contrary to his earlier work Rimé also argues that negative emotions motivates people to seek emotional support to reduce induced distress. Though this belief is mostly based on findings concerned with

verbal sharing it is not unlikely to have the same impact when jointly experiencing the emotion. Especially not since he himself compares sharing an emotion with observing an emotion, using a quote of the Perception-Action model which states that

> *...attended perception of the object's state automatically activates the subject's representations of the state, situation, and object, and that activation of these representations automatically primes or generates the associated autonomic and somatic responses, unless inhibited* [Preston et al. 2002]

Studying the list of most emailed New York Times articles, Berger et al. [2010] found that articles of positive valence was most often shared due to positive self-reflectance. However, articles with a particular type of negative content with a high level of arousal, such as anger, was also very likely to be shared. In line with Rimé, the authors explain the negative-content sharing with the theory that sharing affectively rich content deepens social bonds. This result has later been supported by an investigation of the sentiment in retweets on Twitter. Tweets of positive content was most often retweeted except if the tweet was news-related where negative news are more often shared [L. Hansen et al. 2011].

To summarise the above; visual perception of other humans elicit mirroring neural responses which extends to emotions partly by mirroring facial expressions. The presence of others yet disrupts the presumption of an isolated individual which means that a couple needs to be treated as an interacting unit which challenges the paradigms and execution of many experiments. Hyperscanning is so far the only realistic approach and experiments has shown significant effect of social interaction on neural activation. The results and theory on the response to emotionally laden stimuli in the context of social cognition is, however, still ambiguous.

1.3 Canonical correlation analysis and latent variable models

Neuroinformatic experiments can be divided into two groups with respect to the nature of the stimulus the subjects are exposed to. They can be discrete as the displaying of a photograph or the prompting of an imagined movement of a body part. Here the stimulus can be repeated many times and the specific time of stimulus marked in the recordings, enabling a correlation between the recorded data and times of stimulus.

Continuous stimulus can be used to emphasise the brain as a transient and continuously working organ, where the stimulus can take the form of a film where the viewer is engaged in the plot and the build-up to climax. This manner of stimulus enables tracking of ongoing changes in the cognitive state of the viewer [Dmochowski et al. 2012]. But this manner of stimulus comes at the cost of specific event markers. One

solution is correlating the recordings between subjects experiencing the same stimulus such as in CCA.

1.3.1 Canonical correlation analysis

CCA was introduced in Hotelling [1936]. Given two datasets, $\mathbf{X}^{(1)} \in \mathbb{R}^{D_1 \times N}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{D_2 \times N}$, with $\{D_1, D_2\}$ defining the number of features and N the number time samples, it seeks to estimate weights, $\{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}\}$, which maximise the correlation between two time series vectors $\mathbf{y}_1 = \mathbf{X}^{(1)T}\mathbf{w}_k^{(1)}$ and $\mathbf{y}_2 = \mathbf{X}^{(2)T}\mathbf{w}_k^{(2)}$. At the same time CCA constrains the estimated weights with the condition that $\mathbf{X}^{(1)T}\mathbf{w}_k^{(1)}$ and $\mathbf{X}^{(1)T}\mathbf{w}_{k'}^{(1)}$ are uncorrelated for $k \neq k$. CCA then finds the weights through eigenvalue decompositions, which has the benefit of attaining an analytic solution, but without any statistic measure for the certainty of the result [Hardoon et al. 2004; Klami 2013].

1.3.2 Latent variables

Probabilistic models often include *latent* or hidden variables, $\mathbf{z}$, representing sources or time series "hidden" in the recorded data. In terms of neuroinformatics, $\mathbf{z}$ can be seen as representing the elicited response to a given stimulus. The introduction of latent variables enables the definition of a prior distribution for these hidden sources, which are often kept simple such as

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{1.1}$$

As opposed to a probabilistic model without latent variables, where only the marginal distribution, $p(\mathbf{x})$, is available, a latent variable model also enables the working with the joint distribution, $p(\mathbf{x}, \mathbf{z})$, and through this the conditional distribution $p(\mathbf{z}|\mathbf{x})$ using Bayes' theorem.

1.3.3 Latent variable approach to canonical correlation analysis

Expanding on the probabilistic PCA introduced by Tipping et al. [1999], a probabilistic approach to CCA was presented in Bach et al. [2005] using latent variables. Instead of weights, which maximise the correlation between the linear projections of data, Bach and Jordan used Gaussian distributed common sources, $\mathbf{z}$, mixed in both datasets

$$p(\mathbf{x}^{(m)}) = \mathcal{N}\left(\mathbf{A}^{(m)}\mathbf{z}, \Sigma^{(m)}\right), \tag{1.2}$$

with $m = \{1, 2\}$ signifying the number of datasets, and with $\Sigma^{(m)}$ representing the covariance matrix for the observation noise of dataset m. The noise is often simplified to i.i.d. Gaussian noise with view-specific[1] variance; $\Sigma^{(m)} = \sigma_m^2\mathbf{I}$. $\mathbf{A}$ signifies the

[1] In this thesis the terms *view* and *dataset* are used interchangeably, signifying an entire recording, $\mathbf{X}^{(m)}$, with D channels or features and N samples.

mixing matrix (also known as the forward model)[2], where each column represents the mixing of one source into D observed channels, which means that if one posses prior knowledge of the number of hidden sources the dimension of the estimated mixing matrix can be reduced to $\mathbf{A} \in \mathbb{R}^{D \times K}$. This is an advantage when $K < D$, but presents the problem of choosing the right value for K. To avoid discrete model selection C. Bishop [1999] introduced a hierarchical prior over $\mathbf{A}$ using the automatic relevance determination (ARD) framework

$$p(\mathbf{A}^{(m)}) = \prod_{k}^{K} \mathcal{N}\left(\mathbf{A}_k^{(m)} | \mathbf{0}, \alpha_k^{-1}\right) \tag{1.3}$$

$$p(\boldsymbol{\alpha}) = \prod_{k}^{K} \mathcal{G}a(\alpha_k | a_0, b_0), \tag{1.4}$$

where $\mathbf{A}_k$ signifies the k'th row in $\mathbf{A}$ and $\boldsymbol{\alpha}_k$ is a gamma distributed hyper parameter controlling the precision of $\mathbf{A}_k$.

This approach to CCA has lead to Bayesian CCA [Wang 2007; Klami and Kaski 2007], a hierarchical Bayesian spatio-temporal model [Wu et al. 2011] and latest Group Factor Analysis (GFA) [Virtanen et al. 2011], the first practical multi-view generalization of Bayesian CCA, and its two-view version Bayesian Inter-Battery Factor Analysis (BIBFA) [Klami 2013]. The latest two additions divided the sources into shared and view-specific sources enabling the simplification of $\Sigma^{(m)}$ to a diagonal matrix improving computing time for high dimensional data.

The above mentioned articles where not the first to introduce these concepts to probabilistic CCA, but instead had their focus on how to approximate the posterior distribution for the hidden sources. Instead of the commonly used maximum likelihood or maximum a posteriori solutions through expectation maximisation, the methods focus on a fully Bayesian treatment employing either Gibbs sampling or variational inference. Both approaches have their own advantages and drawbacks but since this thesis employs variational inference, Gibbs sampling will not be discussed further.

[2]Different authors use different letters for the mixing matrix. Most Bayesian models use $\mathbf{W}$, probably stemming from C. Bishop [1999], but as this letter is also used to define the demixing matrix, we have chosen to use $\mathbf{A}$ as Parra et al. [2005].

CHAPTER 2

Machine Learning and Digital Signal Processing

This chapter contains the theoretical background for the methods used to analyse the data in this thesis. As part of this thesis revolves around the development of BCoCA, the majority of this chapter will focus on the concepts of variational inference, how it can be used to derive BCoCA, and the different tests regarding the performance of the resulting algorithm.

2.1 Variational inference

When discussing latent models as the one in (1.2) it can ease understanding to divide the variables into visible variables, $\mathbf{V}$, that can be observed such as $\mathbf{x}$, and hidden or latent variables, $\mathbf{H}$, such as $\mathbf{z}$ and $\mathbf{A}$. The joint distribution can then be expressed as $p(\mathbf{H}, \mathbf{V})$. In many cases this distribution gets so complex that the true posterior distribution, $p(\mathbf{H}|\mathbf{V})$, becomes analytically intractable, in which case a suitable approximation, $q(\mathbf{H})$, can be a better option. Approximation of posterior distributions can be divided into two groups; stochastic or deterministic. An example of a stochastic approach is Markov chain Monte Carlo which through sampling can obtain exact results given infinite computation resources. This approach quickly gets computationally expensive, and is better suited for small-scale problems [C. M. Bishop 2006]. The deterministic approach instead uses analytical approximations of the posterior distribution through simplifying assumptions regarding this distribution.

Before enabling the maximisation of the similarity between the true distribution and its approximation, the measure of similarity has to be decided upon. The next section describes one such measure of similarity between two distributions.

2.1.1 The Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is also known in physics as *relative entropy* and describes the the dissimilarity between a true distribution, $p(\mathbf{H}|\mathbf{V})$, and its approximation, $q(\mathbf{H})$. The KL divergence is defined as

$$\mathrm{KL}(q\|p) = \int q(\mathbf{H}) \ln \frac{q(\mathbf{H})}{p(\mathbf{H}|\mathbf{V})} d\mathbf{H}. \tag{2.1}$$

Note that by definition $\mathrm{KL}(q\|p) \neq \mathrm{KL}(p\|q)$ since

$$\mathrm{KL}(p\|q) = \int p(\mathbf{H}|\mathbf{V}) \ln \frac{p(\mathbf{H}|\mathbf{V})}{q(\mathbf{H})} d\mathbf{H}. \tag{2.2}$$

Using Jensen's inequality and the fact that the function $-\ln x$ is strictly convex, the KL divergence can be proven to be always positive through

$$\begin{aligned} \mathrm{KL}(q\|p) = -\int q(\mathbf{V}) \ln \frac{p(\mathbf{H}|\mathbf{V})}{q(\mathbf{H})} d\mathbf{H} &\geq -\ln \int q(\mathbf{H}) \frac{p(\mathbf{H}|\mathbf{V})}{q(\mathbf{H})} d\mathbf{H} \\ &\geq -\ln \int p(\mathbf{H}|\mathbf{V}) d\mathbf{H} = 0. \end{aligned} \tag{2.3}$$

This means that the KL divergence is zero only when $q(\mathbf{H}) = p(\mathbf{H}|\mathbf{V})$ [Murphy 2012].

The evaluation of the KL divergence, as defined in (2.1), depends on the posterior distribution, but since this is assumed intractable (and the reason for the approximation) the equation as such has no usability. Using the product rule, (2.1) can be rearranged into an expression with distributions that are assumed analytically tractable;

$$\mathrm{KL}(q\|p) = \int q(\mathbf{H}) \ln \frac{q(\mathbf{H}) p(\mathbf{V})}{p(\mathbf{H}, \mathbf{V})} d\mathbf{H} \tag{2.4}$$

$$= \int q(\mathbf{H}) \ln \frac{q(\mathbf{H})}{p(\mathbf{H}, \mathbf{V})} d\mathbf{H} + \int q(\mathbf{H}) \ln p(\mathbf{V}) d\mathbf{H} \tag{2.5}$$

$$= \int q(\mathbf{H}) \ln \frac{q(\mathbf{H})}{p(\mathbf{H}, \mathbf{V})} d\mathbf{H} + \ln p(\mathbf{V}). \tag{2.6}$$

Defining the negative of the first term on the right hand side as

$$\mathcal{L}(q) = -\int q(\mathbf{H}) \ln \frac{q(\mathbf{H})}{p(\mathbf{H}, \mathbf{V})} d\mathbf{H}, \tag{2.7}$$

a relationship between the true log likelihood and the approximation of the posterior distribution can be defined as

$$\ln p(\mathbf{V}) = \mathrm{KL}(q\|p) + \mathcal{L}(q). \tag{2.8}$$

It has been proven that the KL divergence is non-negative, which means that $\mathcal{L}(q)$ cannot exceed the true log likelihood, and is therefore a lower bound for it. So when optimising $q(\mathbf{H})$ through minimisation of the KL divergence, one can instead do it through maximisation of $\mathcal{L}(q)$ [J. M. Winn 2004]. This relationship is illustrated on figure 2.1.

An important aspect of maximising the lower bound is that overfitting of the approximated distribution cannot occur and the cost of using the approximated distribution should be continuously falling when iterating towards an optimal solution. A rising cost points towards bugs in the algorithm. [C. M. Bishop 2006].

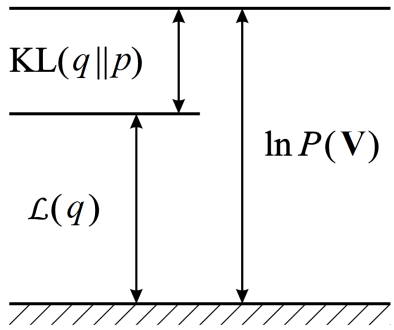


Figure 2.1: The relationship between the true log likelihood, the KL divergence and $\mathcal{L}(q)$. It can be seen that $\mathcal{L}(q)$ cannot exceed the true log likelihood and therefore can be used as a lower bound for it. Modified from C. Bishop [1999].

2.1.2 Maximising the similarity between the true distribution and its approximation

Before maximising $\mathcal{L}(q)$ it is necessary to decide the simplifying assumptions regarding $q(\mathbf{H})$, that makes it easier to work with than the true distribution. A common simplification is to assume that $q(\mathbf{H})$ can be factorised through

$$q(\mathbf{H}) = \prod_i q_i(\mathbf{H}_i), \tag{2.9}$$

meaning that there are no conditional distributions in $q(\mathbf{H})$. This simplification is originally known in physics as *mean field theory* [C. M. Bishop 2006]. Using this factorising assumption the lower bound can then be defined as

$$\mathcal{L}(q) = \int \prod_i q_i(\mathbf{H}_i) \left(\ln p(\mathbf{H}, \mathbf{V}) - \sum_k \ln q_k(\mathbf{H}_k) \right) d\mathbf{H}. \tag{2.10}$$

$\mathcal{L}(q)$ is then optimised with respect to the approximated distribution for each variable, $q_j(\mathbf{H}_j)$, and all terms not dependent on $q_j(\mathbf{H}_j)$ are combined in a constant term, C.

$$\begin{aligned} \mathcal{L}(q_j) = &\int q_j(\mathbf{H}_j) \left(\int \prod_{i \neq j} q_i(\mathbf{H}_i) \ln p(\mathbf{H}, \mathbf{V}) \; d\mathbf{H}_{/j} \right) d\mathbf{H}_j \\ &- \int q_j(\mathbf{H}_j) \left(\int \prod_{i \neq j} q_i(\mathbf{H}_i) \sum_k \ln q_k(\mathbf{H}_k) \; d\mathbf{H}_{/j} \right) d\mathbf{H}_j \end{aligned} \tag{2.11}$$

$$= \int q_j(\mathbf{H}_j) \langle \ln p(\mathbf{H}, \mathbf{V}) \rangle_{\mathbf{H}/j} \, d\mathbf{H}_j - \int q_j(\mathbf{H}_j) \left(\ln q_j(\mathbf{H}_j) \int \prod_{i \neq j} q_i(\mathbf{H}_i) \; d\mathbf{H}_{/j} \right.$$

$$\left. + \int \prod_{i \neq j} q_i(\mathbf{H}_i) \sum_{k \neq j} \ln q_k(\mathbf{H}_k) \; d\mathbf{H}_{/j} \right) d\mathbf{H}_j \tag{2.12}$$

$$= \int q_j(\mathbf{H}_j) \langle \ln p(\mathbf{H}, \mathbf{V}) \rangle_{\mathbf{H}/j} \, d\mathbf{H}_j - \int q_j(\mathbf{H}_j) \ln q_j(\mathbf{H}_j) \; d\mathbf{H}_j + C, \tag{2.13}$$

where $\int d\mathbf{H}_{/j}$ and $\langle \, \rangle_{\mathbf{H}/j}$ signifies the integration and expectation with respect to all variables in $\mathbf{H}$, except $\mathbf{H}_j$.

By defining $\ln \hat{p}(\mathbf{H}_j, \mathbf{V}) = \langle \ln p(\mathbf{H}, \mathbf{V}) \rangle_{\mathbf{H}/j}$, the lower bound can be expressed as a negative KL divergence,

$$\mathcal{L}(q_j) = -\mathrm{KL}(q_j \| \hat{p}) + C, \tag{2.14}$$

which means that the lower bound can be maximised through minimisation of this divergence. A minimisation which occurs when $q_j(\mathbf{H}_j) = \hat{p}(\mathbf{H}_j, \mathbf{V})$. The optimal solution for each distribution, $q_j^*(\mathbf{H}_j)$, can then be found by

$$\ln q_j^*(\mathbf{H}_j) = \langle \ln p(\mathbf{H}, \mathbf{V}) \rangle_{\mathbf{H}/j} + C, \tag{2.15}$$

which is easy to work with when working with exponential distributions [Murphy 2012; C. M. Bishop 2006].

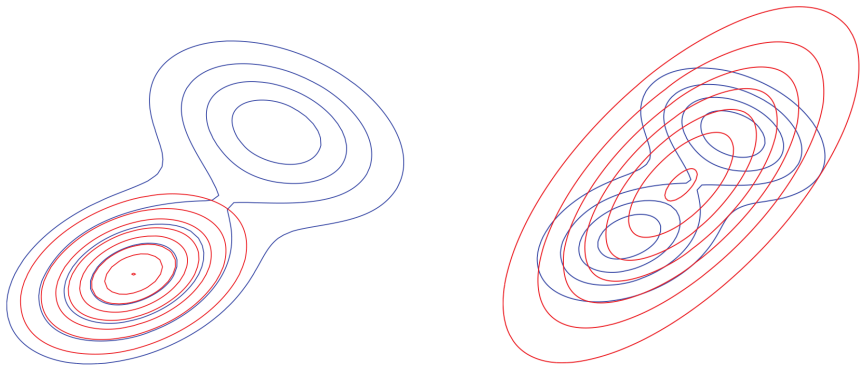
The resulting optimisation is similar to the EM algorithm. After a suitable initialisation, $\mathcal{L}(q)$ is optimised with respect to the approximated distributions for each variable in turn. This process is repeated until convergence has occured, which is guaranteed through the fact that the lower bound is convex with respect to each $q_j(\mathbf{H}_j)$ [C. M. Bishop 2006].

2.1.3 Consequences of estimating the posterior distribution through simplifying assumptions

Everything comes at a cost and so does the simplification of a posterior distribution. Having strong dependencies in the true posterior often results in that $\mathcal{L}(q)$ is not a convex function of the variational distribution when regarding all variables, and that different local maxima may be reached, depending on initialisation. This makes the initialisation important and leads to the usage of other algorithms, like common spatial patterns and maximum likelihood estimation, to find suitable starting values for variational inference [Wu et al. 2011; Wang 2007].

The problem with a multimodal true posterior can be explained in how the KL divergence is minimised. As illustrated on figure 2.2 the minimisation of $\mathrm{KL}(q\|p)$ will lead

to a distribution of $q(\mathbf{H})$, which have its mass in areas where the true distribution has a high probability, but may ignore other areas with a high probability. The reverse is true for the minimisation of $\mathrm{KL}(p\|q)$, which results in $q(\mathbf{H})$ covering all areas where the true distribution has a high probability even if it means covering areas of low probability [Minka 2005]. The minimisation of $\mathrm{KL}(p\|q)$ is not possible through factorised variational inference though, as this would require the expectation with respect to the intractable $p(\mathbf{H}|\mathbf{V})$, when calculating the lower bound [J. M. Winn 2004].



(a) Minimising $\mathrm{KL}(q\|p)$ **(b)** Minimising $\mathrm{KL}(p\|q)$

Figure 2.2: Approximation (red) of the true bimodal distribution (blue) through minimisation of a) $\mathrm{KL}(q\|p)$ and b) $\mathrm{KL}(p\|q)$ Murphy [2012].

2.2 Variational message passing

The factorisation in variational Bayes can be viewed as the decomposition of a large network into a subset of factors that individually can be approximated variationally, creating a message–passing algorithm. Variational message passing (VMP) by J. Winn et al. [2005] is an efficient implementation of this principle in which each factor is only conditioned on variables in the same Markov blanket. Because VMP constrains the factors to be in the exponential family and conjugate with respect to the distributions they are conditioned on (their parents), the variational updates simplifies greatly [Attias 2000]. It is thus possible to write a conditional distribution of the exponential family on a generic form that allows the algorithm to extract sufficient statistics and pass on as a message. The receiving node can then update its posterior belief from all the incoming messages.

VMP is a special case of a larger collection of message–passing algorithms that all rely on minimising the α–divergence. What makes VMP unique is that it, like mean

field theory, seeks the minimisation of the exclusive KL–divergence ($\alpha = 0$), which ensures that minimising the local divergence exactly minimises the global divergence. Expectation Propagation ($\alpha = 0$) is another optimisation scheme, which works for KL($p\|q$). [Minka 2005].

To implement VMP we have used the probabilistic programming framework Infer.NET developed by Minka et al. [2013].

2.2.1 Connection to variational inference

The joint probability distribution $p(\mathbf{S})$ of a directed acyclic graph can be obtained as the product of the distributions of each node S_i [Jordan 1999]

$$p(\mathbf{S}) = \prod_i p(S_i|\mathrm{pa}_i) \tag{2.16}$$

where pa_i is the parents of node i and S_i is the variables in node i. Recall from section 2.1.2 that, using $\mathbf{H}$ as the latent variables, the variational approximation of this can be expressed as

$$q(\mathbf{H}) = \prod_i q_i(H_i)$$

Minimising the Kullback-Leibler divergence we derive the same expression as in (2.15)

$$\ln q_j^*(H_j) = \langle \ln p(\mathbf{H}, \mathbf{V}) \rangle_{\mathbf{H}/j} + C \tag{2.17}$$

$$= \left\langle \sum_i \ln p(S_i|\mathrm{pa}_i) \right\rangle_{\mathbf{H}/j} + C \tag{2.18}$$

Terms that do not depend on H_J is constant which leaves the conditional of H_J and the conditionals of all the children of H_J

$$\ln q_j^*(H_J) = \langle \ln p(H_J|\mathrm{pa}_j) \rangle_{\mathbf{H}/j} + \sum_{k \in \mathrm{ch}_j} \langle p(S_k|\mathrm{pa}_k) \rangle_{\mathbf{H}/j} + C \tag{2.19}$$

The local distribution can be updated with the messages from the connected nodes, but because these nodes have their own dependencies, the updating involves all of the variables in the Markov blanket. These includes parents, children and co-parents as shown in figure 2.3.

2.2.2 Exponential form

A distribution of the exponential family can be written on the form

$$p(H|V) = \exp[\boldsymbol{\phi}(V)^T \mathbf{u}(H) + f(H) + g(V)] \tag{2.20}$$

with $\boldsymbol{\phi}(V)$ as the *natural parameter vector*, $\mathbf{u}(H)$ as the *natural statistic vector* and $g(V)$ as a normalisation function. The conjugacy constraint ensures that distributions has the same functional form as the priors so optimising a distribution only

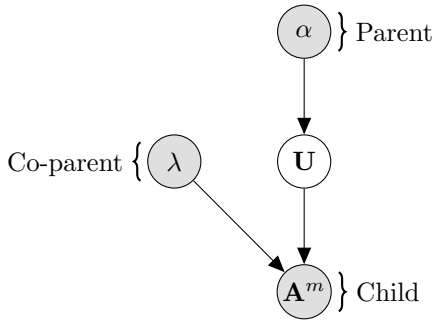


Figure 2.3: Local variational distribution of **U**

changes its parameters, ensuring a multi-linear relationship between the logarithm of a conditional distribution, H, its natural statistic functions, $\mathbf{u}$, and its parents, V.

2.2.3 Example: univariate Gaussian distribution

Using the model in figure 2.3 as an example we will show how the distribution over $\mathbf{U}$ is updated. The columns of $\mathbf{U}$ and $\mathbf{A}$ are independent and can thus be modelled by univariate Gaussians. In the following u_d is an element from the model and $\mathbf{u}$ is the *natural parameter vector* introduced in (2.20). Assuming $\mathbf{a}$ is a column vector, rewriting the log conditional of the model with respect to a_d we get

$$\ln p(a_d|u_d,\lambda) = \ln\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} - \frac{\lambda}{2}(a_d-u_d)^2 \tag{2.21}$$

$$= \frac{1}{2}(\ln\lambda - \ln(2\pi)) - \frac{\lambda}{2}(a_d^2+u_d^2-2w_n u_d) \tag{2.22}$$

$$= \underbrace{\begin{bmatrix} \lambda u_d \\ -\lambda/2 \end{bmatrix}^{\mathrm{T}}}_{\phi_{a_d}(u_d,\lambda)} \underbrace{\begin{bmatrix} a_d \\ a_d^2 \end{bmatrix}}_{\mathbf{u}_{a_d}(a_d)} + \underbrace{\frac{1}{2}(\ln\lambda - \lambda u_d^2)}_{g_{a_d}(u_d,\lambda)} \underbrace{-\frac{1}{2}\ln(2\pi)}_{f_{a_d}(a_d)} \tag{2.23}$$

Separating out the dependencies is done by rewriting (2.21) with respect to these

$$\ln p(a_d|u_d,\lambda) = \begin{bmatrix} -\frac{1}{2}(a_d-u_d)^2 \\ \frac{1}{2} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \lambda \\ \ln\lambda \end{bmatrix} - \ln 2\pi \tag{2.24}$$

$$= \underbrace{\begin{bmatrix} \lambda a_d \\ -\lambda/2 \end{bmatrix}^{\mathrm{T}}}_{\phi_{a_d u_d}} \underbrace{\begin{bmatrix} u_d \\ u_d^2 \end{bmatrix}}_{\mathbf{u}_{a_d}(v_d)} + \frac{1}{2}(\ln\lambda - \ln(2\pi) - \lambda a^2) \tag{2.25}$$

Since the priors are confined to be exponential conjugate, the prior on u_d must be a normal distribution and the prior on λ must be a gamma distribution or alternatively

a normal-gamma can be used over both. From the natural statistics vector in (2.24) and (2.25) it is furthermore possible to see the shape that natural statics vectors must take in the parents in order to uphold the linear relationship. Isolating u_d and λ in the natural statistics vector of the respective distributions we get the expressions

$$\ln p(\lambda|a_\lambda, b_\lambda) = \begin{bmatrix} b_\lambda \\ a_\lambda - 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \lambda \\ \ln\lambda \end{bmatrix} + a_\lambda \ln b_\lambda - \Gamma(a_\lambda) \tag{2.26}$$

$$\begin{aligned} \ln p(u_d|0,\alpha) &= \begin{bmatrix} 0 \\ -\alpha/2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} u_d \\ u_d^2 \end{bmatrix} + \frac{1}{2}(\ln\alpha - \ln(2\pi)) \\ &= \begin{bmatrix} -\frac{1}{2}u_d^2 \\ -\frac{1}{2} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \alpha \\ \ln\alpha \end{bmatrix} - \ln(2\pi) \end{aligned} \tag{2.27}$$

with the expression for α being exactly the same as (2.26) but with α instead of λ. The initialisation of the distributions is important and very model dependent but if unspecified a broad non-informative prior is assumed.

Using (2.19) the distribution $q_{u_d}(u_d)$ can be updated with

$$\begin{aligned} \ln q^*_{u_d}(u_d) &= \langle \ln p(u_d|\alpha)\rangle_{\sim q(u_d)} + \langle \ln p(a_d|u_d,\lambda)\rangle_{\sim q(u_d)} + C \\ &= \left[\langle \phi_{u_d}(\alpha)\rangle_{\sim q(u_d)} + \langle \phi_{a_d u_d}(a_d,\lambda)\rangle_{\sim q(u_d)}\right]^T \mathbf{u}_{u_d}(u_d) + f_{u_d}(u_d) + C \end{aligned} \tag{2.28}$$

Because $q^*_{u_d}$ is a conjugate exponential and thus on the same form as $p(u_d|\alpha)$ it follows that the natural parameter vector

$$\phi^*_{u_d} = \langle \phi_{u_d}(\alpha)\rangle + \langle \phi_{a_d u_d}(a_d,\lambda)\rangle \tag{2.29}$$

is all that is needed to update the posterior and that the natural parameter vectors are multi-linear functions of the natural static vectors. It is then possible to reparameterise the expectations in (2.29) into

$$\widetilde{\phi}_{u_d}(\langle \mathbf{u}_\alpha\rangle) = \langle \phi_{u_d}(\alpha)\rangle \tag{2.30}$$

$$\widetilde{\phi}_{a_d u_d}(\langle \mathbf{u}_{a_d}\rangle, \langle \mathbf{u}_\lambda\rangle) = \langle \phi_{a_d u_d}(a_d,\lambda)\rangle \tag{2.31}$$

It is thus evident that the message from parent to child must be on the form

$$\mathbf{m}_{\alpha\to u_d} = \langle \mathbf{u}_\alpha\rangle \tag{2.32}$$

and from child to parent

$$\mathbf{m}_{a_d\to u_d} = \widetilde{\phi}_{a_d u_d}(\langle \mathbf{u}_{a_d}\rangle, \mathbf{m}_{\lambda\to a_d}) \tag{2.33}$$

So, to update u_d we need a message from a_d in the form of (2.33) but this requires updates from the co-parents of u_d, i.e. λ, in the form of (2.32)

$$\mathbf{m}_{\lambda\to a_d} = \begin{bmatrix} \langle\lambda\rangle \\ \langle\ln\lambda\rangle \end{bmatrix} \tag{2.34}$$

From (2.33) and (2.25) we get the message to u_d from a_d

$$\mathbf{m}_{a_d \to u_d} = \begin{bmatrix} \langle\lambda\rangle\langle a_d\rangle \\ -\langle\lambda\rangle/2 \end{bmatrix} \tag{2.35}$$

and the message from α to u_d

$$\mathbf{m}_{\alpha \to u_d} = \begin{bmatrix} \langle\alpha\rangle \\ \langle\ln\alpha\rangle \end{bmatrix} \tag{2.36}$$

When messages from all parents and children are received we see from (2.29) that the posterior $q^*_{u_d}$ can be updated by updating the natural parameter vector

$$\phi^*_{u_d} = \begin{bmatrix} 0 \\ -\langle\alpha\rangle/2 \end{bmatrix} + \begin{bmatrix} \langle\lambda\rangle\langle a_d\rangle \\ -\langle\lambda\rangle/2 \end{bmatrix} \tag{2.37}$$

The new expectation of $\langle\mathbf{u}_{u_d}\rangle_{q^*_{u_d}}$ can then be computed using

$$\langle\mathbf{u}_{u_d}\rangle_{q^*_{u_d}} = -\frac{d\tilde{g}(\phi)}{d\phi} \tag{2.38}$$

where $\tilde{g}(\phi)$ is a reparameterisation of $g_{u_d}(\alpha)$ with respect to ϕ. By applying (2.38) it is possible to derive the expectation of the natural statics vector $\mathbf{u}(u_d)$

$$\langle\mathbf{u}(u_d)\rangle = \begin{bmatrix} \langle u_d\rangle \\ \langle u_d^2\rangle \end{bmatrix} = \begin{bmatrix} 0 \\ \langle\alpha\rangle^{-1} \end{bmatrix} \tag{2.39}$$

which is passed on as a message to any child node of u_d. Similar computations are required to update the rest of the variables in the model, but since these can be predetermined with respect to parent-child combinations of nodes, it is straightforward to implement in software such as Infer.NET.

2.3 Review of methods for finding hidden correlations in datasets

This section contains a short review of three other models, which are used to find correlated time series in two or more datasets. These models will later be used for comparison when testing the capabilities of BCoCA. First follows a brief explanation of the classical CCA and then how CoCA relates to it. Lastly Bayesian group factor analysis (GFA) [Virtanen et al. 2011], a latent model approach to CCA, will be reviewed.

2.3.1 Canonical correlation analysis

CCA seeks to maximise the correlation between two time series vectors $\mathbf{y}_1 = \mathbf{X}^{(1)T}\mathbf{w}^{(1)}_k$ and $\mathbf{y}_2 = \mathbf{X}^{(2)T}\mathbf{w}^{(2)}_k$. At the same time CCA constrains the estimated weights with the condition that $\mathbf{X}^{(1)T}\mathbf{w}^{(1)}_k$ and $\mathbf{X}^{(1)T}\mathbf{w}^{(1)}_{k'}$ are uncorrelated for $k \neq k$ [Klami 2013].

The maximum correlation is found by maximising the correlation coefficient;

$$\rho = \arg\max_{\mathbf{w}} \frac{\mathbf{y}_1^T \mathbf{y}_2}{\|\mathbf{y}_1\| \|\mathbf{y}_2\|}. \tag{2.40}$$

Introducing the sample covariance matrix, $\mathbf{R}_{ij} = \frac{1}{N}\mathbf{X}^{(i)}\mathbf{X}^{(j)T}$, (2.40) can be rewritten to

$$\rho = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{(1)T}\mathbf{R}_{12}\mathbf{w}^{(2)}}{\sqrt{\mathbf{w}^{(1)T}\mathbf{R}_{11}\mathbf{w}^{(1)}}\sqrt{\mathbf{w}^{(2)T}\mathbf{R}_{22}\mathbf{w}^{(2)}}}. \tag{2.41}$$

CCA then finds the weights analytically through two eigenvalue decompositions [Hardoon et al. 2004];

$$\begin{aligned} \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{w}^{(1)} &= \rho^2\mathbf{w}^{(1)} \\ \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{w}^{(2)} &= \rho^2\mathbf{w}^{(2)}. \end{aligned} \tag{2.42}$$

2.3.2 Correlated component analysis

CoCA is based on the same method as CCA, but differentiates itself by finding *one* set of weights that works for filtering both datasets. This means fewer degrees of freedom and the ability to drop the constraint of orthogonality between weights, which is not meaningful when dealing with EEG where they can be seen as spatial filters [Dmochowski et al. 2012]. (2.41) is then simplified to

$$\rho = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{R}_{12}\mathbf{w}}{\sqrt{\mathbf{w}^T\mathbf{R}_{11}\mathbf{w}}\sqrt{\mathbf{w}^T\mathbf{R}_{22}\mathbf{w}}}. \tag{2.43}$$

To maximise this expression the derivative with respect to $\mathbf{w}$ is set to zero. Introducing the scalar power expression, $\sigma_{ij} = \mathbf{w}^T\mathbf{R}_{ij}\mathbf{w}$, the equation for $\mathbf{w}$ can be expressed as

$$\mathbf{R}_{12}\mathbf{w}\frac{\sigma_{11}\sigma_{22}}{\sigma_{12}} = (\mathbf{R}_{11}\sigma_{22} + \mathbf{R}_{22}\sigma_{11})\,\mathbf{w}. \tag{2.44}$$

Assuming the two datasets have similar levels of power ($\sigma_{11} \approx \sigma_{22}$) the equation can be changed into a generalised eigenvalue equation;

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1}\,\mathbf{R}_{12}\mathbf{w} = \frac{\sigma_{12}}{\sigma_{11}}\mathbf{w} \tag{2.45}$$

To be able to guarantee real eigenvalues, $\mathbf{R}_{12}$ has to be symmetric. This is not likely, but since $\mathbf{R}_{12} = \mathbf{R}_{21}^T$ the matrix $(\mathbf{R}_{12} + \mathbf{R}_{21})$ is symmetric. Before making

this symmetrisation it is first necessary to prove that $\sigma_{12} = \sigma_{21}$ using the fact that $Tr(\mathbf{AB}) = Tr(\mathbf{BA})$ and $\mathbf{a}^T \cdot \mathbf{b} = Tr(\mathbf{b} \cdot \mathbf{a}^T)$ [Petersen et al. 2006]:

$$\sigma_{12} = \mathbf{w}^T \mathbf{X}_1 \mathbf{X}_2^T \mathbf{w} = Tr((\mathbf{X}_1 \mathbf{X}_2^T \mathbf{w}\mathbf{w}^T)^T) \tag{2.46}$$

$$= \mathbf{w}^T \mathbf{X}_2 \mathbf{X}_1^T \mathbf{w} = \sigma_{21} \tag{2.47}$$

Then the cross-covariance matrix can be symmetrised:

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1} \mathbf{R}_{21}\mathbf{w} + (\mathbf{R}_{11} + \mathbf{R}_{22})^{-1} \mathbf{R}_{12}\mathbf{w}$$

$$= \frac{\sigma_{21}}{\sigma_{11}}\mathbf{w} + \frac{\sigma_{12}}{\sigma_{11}}\mathbf{w} \quad \Leftrightarrow \tag{2.48}$$

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1} (\mathbf{R}_{12} + \mathbf{R}_{21})\, \mathbf{w} = 2 \cdot \frac{\sigma_{12}}{\sigma_{11}}\mathbf{w}\ . \tag{2.49}$$

Proof of Real Eigenvalues

Another property that needs to be in place to prove real eigenvalues is that since $\mathbf{R}_{ii} = \frac{1}{T}\mathbf{X}_i\mathbf{X}_i^T$ is symmetric and positive definite, $(\mathbf{R}_{11} + \mathbf{R}_{22})$ is also symmetric and positive definite. This also ensures that the inverse of that matrix as well as $(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1/2}$ exists and that they are symmetric.

Defining equation (2.49) as $(\mathbf{B}^{-1}\mathbf{A}\mathbf{w} = \lambda\mathbf{w})$, the matrix $\mathbf{B}^{-1}\mathbf{A}$ cannot be proved to be symmetric, but writing up the characteristic polynomial and rearranging;

$$|\ \mathbf{B}^{-1}\mathbf{A} - \lambda\mathbf{I}\ | = 0 \quad \Leftrightarrow \tag{2.50}$$

$$|\mathbf{B}^{1/2}|\ |\ \mathbf{B}^{-1}\mathbf{A} - \lambda\mathbf{I}\ |\ |\mathbf{B}^{-1/2}| = 0 \quad \Leftrightarrow \tag{2.51}$$

$$|\ \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} - \lambda\mathbf{I}\ | = 0, \tag{2.52}$$

proves that since $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is symmetric and thereby have real eigenvalues, the same must be the case for $\mathbf{B}^{-1}\mathbf{A}$. This proves that the eigenvalues gained from CoCA, when calculating the weights, are real.

Capability of CoCA when the true weights are dissimilar

Our initial assumption was that CoCA would attain poor results, when the true weigths of each dataset where different from each other. However tests with simulated data, proved only a small drop in performance. This lead to the following analytic investigation of CoCA in the worst case scenario for the two-view situation, where the weights are orthogonal.

The observations are assumed to consist of a single true signal mixed into D dimensions by a vector and Gaussian noise;

$$\mathbf{X}_1 = \mathbf{a}_1\mathbf{z} + \epsilon\ , \quad \mathbf{X}_2 = \mathbf{a}_2\mathbf{z} + \epsilon. \tag{2.53}$$

Given enough samples, the sample covariance matrices can be defined as

$$\mathbf{R}_{11} = P \cdot \mathbf{a}_1 \mathbf{a}_1^T + \sigma^2 \mathbf{I}\,, \quad \mathbf{R}_{12} = P \cdot \mathbf{a}_1 \mathbf{a}_2^T, \tag{2.54}$$

where P signifies the power of $\mathbf{z}$ and σ^2 signifies the noise variance. For simplicity the weight vectors are assumed to have unit length.

The two matrices in (2.49) can now be written as

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1} = \frac{1}{P}\left(\mathbf{a}_1 \mathbf{a}_1^T + \mathbf{a}_2 \mathbf{a}_2^T + \frac{2\sigma^2}{P}\mathbf{I}\right)^{-1} \quad \Leftrightarrow \tag{2.55}$$

$$= \frac{1}{P}\left([\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix} + \frac{2\sigma^2}{P}\mathbf{I}\right)^{-1} \tag{2.56}$$

$$\mathbf{R}_{12} + \mathbf{R}_{21} = P \cdot [\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix} \tag{2.57}$$

using block matrix notation. With $\mathbf{a}_1^T \mathbf{a}_2 = 0$, $\|\mathbf{a}_1\|^2 = \|\mathbf{a}_2\|^2 = 1$ and the Woodbury identity, (2.56) can be expressed as;

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1} = \frac{1}{2\sigma^2}\left(\mathbf{I} - \frac{P}{2\sigma^2}[\mathbf{a}_1\ \mathbf{a}_2]\left(\mathbf{I} - \frac{P}{2\sigma^2}\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix}[\mathbf{a}_1\ \mathbf{a}_2]\right)^{-1}\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix}\right) \tag{2.58}$$

$$= \frac{1}{2\sigma^2}\left(\mathbf{I} - \frac{P}{2\sigma^2}[\mathbf{a}_1\ \mathbf{a}_2]\left(\mathbf{I} - \frac{P}{2\sigma^2}\begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix}\right) \tag{2.59}$$

$$= \frac{1}{2\sigma^2}\left(\mathbf{I} - \frac{P}{2\sigma^2 + P}[\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix}\right). \tag{2.60}$$

The matrix product of (2.56) and (2.57) then gives

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1}(\mathbf{R}_{12} + \mathbf{R}_{21}) = \frac{P}{2\sigma^2}\left(\mathbf{I} - \frac{P}{2\sigma^2 + P}[\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix}\right)[\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix} \tag{2.61}$$

$$= \frac{P}{2\sigma^2}\left(1 - \frac{P}{2\sigma^2 + P}\right)[\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix} \tag{2.62}$$

$$= \frac{P}{2\sigma^2 + P}(\mathbf{a}_1 \mathbf{a}_1^T + \mathbf{a}_2 \mathbf{a}_2^T) \tag{2.63}$$

Using the simplifying assumptions made earlier an eigenvector for (2.63) can be seen to have the form $\alpha\mathbf{a}_1 + \beta\mathbf{a}_2$ since

$$\frac{P}{2\sigma^2 + P}(\mathbf{a}_1 \mathbf{a}_1^T + \mathbf{a}_2 \mathbf{a}_2^T)(\alpha\mathbf{a}_1 + \beta\mathbf{a}_2) = \frac{P}{2\sigma^2 + P}(\alpha\mathbf{a}_2 + \beta\mathbf{a}_1). \tag{2.64}$$

It can be seen that $\mathbf{w} = \alpha\mathbf{a}_1 + \beta\mathbf{a}_2$ is an eigenvector when either $\alpha = \beta$ or $\alpha = -\beta$ with $\pm\frac{P}{2\sigma^2+P}$ as eigenvalues. This means that when the true mixing weights of two datasets are orthogonal CoCA finds a common weight, consisting of the mean of the true weights.

2.3.3 Bayesian group factor analysis

GFA is a Bayesian latent model introduced in Virtanen et al. [2011], which is able to compare multiple datasets at the same time to find common latent sources, as well as view specific sources.[1] It works by concatenating all datasets feature-wise, $\bar{\mathbf{X}}^T = [\mathbf{X}^{(1)T}, \ldots, \mathbf{X}^{(M)T}]$, and treating it as one combined variable;

$$\bar{\mathbf{X}} = \mathbf{AZ} + \mathbf{E}. \tag{2.65}$$

$\mathbf{Z}$ consists of shared sources, view-specific sources, and view-specific structured-noise sources. $\mathbf{E}$ is a diagonal matrix where all the variances, σ_m, are equal for each view. $\mathbf{A}$ consists of the weights for all datasets, where each element is calculated separately and the number of components are controlled through group-wise ARD;

$$p(\mathbf{A}) = \prod_m^M \prod_k^K \prod_d^{D_m} \mathcal{N}\left(\mathbf{a}_{m,k}(d)|0, \alpha_{m,k}^{-1}\right) \tag{2.66}$$

$$p(\boldsymbol{\alpha}) = \prod_m^M \prod_k^K \mathcal{G}a(\alpha_{m,k}|a_0, b_0) \tag{2.67}$$

Figure 2.4 visualises how $\mathbf{A}$ controls which components in $\mathbf{Z}$ that are shared, view-specific or structured noise. GFA like CCA (and unlike CoCA) has the benefit of being able to find correlates in datasets with different number of dimensions.

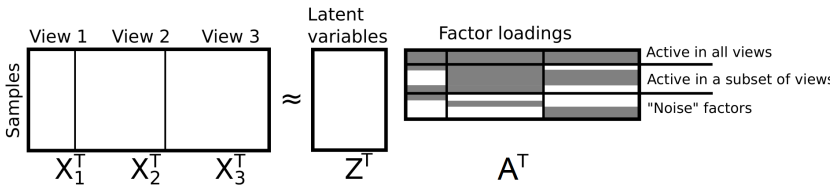


Figure 2.4: Illustration of the concatenation of data and how $\mathbf{A}$ controls which components in $\mathbf{Z}$ that are shared, view-specific or structured noise. Modified from Virtanen et al. [2011].

[1] Note that Virtanen et al. [2011] uses a transposed notation for their data e.g. their observation samples lie in rows of $\mathbf{X}$, where they lie in columns in this thesis. The equations in this section as well as figure 2.4 have been altered to match the notation employed in the rest of this thesis.

2.4 Derivation of Bayesian correlated component analysis

This chapter will present two different prior distributions for a latent model approach to multi-set correlated component analysis and the derivation of their posterior distributions through variational Bayesian inference. In section 2.9 both models will be tested to decide which to use for the remainder of this thesis.

The difference between the two models is only in how they model the relationship between the weights, $\mathbf{A}^{(m)}$. They have the same prior distributions for the rest of the variables, which are inspired by Bayesian CCA (BCCA) [Klami 2013; Wang 2007; Wu et al. 2011];

$$p(\mathbf{X}|\mathbf{Z},\mathbf{A},\boldsymbol{\Psi}) = \prod_m^M \prod_n^N \mathcal{N}\left(\mathbf{x}_n^{(m)}|\mathbf{A}^{(m)}\mathbf{z}_n, {\Psi^{(m)}}^{-1}\right) \tag{2.68}$$

$$p(\mathbf{Z}) = \prod_n^N \mathcal{N}(\mathbf{z}_n|\mathbf{0},\mathbf{I}) \tag{2.69}$$

$$p(\Psi) = \prod_m^M \mathcal{W}(\Psi^{(m)}|\mathbf{S}_0, v_0) \tag{2.70}$$

$$p(\boldsymbol{\alpha}) = \prod_k^K \mathcal{G}a(\alpha_k|a_0,b_0) \tag{2.71}$$

$$p(\lambda) = \mathcal{G}a(a_0,b_0), \tag{2.72}$$

where $\boldsymbol{\alpha}$ is an ARD parameter regularising the number of components as used in BCCA. Where BCoCA differentiates itself from the other BCCA models is by the λ variable. It regularises the similarity between the weights for each dataset and is itself regularised through ARD. Since both α and λ are defined as precisions for Gaussian distributions, they are modelled as belonging to a gamma distribution, $\mathcal{G}a$. For the same reason it was chosen to model the precision matrix, Ψ, instead of the covariance matrix, as its conjugate prior is the Wishart distribution, $\mathcal{W}$.

Below follows a description of the underlying idea of each model, as well as derivations and the final updates. For the purpose of readability, the number of equations in the derivations have been reduced. Instead an extensive example of the derivations for a Bayesian approach to principal component analysis have been supplied in appendix A. The example is simpler and easier to understand compared to BCoCA, but most of the concepts can be directly transferred to the derivations presented in the following.

2.4.1 Bayesian CoCA based on mean weights

The significant aspect of this model is that the relationship among the $\mathbf{A}$s is based on a shared mean weight. This can be seen in the prior distribution for $\mathbf{A}$, which have been expanded to include the latent variable, $\mathbf{U}$, representing the mean weight

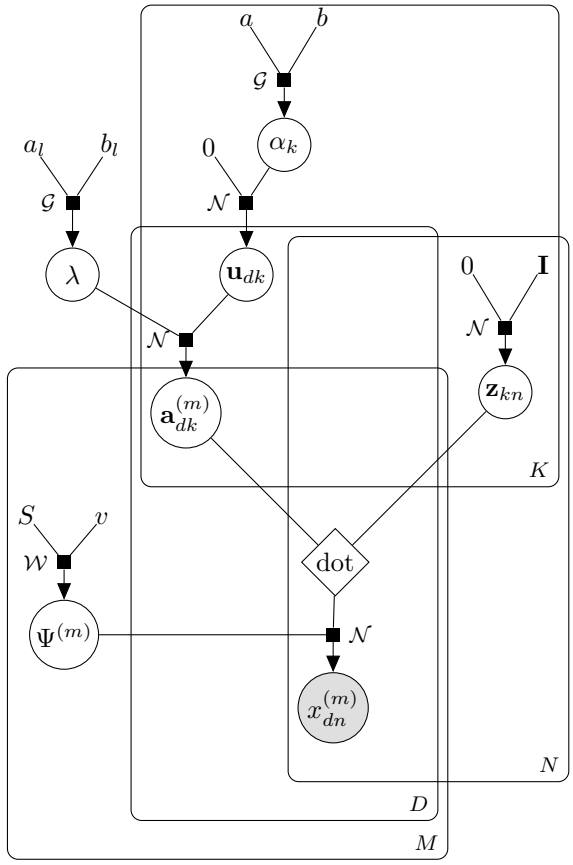


Figure 2.5: BCoCA factor graph showing the relationship between variables and their hyperparameters.

matrix across all datasets. The ARD variable λ now regularises how close the $\mathbf{A}$s lie to $\mathbf{U}$;

$$p(\mathbf{U}|\boldsymbol{\alpha}) = \prod_{k}^{K} \mathcal{N}\left(\mathbf{u}_k|\mathbf{0}, \alpha_k^{-1}\right) \tag{2.73}$$

$$p(\mathbf{A}|\mathbf{U}, \lambda) = \prod_{m}^{M} \prod_{k}^{K} \mathcal{N}\left(\mathbf{a}_k^{(m)}|\mathbf{u}_k, \lambda^{-1}\right) \tag{2.74}$$

The relationship between the variables is also illustrated in figure 2.5. By the having this structure for the model and using only conjugate distributions, the model also fits within the VMP framework and enables a comparison between it and the Matlab implementation developed for this thesis.

The joint probability is then given by

$$\begin{aligned} p(\mathbf{V},\mathbf{H}) &= p(\mathbf{X},\mathbf{Z},\mathbf{A},\mathbf{U},\boldsymbol{\Psi},\boldsymbol{\alpha},\lambda) \\ &= p(\mathbf{X}|\mathbf{Z},\mathbf{A},\boldsymbol{\Psi})p(\mathbf{Z})p(\Psi)p(\mathbf{A}|\mathbf{U},\boldsymbol{\alpha})p(\mathbf{U}|\lambda)p(\boldsymbol{\alpha})p(\lambda). \end{aligned} \tag{2.75}$$

In the rest of this section follows derivations for variational updates using (2.15).

Posterior for Z

The logarithm of the distribution for $\mathbf{Z}$ is approximated by

$$\ln q(\mathbf{Z}) = \langle \ln p(\mathbf{X}|\mathbf{Z},\mathbf{A},\Psi)\rangle_{/\mathbf{z}} + \ln p(\mathbf{Z}) + C \quad \Leftrightarrow \tag{2.76}$$

$$= \sum_m^M \sum_n^N \left\langle -\frac{1}{2}\left(\mathbf{z}_n^T \mathbf{A}^{(m)T}\Psi^{(m)}\mathbf{A}^{(m)}\mathbf{z}_n - 2\mathbf{z}_n^T\mathbf{A}^{(m)T}\Psi^{(m)}\mathbf{x}_n^{(m)}\right)\right\rangle_{/\mathbf{z}} - \sum_n^N \frac{1}{2}||\mathbf{z}_n||^2 + C \quad \Leftrightarrow \tag{2.77}$$

$$= \sum_n^N -\frac{1}{2}\mathbf{z}_n^T\left(\sum_m^M\left\{\left\langle \mathbf{A}^{(m)T}\Psi^{(m)}\mathbf{A}^{(m)}\right\rangle\right\} + \mathbf{I}\right)\mathbf{z}_n + \mathbf{z}_n^T\sum_m^M\left\{\left\langle \mathbf{A}^{(m)T}\right\rangle\left\langle\Psi^{(m)}\right\rangle\mathbf{x}_n^{(m)}\right\} + C. \tag{2.78}$$

A normal distribution for $q(\mathbf{Z})$ can then be expressed as

$$q(\mathbf{Z}) = \prod_n^N \mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_{\mathbf{z},n},\Sigma_\mathbf{z}\right) \tag{2.79}$$

$$\Sigma_\mathbf{z}^{-1} = \sum_m^M\left\{\left\langle \mathbf{A}^{(m)T}\Psi^{(m)}\mathbf{A}^{(m)}\right\rangle\right\} + \mathbf{I} \tag{2.80}$$

$$\Sigma_\mathbf{z}^{-1}\boldsymbol{\mu}_{\mathbf{z},n} = \sum_m^M\left\langle \mathbf{A}^{(m)T}\right\rangle\left\langle\Psi^{(m)}\right\rangle\mathbf{x}_n^{(m)} \quad \Leftrightarrow \tag{2.81}$$

$$\boldsymbol{\mu}_{\mathbf{z},n} = \Sigma_\mathbf{z}\sum_m^M\left\langle \mathbf{A}^{(m)T}\right\rangle\left\langle\Psi^{(m)}\right\rangle\mathbf{x}_n^{(m)}. \tag{2.82}$$

where $\langle . \rangle$ signifies the expectation. The calculation of $\left\langle \mathbf{A}^T\Psi\mathbf{A}\right\rangle$ is not as straightforward as the rest of the expectations in this update. Doing the matrix multiplications first results in a $K \times K$ matrix, where each element is defined by $\left\langle \mathbf{a}_k^T\Psi\mathbf{a}_{k'}\right\rangle$. Using [Petersen et al. 2006, (378)] this can be calculated as $\mathrm{Tr}(\Psi\Sigma_{w_k}) + \left\langle \mathbf{a}_k^T\right\rangle\langle\Psi\rangle\langle\mathbf{a}_{k'}\rangle$.

The covariance matrix, Σ_{w_k}, is assumed diagonal so using the fact that $\mathrm{Tr}(\Sigma_{\mathbf{a}_k}) = \sum_d^D \Sigma_{\mathbf{a}_d}(k,k)$, the elements are then calculated as

$$\left\langle \mathbf{a}_k^T \Psi \mathbf{a}_{k'} \right\rangle = \sum_d^D \psi_{d,d} \Sigma_{\mathbf{a}_d}(k,k) + \left\langle \mathbf{a}_k^T \right\rangle \langle \Psi \rangle \langle \mathbf{a}_{k'} \rangle, \quad \text{for } k = k' \tag{2.83}$$

$$= \left\langle \mathbf{a}_k^T \right\rangle \langle \Psi \rangle \langle \mathbf{a}_{k'} \rangle, \quad \text{for } k \neq k'. \tag{2.84}$$

Note that there is no covariance when $k \neq k'$, since the columns of $\mathbf{A}$ are assumed independent. This means that the expectation of the entire matrix is given by

$$\left\langle \mathbf{A}^T \Psi \mathbf{A} \right\rangle = \sum_d^D \psi_{d,d} \Sigma_{\mathbf{a}_d} + \left\langle \mathbf{A}^T \right\rangle \langle \Psi \rangle \langle \mathbf{A} \rangle, \tag{2.85}$$

where $\mathbf{a}_d$ is a the d'th *row* of $\mathbf{A}$, for a more compact notation.

Posterior for Ψ

The logarithm of the distribution for Ψ is approximated by

$$\ln q(\Psi) = \langle \ln p(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \Psi) \rangle_{/\Psi} + \ln p(\Psi) + C \tag{2.86}$$

Using [Petersen et al. 2006, (15-17)] the approximation can be written as

$$\begin{aligned} \ln q(\Psi) = \sum_m^M \frac{N}{2} \ln \left|\Psi^{(m)}\right| - \frac{1}{2} \sum_n^N \Big\langle \mathrm{Tr}\left(\Psi^{(m)} \mathbf{A}^{(m)} \mathbf{z}_n (\mathbf{A}^{(m)} \mathbf{z}_n)^T\right) \\ + \mathrm{Tr}\left(\Psi^{(m)} \mathbf{x}_n^{(m)} \mathbf{x}_n^{(m)T}\right) - 2 \cdot \mathrm{Tr}\left(\Psi^{(m)} \mathbf{x}_n^{(m)} (\mathbf{A}^{(m)} \mathbf{z}_n)^T\right) \Big\rangle_{/\Psi} \\ + \frac{v_0 - D - 1}{2} \ln \left|\Psi^{(m)}\right| - \frac{1}{2} \mathrm{Tr}(\mathbf{S}_0^{-1} \Psi^{(m)}) + C \quad \Leftrightarrow \end{aligned} \tag{2.87}$$

$$\begin{aligned} = \sum_m^M \frac{1}{2} \ln \left|\Psi^{(m)}\right| (N + v_0 - D - 1) - \frac{1}{2} \mathrm{Tr} \Bigg\{ \Bigg(\Bigg\langle \mathbf{A}^{(m)} \sum_n^N \mathbf{z}_n \mathbf{z}_n^T \mathbf{A}^{(m)T} \Bigg\rangle \\ + \sum_n^N \mathbf{x}_n^{(m)} \mathbf{x}_n^{(m)T} - 2 \cdot \sum_n^N \mathbf{x}_n^{(m)} \left\langle \mathbf{z}_n^T \right\rangle \left\langle \mathbf{A}^{(m)T} \right\rangle + \mathbf{S}_0^{-1} \Bigg) \Psi^{(m)} \Bigg) \Bigg\} + C. \end{aligned} \tag{2.88}$$

A Wishart distribution for $q(\Psi)$ can then be expressed as

$$q(\Psi) = \prod_m^M \mathcal{W}\left(\mathbf{S}_\Psi^{(m)}, v_\Psi\right) \tag{2.89}$$

$$\mathbf{S}_\Psi^{(m)^{-1}} = \left\langle \mathbf{A}^{(m)} \sum_n^N \mathbf{z}_n \mathbf{z}_n^T \mathbf{A}^{(m)T} \right\rangle + \sum_n^N \mathbf{x}_n^{(m)} \mathbf{x}_n^{(m)T} - 2 \cdot \sum_n^N \mathbf{x}_n^{(m)} \left\langle \mathbf{z}_n^T \right\rangle \left\langle \mathbf{A}^{(m)T} \right\rangle + \mathbf{S}_0^{-1} \tag{2.90}$$

$$v_\Psi = N + v_0 \tag{2.91}$$

As with the update for $\mathbf{Z}$ the calculation of $\left\langle \mathbf{A} \sum_n^N \mathbf{z}_n \mathbf{z}_n^T \mathbf{A}^T \right\rangle$ poses some difficulty. Note that the result is now a $D \times D$ matrix with each element calculated as $\left\langle \mathbf{a}_d \sum_n^N \mathbf{z}_n \mathbf{z}_n^T \mathbf{a}_{d'}^T \right\rangle$. Using [Petersen et al. 2006, (378)] again the elements are calculated as

$$\left\langle \mathbf{a}_d \sum_n^N \mathbf{z}_n \mathbf{z}_n^T \mathbf{a}_{d'}^T \right\rangle = \mathrm{Tr}(\Psi \Sigma_{w_d}) + \langle \mathbf{a}_d \rangle \left\langle \sum_n^N \mathbf{z}_n \mathbf{z}_n^T \right\rangle \left\langle \mathbf{a}_{d'}^T \right\rangle, \quad \text{for } d = d' \tag{2.92}$$

$$= \langle \mathbf{a}_d \rangle \left\langle \sum_n^N \mathbf{z}_n \mathbf{z}_n^T \right\rangle \left\langle \mathbf{a}_{d'}^T \right\rangle, \quad \text{for } d \neq d'. \tag{2.93}$$

This calculation is inspired by the R-code supplied for Wang [2007] and assumes that the rows of $\mathbf{A}$ are independent.

Using variational inference results in the following approximated distributions.

Posterior for A

The logarithm of the distribution for $\mathbf{A}$ is approximated by

$$\ln q(\mathbf{A}) = \langle \ln p(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \Psi) + \ln p(\mathbf{A}|\mathbf{U}, \lambda) \rangle_{/\mathbf{A}} + C \quad \Leftrightarrow \tag{2.94}$$

$$= -\frac{1}{2} \sum_m^M \sum_n^N \left\langle (\mathbf{x}_n^{(m)} - \mathbf{A}^{(m)} \mathbf{z}_n)^T \Psi^{(m)} (\mathbf{x}_n^{(m)} - \mathbf{A}^{(m)} \mathbf{z}_n) \right\rangle_{/\mathbf{A}} - \sum_m^M \sum_k^K \left\langle \frac{\lambda}{2} ||\mathbf{a}_k^{(m)} - \mathbf{u}_k||^2 \right\rangle_{/\mathbf{A}} + C \quad \Leftrightarrow \tag{2.95}$$

$$= \sum_d^D \left\langle -\sum_n^N \frac{1}{2} \left(\psi_{dd}^{(m)} \mathbf{a}_d^{(m)} \mathbf{z}_n \mathbf{z}_n^T \mathbf{a}_d^{(m)T} + \sum_{d' \neq d}^D \left\{ \psi_{dd'}^{(m)} \mathbf{a}_d^{(m)} \mathbf{z}_n \mathbf{z}_n^T \mathbf{a}_{d'}^{(m)T} \right\} - 2\mathbf{a}_d^{(m)} \mathbf{z}_n \psi_{(d,:)}^{(m)} \mathbf{x}_n^{(m)} \right) - \frac{\lambda}{2} \mathbf{a}_d^{(m)T} \mathbf{a}_d^{(m)} + \lambda \mathbf{a}_d^{(m)T} \mathbf{u}_d \right\rangle_{/\mathbf{A}} + C. \tag{2.96}$$

Note again that $\mathbf{a}_d$ is a the d'th *row* of $\mathbf{A}$, for a more compact notation. A normal distribution for $q(\mathbf{A})$ can then be expressed as

$$q(\mathbf{A}) = \prod_m^M \prod_d^D \mathcal{N}\left(\hat{\mathbf{a}}_d^{(m)} |\ \boldsymbol{\mu}_{\mathbf{a}_d}^{(m)}, \Sigma_{\mathbf{a}_d}^{(m)}\right) \tag{2.97}$$

$$\Sigma_{\mathbf{a}_d}^{(m)^{-1}} = \left\langle \psi_{dd}^{(m)} \right\rangle \sum_n^N \left\langle \mathbf{z}_n \mathbf{z}_n^T \right\rangle + \langle \lambda \rangle \mathbf{I} \tag{2.98}$$

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{a}_d}^{(m)} = \Sigma_{\mathbf{a}_d}^{(m)} \Bigg(& \sum_n^N \langle \mathbf{z}_n \rangle \left\langle \psi_{(d,:)}^{(m)} \right\rangle \mathbf{x}_n^{(m)} + \langle \lambda \rangle \langle \mathbf{u}_d \rangle \\ & - \sum_{d' \neq d}^D \left\langle \psi_{dd'}^{(m)} \right\rangle \sum_n^N \left\langle \mathbf{z}_n \mathbf{z}_n^T \right\rangle \left\langle \mathbf{a}_{d'}^{(m)T} \right\rangle \Bigg), \end{aligned} \tag{2.99}$$

where $\hat{\mathbf{a}}_d^{(1)}$ is a column vector corresponding to the d'th row of $\mathbf{A}$.

Posterior for U

The logarithm of the distribution for $\mathbf{U}$ is approximated by

$$\ln q(\mathbf{U}) = \langle \ln p(\mathbf{A}|\mathbf{U}, \lambda) + \ln\, p(\mathbf{U}|\boldsymbol{\alpha}) \rangle_{/\mathbf{U}} + C \quad \Leftrightarrow \tag{2.100}$$

$$\begin{aligned} &= \sum_k^K \sum_m^M \left\{ \langle \lambda \rangle\, \mathbf{u}_k^T \left\langle \mathbf{a}_k^{(m)} \right\rangle - \frac{\langle \lambda \rangle}{2} \mathbf{u}_k^T \mathbf{u}_k \right\} \\ &\quad - \frac{\langle \alpha_k \rangle}{2} \mathbf{u}_k^T \mathbf{u}_k + C. \end{aligned} \tag{2.101}$$

A normal distribution for $q(\mathbf{U})$ can then be expressed as

$$q(\mathbf{U}) = \prod_{k=1}^K \mathcal{N}\left(\mathbf{u}_k |\ \boldsymbol{\mu}_{\mathbf{u}_k}, \sigma_{\mathbf{u}_k}^2 \mathbf{I}\right) \tag{2.102}$$

$$\sigma_{\mathbf{u}_k}^{-2} = M \langle \lambda \rangle + \langle \alpha_k \rangle \tag{2.103}$$

$$\boldsymbol{\mu}_{\mathbf{u}_k} = \sigma_{\mathbf{u}_k}^2 \langle \lambda \rangle \sum_m^M \left\langle \mathbf{a}_k^{(m)} \right\rangle. \tag{2.104}$$

Posterior for $\boldsymbol{\alpha}$

The logarithm of the distribution for $\boldsymbol{\alpha}$ is approximated by

$$\ln q(\boldsymbol{\alpha}) = \langle \ln\, p(\mathbf{U}|\boldsymbol{\alpha}) \rangle_{/\boldsymbol{\alpha}} + \ln p(\boldsymbol{\alpha}) + C \quad \Leftrightarrow \tag{2.105}$$

$$= \sum_k^K \frac{D}{2} \ln \alpha_k - \frac{\alpha_k}{2} \left\langle ||\mathbf{u}_k||^2 \right\rangle + (a_0 - 1) \ln \alpha_k - b_0 \alpha_k + C. \tag{2.106}$$

A gamma distribution for $q(\boldsymbol{\alpha})$ can then be expressed as

$$q(\boldsymbol{\alpha}) = \prod_{k}^{K} \mathcal{G}a(\alpha_k | a_\alpha, b_{\alpha_k}) \tag{2.107}$$

$$a_\alpha = a_0 + \frac{D}{2} \tag{2.108}$$

$$b_{\alpha_k} = b_0 + \frac{\left\langle \mathbf{u}_k^T \mathbf{u}_k \right\rangle}{2}. \tag{2.109}$$

Posterior for λ

The logarithm of the distribution for λ is approximated by

$$\ln q(\lambda) = \left\langle \ln p(\mathbf{A}|\mathbf{U}, \boldsymbol{\alpha}) \right\rangle_{/\lambda} + \ln p(\lambda) + C \quad \Leftrightarrow \tag{2.110}$$

$$= -\frac{\lambda}{2} \sum_{k}^{K} \sum_{m}^{M} \left\{ \left\langle \mathbf{u}_k^T \mathbf{u}_k \right\rangle + \left\langle \mathbf{a}_k^{(m)T} \mathbf{a}^{(m)} \right\rangle - 2 \left\langle \mathbf{a}_k^{(m)T} \right\rangle \left\langle \mathbf{u}_k \right\rangle \right\} + (a_0 - 1) \ln \lambda - b_0 \lambda + C. \tag{2.111}$$

A gamma distribution for $q(\lambda)$ can then be expressed as

$$q(\lambda) = \mathcal{G}a(\lambda | a_\lambda, b_\lambda) \tag{2.112}$$

$$a_\lambda = a_0 + \frac{MKD}{2} \tag{2.113}$$

$$b_\lambda = b_0 + \sum_{k}^{K} M \frac{\left\langle \mathbf{u}_k^T \mathbf{u}_k \right\rangle}{2} + \sum_{m}^{M} \left\{ \frac{\left\langle \mathbf{a}_k^{(m)T} \mathbf{a}_k^{(m)} \right\rangle}{2} - \left\langle \mathbf{a}_k^{(m)T} \right\rangle \left\langle \mathbf{u}_k \right\rangle \right\}. \tag{2.114}$$

Note that v_Ψ, a_α and a_λ are constants and can be defined before iterating over the other updates.

2.4.2 Bayesian CoCA based on pair-wise similarity

The priors presented in this section is based on a two-view conditional relationship between the $\mathbf{A}^{(m)}$'s described by:

$$\mathbf{A}^{(m)} \sim \prod_{k}^{K} \mathcal{N}\left(\mathbf{a}_k^{(m)} | \mathbf{0}, \alpha_k^{-1} \right) \tag{2.115}$$

$$p\left(\mathbf{A}^{(1)} | \mathbf{A}^{(2)}, \lambda \right) = \prod_{k}^{K} \sqrt{\frac{\lambda}{2\pi}}^{D} \exp\left\{ -\frac{\lambda}{2} || \mathbf{a}_k^{(1)} - \mathbf{a}_k^{(2)} ||^2 \right\}. \tag{2.116}$$

It can be seen that high values of λ forces the two mixing matrices to be similar, resulting in a Bayesian version of CoCA [Dmochowski et al. 2012]. On the other hand a value of λ close to zero will remove the influence of the conditional probability and the mixing matrices are free to be dissimilar. This will result in a bayesian CCA as proposed by [Klami 2013; Wang 2007], though with the constraint that the mixing matrices can only find the same shared sources.

Conditional distribution between A's

The regulation of the similarity between the weights, $\mathbf{A}$, through their variance in (2.116) can be seen as

$$\ln(p(\mathbf{A})) \propto -\frac{1}{2}\sum_{k}^{K}\left((\boldsymbol{\alpha}+\lambda)||\mathbf{a}_k^{(1)}||^2+(\boldsymbol{\alpha}+\lambda)||\mathbf{a}_k^{(2)}||^2-2\lambda\mathbf{a}_k^{(1)T}\mathbf{a}_k^{(2)}\right) \tag{2.117}$$

$$= -\frac{1}{2}\sum_{k}^{K}\begin{bmatrix}\mathbf{a}_k^{(1)T} & \mathbf{a}_k^{(2)T}\end{bmatrix}\begin{bmatrix}(\boldsymbol{\alpha}+\lambda) & -\lambda \\ -\lambda & (\boldsymbol{\alpha}+\lambda)\end{bmatrix}\begin{bmatrix}\mathbf{a}_k^{(1)} \\ \mathbf{a}_k^{(2)}\end{bmatrix}. \tag{2.118}$$

With this matrix notation in mind, an expansion for multiple datasets could then be expressed as

$$\ln(p(\mathbf{A})) \propto -\frac{1}{2}\sum_{k}^{K}\begin{bmatrix}\mathbf{a}_k^{(1)} \\ \vdots \\ \mathbf{a}_k^{(m)}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\alpha}+(M-1)\lambda & \cdots & -\lambda \\ \vdots & \ddots & \vdots \\ -\lambda & \cdots & \boldsymbol{\alpha}+(M-1)\lambda\end{bmatrix}\begin{bmatrix}\mathbf{a}_k^{(1)} \\ \vdots \\ \mathbf{a}_k^{(m)}\end{bmatrix} \tag{2.119}$$

$$= -\frac{1}{2}\sum_{m}^{M}\sum_{k}^{K}\left((\boldsymbol{\alpha}+(M-1)\lambda)||\mathbf{a}_k^{(m)}||^2-2\sum_{m'=m+1}^{M}\lambda\mathbf{a}_k^{(m)T}\mathbf{a}_k^{(m')}\right) \tag{2.120}$$

$$= \sum_{m}^{M}\left\{\ln p\left(\mathbf{A}^{(m)}|\boldsymbol{\alpha}\right)+\sum_{m'=m+1}^{M}\ln p\left(\mathbf{A}^{(m)}|\mathbf{A}^{(m')},\lambda\right)\right\}. \tag{2.121}$$

The $(M-1)$ scaling for λ in eq. (2.119) is necessary to still use the gaussian expression for the conditional distribution between the weights in eq. (2.121). The symmetric nature of $p\left(\mathbf{A}^{(m)}|\mathbf{A}^{(m')},\lambda\right)$ makes it possible to calculate all relationships between the $\mathbf{A}$'s in this manner.

The joint probability is then given by

$$p(\mathbf{V},\mathbf{H}) = p(\mathbf{X}|\mathbf{Z},\mathbf{A},\Psi)p(\mathbf{Z})p(\Psi)p(\boldsymbol{\alpha})p(\lambda) \cdot\prod_{m}^{M}\left\{\ln p\left(\mathbf{A}^{(m)}|\boldsymbol{\alpha}\right)\prod_{m'=m+1}^{M}\ln p\left(\mathbf{A}^{(m)}|\mathbf{A}^{(m')},\lambda\right)\right\}. \tag{2.122}$$

Variational approximation of posterior distributions

As mentioned earlier the two models for BCoCA are equal in most areas, which also expresses itself in the similarity in the variational updates for the models. In fact, the updates for $q(\mathbf{Z})$ and $q(\Psi)$ are identical for both models, which in message passing terminology can be explained with the children and parents for both variables being the same. This is not the case for the rest of the variables, but since the principles for the derivations of their updates are similar between the models they are omitted in this section. Below are the updates for $\mathbf{A}$, λ, and $\boldsymbol{\alpha}$.

$$q(\mathbf{A}^{(m)}) = \prod_{d=1}^{D} \mathcal{N}\left(\hat{\mathbf{a}}_d^{(m)} | \boldsymbol{\mu}_{\mathbf{a}_d}^{(m)}, \Sigma_{\mathbf{a}_d}^{(m)}\right) \tag{2.123}$$

$$\Sigma_{\mathbf{a}_d}^{(m)^{-1}} = \left\langle \psi_{dd}^{(m)} \right\rangle \sum_{n}^{N} \left\langle \mathbf{z}_n \mathbf{z}_n^T \right\rangle + (M-1) \left\langle \lambda \right\rangle \mathbf{I} + \mathrm{diag}(\left\langle \boldsymbol{\alpha} \right\rangle) \tag{2.124}$$

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{a}_d}^{(m)} = \Sigma_{\mathbf{a}_d}^{(m)} \Bigg(& \sum_{n}^{N} \left\langle \mathbf{z}_n \right\rangle \left\langle \psi_{(d,:)}^{(m)} \right\rangle \mathbf{x}_n^{(m)} + \sum_{m' \neq m}^{M} \left\langle \lambda \right\rangle \left\langle \mathbf{a}_d^{(m')T} \right\rangle \\ & - \frac{1}{2} \sum_{d' \neq d}^{D} \left\{ \left\langle \psi_{dd'}^{(m)} \right\rangle \sum_{n}^{N} \left\langle \mathbf{z}_n \mathbf{z}_n^T \right\rangle \left\langle \mathbf{a}_{d'}^{(m)T} \right\rangle \right\} \Bigg) \end{aligned} \tag{2.125}$$

$$q(\boldsymbol{\alpha}) = \prod_{k}^{K} \mathcal{G}a(\alpha_k | a_\alpha, b_{\alpha_k}) \tag{2.126}$$

$$a_\alpha = a_0 + \frac{DM}{2} \tag{2.127}$$

$$b_{\alpha_k} = b_0 + \sum_{m}^{M} \frac{\left\langle \mathbf{a}_k^{(m)T} \mathbf{a}_k^{(m)} \right\rangle}{2} \tag{2.128}$$

$$q(\lambda) = \mathcal{G}a(\lambda | a_\lambda, b_\lambda) \tag{2.129}$$

$$a_\lambda = a_0 + \frac{(M-1)MKD}{4} \tag{2.130}$$

$$b_\lambda = b_0 + \frac{1}{2} \sum_{m}^{M} \sum_{k}^{K} (M-1) \left\langle \mathbf{a}_k^{(m)T} \mathbf{a}_k^{(m)} \right\rangle - 2 \sum_{m'=m+1}^{M} \left\langle \mathbf{a}_k^{(m)T} \mathbf{a}_k^{(m')} \right\rangle, \tag{2.131}$$

where $\hat{\mathbf{a}}_d^{(1)}$ is a column vector corresponding to the d'th row of $\mathbf{A}$.

2.5 Lower bound for BCoCA

In section 2.1.1 the lower bound, $\mathcal{L}(q)$ was introduced as an expression to maximise instead of minimising the KL divergence, as the sum of these two equals the logarithm to the true likelihood function. It was also explained that the $\mathcal{L}(q)$ is a good measure for estimating time of convergence, which is why this section will concern the calculation of this measure.

$\mathcal{L}(q)$ is often calculated to estimate the time of convergence, by setting a threshold for the relative change wrt. previous iteration. It is usually derived as the sum of the expectations of each variable in $q(\mathbf{H})$ and $p(\mathbf{H}, \mathbf{V})$ wrt. $q(\mathbf{H})$ calculated independently. Inspired by Murphy [2012] we have chosen to combine the expectations into one equation and let terms containing the same variables cancel each other out, where applicable. Since it is the change of the lower bound that is of interest, we also combined all constant terms into the common constant, C. As can be seen later in this section, the result is a much simpler expression for $\mathcal{L}(q)$, compared to the one presented in Wu et al. [2011] and in the R-code supplied for Wang [2007].

Expanding the expression for the lower bound as defined in (2.7) gives

$$\begin{aligned}\mathcal{L}(q) = &\langle \ln p(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \Psi)\rangle_{q(\mathbf{H})} + \langle \ln p(\mathbf{Z})\rangle_{q(\mathbf{H})} + \langle \ln p(\Psi)\rangle_{q(\mathbf{H})} + \langle \ln p(\mathbf{A}|\mathbf{U}, \lambda)\rangle_{q(\mathbf{H})} \\ &+ \langle \ln p(\mathbf{U}|\boldsymbol{\alpha})\rangle_{q(\mathbf{H})} + \langle \ln p(\boldsymbol{\alpha})\rangle_{q(\mathbf{H})} + \langle \ln p(\lambda)\rangle_{q(\mathbf{H})} + \mathrm{H}[q(\mathbf{Z})] + \mathrm{H}[q(\Psi)] \\ &+ \mathrm{H}[q(\mathbf{A})] + \mathrm{H}[q(\mathbf{U})] + \mathrm{H}[q(\lambda)] + \mathrm{H}[q(\boldsymbol{\alpha})], \end{aligned} \tag{2.132}$$

where $\mathrm{H}[q(\mathbf{H})]$ signifies the entropy of $q(\mathbf{H})$, and we have used the fact that $\langle -\ln q(\mathbf{H})\rangle_{q(\mathbf{H})} = \mathrm{H}[q(\mathbf{H})]$ [Murphy 2012]. To calculate each term the following relations for Gaussian, gamma, and Wishart distributions are used;

$\mathcal{N}(\mathbf{x}|\mu, \Sigma)$:

$$\langle \mathbf{x}\rangle = \mu \tag{2.133}$$

$$\langle \mathbf{x}^T\mathbf{x}\rangle = \mathrm{Tr}(\Sigma) + \mu^T\mu \tag{2.134}$$

$$\mathrm{H}[\mathbf{x}] = \frac{1}{2}\ln|\Sigma| + \frac{D}{2}(1 + \ln 2\pi) \tag{2.135}$$

$\mathcal{G}a(\lambda|a, b)$:

$$\langle \lambda\rangle = \frac{a}{b} \tag{2.136}$$

$$\langle \ln\lambda\rangle = \psi(a) - \ln b \tag{2.137}$$

$$\mathrm{H}[\lambda] = \ln\Gamma(a) - (a-1)\psi(a) - \ln b + a \tag{2.138}$$

$$\mathcal{W}(\Psi|\mathbf{S}, v):$$

$$\langle \Psi \rangle = v\mathbf{S} \tag{2.139}$$

$$\langle \ln|\Psi| \rangle = \sum_{i=1}^{D} \psi\left(\frac{v+1-i}{2}\right) + D\ln 2 + \ln|\mathbf{S}| \tag{2.140}$$

$$\mathrm{H}[\Psi] = -\ln B(\mathbf{S}, v) - \frac{v-D-1}{2}\langle \ln|\Psi| \rangle - \frac{vD}{2} \tag{2.141}$$

$$B(\mathbf{S}, v) = |\mathbf{S}|^{-\frac{v}{2}} \left(2^{\frac{vD}{2}} \pi^{\frac{D(D-1)}{4}} \prod_{i=1}^{D} \Gamma\left(\frac{v+1-i}{2}\right)\right)^{-1}, \tag{2.142}$$

where $\psi(a)$ and $\Gamma(a)$ are the digamma and gamma function, respectively [C. M. Bishop 2006].

Each term in (2.132) are derived using the updates shown in 2.4.1 and constant terms are absorbed into the constant C. The derivations in 2.4.1 are also used to exchange some expressions with some of the variables already calculated in the updates.

$$\langle \ln p(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \Psi) \rangle_{q(\mathbf{H})} = \sum_{m}^{M}\sum_{n}^{N} \frac{1}{2}\left\langle \ln\left|\Psi^{(m)}\right| \right\rangle - \frac{1}{2}\mathrm{Tr}\left\{\left(\mathbf{S}_{\Psi}^{(m)-1} - \mathbf{S}_{0}^{-1}\right)\left\langle \Psi^{(m)} \right\rangle\right\} \quad \Leftrightarrow \tag{2.143}$$

$$= \frac{1}{2}\sum_{m}^{M}\left\{N \ln\left|\mathbf{S}_{\Psi}^{(m)}\right| + v_{\Psi}\mathrm{Tr}\left(\mathbf{S}_{0}^{-1}\mathbf{S}_{\Psi}^{(m)}\right)\right\} + C \tag{2.144}$$

$$\langle \ln p(\mathbf{Z}) \rangle_{q(\mathbf{H})} = -\sum_{n}^{N}\left\{\frac{K}{2}\ln 2\pi + \frac{1}{2}\left\langle \mathbf{z}_n^T \mathbf{z}_n \right\rangle\right\} \quad \Leftrightarrow \tag{2.145}$$

$$= -\frac{1}{2}\left(N \cdot \mathrm{Tr}(\Sigma_{\mathbf{z}}) + \sum_{n}^{N} \mu_{\mathbf{z},n}^T \mu_{\mathbf{z},n}\right) + C \tag{2.146}$$

$$\langle \ln p(\Psi) \rangle_{q(\mathbf{H})} = \sum_{m}^{M} \ln B(\mathbf{S}_0, v_0) + \frac{v_0 - D - 1}{2}\left\langle \ln\left|\Psi^{(m)}\right| \right\rangle - \frac{1}{2}\mathrm{Tr}\left(\mathbf{S}_0^{-1}\left\langle \Psi^{(m)} \right\rangle\right) \quad \Leftrightarrow \tag{2.147}$$

$$= \frac{1}{2}\sum_{m}^{M}(v_0 - D - 1)\ln\left|\mathbf{S}_{\Psi}^{(m)}\right| - v_{\Psi}\mathrm{Tr}\left(\mathbf{S}_0^{-1}\mathbf{S}_{\Psi}^{(m)}\right) + C \tag{2.148}$$

$$\langle \ln p(\mathbf{A}|\mathbf{U},\lambda)\rangle_{q(\mathbf{H})} = \sum_m^M \sum_k^K -\frac{\langle\lambda\rangle}{2}\left(\langle \mathbf{u}_k^T\mathbf{u}_k\rangle + \left\langle \mathbf{a}_k^{(m)T}\mathbf{a}^{(m)}\right\rangle - 2\left\langle \mathbf{a}_k^{(m)T}\right\rangle\langle \mathbf{u}_k\rangle\right) + \frac{D}{2}\langle\ln\lambda\rangle + C \quad \Leftrightarrow \tag{2.149}$$

$$= -\frac{a_\lambda}{b_\lambda}(b_\lambda - b_0) + \frac{DMK}{2}(\psi(a_\lambda) - \ln b_\lambda) + C \quad \Leftrightarrow \tag{2.150}$$

$$= b_0\frac{a_\lambda}{b_\lambda} - \frac{DMK}{2}\ln b_\lambda + C \tag{2.151}$$

$$\langle \ln p(\mathbf{U}|\boldsymbol{\alpha})\rangle_{q(\mathbf{H})} = \sum_k^K \frac{D}{2}\langle\ln\alpha_k\rangle - \frac{\langle\alpha_k\rangle}{2}\langle \mathbf{u}_k^T\mathbf{u}_k\rangle \quad \Leftrightarrow \tag{2.152}$$

$$= \sum_k^K -\frac{D}{2}\ln b_{\alpha_k} + b_0\frac{a_\alpha}{b_{\alpha_k}} \tag{2.153}$$

$$\langle \ln p(\boldsymbol{\alpha})\rangle_{q(\mathbf{H})} = \sum_k^K (a_0-1)\langle\ln\alpha_k\rangle - b_0\langle\alpha_k\rangle + C \quad \Leftrightarrow \tag{2.154}$$

$$= \sum_k^K -(a_0-1)\ln b_{\alpha_k} - b_0\frac{a_\alpha}{b_{\alpha_k}} + C \tag{2.155}$$

$$\langle \ln p(\lambda)\rangle_{q(\mathbf{H})} = (a_0-1)\langle\ln\lambda\rangle - b_0\langle\lambda\rangle + C \quad \Leftrightarrow \tag{2.156}$$

$$= -(a_0-1)\ln b_\lambda - b_0\frac{a_\lambda}{b_\lambda} + C \tag{2.157}$$

$$\mathrm{H}[q(\mathbf{Z})] = \frac{1}{2}\ln|\Sigma_z| + C \tag{2.158}$$

$$\mathrm{H}[q(\Psi)] = \sum_m^M -\ln B(\mathbf{S}_\Psi^{(m)}, v_\Psi) - \frac{v_\Psi - D - 1}{2}\left\langle \ln\left|\Psi^{(m)}\right|\right\rangle + \frac{v_\Psi D}{2} \quad \Leftrightarrow \tag{2.159}$$

$$= \sum_m^M -\frac{v_\Psi}{2}\ln\left|\mathbf{S}_\Psi^{(m)}\right| - \frac{v_\Psi - D - 1}{2}\ln\left|\mathbf{S}_\Psi^{(m)}\right| + C \quad \Leftrightarrow \tag{2.160}$$

$$= \sum_m^M -\frac{D+1}{2}\ln\left|\mathbf{S}_\Psi^{(m)}\right| + C \tag{2.161}$$

$$\mathrm{H}[q(\mathbf{A})] = \sum_m^M \sum_d^D \frac{1}{2}\ln\left|\Sigma_{\mathbf{a}_d}^{(m)}\right| + C \tag{2.162}$$

$$\mathrm{H}[q(\mathbf{U})] = \sum_k^K \frac{1}{2}\ln\left|\sigma_{\mathbf{u}_k}^2\mathbf{I}\right| + C \quad \Leftrightarrow \tag{2.163}$$

$$= \sum_k^K \frac{D}{2}\ln\sigma_{\mathbf{u}_k}^2 + C \tag{2.164}$$

$$\mathrm{H}[q(\lambda)] = -\ln b_\lambda + C \tag{2.165}$$

$$\mathrm{H}[q(\boldsymbol{\alpha})] = -\sum_{k}^{K} \ln b_{\alpha_k} + C \tag{2.166}$$

Note that v_Ψ, a_α and a_λ are constant over the iterations and therefore absorbed into the constant C. Setting all these equations into (2.132) and letting terms with the same variables cancel each other out results in a much simpler expression for the lower bound, as compared to calculating each equation and summing over these afterwards;

$$\begin{aligned}\mathcal{L}(q) = \frac{1}{2}\sum_{m}^{M}\left\{v_\Psi \ln\left|S_\Psi^{(m)}\right| + \sum_{d}^{D}\ln\left|\Sigma_{\mathbf{a}_d}^{(m)}\right|\right\} + \sum_{k}^{K}\left\{-a_\alpha \ln b_{\alpha_k} + \frac{D}{2}\ln\sigma_{\mathbf{u}_k}^2\right\} \\ - a_\lambda \ln b_\lambda + \frac{1}{2}\ln|\Sigma_\mathbf{z}| - \frac{1}{2}\left(N\cdot\mathrm{Tr}(\Sigma_\mathbf{z}) + \sum_{n}^{N}\mu_{\mathbf{z},n}^T\mu_{\mathbf{z},n}\right) + C.\end{aligned} \tag{2.167}$$

This expression only calculates how the variables, that change between iterations, influence the lower bound. Therefore it cannot be used to directly compare with other models based on other priors. It can however be used for estimating a time of convergence and as measure to decide on the best result among multiple runs on the same data, which is also what it will be used for in this thesis.

2.6 Independent component analysis

Eye movement and eye blinks cause major artefacts in EEG recordings and it is often necessary to remove these in order to further process the data. Since the idea of independent component analysis (ICA; Molgedey et al. [1994] and Bell et al. [1995]) it has been used to separate artefactual components from EEG data [Makeig et al. 1996; Jung et al. 2000] and the research into this field continues rigorously in order to remove as much noise as possible, preferably automatically, while retaining the signal [Mammone et al. 2012].

ICA assumes that a set of recorded signals $\mathbf{x} = [x_1 \dots x_N]^T$ are a linear mixture of the sources $\mathbf{z} = [z_1 \dots z_N]^T$ by the square mixing $\mathbf{x} = \mathbf{A}\mathbf{z}$. ICA tries to estimate the spatial filter that inverts the mixing process and thus recovers the sources $\hat{\mathbf{z}} = \mathbf{W}\mathbf{x}$ with the constraint that the sources are statistically independent.

2.7 Correlation permutation test

CoCA was developed by Dmochowski et al. [2012] to find shared signals in two EEGs from subjects experiencing the same stimulus. In their article they used the averaged correlation coefficient [Pearson 1896] calculated pair-wise between a group of subjects as a measure to find time intervals of high correlation in the group and a permutation test to establish a level of significance. As we will employ the same method on the

EEG recordings presented in this thesis, the concept of a permutation test will be explained here.

A permutation, or randomisation, test between two groups of data has the advantage that it does not attempt to make assumptions for the true distributions of the datasets, and therefore refrains from estimating parameters for them. Instead it uses repeated shuffling of the members of each group to estimate whether or not there is a difference between the original two groups and these random groupings of data.

After estimating a statistical measure between two groups of data, the permutation test shuffles the data, randomly assigns each observation to one of the two groups, and estimates the statistical measure again. This process is repeated until the statistical measure is calculated on all or a fixed amount of the permutation possibilities. One can define a null hypothesis that the original measure is no different from the statistical measures stemming from the permuted data. This null hypothesis can then be disproved if the original measure lies outside the distribution of permuted data by some critical value. If the null hypothesis can be disproved, the original statistical measure between the two groups of data is deemed significant [Manly 2007].

As the null hypothesis cannot be proved, but only disproved, the validity of the permutation test is increased when it is able to disprove the null hypothesis more readily. This ability is improved when the amount of permutations are increased, at the cost of additional calculations [Fisher et al. 1949].

In this thesis the permutation test will be employed on the correlation between two time series stemming from EEG filtered by the weights from either CoCA or BCoCA. The correlation between the time series is calculated on windows of the data with a fixed length and overlap, which defines the temporal resolution of the test. As in Dmochowski et al. [2012] the order of one of the time series will be unchanged, and the permutations will only be conducted on the other time series. However, Dmochowski et al. [2012] does not state how they calculate p-values for the correlation coefficient averaged over all pairs of subjects in a given group.

In this thesis it was therefore chosen to employ a permutation test for the averaged correlation coefficient from the pair-wise correlations between a group of subjects. For each permutation of a specific window of the time series, the ordering of samples in each of the permuted correlations were saved and used for the corresponding window in the other time series. This way a permutations test could be conducted for the average of the correlations using the average of each correlation for a specific permutation of the time series. However, taking the average over the permuted correlations lowered their correlation coefficients and produced low p-values for all windows. It was therefore chosen to calculate the p-values for each pair-wise correlation and use these in two ways to test for differences. The first way was to use the average critical correlation value for a p-value of 0.01, to decide when a window of the average correlation coefficient was significant. The other method of testing was to test for the total number of significant windows in all pair-wise correlations, using their p-values to

correct for multiple comparisons by controlling the false discovery rate, as explained in the following section.

The correlation permutation test can be explained in the following steps:

1. Calculate the correlation coefficient for the first window between the two time series.

2. Randomly reorder the second time series N_P times for this window, gaining N_P new time series.

3. Calculate the correlation coefficient for each permuted time series.

4. Sort the correlation coefficients for the window after size and calculate the p-value as the number of correlation coefficients higher or equal the original (including the original), divided by the total number of correlation coefficients (including the original).

5. Move the window and repeat step 1 to 4 for the entire length of the pair of time series.

The implementation in Matlab was utilised using a modified version of the script supplied by Groppe et al. [2011]. The number of permutations, N_P, was set equal to 5000 to ensure a test with a alpha level of significance equal to 0.01 [Manly 2007].

2.8 Controlling the false discovery rate

When conducting multiple tests there will by definition be a number of false positives based on the significance level, α, for the test. The correlation permutation test is conducted for each window of the time series, so a test with 300 windows and a significance level of 0.01 should produce 3 false positives on average. To control for this effect a number of correction schemes exists for multiple comparisons, some more conservative than others. In this thesis the control for false discovery rate (FDR) will be employed.

Having N tests with N null hypotheses, H_n, and N p-values, p_n, FDR orders the p-values after size so that $p_1 \leq ... \leq p_i \leq ...$ for $i = \{1, ..., N\}$. It then finds the highest number of i for which

$$p_i \leq \frac{i}{N}\alpha \tag{2.168}$$

is true. The null hypotheses with p-values below p_i can then be rejected (H_i included).

2.9 Testing BCoCA on simulated data

In this section the performance of BCoCA will be evaluated on simulated data. First the two versions of BCoCA will be tested and compared, where the one with the highest test scores will be used for the remainder of this thesis. This is followed by a test of implementation, where the same BCoCA model will be implemented with VMP using Microsoft's Infer.NET and with the updates derived in section 2.4 implemented in Matlab. Finally follows a comparison with GFA, CoCA and CCA, tested under varying conditions.

2.9.1 Simulation design

To measure the performance between the different algorithms, data is generated from the BCoCA model with a varying λ parameter. This approach generates data, with equal true weights for all datasets, when $\lambda \gg 1$ and i.i.d. true weights when $\lambda \ll 1$. In the two-view situation $\lambda \gg 1$ should be ideal for CoCA, and CCA better suited for $\lambda \ll 1$. The data is simulated with the following model

$$\mathbf{X}^{(m)} = \mathbf{A}^{(m)}_{\text{true}}\mathbf{Z} + \boldsymbol{\epsilon} \tag{2.169}$$

where $\mathbf{Z}$ is a $K \times N$ source matrix containing K time series. The added noise, $\boldsymbol{\epsilon}$, is i.i.d. gaussian with zero mean and a variance, σ^2_ϵ, that is varied to obtain the desired signal-to-noise ratio (SNR). The SNR is calculated as

$$\text{SNR} = 10\log_{10}\left(\frac{\mathbb{E}[\mathbf{s}^2]}{\mathbb{E}[\mathbf{n}^2]}\right) \tag{2.170}$$

Since the noise has zero mean its power expression can be exchanged by its variance, which can then be isolated;

$$\sigma^2_n = \mathbb{E}[\mathbf{s}^2] \cdot 10^{-\text{SNR}/10}. \tag{2.171}$$

$\mathbf{A}$ is formulated as

$$\mathbf{A}^{(m)}_{\text{true}} = \mathbf{U} + \boldsymbol{\delta}^{(m)} \tag{2.172}$$

with $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{\alpha}^{-1})$ and $\boldsymbol{\delta}^{(m)} \sim \mathcal{N}(0, \lambda^{-1})$. The variance across views are hence only modelled by λ.

Choosing hidden sources

We have used up to four hidden sources, generated in the same manner as in Klami [2013], for comparability with their results. The tests will be conducted with the simple case of one hidden source corresponding to $K = 1$ in (2.169), meaning that the data is generated from one sinusoid and additive noise, and a more complex case with all four components.

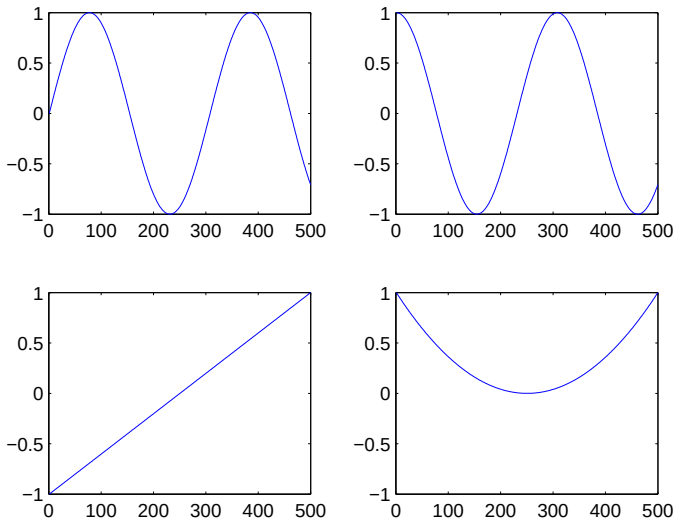


Figure 2.6: The four signals used as true sources in simulated data used for testing BCoCA, GFA, CoCA and CCA.

Measure of performance

The correlation coefficient between the inferred sources and the true source was chosen as the measure of performance. Since the latent models infer a common $\mathbf{Z}$ for all datasets, the mean of the view specific $\mathbf{y}_1$ and $\mathbf{y}_2$ was used as the inferred sources for CCA and CoCA. This improved their performance by approximately 10 - 20% compared to only using $\mathbf{y}_1$. For each condition 20 datasets were randomly generated from the distributions described in section 2.9.1 and each algorithm was tested on the same data. The mean and standard error of the mean for the 20 datasets were calculated and used to compare the performance between the algorithms. In the tests with four hidden sources all correlation combinations between the inferred sources and the true ones were calculated, where each inferred source was only allowed to correlate with one true source and vice versa. The combination with the highest mean correlation was then chosen.

CoCA and CCA on multiple datasets

CoCA and CCA can only compare two datasets at a time. In case of multiple dataset comparison this thesis will follow the same method as in Dmochowski et al. [2012],

where the datasets are concatenated sample-wise into

$$\bar{\mathbf{X}}^{(1)} = [\mathbf{X}^{(1)}, \mathbf{X}^{(1)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}],$$
$$\bar{\mathbf{X}}^{(2)} = [\mathbf{X}^{(2)}, \mathbf{X}^{(3)}, \mathbf{X}^{(4)}, \mathbf{X}^{(3)}, \mathbf{X}^{(4)}, \mathbf{X}^{(4)}] \tag{2.173}$$

so that all combinations of datasets will be compared. As both CoCA and CCA use eigenvalue decomposition on the sample covariance matrices, using $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ corresponds to using the average pair-wise sample covariance matrices. This way the eigenvalue decomposition has to be calculated only once. However using this method the number of samples in $\bar{\mathbf{X}}_1$ scales by $M(M-1)/2$, with (2.173) showing the case of concatenating with four datasets.

Testing conditions

The algorithms were tested at varying levels of SNR, number of datasets, M, and similarity between the true weights of each dataset. In each test the dataset had six dimensions and the number of observations was set to 500, except when varying the number of datasets. This test was conducted on a total of 5.000 samples spread out equally among the datasets, so that each contained 2.500 samples for $M = 2$ and 500 samples for $M = 10$. All the conditions were tested with one and four hidden sources.

2.9.2 Results

Testing the implementation of BCoCA

Figure 2.7 shows the performance of the two approaches to BCoCA presented in 2.4 under varying number of datasets. The datasets were set to be dissimilar with $\lambda = 0.001$ and two levels of $SNR = \{-6, 0\}$ were used. It can be seen that the approach with pair-wise similarity between the weights has difficulties with higher number of dissimilar datasets, in the same manner as will be seen for CCA and CoCA later in this section. This might be a result of the view-specific $\mathbf{A}$s being calculated in a pair-wise manner, as seen in (2.121), similar to the method used for CCA and CoCA explained in (2.173). All of the tests with varying conditions for comparison with CCA, CoCA, and GFA, were tested on both approaches to BCoCA. The approach with a shared mean for the weights achieved the most consistent high performance and it was therefore decided to use this as the model for BCoCA for the remainder of this thesis.

Figure 2.8 shows the performance of BCoCA implemented in Matlab using the derivations presented in 2.4, and a version implemented with VMP at varying levels of SNR, and with the same initialisation. The two mean values are seen to be nearly equal across all values of SNR. The two implementations were therefore deemed to have equal performance, with deviations that could be explained by different order of updates as well as the computational differences by different software. The implementation in Matlab based on a common mean for the weights is therefore used for the remainder of this thesis.

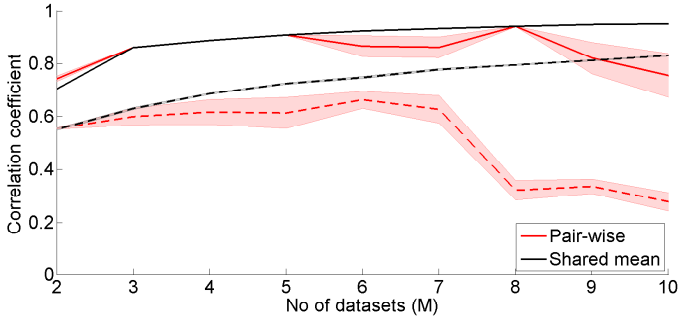


Figure 2.7: Comparison between the two approaches to BCoCA presented in section 2.4, in the case of one hidden source with a varying number of datasets and SNRs equal to 0 and -6. λ was set equal to 10^{-3} making the true weights of the datasets i.i.d. The shown correlation coefficient is calculated as the mean of 20 simulations at each condition, with the standard error of the mean illustrated as the opaque area.

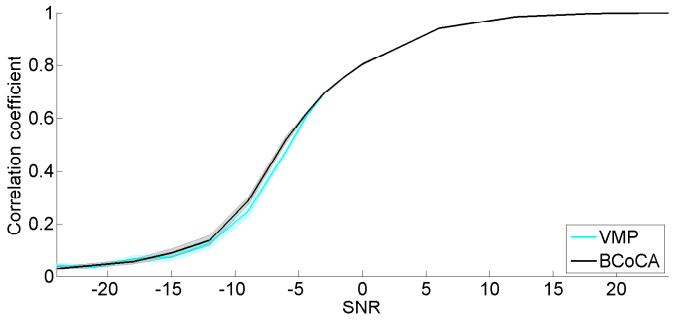
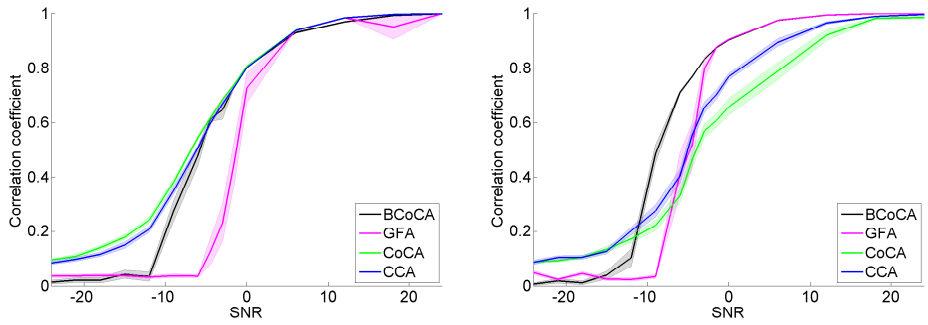


Figure 2.8: Comparison between BCoCA implemented in Matlab using the derivations with a shared mean for the weights, and a version implemented with VMP using Infer.NET at varying levels of SNR and one hidden source. λ was set equal to 10^3 making the true weights of both datasets nearly equal. The shown correlation coefficient is calculated as the mean of 20 simulations at each condition, with the standard error of the mean illustrated as the opaque area.



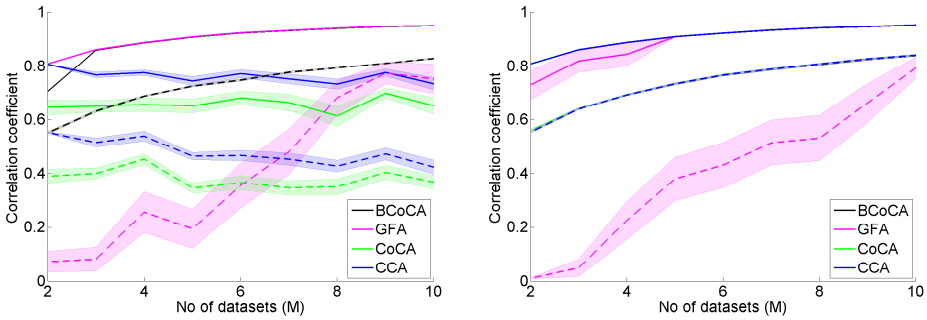
(a) $M = 2$, $\lambda = 1000$ **(b)** $M = 5$, $\lambda = 0.001$

Figure 2.9: Performance of BCoCA, GFA, CoCA and CCA on simulated data measured by mean correlation coefficient and standard error of the mean with respect to the true source calculated over 20 repetitions. The performance is tested under different levels of SNR and **(a)** 2 datasets and similar true weights ($\lambda = 1000$) as well as **(b)** and dissimilar true weights ($\lambda = 0.001$) and 5 datasets.

Performance on varying conditions

Selected results from the tests on the simulated data can be seen in figures 2.9 to 2.11. Additional figures with other combinations of the conditions can be seen in appendix B. Figures 2.9(a) and 2.9(b) show the performance on increasing values of SNR for one hidden sources and $\lambda = 10^3$ and $\lambda = 10^{-3}$, signifying similar and non-similar weights, respectively. In the two view situation it can be seen that for high levels of SNR the algorithms perform equally well, but as the noise levels increase the latent models quickly drop towards zero correlation, though BCoCA do so less steeply and can perform at lower levels of SNR compared to GFA. This quick drop is due to the models choosing the zero-source solution as the cost of a poor estimation gets too high. BCoCA comes closer to zero as this algorithm seemingly choose a source of constant zeros, as opposed to what appears to be low amplitude noise. Better initialisation and basing the prior hyper parameters on the observed data might improve the performance on data with high levels of noise for the latent models. It can also be seen that increasing the number of datasets to five increases the performance of the latent models at the low levels of SNR, and that the opposite is true for CCA and CoCA, in the case of dissimilar true weights.

The impact of increasing the number of datasets are further explored in figures 2.10(a) and 2.10(b). Here two things are evident for the latent models; That BCoCA again outperforms GFA at low levels of SNR and that increasing the number of datasets increases the correlation even though the number of observations do not increase. Some of this effect could stem from averaging out the random noise, when calculating the inferred source as the mean of the sources of estimated on each datasets. The figures also show that CCA and CoCA only benefits for the increased number of



(a) $\lambda = 0.001$ **(b)** $\lambda = 1000$

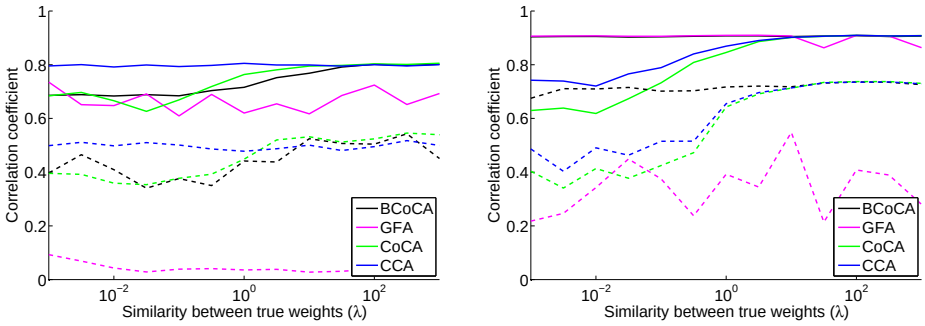
Figure 2.10: Performance of BCoCA, GFA, CoCA and CCA on simulated data measured by mean correlation coefficient and standard error of the mean with respect to the true source calculated over 20 repetitions. The performance is tested with a varying number of datasets and two levels of SNR (solid: SNR = 0, dashed: SNR = -6) and **(a)** dissimilar true weights ($\lambda = 0.001$) as well as **(b)** similar true weights ($\lambda = 1000$).

datasets, when their true weights are equal (or just similar as seen on figure B.9). With completely dissimilar true weights, it actually seems to have a negative effect, when the observations are spread out over more datasets. When considering that CCA and CoCA deal with increasing datasets by concatenating them into two datasets, the importance of having equal weights makes sense as this corresponds to actually just having two datasets. The increased performance must then stem from having more instances of the signal and then be able to average the noise out.

All tests were run at different levels of similarity between the true weights of each dataset by varying λ. Figure 2.11(a) shows that in case of two datasets and one hidden source. As expected the effect can only be seen on CoCA and BCoCA, but CoCA handles the datasets with different true weight better than initially anticipated. An explanation to this is discussed in 2.3.2. Figure 2.11(b) illustrates that for multiple datasets it is CCA and CoCA and that benefits from similar true weights, where the two latent models are indifferent to the change. This, however, is only true when hidden sources only consists of a sinusoid. Figures B.1 and B.2 show that with four hidden sources only CoCA have a slight advantage from increasing λ.

2.10 Model validation on real EEG data

To validate the model beyond testing on artificial data, the performance of BCoCA will be evaluated on EEG from two separate experiments. Both datasets are recorded using event-related paradigms, but for processing with BCoCA, the data is treated as being continuous.



(a) $M = 2$ **(b)** $M = 5$

Figure 2.11: Performance of BCoCA, GFA, CoCA and CCA on simulated data measured by mean correlation coefficient with respect to the true source calculated over 20 repetitions. The similarity between the true weights are varied by the λ parameter and shows the correlation with two levels of SNR (solid: SNR = 0, dashed: SNR = -6).

2.10.1 Face-evoked response

A widely accepted theory of face recognition is the multi-component model of face-processing [Bruce et al. 1986] in which the brain derives details about a person from physical aspects. These are used to create a structural model that is passed on to other processes that are responsible for recognition, identification, expression analysis, etc. Henson et al. [2003] conducted an experiment in which subjects where exposed to a series of images of faces or scrambled faces. The hypothesis from earlier [Bentin et al. 1996] was that a negative peak around 170 ms (N170) post stimulus in the posterior region is greater when the subject is shown a face compared to a scrambled face.

Paradigm and pre-processing

Based on Phase 1 in the study by Henson et al. [2003] a subject was over two trials presented with 86 images of faces and 86 images of scrambled faces. The data was bandpass filtered (2-100 Hz), down-sampled to 200 Hz and epoched using the software package *statistical parametric mapping* (SPM12). The dataset is available online from the SPM website[2] [Henson n.d.].

ERP Analysis

The epoched data was for each trial concatenated and tested using BCoCA, CoCA and CCA. For the latter two the filters corresponding to the maximally correlated components where used. Epochs from both conditions were processed at the same

[2]http://www.fil.ion.ucl.ac.uk/spm/software/spm12/

time, yielding a single component that was then divided into epochs corresponding to the raw data. In figure 2.12a the averaged epochs for the two conditions are illustrated for the raw EEG and from this it is not possible to distinguish the events of interest from noise. In 2.12b the averaged epoch of the first components for each condition shows that all the algorithms have extracted a coherent signal and that the results are very similar. In figure 2.12c the difference signal of the averaged epochs shows as expected that the negative peak at N170 is greater for the face condition. All algorithms locate the time of this occurrence as around 190 ms which corresponds well with the literature [Henson et al. 2003]. To localize the neurons that are responsible for the face processing, the average of the epochs for the face condition at 170 ms was subtracted from the average scrambled condition. Projecting the channels onto a 2D scalp map as in figure 2.13(d) illustrates clearly how the posterior regions in the occipital lobe contribute more negatively in the face condition. Projecting the weights from BCoCA the illustration in figure 2.13(a) depicts the correlated neural activity. The result shows that the signals in the posterior region are highly correlated. The BCoCA algorithm has thus effectively extracted the component from the datasets that exactly depicts the neural activity of interest. Using the forward model [Parra et al. 2005] to plot the projection of CoCA in figure 2.13(b), it shows a very similar result to BCoCA as expected. The CCA weights in figure 2.13(c) is, however, much more localised to specific channels. This may be more accurate in determining precise neural activation of dominant areas, but by comparison to the projection of the actual activity, BCoCA and CoCA clearly derives more anatomically correct results.

2.10.2 Synonym/non-synonym EEG

Using a cohort of 5 subjects, the data was recorded by speaking two words to a person who then responded whether they were synonyms. The dataset was then separated into synonyms and non-synonyms by independent component analysis.

Pre-processing

The data was bandpass filtered to 0.5 Hz – 100 Hz, re-sampled to 200 Hz and divided into epochs with the latency of the second word as zero. To reduce noise from eye movement the independent component with most activity in the eye region was used as a template to find similar components, using the function CORRMAP in EEGLAB [Delorme et al. 2004], which were extracted from all datasets. It has previously been shown that alpha band de-synchronisation is linked with tasks that require the subjects attention [Klimesch et al. 1998] so the band power of the alpha band (7–15Hz) was used as test data. The variation in response time is quite significant and is probably due to difference in level of familiarity with the words. To remove outliers the epochs where ordered with respect to latency of response time and only epochs 21–160 where used.

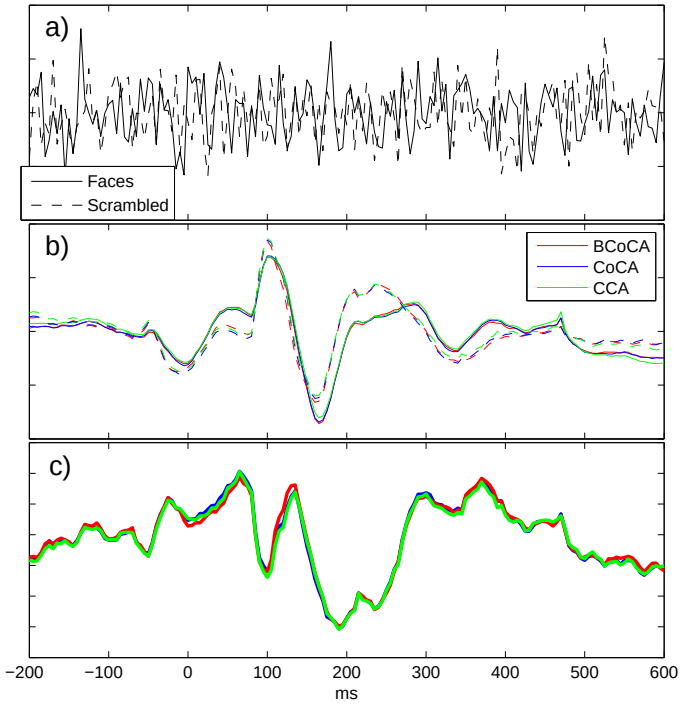
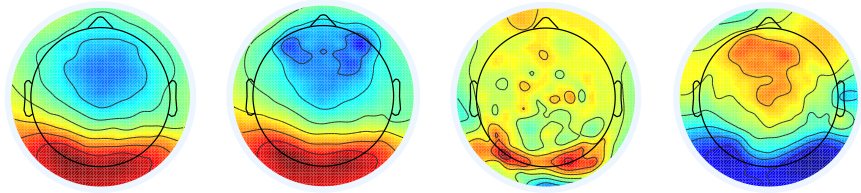


Figure 2.12: a) Averaged epochs across all channels of the raw EEG for faces and scrambled. **b)** Averaged epochs in component space found by BCoCA, CoCA and CCA for face and scrambled condition **c)** Difference between average epoch for face and scrambled condition in component space



(a) BCoCA weights **(b)** CoCA weights **(c)** CCA **(d)** Difference

Figure 2.13: (**a**) Scalp projections of weights from the BCoCA algorithm. (**b**) Scalp projections of weights from the CoCA algorithm using the forward model [Parra et al. 2005] (**c**) Scalp projections of the average of the two spatial filters from CCA (**d**) Scalp projections of the average difference between epochs of faces and scrambled images at 170 ms. The blue colour in the posterior regions depicts a negative value.

Intra-subject correlation

Intra-subject correlation (IaSC) is tested by using BCoCA on the five datasets in each condition respectively. The resulting filters are then used to find the components with maximum mutual correlation from each of the datasets and the coincidence in neural activity is measured by computing the correlation coefficient on an intra subject basis. The correlation is computed in a window equal to the sample-length of one epoch and the step size is 25% of this. The population IaSC is the average of all the individual IaSC. A two-sided permutation test shows that 91% of the windows are significantly correlated.

Inter-subject correlation

The inter-subject correlation (ISC) is found by pooling all the datasets and using BCoCA. The components from each dataset is correlated with all of the others and then averaged to get the population correlation. In figure 2.14 the average over all the epochs of the combined component clearly show the decrease of alpha activity after mention of the first word and almost immediate increase after the response. A two-sided permutation test shows that 90% of the windows are significantly correlated.

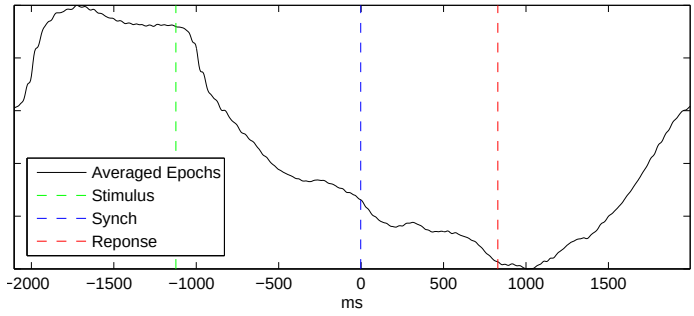


Figure 2.14: Averaged epoch of the first component from the inter-subject paradigm

CHAPTER 3

Recording EEG on One or Multiple Subjects

3.1 Hardware

Research grade EEG equipment is often very expensive, time-consuming to equip, and immobile. Using smaller consumer grade hardware has thus many advantages if it is able to measure the required signals adequately. Part of this experiment is to validate if the hardware used is sufficient for this paradigm, in what areas it may be advantageous, and in what areas it is lacking.

3.1.1 Emocap

To conduct the experiments the mobile 14 channel consumer EEG headset Emotiv EPOC has been rebuild to a wireless cap based on EasyCap, the Emocap. The sampling frequency of the ADC is 2048Hz but since the EPOC only have one ADC the data is sampled sequentially which means that the effective sample frequency of each channel is 128Hz (including Common Mode Reference and Driven-Right-Leg electrodes) [Emotiv 2012]. Each sample is assigned a number from 0 - 128 in the EPOC in order to ensure detection of packet loss on a sub-second time-scale. The EPOC has previously been validated against the Biosemi Active-II device with 64 channels using an imagined finger tapping paradigm [Stopczynski et al. 2013].

The electrode placement of the Emocap follows the 10-20 system in naming and placement but with only 16 channels the configuration is specific to this setup. The placement of the 14 measurement electrodes is illustrated in figure 3.1.

Noise reduction

Biopotential amplifiers usually amplify very low amplitude signals and is therefore very affected by most sources of noise. To reject external noise a method called Driven-Right-Leg (DRL) is applied [Nagel 2000]. The DRL circuit is a negative feedback loop of the common mode reference which effectively reduces the voltage in the common mode reference and resulting interference from e.g. 50Hz power line noise [Webster 1984].

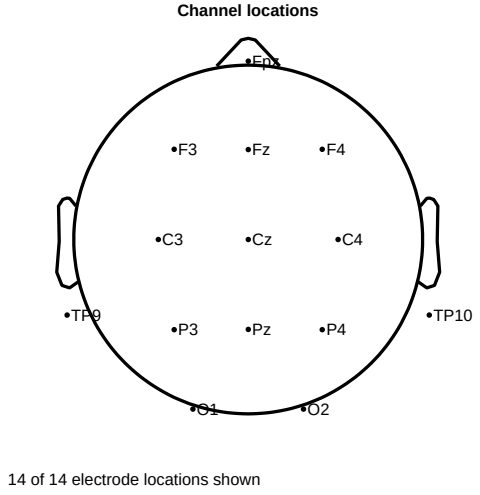


Figure 3.1: Channel locations of the Emocap

To further suppress noise from electrical power lines the EPOC uses digital notch filters at 50/60Hz and the harmonics. The bandwidth is, however, only reported to be 0.2 - 45Hz which must originate from applying the reported digital 5th order Sinc filter.

Conductance

To measure the conductance, inverse impedance, the Emotiv EPOC superimposes a 128Hz square wave on the DRL feedback signal. In each electrode the amplitude of the wave is measured on a 2Hz basis and from that the electrode *Contact Quality* is derived and presented as an arbitrary value relative to the true conductance [Delic et al. 2008].

To estimate the relationship between the conductance and the *Contact Quality* a range of resistors with varying resistance was connected to the DRL and an electrode. Figure 3.2 shows that the measured *Contact Quality* is about 530 when the connection is shorted and then a near exponential decrease (fitted curve) with increased resistance. It should be noted that the measurement points are sparse and the test thus only gives a rough estimate. A more thorough test would likewise need to include all the electrodes and would still only be valid for the device which the test was conducted on. In fact, having measured only one electrode may default this test completely.

Signal strength

The maximal distance between transmitter and receiver proved to vary a great deal among the transmitters. The best distance achieved is about five meters and the

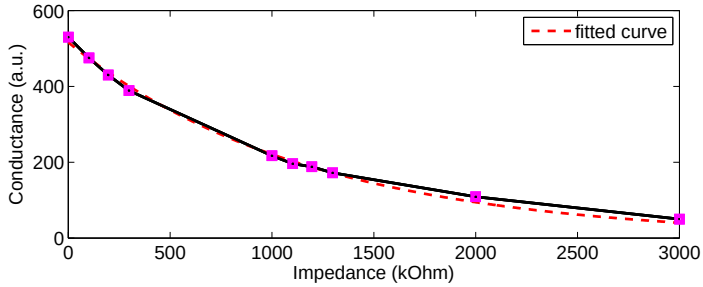


Figure 3.2: Contact Quality versus known resistance

worse less than one meter. It is unclear whether this problem is inherited from the EPOC or from the rebuild to Emocap.

Crosstalk

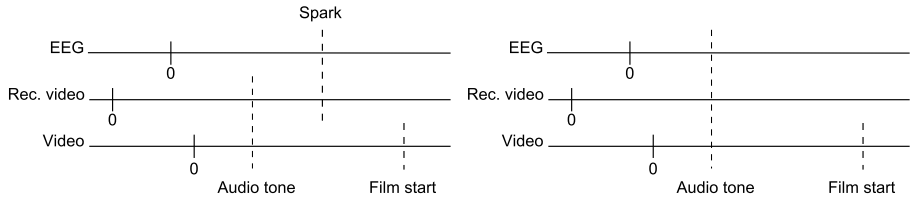
During experiments with more than one transmitter-receiver pair activated a high amplitude noise was observed in both receivers for particular transmitter-receiver configurations. The noise is clearly caused by crosstalk but it is unclear whether the problem lies in the transmitter, receiver or both. When a connection is established the receiver presumable lock on to the transmitter using a fixed frequency channel that is specific to the pair. If the hardware has difficulties determining the already used frequencies they could end up using the same. It should be noted that the environment the tests are conducted in are very polluted in the 2.4GHz range which is also used by the EPOC. By testing it was possible to find a configuration of nine pairs that did not crosstalk. See appendix D.1 for further details.

3.1.2 Tablet

To acquire and record data from the Emocap it is possible to use a computer or a mobile device supporting direct access to the USB port. In this experiment tablets of the model Asus Nexus 7 were used. The processing power of the device is much greater than needed for this application and since the tablet has previously been shown to work well with the EPOC [Stopczynski et al. 2013], the device has not been tested further in this regard. When the tablet went into sleep mode while recording EEG the connection between transmitter and receiver was lost after a short period of time but the application continued recording. To counter this problem a Wake Lock application was installed on the tablets so they did not go into sleep mode.

3.1.3 Synchronisation

Experiments involving a stimulus are highly dependent on temporal alignment if the objective is to compare the results across modalities or recordings. To synchronise



(a) Single Viewing **(b)** Joint Viewing

Figure 3.3: Illustration of synchronisation

EEG recordings with the film, two ideas were pursued.

Single viewing

The first idea was based on the theory that the electro magnetic wave generated by creating a powerful spark would induce a small current in the wires from the electrodes. This was confirmed using a piezoelectric spark generator normally used to ignite a Bunsen burner. Based on the length of the spark it was estimated that the spark was around 2 kV and very low amperage. When used approximately 2 centimetres from the electrodes a spike with much higher amplitude than the surrounding artefact free EEG was observed. Generation of the spark also emitted a noise that was distinguishable in the audiotrack of the recorded video. This method was hence used to synchronise the EEG with the recorded video in single subject experiments. To synchronise the recorded video with the film showed on the tablet, the audio output from the tablet was connected to the input of the camera in parallel with a microphone. At a fixed time before the first film clip a 43Hz sinus tone was played to make this part of the synchronisation easier, as illustrated in figure 3.3(a).

Joint viewing

The small spark generator was too weak to induce a sufficient current more than 30 cm away which was required in the joint viewing experiment. To increase the distance it is necessary to increase the field strength and thus the power used in the spark generation. A circuit similar to the one in figure 3.4 was created using the circuit from an electric fly swatter with increased capacitance and a flyback transformer from an old CRT television. The gap provides a high resistance which allows energy to build up in the capacitor until the potential across the gap is high enough to ionise the air and the spark is generated. The field from the circuit was not visible in the EEG so either the signal strength was not adequate or the DRL circuit in the Emocap managed to suppress the signal.

As an alternative method of synchronisation the audio used to synchronise film and recorded video in the single viewings was expanded to include the EEG recordings directly, as illustrated in figure 3.3(b). The audio output from the tablet was too weak

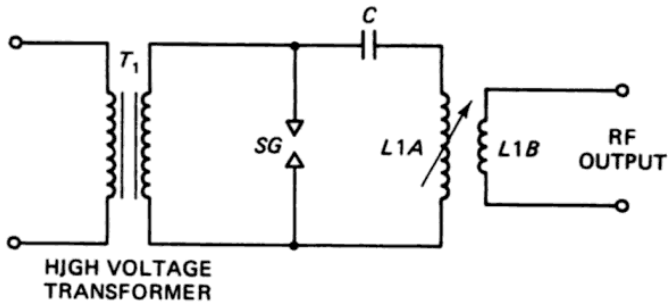


Figure 3.4: Spark Gap Generator [Carr 1997]

to register in the EEG recordings so a pair of computer speakers were disassembled and the amplifier extracted. To distribute the audio signal nine thin speaker wires were soldered together in one end, positive and negative separately, and in the other end each wire connected to a crocodile clip. In the caps one wire was connected to the reference electrode and the other to an electrode in the posterior region. Using this approach the signal was clear even with the amplifier in its lowest setting. The amplifier power supply is an AC-AC 15 volts $\sim$ 800 mA transformer and ensures galvanic separation between power lines and amplifier. The AC voltage at the used setting is measured to 61.4 mV and at maximum gain to 2.6 V. Assuming a low but realistic electrode-scalp impedance of 10 kOhms the current would be

$$\frac{61.4 \cdot 10^{-3} V}{10 \cdot 10^{3} \Omega} = 6.14 \mu A$$

Transcranial direct current stimulation (tDCS) uses currents of 1 mA, which is within the safety range, why the theoretical maximum current of this study is considered safe as well [Poreisz et al. 2007]. With ourselves as test subjects the amplifier was tested with maximum gain in which case it was possible to feel a small tingling between the electrodes.

3.2 Software: SBS2 DataRecorder

The application to record EEG from the Emocap is based on the multi-platform *Smartphone Brain Scanner* (SBS2) framework by Stopczynski et al. [2013]. The software is developed in the cross-platform environment Qt which is based on standard C++. The core of the system is created like a pipeline in which the data from device to application flows and the modular framework ensures easy accessibility into the pipeline at any point. The pipe consists of minimum three layers where the core of the framework is build from the first two. The first is data acquisition in which the application applies low-level functionalities that are specific to the hardware and operating system. Data is read directly from the USB mounting point which requires

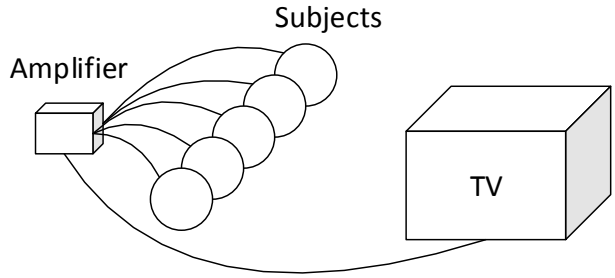


Figure 3.5: Joint viewing setup

a custom kernel and root privileges. The raw data is encrypted from the Emotiv hardware and needs to be decrypted before it is packed into a well-defined EEG packet object that can be handled further down the line.

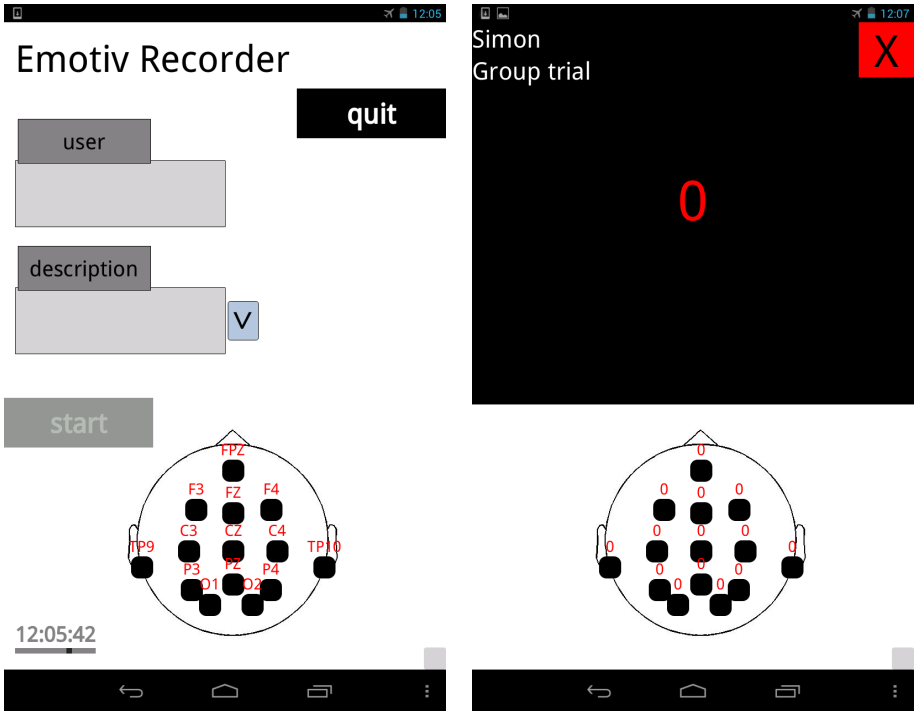
The second layer handles data processing and in the core a number of functionalities including, filtering, FFT and classifying is already implemented, however, the DataRecorder application only employs the recording functionality.

The third layer consists of the user interface and application specific functions to handle user inputs and data. The DataRecorder is a simple two-page interface with a setup screen (figure 3.6(a)) for text-based user input for the *Name* and *Description* of the forthcoming recording and a recording view (figure 3.6(b)) showing the sample frequency. In both views the lower third of the screen shows the electrode positions of either the Emotiv Epoc or Emocap (depending on hardware). Above the electrodes it is possible to see the conductance value received from the hardware or the name of the electrode.

3.3 Experimental setup

This section contains information regarding how both the solo viewing and joint viewing EEG recordings were conducted. To avoid gender playing a part in the results all 42 subjects were female with an average age of 22.4 years, distributed with minimum, median, and maximum ages of 18, 22, and 32 respectively. All subjects signed a consent for the use of data, video and image. Further information regarding the subjects can be seen in table D.2.

The subjects were divided into two groups, with one group of 24 subjects watching the films alone (single viewing) and another group of 18 subjects subdivided into groups of nine who watched the films together (joint viewing). There were taken precautions to ensure that the subjects participating in the same joint viewing, did not know each other beforehand. The group with single viewings were additionally evenly divided



(a) Setup **(b)** Recording

Figure 3.6: Graphical interface of DataRecorder application

into a group watching the films with the order of the scenes scrambled and a group watching the film clips normally.

3.3.1 Stimulus

One of the goals of the experiment was to recreate the results presented in Dmochowski et al. [2012], where the subjects were shown clips from three different films; *Bang! You're Dead* (1961) directed by Alfred Hitchcock, *The Good, the Bad, and the Ugly* (1966), a western directed by Sergio Leone, and a control film of a natural outdoor scene on a college campus. The Hitchcock film produced great results and the same clip was therefore included in the experiment for this thesis. The western, however, did not produce as many significant times of correlation, and it was decided to replace this clip with one from *Sophie's Choice* (1982) directed by Alan J. Pakula.

The clip from *Sophie's Choice* depicts a young Polish mother on her way to concentration camp during World War II, with her two children. She is accosted by a German officer, who forces her to choose which of her children lives or dies. The dialogue in

(a) Bang! You're Dead

(b) Sophie's Choice

(c) Control

Figure 3.7: Stills from the three film clips shown to the subjects.

the film clip is in German. The same film clip was used by Raz, Winetraub, et al. [2012], where the subjects were investigated for emotion-related changes using fMRI and viewer feedback rating. The study found a monotonic increasing response with the highest scoring emotions being "horror", "hate", "fear", and "anger".

To act as a control, a video was recorded of the escalators Kgs. Nytorv metro station in Copenhagen. This setting was chosen to eliminate the argument that the joint engagement is found for vision of a body versus non-body stimulus. The metro station was chosen as it was rationalised that the passengers getting on the metro in this station, were in less of a hurry compared to other stations, thereby reducing any excitement of people running to catch their train. Figure 3.7 shows stills from the three film clips.

Each clip had a length of approximately 6 minutes and were shown twice to each subject. For each viewing the order was randomised, but the same order was used the second time the clips were shown. A combined video was created for each of the six possible permutations of the order of the clips, starting with a 10 second 43 Hz tone for use in post processing synchronisation, and 20 seconds black screen between each film clip. At the end of the video the subject was presented with a text announcing that the video was over, to avoid the subject wondering if they just saw the last clip, between each clip. The total length of the video amounted to 39

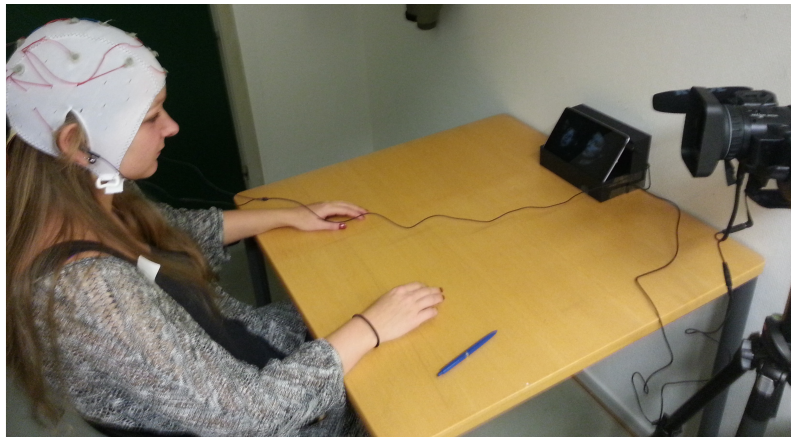


Figure 3.8: Experimental setup for solo viewings.

minutes.

In Dmochowski et al. [2012] the order of the scenes in *Bang! You're Dead* were scrambled to investigate the response when the meaning of the film was lost. The same approach was used in this thesis for both *Bang! You're Dead* and *Sophie's Choice*. Since the control video was intended not to carry any meaning, this was left out of the video with scrambled scenes resulting in only two permutations and a length of 23 minutes.

3.3.2 Solo viewings

24 subjects were used for the solo viewings, which were conducted in a small office as seen on figure 3.8. The film was shown on a Google Nexus 7 (2012) tablet, with a 7" (17.78 cm) screen with the subject hearing the films through in-ear headphones to avoid wires crossing the head. The headphones had a noise dampening effect which was important due to some of the recordings being made in office hours. The subject was instructed to sit straight, and avoid movements which can cause artefacts in the EEG, such as chewing, heavy breathing, and limb movement. The subject was instructed to keep the eyes inside the screen to reduce eye artefacts, but was also told to relax and follow the film.

Before the viewing started each subject drew without replacement for whether the films should be scrambled or not, and afterwards used a dice to decide the order of the film clips. Due to some initial technical difficulties, only the non-scrambled clips where available the first day of recording.

The subject was filmed with a camera receiving sound input from the tablet (which had its sound output split in two) as well as from an external microphone. An electric spark was used for post processing synchronisation between the spark showing in the

EEG and its clicking sound on the camera recording. As the camera also recorded the sound output from the tablets, the time interval between the spark and the time of the 43 Hz tune could be calculated, and from this the time of start for each film clip.

The lighting in the room was controlled by blacking out the office window and only having an architect lamp on, so the subject was visible on the camera in the dim light.

3.3.3 Joint viewing

The joint viewing experiment is an expansion of the solo viewing experiment presented by Dmochowski et al. [2012]. Since recording on nine subjects simultaneously is relatively new territory and presents new obstacles, the experimental setup deviates from the one in the solo viewing in some areas.

A lot of thought was put into the placement of the subjects in the room in relation to the screen and to each other. It was decided to go for a ”cinema experience”, with all nine subjects sitting on a line of chairs. By instructing the subjects to keep their eyes within the screen, as in the solo viewings, they were not able to directly see the facial expressions of one another. As the films were watched on a projector it was possible to both regulate the distance from the subject to the screen and the length of the diagonal of the picture projected on the screen. It was decided to keep the viewing angle from one corner of the screen to the opposite corner similar to the one in the joint viewing. By assuming the line of sight was orthogonal to the screen the relation

$$\text{angle} = \tan^{-1}\frac{\text{screen diagonal}}{\text{distance to screen}} \tag{3.1}$$

was used to find the maximal angle the eye could move while still viewing the screen. In the solo viewings the distance from head to screen varied from 70-90 cm, giving angles of maximal eye movement of 11.2° to 14.3°. The distance from the subject in the centre chair to the screen was measured to be 450 cm and 490 cm for the outermost placed subjects. With a screen diagonal of 102 cm this resulted in angles of maximal eye movement between 11.8° and 12.8°.

The recordings were done in a larger room, to accommodate all the subjects, and the sound from the films was played through loudspeakers, to avoid the emotional distance which noise dampening headphones might produce. On the basis of creating similar lighting as in the solo viewings the windows were blacked out and four lamps placed strategically to avoid shining a light in the eyes of the subjects, but still illuminating them for the purpose of filming them. Unfortunately the camera used in the solo viewings had to be used for another project, and was replaced by a GoPro Hero 2. The image and sound quality of the GoPro was not as good as the original camera, but it had the benefit of being unobtrusive, and could be placed directly in front of the subjects. The recording tablets were placed on tables directly behind the subjects to

Figure 3.9: Experimental setup for joint viewings. **Left:** Picture of the subjects seen from the front before viewing the films. All subjects were placed on a line to induce a "cinema experiences". **Right:** Subjects seen from the back before viewing the films. The recording tablets were placed on tables directly behind the subjects to avoid loss of connection from transmitters with poor transmitting distance. Cables were connected to the reference electrode and Cz on each subject to induce the 43 Hz tune directly into the measured EEG. The cables were removed before the film started.

avoid loss of connection from transmitters with poor transmitting distance. Cables were connected to the reference electrode and one of the other electrodes on each subject to induce the 43 Hz tune directly into the measured EEG, for later post processing synchronisation. The cables were removed before the film started.

3.3.4 Questionnaires and general information about the subjects

Before the EEG recordings all subjects were asked to fill out a questionnaire. Apart from asking relevant physiological questions, it was also chosen to ask the subjects to evaluate their level of proficiency in German, because of the German dialogue in *Sophie's Choice*. This was done to enable future subdivision of the subjects based on their self proclaimed understanding of German.

After viewing the films the subjects were asked to answer another questionnaire regarding whether they knew the scenes beforehand and which scenes had the biggest impact on them. Subjects viewing the films with scrambled scenes were also asked to describe the plot in the two films. This was both done to evaluate and possibly subdivide the subjects based on their understanding as well as for a comparison with the results gained from the EEG.

Appendix D contains English versions of the questionnaires the subjects where presented with before and after the EEG recording. As most subjects were Danes, they were presented with a Danish version. A summery of their answers can be seen in table D.2.

3.4 Pre-processing

Because of temporary loss of connection during recording, data from subject 3, 10 (single) and 26 (joint) has been discarded. For subject 30 in the joint viewing the synchronisation stimulus is not registrable in the EEG and has therefore also been discarded.

Using the assigned number to each packet between 0 and 128, loss of packets were detected and corrected for by inserting zeros in their place. However, since the sequence restarts at 128 the procedure is only precise in the sub-second time-scale. Any gap longer than 1 second is undetectable.

In some of the recordings very low frequency components were observed and since the amplitude was too high for the component to originate from the desired EEG signals, this was considered baseline drift. To remove the DC component and high frequency artefacts, the data was bandpass filtered using a linear phase windowed sinc FIR filter between 0.5 and 45 Hz and shifted to adjust for group delay [Widmann et al. 2012].

Using the Extended Infomax ICA algorithm implemented in EEGLAB in Matlab, the data was decomposed into statistically independent components. Ideally this would isolate the artefacts pertaining to eye blinks and these would then be easily removed by excluding the component when reconstructing the data. If the data contains many recordings with many channels it can become a challenge to manually pick the correct components to exclude. To alleviate this the CORRMAP plugin for EEGLAB can correlate a chosen scalp map with the components in all of the datasets and thus semi-automatically identify artefactual components [Viola et al. 2009]. From figure 3.10 it can be seen that the chosen template is very positive in the anterior region and negligible elsewhere, which is typical for eye artefact components. From a total of 38 datasets the algorithm found 36 components from 36 individual sets with a mean correlation of 0.9953.

To remove outliers, samples whose power was 4 standard deviations above the mean power of the respective channels were replaced by zeros. Removing artefacts in EEG is usually done by removing a segment around the artefact but since temporal synchronisation across multiple subjects is necessary for the experimental paradigm, that is not possible.

Normalisation of power levels across datasets is performed by dividing each sample with the root mean square of the collected dataset.

3.4.1 Additional synchronisation using CoCA components

Correction of intra subject synchronisation

The initial results from intra subject analysis did not prove to be as good as expected. It was suspected that package loss in the recordings or incorrect times of synchronisation for the films could cause a misalignment between the recordings from the first and the second viewing of the films. Since CoCA and BCoCA as well as the permuted

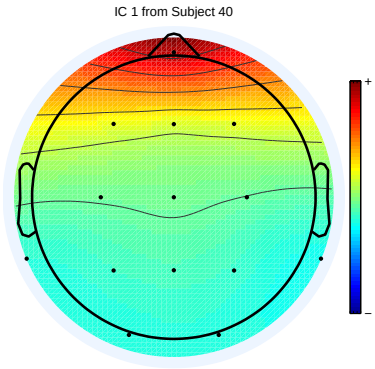


Figure 3.10: Scalp map of independent component used as template in CORRMAP (IC 1 from subject 40)

correlation test work by instantaneous correlation, a time shift could cause problems. To investigate this and find possible time shifts the weights for the first component, found by using CoCA on the joint inter subject tests, were used to filter the data for the intra subject analysis. The correlations between the first and second viewing were then calculated where the second viewing was shifted from -2 seconds to 2 seconds, ones sample at a time. With a sampling frequency of 128 Hz, this meant 257 correlations for each subject. The correlations were calculated in 5 second windows as in the ”real” analysis, but the permutations were left out to speed up computation times. Instead a value of 0.092 were used as the critical correlation, as earlier tests had attained values close to this as the critical correlations for a p-value of 0.01. The percentage of windows above this level were then used to decide which time shift to use for each subject and each film.

As can be seen on figure 3.11 this approach had mixed success depending on the subjects. The intra subject analysis were used both on data shifted in time according to the lags found, and on data which was not shifted in time. The outcome using CoCA was a large improvement of the solo viewings and the second joint viewing group and more moderate results from the first joint viewing group. Using BCoCA only an improvement of the second joint viewing group could be seen. In all cases however, the scalp maps were more similar to the ones found in Dmochowski et al. 2012, when using the time shifted data.

Correction of inter subject synchronisation

Synchronisation of viewings across subjects is precarious if all conditions in this regard are not tightly controlled. To adjust for some of the misalignment a method similar to the intra-subject correction can be applied but a search for the lag with highest

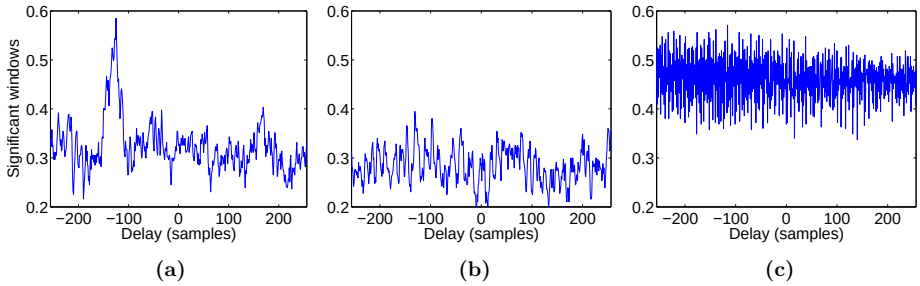


Figure 3.11: The percentage of significant windows when the second viewing have been delayed a varying amount of samples resulting in a time shift from -2 seconds to 2 seconds. **(a)** shows subject 38 which is an example were the right time shift is easy to decide. **(b)** shows subject 29 where the right delay is not as clear. **(c)** shows subject 24 which shows a suspicious indifference to a shift in time.

proportion of significant windows is highly resource demanding and thus infeasible when adjusting up 16 signals at the same time. Regular cross-correlation can take advantage of Fast Fourier Transform and is thus many orders of magnitude faster than the intra-subject approach, but with the disadvantage that correlations below the significant threshold are included.

It is difficult to determine the most accurately synchronised time series in a group so the first one was chosen as the initial template. To acquire the time series in component space the previously mentioned filter was used. With a two-second margin in both ends the template is correlated with the time series from the second subject. The maximum correlation coefficient is found and the second time series is adjusted with the lag of this coefficient. The template is then updated as the mean of the template and the adjusted second time series. This procedure iterates over every subject in the group and is repeated for the whole group until all the lags converges to zero or are equal. The scheme is very fast and allows adjustment for randomly permuted surrogate groups.

CHAPTER 4

Analysis of Recorded EEG

This chapter contains the results from the analysis of the EEG experiments described in chapter 3. The data obtained have been analysed using CoCA introduced in Dmochowski et al. [2012] and the Bayesian expansion, BCoCA, presented in this thesis. The analysis consists of two parts; intra and inter subject analysis. Intra subject analysis is also known as within subject analysis and compares the EEG between the first and second viewing of the films for each subject separately. With CoCA this is done by concatenating the datasets sample-wise as explained in Dmochowski et al. [2012]

$$\begin{aligned}\bar{\mathbf{X}}^{(1)} &= [\mathbf{X}_1^{(1)}, \mathbf{X}_1^{(2)}, \mathbf{X}_1^{(3)}, \mathbf{X}_1^{(4)}], \\ \bar{\mathbf{X}}^{(2)} &= [\mathbf{X}_2^{(1)}, \mathbf{X}_2^{(2)}, \mathbf{X}_2^{(3)}, \mathbf{X}_2^{(4)}]\end{aligned} \tag{4.1}$$

for $\mathbf{X}_i^{(m)}$, where i signifies the first or second viewing and m signifies subject number. BCoCA is not directly designed for pairwise comparisons between datasets. For the intra subject analysis presented in this thesis BCoCA will use the same concatenated data as CoCA. This might cause some difficulties for BCoCA when estimating the the noise covariance matrices for each of the concatenated datasets, as the covariance of their noise will be likely to vary between subjects.

Inter subject analysis is also known as between subject analysis and compares the EEG between all subjects for the same viewing of the films. Here BCoCA is better suited as it can do the comparisons between all datasets at the same time. CoCA however can only compare two datasets at the same time. For this analysis the data is concatenated for CoCA as explained in (2.173). As mentioned will the length of these concatenated datasets scale with the number of datasets squared. As an example will a comparison between 16 subjects result in a concatenated dataset with a length 120 times that of one dataset.

In this chapter the time series obtained from using the filters on the observed data will be mentioned as components. While the order in which these are generated from CoCA depends on their magnitude in a given dataset, the order of the three components presented in Dmochowski et al. [2012] will be used as a ”golden standard” for comparison. As a means of finding times of high intra- or inter-subject synchroni-

sation these components will be correlated with each other, to obtain the intra- and inter-subject correlations (IaSC, ISC).

The inter- and intra-subject analysis will be conducted where the subjects have been subdivided into groups depending on under which conditions they saw the films. These groups count: *single* for the subjects who saw the films alone, *scrambled* for the subjects who were also alone and saw the films with the order of scenes scrambled, as well as *joint 1* and *joint 2* for the first and second group of subjects that saw the films together.

As thesis have expanded on both CoCA and the experimental design, the ways in which to analyse and compare the data have expanded as well. To maintain an overview of the results only selected figures will be presented in this chapter, with appendix C containing additional relevant figures. Furthermore has the chapter been divided into the following sections:

1. A section to briefly illustrate the effects of using the time shifts explained in 3.4.1.
2. Then a presentation of the results, from using CoCA on the recorded EEG data, in the same manner as in Dmochowski et al. [2012] and which of their results it was possible to reproduce.
3. A section to present the intra and inter subject analysis using CoCA components with focus lying on differences between the results from the different groups of subjects.
4. Selected results using BCoCA will be presented here, and compared to the results from using CoCA on the same data.

4.1 Effect from additional synchronisation using CoCA components

As described in 3.4.1 did the initial results not prove as good as expected, but they were improved with relative shifts in time found through cross correlations.

The intra subject analysis were used both on data shifted in time according to the lags found and data which was not shifted in time. The results from the intra subject analysis can be seen on figure 4.1. The outcome using CoCA was a large improvement of the solo viewings and the second joint viewing group and more moderate results from the first joint viewing group. Using BCoCA only an improvement of the second joint viewing group could be seen. In all cases however, the scalp maps were more similar to the ones found in Dmochowski et al. [2012], when using the time shifted data. Figure 4.2 shows the scalp maps for the first joint viewing group using CoCA which, based on significant windows, did not show as large an improvement. It can be seen that without the time delay the resulting first component more or less corresponds to taking the average of all channels.

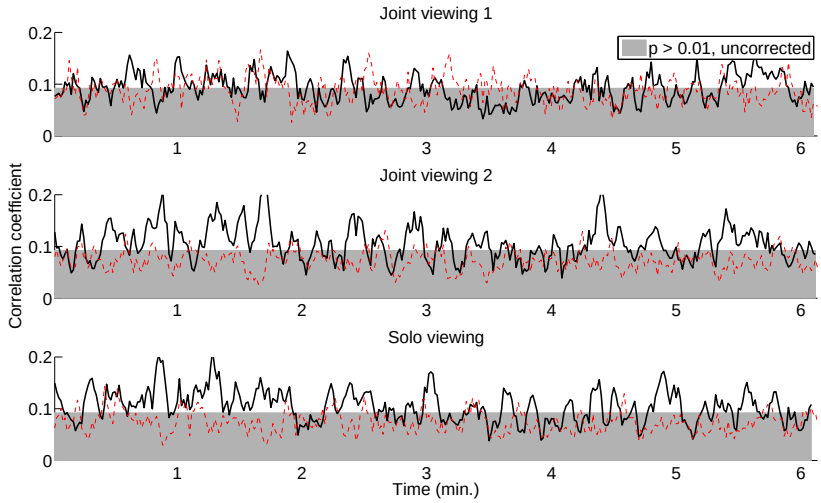
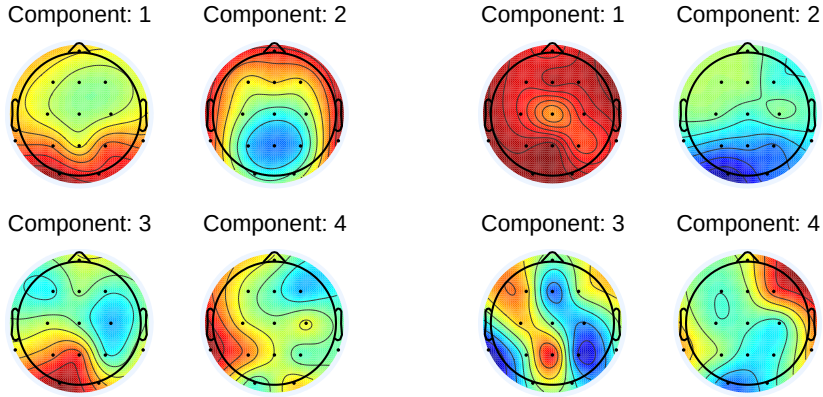


Figure 4.1: Intra subject analysis performed on EEG from the film *Bang! You're Dead* using CoCA. The black line shows the results with data shifted in time, and the red dashed line shows the results when data where not shifted in time.



(a) **(b)**

Figure 4.2: Scalp maps for intra subject analysis on the first joint viewing group using CoCA on **(a)** time shifted data and **(b)** and data without timeshifting. It can be seen that the first component in the latter corresponds to just taking the average of all channels.

4.2 Reproduction of results

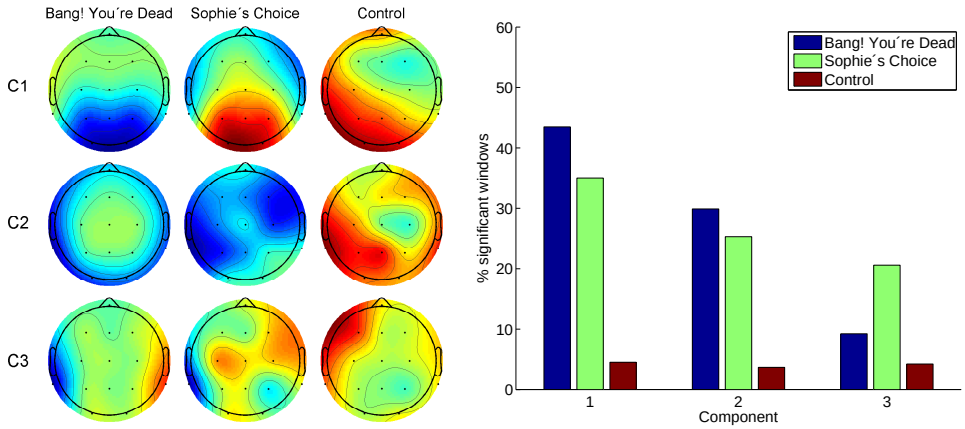
Reproduction of the results by Dmochowski et al. [2012] is important to establish whether the experiment has been successful in measuring brain activity related to the stimulus. The results are summarised in three ways; through a visual representation of the spatial filters in the form of scalp projections, the correlation coefficients viewed as a time series, and the percentage of significant correlations. In this section only the single viewings and the CoCA algorithm are considered.

4.2.1 Intra-subject correlations (IaSC) and arousing moments in the film

Figure 4.3(a) depicts the spatially distributed neural activity of the first three components from the single viewing cohort of non-scrambled films (n=10). Due to permutation ambiguity of the filters, it is not important whether a filter value is positive or negative, only the pattern of the filter is of interest. The neural activation pattern is very similar for the first components from the two films and the control to a lesser degree. Figure 4.3(b) summarises the proportion of significant correlations for each film in each of the first three components. The level drops with each component for *Bang! You're Dead* but fails to do so for *Sophie's Choice* though the scalp projections are very dissimilar. The level of correlated activity for the control film is as anticipated very low. Using a two-proportion z-test it was found that the proportions of significant windows in the two films are significantly different from the control with the p-values < 0.001 except for *Bang! You're Dead* in the third component with $p = 0.0038$. Figure 4.4(a) illustrates the correlation coefficient from *Sophie's Choice* and in the grey area the level of significance required for a correlation to be significant. Some of the peaks are represented in more than one component, but the large peaks are mostly unique to one. The peaks generally coincide with scenes of close-up shots of faces and high tension. Labels going from (b) to (g) mark the spots of peaks with large mass in different scenes. The first scene (b) is of a panoramic view over a queue of people waiting to get through the control. At this time the viewer might realise the place could be a World War II concentration camp. Scenes (c) and (d) is of encounters between Sophie and the German officer with increasing level of tension and anticipation. The culminating moment in (f) when the German officer commands one of her children taken away and (g) when the daughter is forcefully removed, though by her choice of what child to remove.

4.2.2 Comparison to films scrambled in time

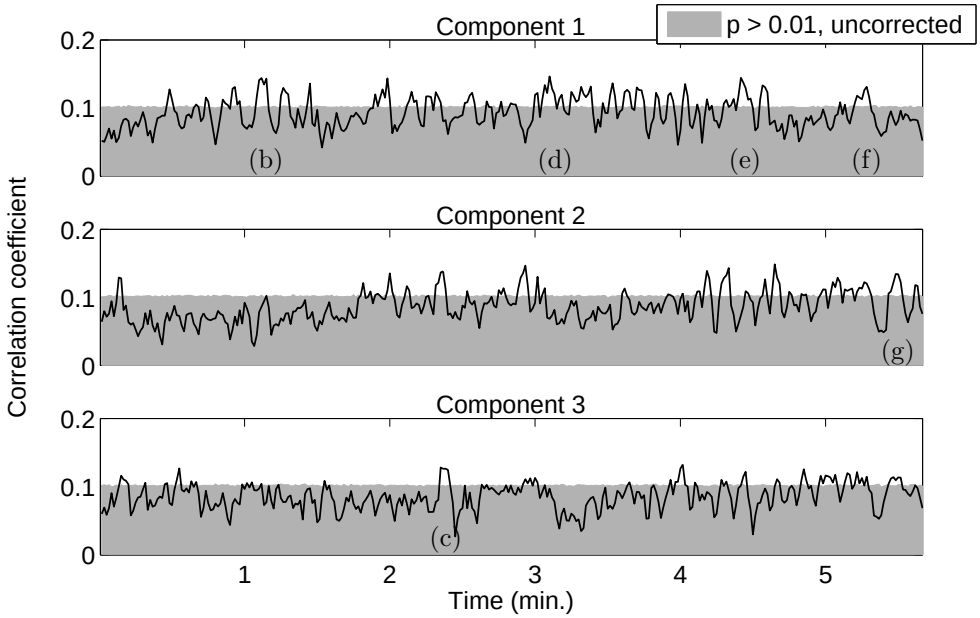
Visual perception of other humans activates low-level cognitive networks, as seen in section 2.10.1, so in order to control for this a scrambled version of the films were shown to a separate group of n = 12 subjects. In figure 4.5(a) it is shown that the proportion of statistically significant windows are reduced when the contextual meaning is removed from the film. Using the hypothesis test of proportions, both films



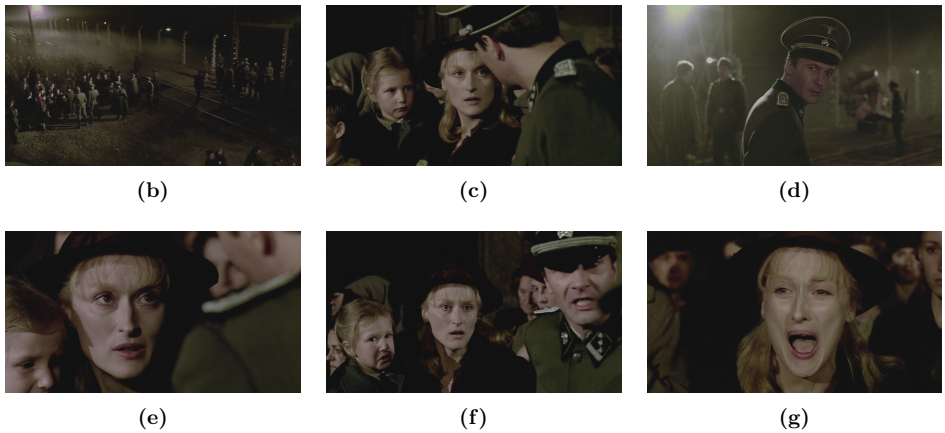
(a) Scalp projections **(b)** Significant correlations

Figure 4.3: Single viewing IaSC (a) Scalp projections of the first three components for each film and (b) the percentage of statistically significant correlations

were found to significantly different from the scrambled version in all components. However, even though the scenes were scrambled in random order, 10 of the subjects reported to have understood both films (appendix D). The order of the scenes were random so this particular order may elicit a larger response than others.



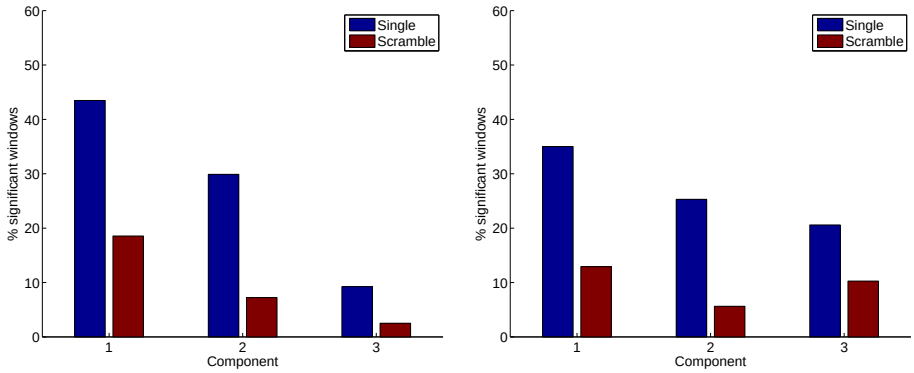
(a) Intra subject correlation of *Sophie's Choice* from the three first components



(b) **(c)** **(d)**

(e) **(f)** **(g)**

Figure 4.4: Inter-subject correlation and matching stills from *Sophie's Choice* at particular arousing moments. Peaks are chosen based on their mass and is shown in chronological order



(a) *Bang! You're Dead* **(b)** *Sophie's Choice*

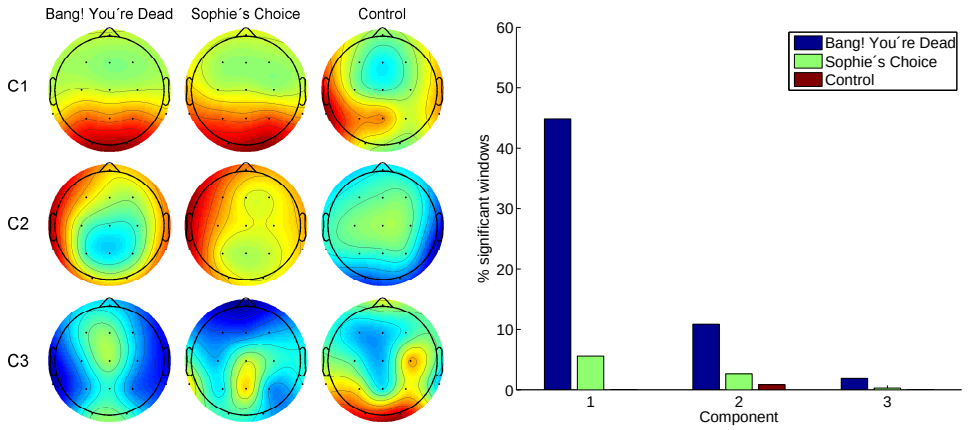
Figure 4.5: Comparison of significant correlations in the Single viewing IaSC between the original film and the scrambled version of *Bang! You're Dead* and *Sophie's Choice*

4.2.3 Inter-subject correlations (ISC) show decreasing correlation in the second viewing

We hypothesise that the neural activation pattern represents inherently low-level cognitive mechanisms why it is logical to assume similar patterns across individuals. To test for this a similar analysis to IaSC is carried out but across subjects and only for the first viewing. In figure 4.6(a) we see that the projections are similar for the films in the first and second component and similar to the ones in the intra-subject analysis in the sense they show high occipital activation. In this analysis the control is not similar at all, however. The proportion of significant windows shown in figure 4.6(b) are very low for *Sophie's Choice* though high for *Bang! You're Dead*. This might point to a difficulty in synchronising the segments for the former.

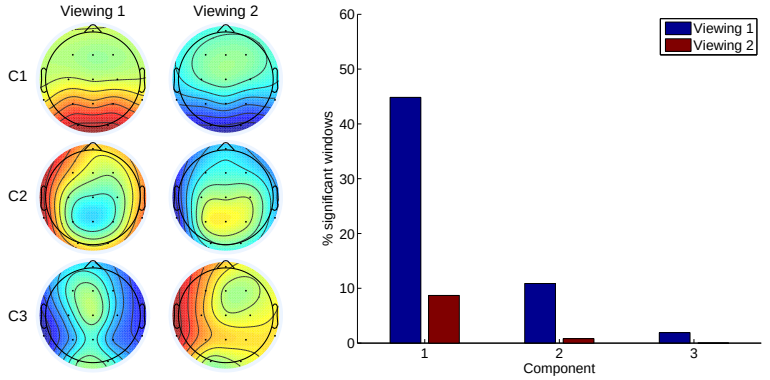
To test for the effect of watching a film a second time we compute the inter-subject correlation for the first and second viewing respectively. In figure 4.8(a) the correlation time series for both viewings of *Bang! You're Dead* show how the correlation of the second viewing is significant in some of the same peaks as viewing 1, but generally lower. This is also illustrated in figure 4.7(b) were the proportion of significant windows are seen to be much lower. The scalp projections (figure 4.7(a)) are, however, very similar so the main difference could be the order of magnitude of the spatial filters. The Wilcoxon signed rank test was performed to test the null hypothesis that the difference between viewings could originate from a distribution with zero median with the result that $p < 0.001$ for all components.

The peaks of *Bang! You're Dead* coincide with scenes of either the revolver or the bullets and the handling of these (b - e and i), scenes in which the boy pretend to trigger (f and h) and the scene in which the uncle discovers that the boy must have a real gun.



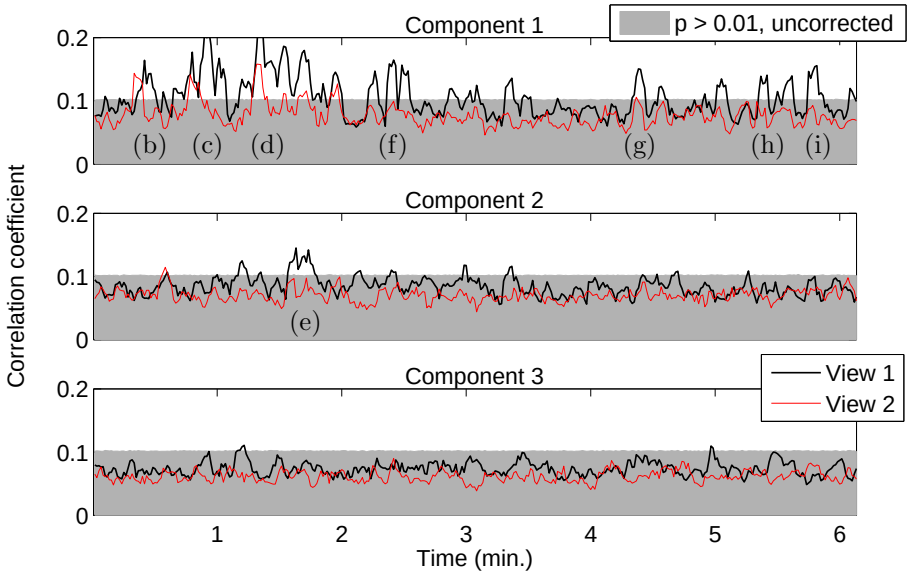
(a) Scalp projections **(b)** Significant correlations

Figure 4.6: Single viewing ISC (a) Scalp projections of the first three components for each film and (b) the percentage of statistically significant correlations

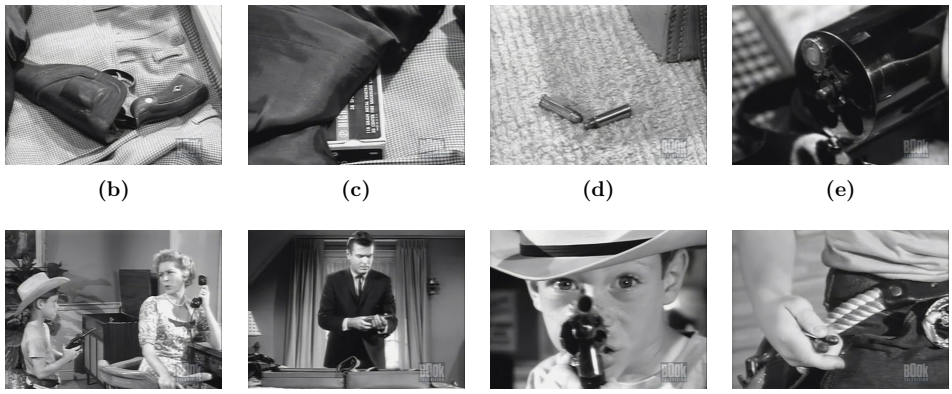


(a) Scalp projections **(b)** Significant correlations

Figure 4.7: Inter subject scalp projections (a) and significant correlations (b) of viewing 1 and 2 for *Bang! You're Dead*



(a) Inter subject correlation of *Bang! You're Dead* from the three first components of viewing 1 and 2 in the Single condition and indices corresponding to stills from the film



(b) **(c)** **(d)** **(e)**

(f) **(g)** **(h)** **(i)**

Figure 4.8: Inter-subject correlation and matching stills from *Bang! You're Dead* at particular arousing moments. Peaks are chosen based on their mass and shown in chronological order

4.3 Analysis regarding effect of viewing films in groups

This section presents the intra subject analysis using CoCA components with a focus on the differences in the conditions under which the groups of subjects watched the films. In this section the focus lies in the differences between watching the films alone or in a group, and if there are differences between the two joint viewings.

4.3.1 Intra subject analysis

Figures 4.9, 4.10, and 4.11 show the population IaSCs for the first component attained with CoCA for *Bang! You're Dead*, *Sophie's Choice* and the video recorded in a metro acting as a baseline. Comparing the single viewing with both groups for *Bang! You're Dead* it can be seen that peaks of correlation occuring in the single viewing also occurs in either joint viewing or both. Comparing the two joint groups it can be seen that the times of high correlation rarely occurs in the same places of the film.

A peak occurring in either group can generally be traced back to a similar peak in the single viewing group and often with a higher amplitude. This suggests that a group focusing on the same arousing stimulus can obtain a higher synchrony compared to watching the film separately. But a peak in the single viewings cannot always be traced back to a specific joint viewing, which might indicate that interesting moments can be suppressed when viewing the film in a group.

It can be seen on figure 4.12 that the scalp map for the first CoCA component stemming from the first joint viewing of *Sophie's Choice* differs a lot from the corresponding scalp maps for the second joint viewing and the single viewings. On figure C.7 the

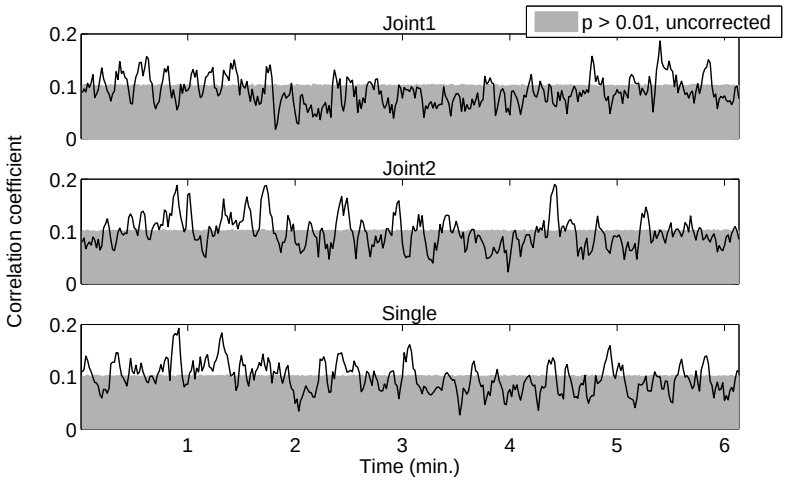


Figure 4.9: Population IaSC for the first CoCA component for the viewing of *Bang! You're Dead* in the first and second joint group as well as the single viewings.

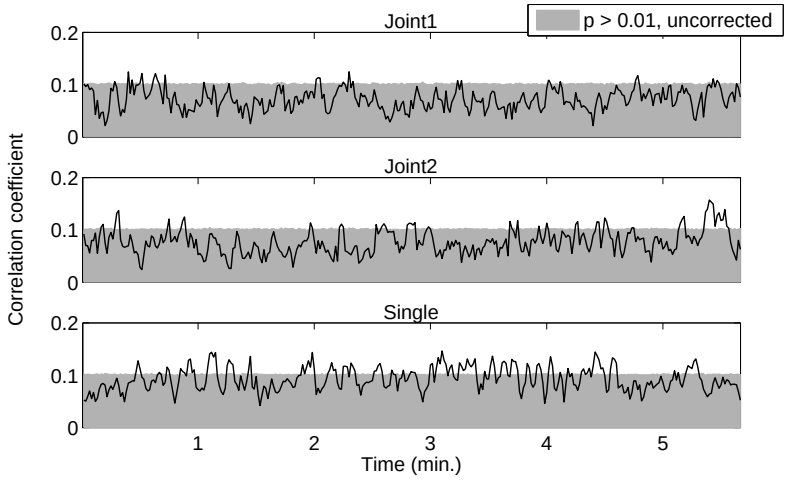


Figure 4.10: Population IaSC for the first CoCA component for the viewing of *Sophie's Choice* in the first and second joint group as well as the single viewings. Note that for joint 1 the component has been switched with what was estimated as being component 3, since its scalp map similar to the first scalp map in Dmochowski et al. [2012].

third component for *joint 1* is seen to be similar to the first scalp map in Dmochowski et al. [2012]. For this reason figure 4.10 contains the third component from *joint 1* and the first component from *joint 2* and *single*. The first component for *joint 1* can instead be seen on figure C.7, and shows many points of high correlation, which is probably the reason it was estimated as the first component.

After having done this rearrangement the same trend as seen on the IaSCs for *Bang! You're Dead* can be seen for the population IaSCs for *Sophie's Choice*, where the peaks of correlation for *joint 1* and *joint 2* rarely happens at the same time, but their peaks usually align with similar areas in the IaSC for *single*. For this film though, the viewing groups does not attain as many or as high peaks in correlation they did in *Bang! You're Dead*. Curiously the trend with aligned peaks of significant correlation can also be hinted in the IaSCs for the control video, though no interesting things are supposed to happen in this video.

Apart from *joint 1* for *Sophie's Choice* the scalp maps for the films are similar to the ones attained in Dmochowski et al. [2012], but the scalp maps stemming from the control video are not as similar.

Figure 4.13 shows a comparison of significant correlations for the average IaSCs for the different groups of subjects. A critical value of correlation corresponding to $p = 0.01$ have been obtained from a permutation test using 5000 permutations for each window. If the averaged IaSC had a correlation coefficient higher than this critical level, the window was classified as significant. Another method of comparison can be seen on figure C.11 in the appendix. Here all windows for each pair-wise

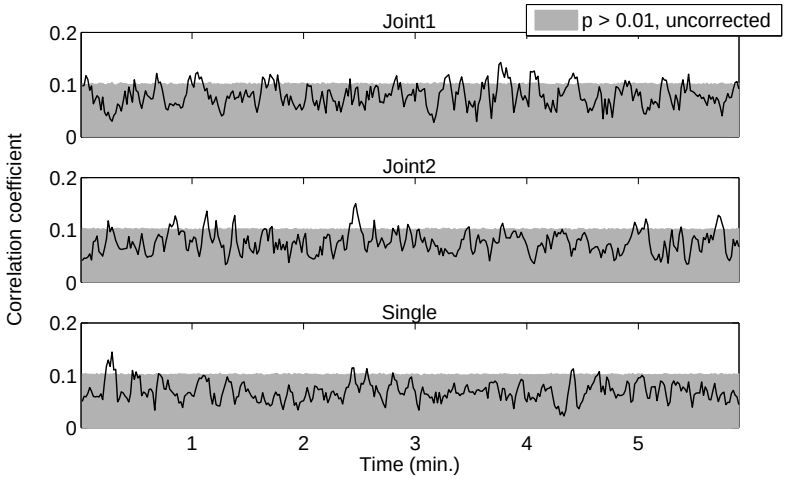


Figure 4.11: Population IaSC for the first CoCA component for the viewing of the control video in the first and second joint group as well as the single viewings.

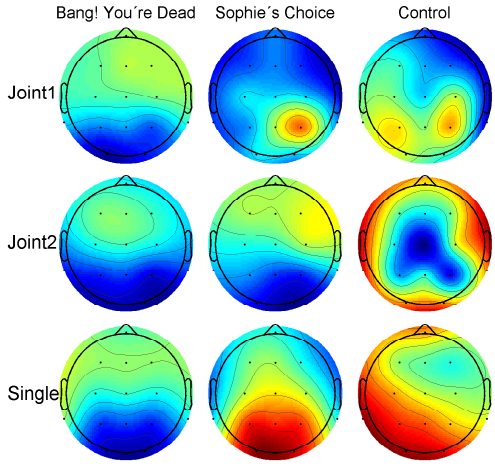
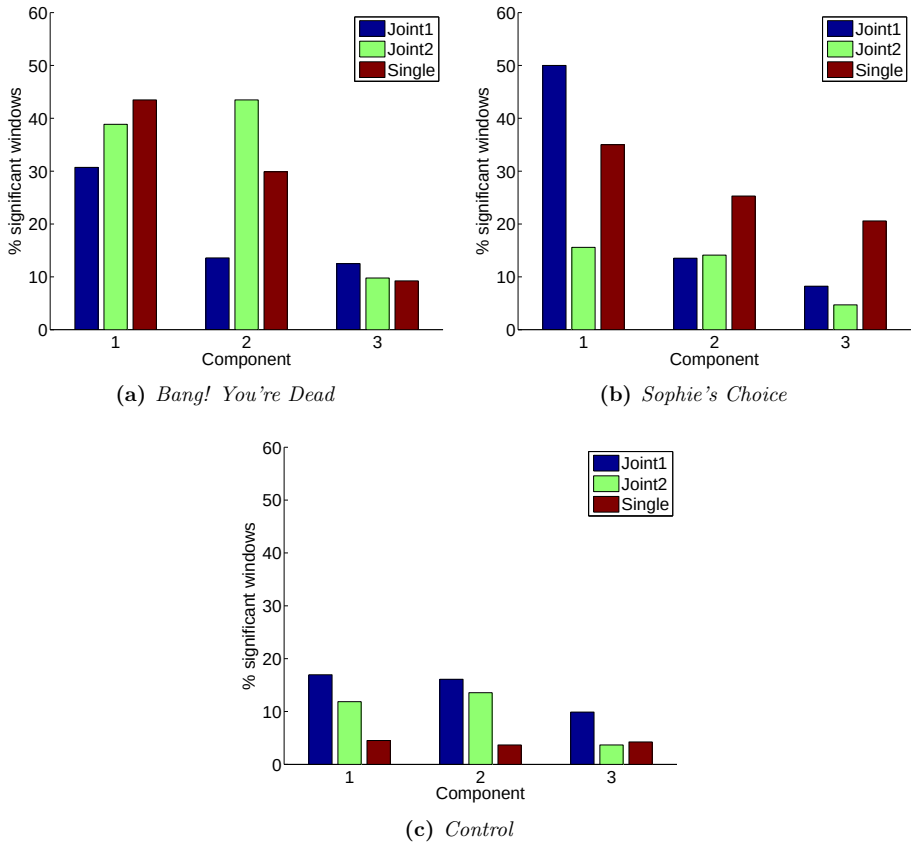


Figure 4.12: Intra subject scalp projections for the first component recorded for all three films and the *joint 1*, *joint 2* and *single* viewing groups.



(a) *Bang! You're Dead*

(b) *Sophie's Choice*

(c) *Control*

Figure 4.13: Comparison of significant correlations for the average IaSCs for different groups of subjects. The critical value of correlation corresponding to $p = 0.01$ have been obtained from a permutation test.

correlation for a group is tested for significance using the calculated p-values and controlled for multiple comparisons using FDR with a alpha level of 0.01. Though there are significant differences to be seen on figure 4.13 they are not consistent across the film or components.

4.3.2 Inter subject analysis

As in the previous section this section will investigate the differences in the conditions under which the groups of subjects watched the films. The focus will also lie on the differences between watching the films alone or in a group, and whether there is a difference between the two joint viewings, but it will be carried out through inter subject analysis. As an addition this analysis will introduce twelve *surrogate* joint

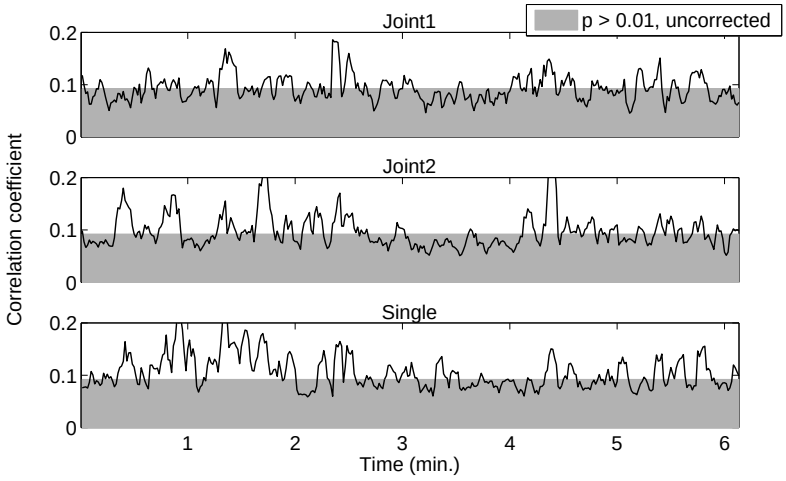


Figure 4.14: Population ISCs for the first CoCA component for the viewing of *Bang! You're Dead* in the first and second joint group as well as the single viewings.

groups each consisting of eight subjects chosen at random from each joint viewing group, with a distribution of either 3-5 or 4-4 in *joint 2*'s favor (on account of this group containing nine good datasets as opposed to *joint 1*'s seven).

Figure 4.14 shows the ISCs for the first CoCA component from the inter subject analysis of the *joint 1*, *joint 2* and *single* groups of subjects watching *Bang! You're Dead*. As could also be seen on figure 4.9 there are scenes in the film which generates high peaks of correlation for *single* and one of the joint groups, but not the other. As a comparison figure 4.15 shows the ISCs for three surrogate joint groups chosen for having the maximum, median, and minimum amount of significantly correlated windows (corrected with FDR) for *Bang! You're Dead*. All three have similar performance for *Sophie's Choice*, which can be seen in appendix C. It can be seen that these surrogate groups have peaks in all the areas that both *joint 1* and *joint 2* attain high correlations. It can also be seen that even the surrogate group with the lowest amount of significant windows still attain times with high correlation.

The scalp maps on figure 4.16 show both the "real" groups and the surrogates attain weights as expected for the first component, with the exception of *joint 1* which have a different scalp map for *Sophie's Choice*. This was also the case for the intra subject analysis though the estimated scalp maps are not the same. It can be seen that the surrogate group with the maximum number of significant windows have "inherited" this problem as well.

Figure 4.17 shows a comparison of significant correlations for the average ISCs for the different groups of subjects. This is calculated in the same manner as in the case of the intra subject analysis and figure C.14 shows a FDR corrected comparison for each pair-wise correlation. Like in the intra subject analysis no consistent conclusion can

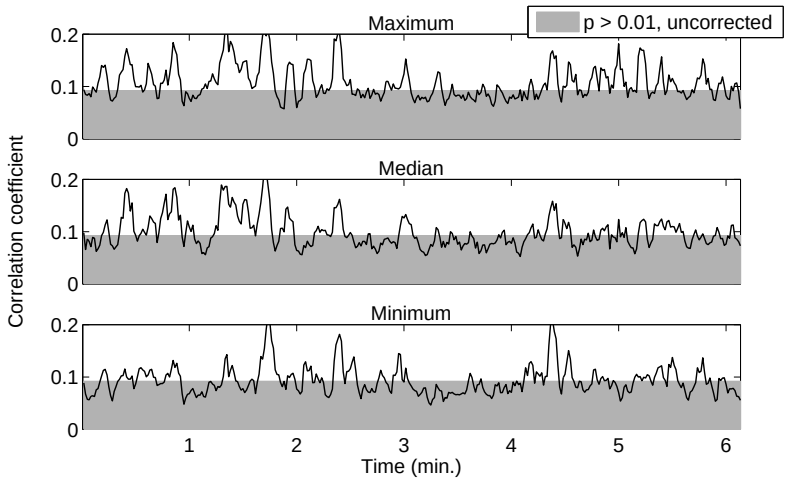
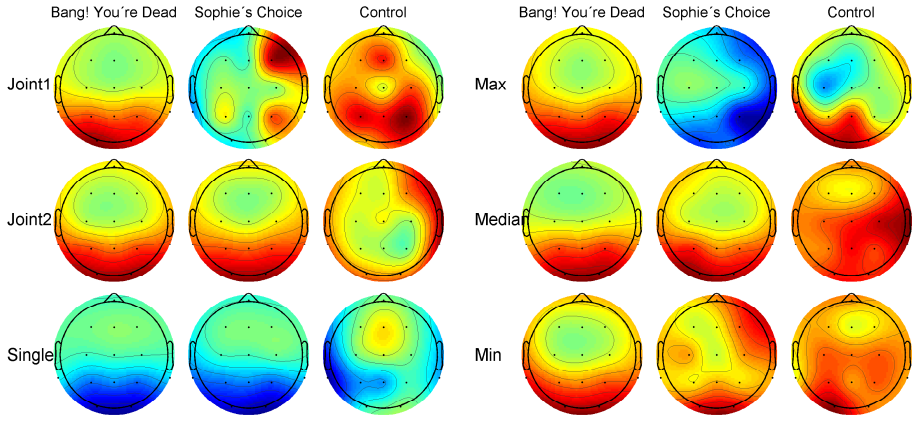
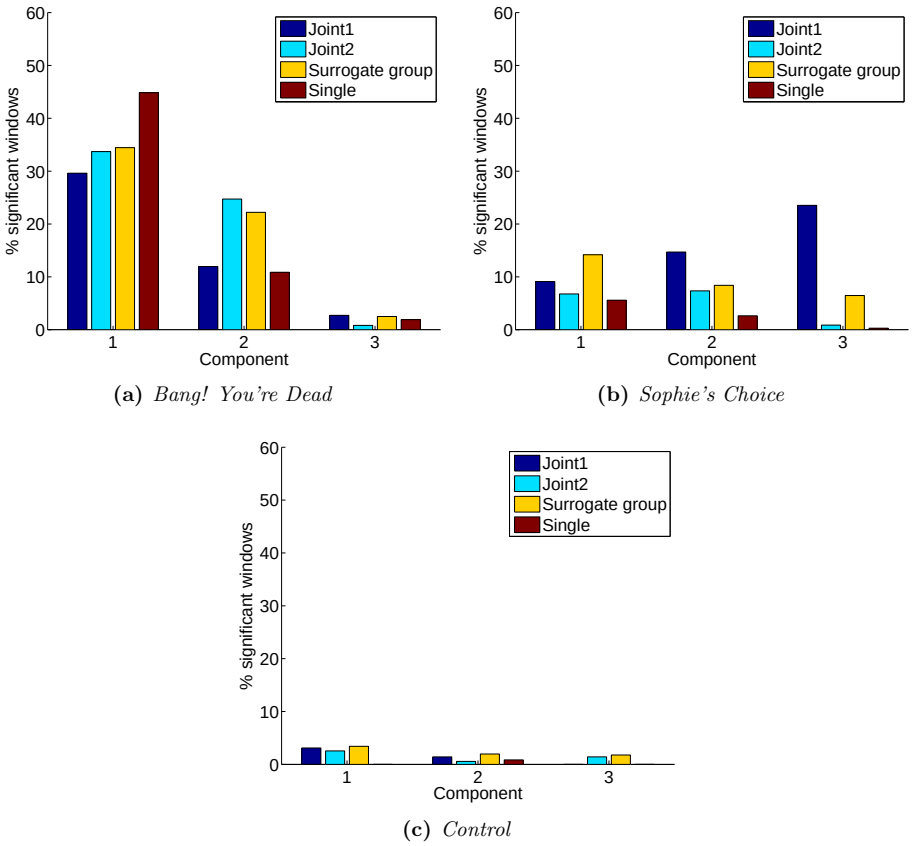


Figure 4.15: Population ISCs for the first CoCA component for the viewing of *Bang! You're Dead*. Each ISC is stemming from a surrogate group of eight subjects picked at random from *joint 1* and *joint 2*. The three ISCs seen in this figure attained the maximum, median and minimum number of significant windows.



(a) **(b)** Surrogate groups

Figure 4.16: Inter subject scalp projections, for all three films calculated from **(a)** the *joint 1*, *joint 2* and *single* viewing groups and **(b)** the three surrogate groups attaining the maximum, median and minimum number of significant windows.



(a) *Bang! You're Dead* **(b)** *Sophie's Choice*

(c) *Control*

Figure 4.17: Comparison of significant correlations for the average ISCs for different groups of subjects. The critical value of correlation corresponding to $p = 0.01$ have been obtained from a permutation test. For the surrogate groups the mean number of significant windows across all surrogate groups is shown.

be drawn regarding the relationship between the relationship between components and films across groups. This is also due to a large standard deviation in the mean number of significant windows for the surrogate groups ranging from 2.8 for the third component in *Bang! You're Dead* to 11.2 for the first component in *Sophie's Choice*.

Correlating joint viewing subjects from different groups

As an alternative to the surrogate groups, a surrogate joint ISC was created. Weights for the first component was calculated using CoCA on a combined group, containing all subjects viewing *Bang! You're Dead* in a group. Figure 4.18 shows the scalp map for this component. The surrogate population ISC was then calculated as the average

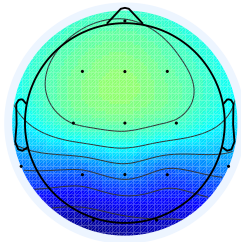


Figure 4.18: Scalp map for the first CoCA component estimated on a combined group containing all subjects from both groups viewing *Bang! You're Dead* jointly.

of all correlations between pairs of joint viewing subjects, which did not watch the film together. It was in other words calculated using pairs consisting of one subject from *joint 1* and one subject from *joint 2*. Figure 4.19 shows the population ISCs for the combined joint group, *joint 1*, *joint 2*, and the surrogate ISC using the CoCA component attained from the combined group.

It is interesting to see how the population ISCs for *joint 1* and *joint 2* are almost identical to the ones seen in figure 4.14 even though the weights for the former are calculated on the combined joint group. Looking closely, subtle differences can be seen but the peaks in correlation occur in the same places. But the most interesting aspect of figure 4.19 is how similar the combined joint ISC is to the surrogate joint ISC.

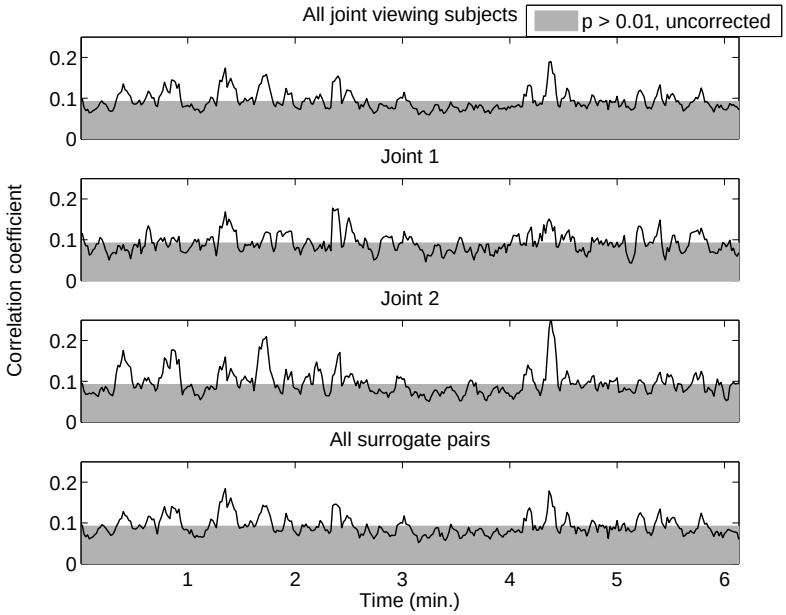


Figure 4.19: Population ISCs using the same CoCA component stemming from combining both groups viewing *Bang! You're Dead* jointly. The top figure shows the population ISC for this combined joint group. The two middle figures show *joint 1* and *joint 2* ISCs using the CoCA component attained from the combined group. The bottom plot shows an alternative surrogate ISC consisting only of the correlations between all pairs of joint subjects which were not in the same group.

4.4 Comparison of CoCA with Bayesian CoCA

This section will compare the performance of BCoCA to CoCA. It would be too extensive to present all the results so instead focus will be on selected scenarios. BCoCA does not necessarily return the maximally correlated components ordered by correlation so it is for every analysis important to investigate them all. For these results 3 components are calculated.

Recall from figure 4.4 that the population IaSC for the single viewing of *Sophie's Choice* had a large number of significant windows. Figure 4.21(a) is a duplicate of that illustration but now includes the correlations from the components computed using BCoCA. Many of the peaks are found by both algorithms, but the response from BCoCA is generally higher and more importantly, correlates with arousing moments in the film. The arrows on top of component two and three points to a large peak followed by a large decrease of correlation. In figure 4.20 it is seen that the same very distinctive curve is found to occur in the *Theory of Mind* but not the *embodied simulation* network, in the study by Raz, Jacob, et al. [2013].

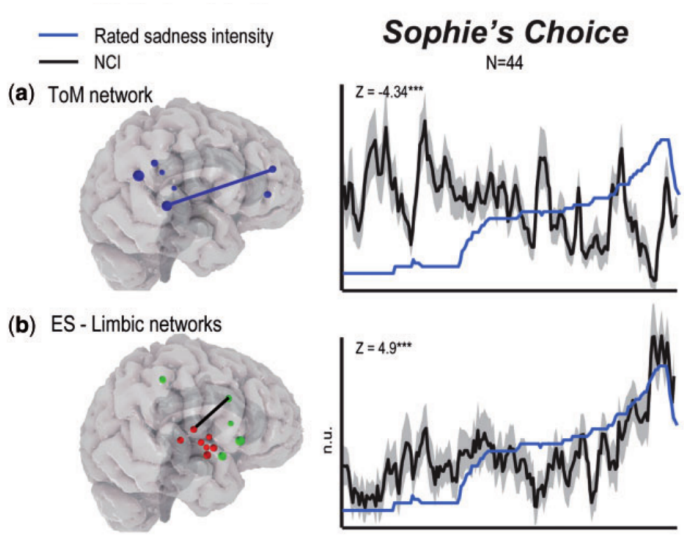
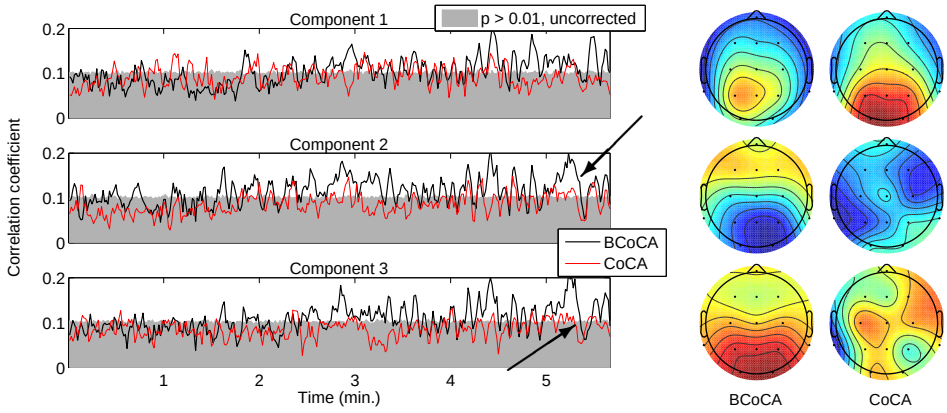


Figure 4.20: Illustration depicting the fMRI measured *network-cohesion index* in the *Theory of Mind* (ToM) and *embodied simulation* (ES) networks while watching the clip from *Sophie's Choice* [Raz, Jacob, et al. 2013]

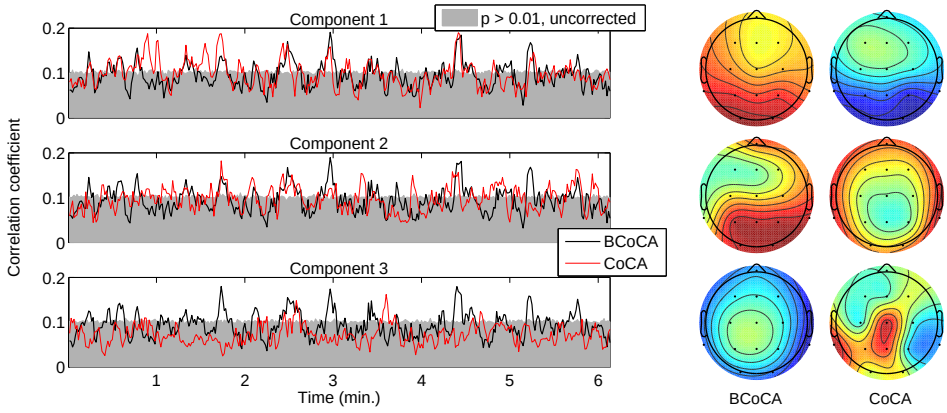
In figure 4.22 we see that BCoCA share many significant peaks with CoCA for *Bang! You're Dead* but the three components do not differ much from each other even though the scalp projections clearly do. Since CoCA finds peaks at different temporal locations in different components, the BCoCA algorithm could miss peaks in components with less correlation.

In the inter-subject condition BCoCA returns similar correlation results as CoCA, but numerically lower (figure 4.23 and 4.24). The algorithm also return similar correlations in all the components for this condition.



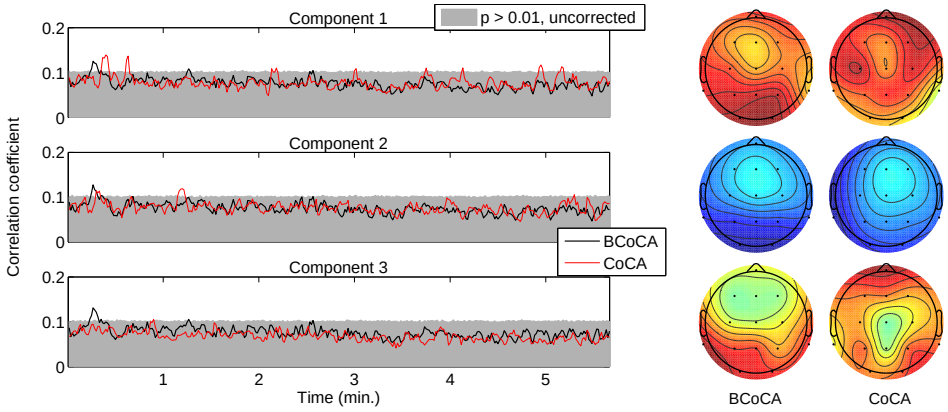
(a) Intra subject correlation **(b)** Scalp projections

Figure 4.21: Population intra subject analysis of *Sophie's Choice* in the *single* viewing group. The large peak and following dip in BCoCA component 2 and 3 are noteworthy because they correlate with arousing events in the film. The scalp projections (b) also show a more homogeneous pattern in BCoCA that extends further into the temporo-parietal region



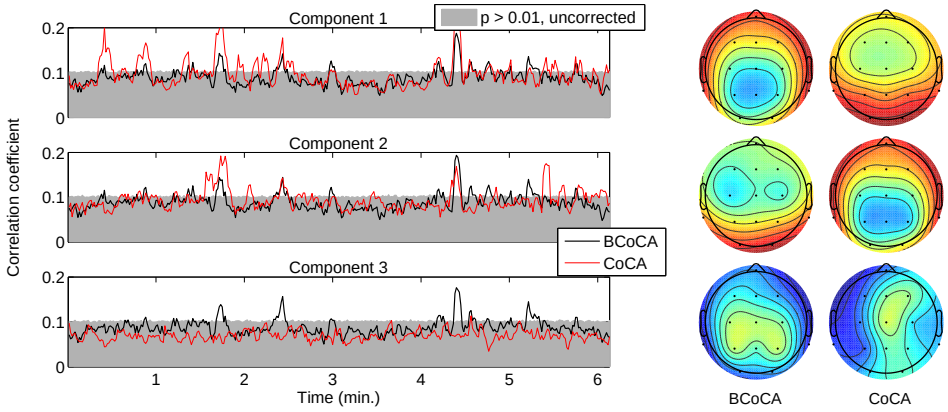
(a) Intra subject correlation **(b)** Scalp projections

Figure 4.22: Population intra subject analysis of *Bang! You're Dead* in the *joint 1* group. BCoCA seems to find the same filters but lower correlations



(a) Inter subject correlation **(b)** Scalp projections

Figure 4.23: Population inter subject analysis of *Sophie's Choice* in the *single* group. Both algorithms have difficulties finding correlated components and BCoCA seem to find the same component.



(a) Inter subject correlation **(b)** Scalp projections

Figure 4.24: Population inter subject analysis of *Bang! You're Dead* in the *joint 2* group. Some similar peaks between the algorithms, but very similar filters.

CHAPTER 5

Discussion

This thesis had as its starting point the article *Correlated components of ongoing EEG point to emotionally laden attention - a possible marker of engagement?* published by Dmochowski et al. in 2012, and had the goal to expand on it in two ways. The first goal was to create an algorithm representing a Bayesian approach to their signal decomposition method, CoCA. The second was to reproduce their experiment with having subjects view six minute film clips while recording EEG, as well as expand it by including a joint viewing experiment. This chapter discusses the results of our efforts to reach these two goals as well as ideas for how to improve and elaborate on them.

5.1 Bayesian correlated component analysis

Chapter 2 presented the theoretical background and the derivation of BCoCA, a Bayesian expansion to CoCA using variational inference. The derivations included a cost effective calculation of the lower bound for the purpose of estimating the time of convergence, by focusing on the variables that changes between iterations and letting the terms cancel each other out were applicable. The chapter also included two types of tests, one on data simulated for this purpose and one consisting of analysis on two real EEG datasets using BCoCA.

The first things tested for using the simulated data tests were which of two BCoCA models to use, and to see if there were differences between implementing the derived updates in Matlab or a VMP implementation using Infer.NET. The result was a BCoCA model based on a shared mean for the weights, $\mathbf{U}$, and an indication that the Matlab and Infer.NET implementations performed equally. The tests on simulated data were conducted on data where the true sources, their mixing matrices, and the added noise were all created from the same distributions, which BCoCA was modelled on. The true sources were identical to the ones used in Klami [2013] for increased comparability with their results, though other parameters where chosen differently, with the addition of noise probably being the most important. Under these circumstances BCoCA thrived and attained better or nearly just as good results compared to another latent variable model, GFA, as well as CCA and CoCA, which solves the de-mixing problem analytically through eigenvalue decomposition. BCoCA showed its strengths when working with more than two datasets where it had higher or equal performance compared to the other algorithms. In the two-view situation CCA and CoCA proved to attain better results at low values of SNR with BCoCA tending

to chose the ”zero-solution” when the noise level got too high, though this threshold happened at higher noise levels compared to GFA. By implementing a greater control of the covariance matrix it might prevent BCoCA from choosing the zero solution, which is rarely an optimal solution. The ability to discern multiple mixed sources with high levels of added noise is good qualities for an algorithm intended to find hidden responses in EEG, since this monitors all brain activity simultaneously, have low signal amplitude due to physiological dampening of the signal, and a resulting susceptibility to unwanted noise.

The tests on real EEG datasets proved that BCoCA, as well as CoCA and CCA, were able to find a reported N170 response in EEG from a face recognition experiment. The corresponding scalp maps showed that CoCA and BCoCA achieved more anatomically correct results when compared to the projection of the actual activity. CCA managed to find the same signals, but with a scalp map focusing on specific channels. This difference in resulting scalp maps between CCA and CoCA, was also reported in Dmochowski et al. [2012]. When it comes to extracranial EEG research show that the underlying cortical area influencing the EEG can be up to 45 cm^2 [Duun-Henriksen et al. 2012], which advocates for scalp maps as the ones found by BCoCA and CoCA as opposed to the channel specific ones by CCA. Another EEG dataset stemming from an experiment with a semantic stimulus was tested with an approach similar to the one presented in Dmochowski et al. [2012], and the one employed on the data from the experiments presented in this thesis. The result was a 91 % of significantly correlated windows for IaSC and 90 % for ISC as well as a clearly shown decrease of alpha activity after stimulus, when averaging over all the epochs of the combined component.

In the analysis of the EEG recorded for this thesis the majority of the presented results stemmed from using CoCA only. This approach was chosen for comparability with the results in Dmochowski et al. [2012], and to avoid cluttering the results with duplicate figures. However these analyses were also conducted using BCoCA and a few chosen figures with comparisons to CoCA were presented. These showed that even though BCoCA could achieve higher correlations in some situations, the components it found were not as separated as the ones found with CoCA. This even though the illustrated scalp maps showed higher individuality than the one manifested in the average IaSCs and ISCs. A reason for this could lie in the fact that BCoCA estimates a full covariance matrix for the noise in each dataset, which are used in the calculation of the hidden sources. A dominating covariance matrix could explain the increased similarity in the components and warrants further investigation in this area. It should be mentioned that BCoCA performed equally well in separating the sources in the simulated data with four hidden sources, as compared to the other algorithms. Experimenting with having BCoCA output more or fewer components might change the output.

5.2 Recording and comparing EEG on multiple subjects

To establish the significance of coincidences between neural correlation and stimulus, the experiment included manipulation of the stimulus with viewings of films with scenes scrambled in time, repeated viewings and a control film. The control generally elicited a low correlation within and between subjects which can be accredited to its monotonous, to the extent of inducing sleep in a few subjects, content. The statistically significant difference in the proportion of correlation between the films and the control show that the neural correlation is not merely by chance or a product of synchronisation of the default mode network. Contrary to Dmochowski et al. [2012] the scalp projection of the control does not correspond well to those of the films which could mean that the algorithm found noisy components that were more correlated than the neural response to the films.

Viewing the films scrambled in time removes most of the contextual meaning and deflates the tension and suspension created using cinematographic and sound effects. Though statistically different from the original film the scrambled ones elicited a higher response than anticipated with almost 20% significant windows in the first component of *Bang! You're Dead*. As mentioned, this may be the result of an unfortunate random shuffling of the scenes in which contextually meaningful scenes were placed early and thus provided sufficient context to decode from other scenes that the gun is real and Sophie is about to loose a child. Furthermore, a cinematographic method to surprise an audience is the use of rapid and unexpected scene changes. By cutting up the scenes and shuffling them this effect may induce a mild shock in the viewer if they at one point are looking at a child's face and in the next down the barrel of a gun.

Provided that the measured effect in neural activation is caused by tension, suspension and "emotionally laden attention", the effect of already knowing the content of a film will be a lowered neural response. Comparing inter-subject correlations for the first and second viewing in the single group, this is exactly the result we got. However, the proportion of significant correlations for the second viewing is very low, 45% vs 9%, which may indicate issues with synchronising the second viewing across subjects.

By investigating the coincidence of scenes with peaks of large mass it was discovered that these often occur at times of arousing moments, close-ups of faces and objects, and immediately following a scene change. In *Bang! You're Dead* the discovery of the gun and bullets elicits the largest responses. This was surprising since the scenes in which the boy triggers the gun are, in the authors perspective, more intense and induces a high level of anticipation and suspense. From earlier mentioned studies it was shown that the perception of a face elicits a neural response approximately 170 ms post stimulus but this phenomenon is not restricted to faces only. The occipito-temporal cortex is activated when objects and faces are perceived and even higher activation occurs when the object is recognised [Grill-Spector 2003]. The neural correlation may thus stem from bottom-up processes of object and face perception and abrupt changes in the viewers perspective. This is supported by the scalp projections

of the most correlated components that show high neural activation in these areas and similar results reported by Hasson, Nir, et al. [2004]. The neural activation pattern for *Sophie's Choice* was consistently similar to that of *Bang! You're Dead* and from the scenes with high correlation it is evident that close-ups of faces may cause much of the neural response. However, scalp projections of the former (see figure 4.3(a) and 4.21(b)) extends farther into the parietal lobe which may indicate activation of the posterior cingulate cortex, known to be involved in emotion processing [Maddock et al. 2003]. *Sophie's Choice* contains elements of intense sadness which has been shown to be a powerful emotional stimulant activating many areas in the posterior region of the brain, including the posterior cingulate [Goldin et al. 2005]. Recall that empathy is modulated by higher order processes and potentially dialled down if the inferred state of another becomes so powerful that it threatens to confuse the perception of self and other. During the last scene of *Sophie's Choice*, when the girl is removed, a large extended period of neural correlation followed by a large drop is observed (see figure 4.21(a)). Exactly the same result was produced in a fMRI study using the same film as shown in figure 4.20 [Raz, Jacob, et al. 2013]. As previously mentioned, ES is a low-level representation of another's state linked to the MNS while ToM is a higher level cognitive representation. We know that more intensive emotional stimulus should induce a higher neural response, as figure 4.20(b), which should lead to higher correlation between the subjects, which happens just before the drop. It is thus very plausible that BCoCA in this situation has found components of neural activity that mainly originates from emotional representations and that the drop is caused by regulation of emotions. Emotional activation may become clearer by contrasting emotionally sad scenes with emotionally neutral scenes depicting faces of the same actors.

5.2.1 Comparing EEG from subject viewing films together

Experiments regarding the influences of viewing films jointly was conducted with two groups of nine subjects. Unfortunately due to synchronisation and connection issues, the data from two subjects in the first joint viewing were not usable. Unlike the results from the single viewing subjects, the results were not as clear. Even though significant differences could be seen in the number of significant windows, the manner in which the groups differed from each other was not consistent across the three components or the films seen. Though being a more qualitative comparison, the times of high correlation of the average population correlation indicated a consistent difference. The two joint viewing groups did not obtain high correlation in their average IaSC and ISC in all of the scenes for which the single viewing group obtained high correlation. Furthermore the scenes, that the two viewing groups did attain high correlations for, were not always the same scenes. In turn the joint viewing groups obtained higher values of correlation for their significant scenes compared to the *single* group. A conclusion from this could be that viewing the films in a large group makes it harder to obtain a common synchronised experience, but when it happens the synchronisation is increased. However the joint viewing experiments

consisted of only two groups, and even though these amounted to 16 usable datasets, more joint viewing groups are needed to obtain statistical significant results.

5.3 Future work

As a conclusion to the discussion of the results we will explain some of the ideas and areas for further exploration that occurred to us during the project and the review of the results, but unfortunately had to be down prioritised due to time limitations.

5.3.1 Improving BCoCA

Starting with BCoCA, there are tests for the implementation that could have been interesting and have potential for increasing both the performance and computation efficiency, the first of these being the initialisation of the algorithm.

Initialisation

In the present state of BCoCA only the weights, $\mathbf{A}$, are initialised based on the data from a zero mean Gaussian distribution with a standard deviation equal to the average standard deviation across all channels in the dataset. With more knowledge regarding the datasets the variables could be initialised closer to a likely solution. This could be done through simple analysis of the data as in the case of $\mathbf{A}$, by using the results of another algorithm such as done by Wang [2007] and Wu et al. [2011], or by using prior knowledge of the data. The prior knowledge could be specific to the subject from which the EEG is recorded or to the experiment. There was a high degree of similarity between the scalp maps attained from the experiments conducted in this thesis, and an even higher similarity in Dmochowski et al. [2012], which suggests that initialising $\mathbf{A}$ as having these values might help BCoCA find solutions with weights close to these, as well as decrease the amount of iterations required to reach them. A related area is the question of how to model the hyperparameters, $\boldsymbol{\alpha}$ and λ, and the noise covariance, $\Psi^{(-1)}$. In BCoCA they have been modelled using constant parameters, a_0, b_0, $\mathbf{S}_0$, and v_0, which were set close to zero, as seem to be the general consensus [C. Bishop 1999; Klami 2013; Wang 2007; Wu et al. 2011], but they could be modelled using inference for another layer of hyperparameters. This would make the algorithm more flexible, but also asks the question of when to stop.

Variable updating scheme for the variables

The question of computational speed could be addressed by using a relaxed expectation maximisation scheme, where the change in the variables are increased by a parameter that varies in size to avoid negative changes in the lower bound. Another possibility is to monitor the contribution of each variable to the change in lower bound, and have the ones with low change skip some of the iterations. Initial investigations hint that the noise covariance matrix have high rates of change in few

iterations reaching a plateau afterwards, though the other variables still contribute to large changes. As the update of the noise covariance involves the inversion of a $D \times D$ matrix, with D being the number of channels, only updating it when necessary could result in a significant lower computation cost.

Management of components

A noteworthy area of improvement lies in the management of components. In its present state BCoCA regularises the number of components, and their variance, through ARD, but the performance of this feature have not been thoroughly tested. Similar implementations have been tested with great success [C. Bishop 1999; Wang 2007; Wu et al. 2011; Klami 2013], but the addition of the shared $\mathbf{U}$ and the new ARD variable, λ, poses a change significant enough that further testing is warranted. Where CoCA is able to sort its components through the size of the calculated eigenvalues, BCoCA does not have this ability. A possible solution could be to sort the components by the product between the average power of the component and the average power of the corresponding weights.

5.3.2 Further analysis and expansion of EEG recording experiment

Pre-processing of EEG data can have a major effect on the results because of the typically low signal-to-noise ratio. In this thesis noise and eye-artefact reduction has been applied but a thorough investigation into other sources of noise that are less obvious may improve the results. It has come to our attention too late that the artefacts in a few subjects alter the result of the entire group why these should be removed from the dataset altogether and the processing repeated.

Synchronisation of EEG and films has proven difficult and from the large difference in significant windows between intra- and inter-subject correlations for some of the groups, it is evident that synchronisation remains a problem. Though the current method improved the results considerably it only allows correction for up to two seconds. Increasing this interval and applying a more intelligent way of determining the optimal correction is expected to alter some of the results significantly.

Further processing

As previously mentioned initialising BCoCA with the solution of another, faster, algorithm may produce better results or at least faster convergence. To initialise BCoCA it would be obvious to use the filters from CoCA or stationary filters of the ”expected” solution.

When computing the inter-subject correlations the current paradigm correlates every subject with the rest. As part of its solution BCoCA returns the inferred source $\mathbf{Z}$ which is the maximally correlated component for all of the datasets. Instead of

correlating on an inter-subject basis it would be interesting to correlate each subject with just the shared source.

Oscillation of brain waves hold a wealth of information in most aspects of EEG analysis and would thus be natural to investigate further. Desynchronisation of the alpha band has long been associated with increased attentional demand [Klimesch et al. 1998] and reduction of beta activity with tasks related to e.g. processing of external emotional stimuli [Dmochowski et al. 2012]. As proposed by the latter the instantaneous power of different frequency bands in windows of high correlation could be compared to windows of low correlation to test for significant differences in frequency suppression/synchronisation.

Negative valence and high arousal in films has shown to synchronise areas in the emotion-processing and default-mode networks [Nummenmaa et al. 2012], in line with our results, but the synchronisation can be investigated further by considering the phase of different frequency bands. To this purpose the phase-locking method has been successfully applied to discover inter- and intra-brain connectivity in a dual EEG study [Yun et al. 2012], which implies that the method could be used as an alternative way of processing our data.

Another way of expanding on the results of the joint viewing experiments would be to conduct additional experiments, with two alterations of the original setup being of particular interest. The first change would be to conduct a series of joint viewing experiments in smaller groups of twos and fives. This would enable an investigation of the significance of group size and, especially with the two subject setup, it would be possible to better control the testing environment and conduct enough tests to obtain statistical significance. A second change could be to conduct single viewing experiments in the same large room with the same setup as for the joint viewings. This test would prove if and how the viewing environment influences the results.

CHAPTER 6

Conclusion

This thesis has described and derived a Bayesian approach to CoCA, a novel signal decomposition method introduced in Dmochowski et al. [2012]. With BCoCA the method is generalised to enable comparisons between more than two subjects at the same time, and relaxes the constraint of equal weights with an adaptable parameter controlling the similarity between the weights for each dataset. This gives it applications in multiple subject experiments, with the purpose of locating neural activations that are synchronised within and between brains. The algorithm has proven its usability compared to similar methods using simulated and real EEG data. In the simulated data tests BCoCA was proven to have equal or better performance when handling multiple datasets, while the tests on real EEG showed that the algorithm shared CoCA's ability to obtain anatomically correct scalp maps.

The second part of this thesis consists of an EEG experiment with a cohort of 42 subjects who either viewed a film alone or in a group. Experiments of this kind has many variables that need to align for the experiment to be successful and throughout the entire duration of the planning and execution, new variables were discovered. The first challenge was, in order to record data, to write a new application based on an unknown framework in an unfamiliar programming language. The second challenge, and by far the most time-consuming, was to establish a method to synchronise not only one but a whole group of wireless EEG recordings with a film showing on a tablet and a camera recording the session. Though the synchronisation worked for the most part, a software based method is highly encouraged whenever possible. The last of the many challenges was of course the logistics and execution of the experiments. The Emocap was thankfully easy to equip and most of the recordings went, to our knowledge, trouble-free, but a number of technological issues will need further attention if the experiment is repeated.

A study was conducted on the neural response to a known stimuli and the correlation of this among subjects. It was discovered that neural correlation is detectable using consumer-grade hardware and that there is a significant difference between neural correlation originating from emotionally arousing and neutral films, respectively. It was shown that although object perception processes are responsible for some of the neural activity, contextual meaning and emotion are a highly significant components as well. This was further established by comparing scenes with periods of significant correlation and scalp projections of the neural activity. The latter showed higher activation in areas related to emotion for the emotionally intense *Sophie's Choice* compared to the suspenseful but otherwise emotionally indifferent *Bang! You're Dead*.

It was unfortunately not possible to determine whether the effect of experiencing an emotionally laden stimulus in a group is significantly different to experiencing it alone. We maintain the belief that there is a difference, but further processing is needed to reveal it.

This thesis has contributed to the field of neuroscience with a new algorithm, by validating important results of another study, validating the use of an affordable EEG monitor in research, conducting the, to our knowledge, largest simultaneous EEG experiment, and possibly producing new results regarding emotion regulation in social circumstances.

APPENDIX A

Worked Through Example: Variational Principal Components

This appendix contains a worked through example of Variational approximation of the Bayesian PCA proposed by C. Bishop [1999]. The notation is mostly the same as the one used in the article. For a simpler example see C. M. Bishop [2006] or Murphy [2012] for a work through of a unimodal Gaussian.

The prior (and $\mathbf{t}$'s conditional) distributions are given by

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q) \tag{A.1}$$

$$\mathbf{t} \sim \mathcal{N}(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \tau^{-1}\mathbf{I}_d) \tag{A.2}$$

$$\mathbf{W} \sim p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{q} \left(\frac{\alpha_i}{2\pi}\right)^{d/2} \exp\left\{-\frac{\alpha_i}{2}||\mathbf{w}_i||^2\right\} \tag{A.3}$$

$$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I}_d) \tag{A.4}$$

$$\tau \sim \mathcal{G}a(a_0, b_0) \tag{A.5}$$

$$\boldsymbol{\alpha} \sim \prod_{i=1}^{q} \mathcal{G}a(\alpha_i|a_0, b_0) \tag{A.6}$$

The joint probability is then given by

$$p(\mathbf{X},\mathbf{t}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}) = p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau})p(\mathbf{X}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}) \quad \Leftrightarrow \tag{A.7}$$

$$= p(\mathbf{X})p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\tau})p(\mathbf{W}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\boldsymbol{\mu})p(\boldsymbol{\tau}) \quad \Leftrightarrow \tag{A.8}$$

$$= \prod_{n=1}^{N} \left(\frac{1}{2\pi}\right)^{q/2} e^{-\frac{1}{2}||\mathbf{x}_n||^2} \prod_{n=1}^{N} \left(\frac{\tau}{2\pi}\right)^{d/2} e^{-\frac{\tau}{2}||\mathbf{t}_n - (\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu})||^2} \prod_{i=1}^{q} \left(\frac{\alpha_i}{2\pi}\right)^{d/2} e^{-\frac{\alpha_i}{2}||\mathbf{w}_i||^2}$$

$$\prod_{i=1}^{q} b_0^{a_0} \alpha_i^{a_0 - 1} e^{-b_0 \alpha_i} \frac{1}{\Gamma(a_0)} \cdot \left(\frac{\beta}{2\pi}\right)^{d/2} e^{-\frac{\beta}{2}||\boldsymbol{\mu}||^2} \cdot b_0^{a_0} \tau^{a_0 - 1} e^{-b_0 \tau} \frac{1}{\Gamma(a_0)} \tag{A.9}$$

Posterior for $q(\mathbf{X})$

The logarithm of the distribution for $\mathbf{X}$ is approximated by

$$\ln q(\mathbf{X}) = \mathbb{E}_{/\mathbf{X}} \left[\ln p(\mathbf{t}_n|\mathbf{W}, \mathbf{x}_n, \boldsymbol{\mu}, \tau) + \ln p(\mathbf{X})\right] + C \quad \Leftrightarrow \tag{A.10}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{/\mathbf{X}} \left[\frac{d}{2}\left(\ln \tau - \ln 2\pi\right) - \frac{\tau}{2}||\mathbf{t}_n - (\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu})||^2 - \frac{q}{2}\ln 2\pi - \frac{1}{2}||\mathbf{x}_n||^2\right] + C \quad \Leftrightarrow \tag{A.11}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{/\mathbf{X}} \left[-\frac{\tau}{2}\left((\mathbf{W}\mathbf{x}_n)^T\mathbf{W}\mathbf{x}_n + ||\mathbf{t}_n - \boldsymbol{\mu}||^2 - 2(\mathbf{W}\mathbf{x}_n)^T(\mathbf{t}_n - \boldsymbol{\mu})\right) - \frac{1}{2}||\mathbf{x}_n||^2\right] + C, \tag{A.12}$$

where the parts of the equation which are independent of $\mathbf{x}_n$ are absorbed into the constant, C. The expectation is taken with respect to all variables with a defined prior distribution, except for $\mathbf{x}_n$. Elements in the equation which are constant with regard to these are left out of the expectation. Since $\mathbf{x}_n$ has a gaussian as its prior distribution, the goal is to arrange the elements of the equation to resemble a log gaussian. Then by "completing the square" [C. M. Bishop 2006, p.86] expressions for a new mean and variance can be found.

$$\ln q(\mathbf{X}) = \sum_{n=1}^{N} -\frac{1}{2}||\mathbf{x}_n||^2 - \frac{1}{2}\langle\tau\rangle\,\mathbf{x}_n^T\langle\mathbf{W}^T\mathbf{W}\rangle\mathbf{x}_n + \mathbf{x}_n^T\langle\tau\rangle\left\langle\mathbf{W}^T\right\rangle(\mathbf{t}_n - \langle\boldsymbol{\mu}\rangle) + C \quad \Leftrightarrow \tag{A.13}$$

$$= \sum_{n=1}^{N} -\frac{1}{2}\mathbf{x}_n^T\left(\mathbf{I} + \langle\tau\rangle\langle\mathbf{W}^T\mathbf{W}\rangle\right)\mathbf{x}_n + \mathbf{x}_n^T\langle\tau\rangle\left\langle\mathbf{W}^T\right\rangle(\mathbf{t}_n - \langle\boldsymbol{\mu}\rangle) + C \tag{A.14}$$

$$q(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{m}_{\mathbf{x},n}, \Sigma_{\mathbf{x}}) \tag{A.15}$$

$$\Sigma_{\mathbf{x}}^{-1} = \mathbf{I} + \langle\tau\rangle\langle\mathbf{W}^T\mathbf{W}\rangle \tag{A.16}$$

$$\Sigma_{\mathbf{x}}^{-1}\mathbf{m}_{\mathbf{x},n} = \langle\tau\rangle\left\langle\mathbf{W}^T\right\rangle(\mathbf{t}_n - \langle\boldsymbol{\mu}\rangle) \quad \Leftrightarrow \tag{A.17}$$

$$\mathbf{m}_{\mathbf{x},n} = \Sigma_{\mathbf{x}}\langle\tau\rangle\left\langle\mathbf{W}^T\right\rangle(\mathbf{t}_n - \langle\boldsymbol{\mu}\rangle) \tag{A.18}$$

where $\langle.\rangle$ signifies the expectation.

Posterior for $q(\boldsymbol{\alpha})$

The logarithm of the distribution for $\boldsymbol{\alpha}$ is approximated in the same manner as for $\mathbf{X}$

$$\begin{aligned}
\ln q(\boldsymbol{\alpha}) &= \mathbb{E}_{/\boldsymbol{\alpha}}\left[\ln p(\mathbf{W}|\boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha})\right] + \text{const.} \quad \Leftrightarrow && \text{(A.19)}\\
&= \mathbb{E}_{/\boldsymbol{\alpha}}\left[\ln p(\mathbf{W}|\boldsymbol{\alpha})\right] + \ln p(\boldsymbol{\alpha}) + \text{const.} \quad \Leftrightarrow && \text{(A.20)}\\
&= \sum_{i=1}^{q} \mathbb{E}_{/\boldsymbol{\alpha}}\left[\frac{d}{2}\left(\ln \alpha_i - \ln 2\pi\right) - \frac{\alpha_i}{2}||\mathbf{w}_i||^2\right] + a_0 \ln b_0 + (a_0 - 1)\ln \alpha_i \\
&\qquad - b_0\alpha_i - \ln \Gamma(a_0) + \text{const.} \quad \Leftrightarrow && \text{(A.21)}\\
&= \sum_{i=1}^{q} \frac{d}{2}\ln \alpha_i - \frac{\alpha_i}{2}\left\langle \mathbf{w}_i^T\mathbf{w}_i\right\rangle + (a_0 - 1)\ln \alpha_i - b_0\alpha_i + \text{const.} \quad \Rightarrow && \text{(A.22)}
\end{aligned}$$

$$\begin{aligned}
q(\boldsymbol{\alpha}) &= \prod_{i=1}^{q} \mathcal{G}a(\alpha_i|a_\alpha, b_{\alpha,i}) && \text{(A.23)}\\
a_\alpha - 1 &= \frac{d}{2} + a_0 - 1 \quad \Leftrightarrow && \text{(A.24)}\\
a_\alpha &= a_0 + \frac{d}{2} && \text{(A.25)}\\
b_\alpha &= \frac{1}{2}\left\langle \mathbf{w}_i^T\mathbf{w}_i\right\rangle + b_0 && \text{(A.26)}
\end{aligned}$$

Posterior for $q(\mathbf{W})$

The logarithm of the distribution for $\mathbf{W}$ is approximated in the same manner as for $\mathbf{X}$

$$\begin{aligned}
\ln q(\mathbf{W}) &= \mathbb{E}_{/\mathbf{w}}\left[\ln p(\mathbf{t}_n|\mathbf{W}, \mathbf{x}_n, \boldsymbol{\mu}, \tau) + \ln p(\mathbf{W})\right] + \text{const.} \quad \Leftrightarrow && \text{(A.27)}\\
&= \mathbb{E}_{/\mathbf{w}}\left[\sum_{n=1}^{N} \frac{d}{2}\left(\ln \tau - \ln 2\pi\right) - \frac{\tau}{2}||\mathbf{t}_n - (\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu})||^2 \right.\\
&\qquad \left. + \sum_{i=1}^{q} \frac{d}{2}\left(\ln \alpha_i - \ln 2\pi\right) - \frac{\alpha_i}{2}||\mathbf{w}_i||^2\right] + C \quad \Leftrightarrow && \text{(A.28)}\\
&= \sum_{k=1}^{d} \mathbb{E}_{/\mathbf{w}}\left[\sum_{n=1}^{N} -\frac{\tau}{2}\left((\mathbf{w}_k\mathbf{x}_n)^T(\mathbf{w}_k\mathbf{x}_n) - 2(\mathbf{w}_k\mathbf{x}_n)(\mathbf{t}_{n,k} - \boldsymbol{\mu}_k)\right)\right.\\
&\qquad \left. -\frac{1}{2}\mathbf{w}_k \text{diag}(\boldsymbol{\alpha})\mathbf{w}_k^T\right] + C && \text{(A.29)}
\end{aligned}$$

Note that $\mathbf{w}_k$ is a row vector. $(\mathbf{w}_k\mathbf{x}_n)^T$ and $(\mathbf{w}_k\mathbf{x}_n)$ are therefore scalars, and their order can be changed freely.

$$\ln q(\mathbf{W}) = \sum_{k=1}^{d}\sum_{n=1}^{N} -\frac{1}{2}\left(\mathbf{w}_k \langle\tau\rangle \left\langle \mathbf{x}_n\mathbf{x}_n^T\right\rangle \mathbf{w}_k^T + \mathbf{w}_k \mathrm{diag}(\langle\boldsymbol{\alpha}\rangle)\mathbf{w}_k^T\right) + \mathbf{w}_k \langle\tau\rangle \langle\mathbf{x}_n\rangle (\mathbf{t}_{n,k} - \langle\boldsymbol{\mu}_k\rangle) + C \Rightarrow \tag{A.30}$$

$$q(\mathbf{W}) = \prod_{k=1}^{d} \mathcal{N}(\hat{\mathbf{w}}_k|\mathbf{m}_{\mathbf{w},k}, \Sigma_{\mathbf{w}}) \tag{A.31}$$

$$\Sigma_{\mathbf{w}}^{-1} = \langle\tau\rangle \sum_{n=1}^{N} \left\langle \mathbf{x}_n\mathbf{x}_n^T\right\rangle + \mathrm{diag}(\langle\boldsymbol{\alpha}\rangle) \tag{A.32}$$

$$\Sigma_{\mathbf{w}}^{-1}\mathbf{m}_{\mathbf{w},k} = \langle\tau\rangle \sum_{n=1}^{N} \langle\mathbf{x}_n\rangle (\mathbf{t}_{n,k} - \langle\boldsymbol{\mu}_k\rangle) \quad \Leftrightarrow \tag{A.33}$$

$$\mathbf{m}_{\mathbf{w},k} = \Sigma_{\mathbf{w}} \langle\tau\rangle \sum_{n=1}^{N} \langle\mathbf{x}_n\rangle (\mathbf{t}_{n,k} - \langle\boldsymbol{\mu}_k\rangle) \tag{A.34}$$

Posterior for $q(\boldsymbol{\mu})$

The logarithm of the distribution for $\boldsymbol{\mu}$ is approximated in the same manner as for $\mathbf{X}$

$$\ln q(\boldsymbol{\mu}) = \mathbb{E}_{/\boldsymbol{\mu}}[\ln p(\mathbf{t}_n|\mathbf{W}, \mathbf{x}_n, \boldsymbol{\mu}, \tau) + \ln p(\boldsymbol{\mu})] + \text{const.} \quad \Leftrightarrow \tag{A.35}$$

$$= \mathbb{E}_{/\boldsymbol{\mu}}[\ln p(\mathbf{t}_n|\mathbf{W}, \mathbf{x}_n, \boldsymbol{\mu}, \tau)] + \ln p(\boldsymbol{\mu}) + \text{const.} \quad \Leftrightarrow \tag{A.36}$$

$$= \mathbb{E}_{/\boldsymbol{\mu}}\left[\sum_{n=1}^{N} -\frac{\tau}{2}||\mathbf{t}_n - (\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu})||^2\right] - \frac{\beta}{2}||\boldsymbol{\mu}||^2 + C \quad \Leftrightarrow \tag{A.37}$$

$$= \mathbb{E}_{/\boldsymbol{\mu}}\left[\sum_{n=1}^{N} -\frac{\tau}{2}(||\boldsymbol{\mu}||^2 - \boldsymbol{\mu}^T(\mathbf{t}_n - \mathbf{W}\mathbf{x}_n)\right] - \frac{\beta}{2}||\boldsymbol{\mu}||^2 + C \quad \Leftrightarrow \tag{A.38}$$

$$= -\frac{1}{2}||\boldsymbol{\mu}||^2(N\langle\tau\rangle + \beta)\mathbf{I} - \boldsymbol{\mu}^T\langle\tau\rangle \sum_{n=1}^{N}(\mathbf{t}_n - \langle\mathbf{W}\rangle\langle\mathbf{x}_n\rangle) + C \quad \Rightarrow \tag{A.39}$$

$$q(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_{\boldsymbol{\mu}}, \Sigma_{\boldsymbol{\mu}}) \tag{A.40}$$

$$\Sigma_{\boldsymbol{\mu}}^{-1} = (N\langle\tau\rangle + \beta)\mathbf{I} \tag{A.41}$$

$$\mathbf{m}_{\boldsymbol{\mu}} = \Sigma_{\boldsymbol{\mu}}\langle\tau\rangle \sum_{n=1}^{N}(\mathbf{t}_n - \langle\mathbf{W}\rangle\langle\mathbf{x}_n\rangle) \tag{A.42}$$

Posterior for $q(\tau)$

The logarithm of the distribution for τ is approximated in the same manner as for $\mathbf{X}$

$$\ln\, q(\tau) = \mathbb{E}_{/\tau}\left[\ln\, p(\mathbf{t}_n|\mathbf{W},\mathbf{x}_n,\boldsymbol{\mu},\tau) + \ln\, p(\tau)\right] + \text{const.} \quad \Leftrightarrow \tag{A.43}$$

$$= \mathbb{E}_{/\tau}\left[\ln\, p(\mathbf{t}_n|\mathbf{W},\mathbf{x}_n,\boldsymbol{\mu},\tau)\right] + \ln\, p(\tau) + \text{const.} \quad \Leftrightarrow \tag{A.44}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{/\tau}\left[\frac{d}{2}\,\ln\,\tau - \frac{\tau}{2}||\mathbf{t}_n - (\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu})||^2\right] + (a_0 - 1)\ln\tau - b_0\tau + C \quad \Leftrightarrow \tag{A.45}$$

$$= \ln\,\tau\,\left(\frac{Nd}{2} + a_0 - 1\right) - \tau\frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_{/\tau}\left[||\mathbf{t}_n - (\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu})||^2\right] - \tau b_0 + C \quad \Leftrightarrow \tag{A.46}$$

$$= \ln\,\tau\,\left(\frac{Nd}{2} + a_0 - 1\right) - \tau\frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_{/\tau}\left[(\mathbf{W}\mathbf{x}_n)^T(\mathbf{W}\mathbf{x}_n) + ||\mathbf{t}_n - \boldsymbol{\mu}||^2 - 2(\mathbf{W}\mathbf{x}_n)(\mathbf{t}_n - \boldsymbol{\mu})\right] - \tau b_0 + C. \tag{A.47}$$

Using [Petersen et al. 2006, (16-17)] it can be shown that

$$(\mathbf{W}\mathbf{x}_n)^T(\mathbf{W}\mathbf{x}_n) = \text{Tr}(\mathbf{W}^T\mathbf{W}\mathbf{x}_n\mathbf{x}_n^T) \tag{A.48}$$

$$\ln\, q(\tau) = \ln\,\tau\,\left(\frac{Nd}{2} + a_0 - 1\right) - \tau\Bigg(b_0 + \frac{1}{2}\sum_{n=1}^{N}\text{Tr}(\left\langle\mathbf{W}^T\mathbf{W}\right\rangle\left\langle\mathbf{x}_n\mathbf{x}_n^T\right\rangle) + ||\mathbf{t}_n||^2 + \left\langle||\boldsymbol{\mu}||^2\right\rangle + 2\left(\left\langle\boldsymbol{\mu}^T\right\rangle\left\langle\mathbf{W}\right\rangle\left\langle\mathbf{x}_n\right\rangle - \mathbf{t}_n^T\left\langle\mathbf{W}\right\rangle\left\langle\mathbf{x}_n\right\rangle - \mathbf{t}_n^T\left\langle\boldsymbol{\mu}\right\rangle\right)\Bigg) + C \tag{A.49}$$

The distribution can then be approximated by

$$q(\tau) = \mathcal{G}a(\tau|a_\tau, b_\tau) \tag{A.50}$$

$$a_\tau = \frac{Nd}{2} + a_0 \tag{A.51}$$

$$b_\alpha = b_0 + \frac{1}{2}\sum_{n=1}^{N}\text{Tr}(\left\langle\mathbf{W}^T\mathbf{W}\right\rangle\left\langle\mathbf{x}_n\mathbf{x}_n^T\right\rangle) + ||\mathbf{t}_n||^2 + \left\langle||\boldsymbol{\mu}||^2\right\rangle + 2\left(\left\langle\boldsymbol{\mu}^T\right\rangle\left\langle\mathbf{W}\right\rangle\left\langle\mathbf{x}_n\right\rangle - \mathbf{t}_n^T\left\langle\mathbf{W}\right\rangle\left\langle\mathbf{x}_n\right\rangle - \mathbf{t}_n^T\left\langle\boldsymbol{\mu}\right\rangle\right) \tag{A.52}$$

APPENDIX B

Performance on Simulated Data

B.1 Varying similarity between true weights

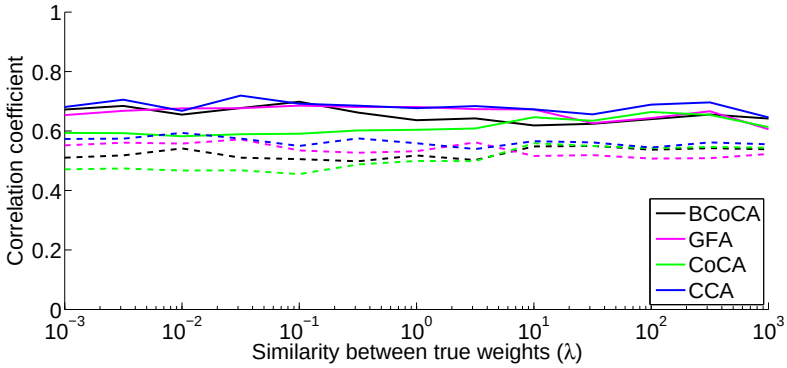


Figure B.1: $M = 2$, $SNR = \{-6, 0\}$, K $= 4$

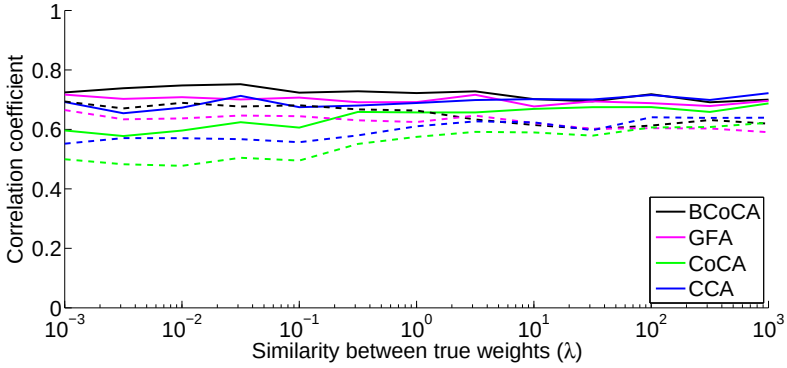


Figure B.2: $M = 5$, $SNR = \{-6, 0\}$ K $= 4$

B.2 Varying SNR

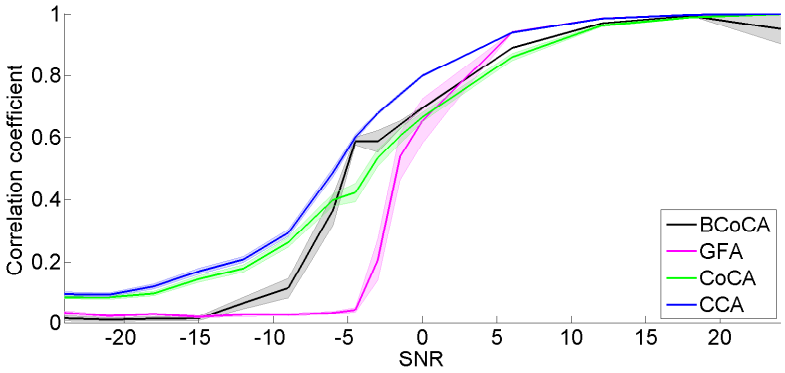


Figure B.3: $M = 2$, $\lambda = 0.001$

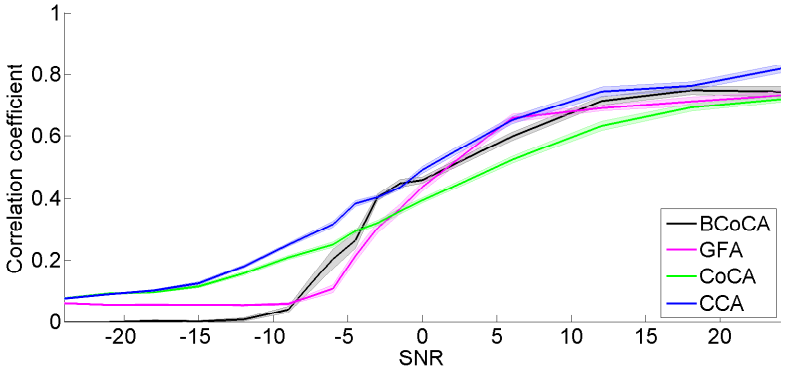


Figure B.4: $M = 2$, $\lambda = 0.001$, K $= 4$

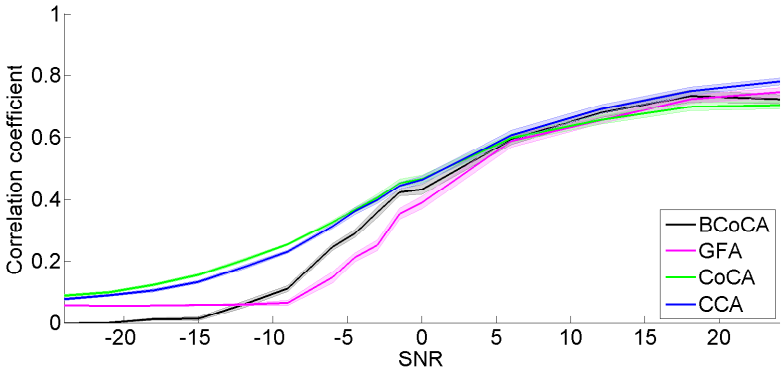


Figure B.5: $M = 2$, $\lambda = 1000$, $\mathrm{K} = 4$

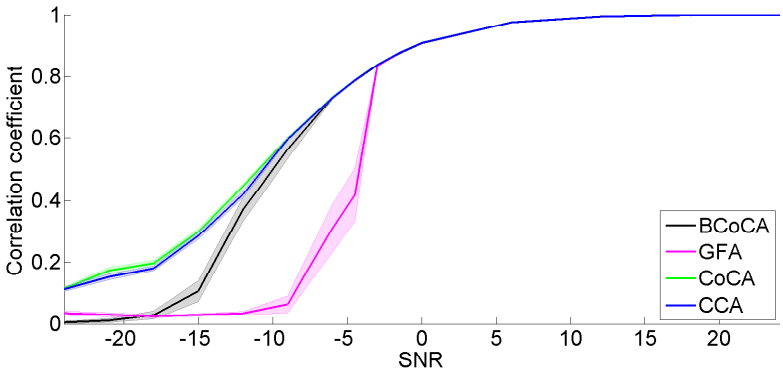


Figure B.6: $M = 5$, $\lambda = 1000$, $\mathrm{K} = 1$

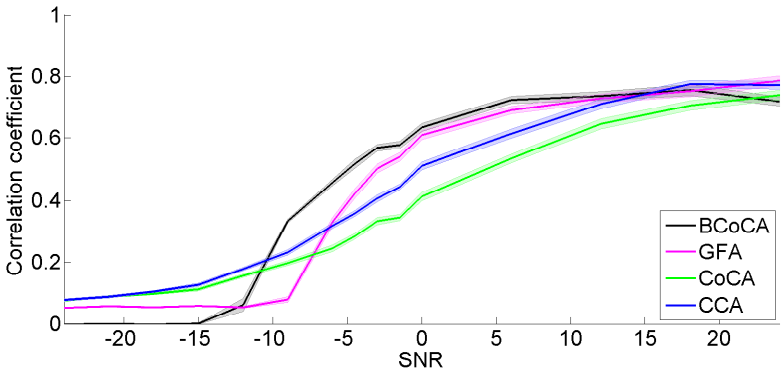


Figure B.7: $M = 5$, $\lambda = 0.001$, $\mathrm{K} = 4$

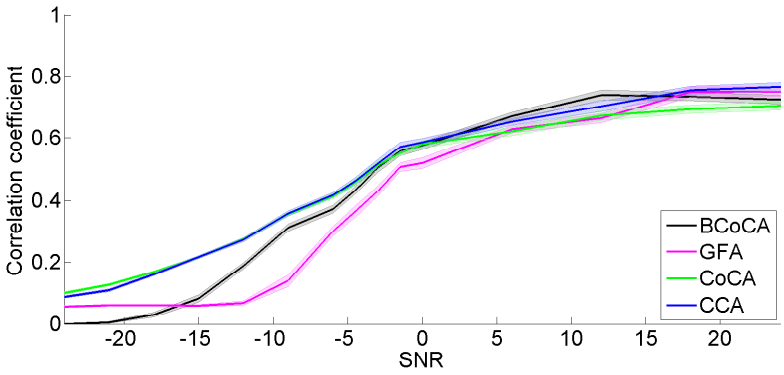


Figure B.8: $M = 5$, $\lambda = 1000$, $\mathrm{K} = 4$

B.3 Varying number of datasets

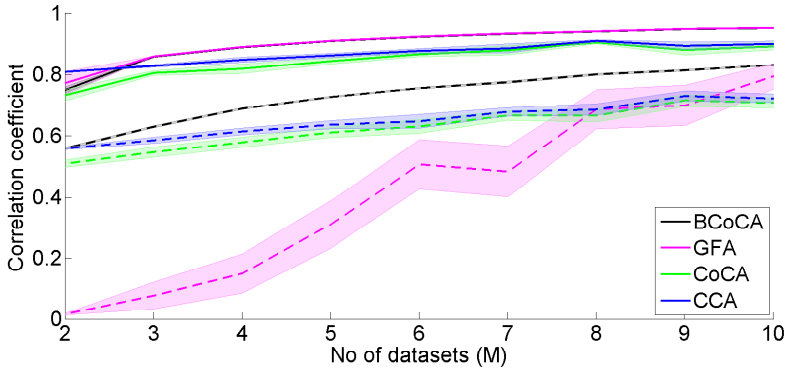


Figure B.9: $\lambda = 1$, K $= 1$

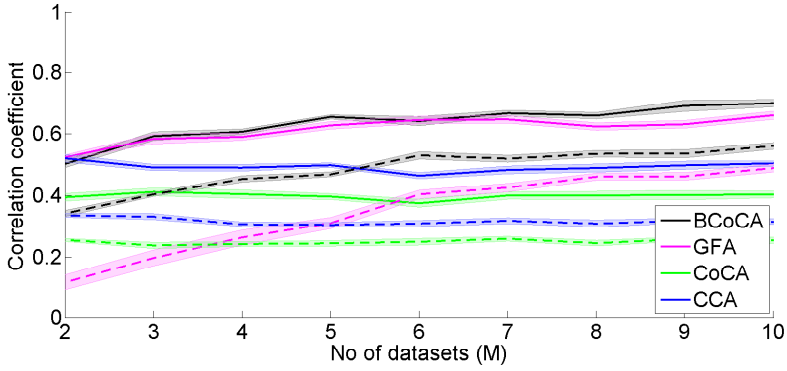


Figure B.10: $\lambda = 0.001$, K $= 4$

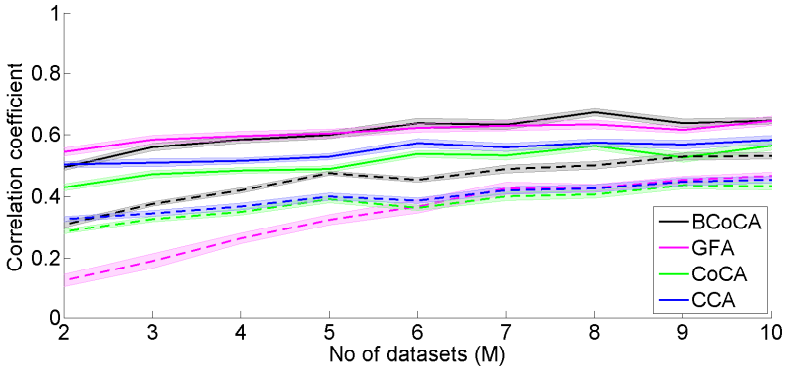


Figure B.11: $\lambda = 1$, K $= 4$

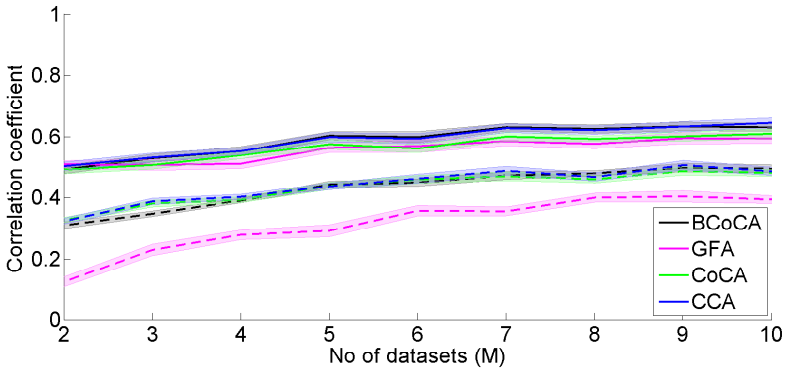


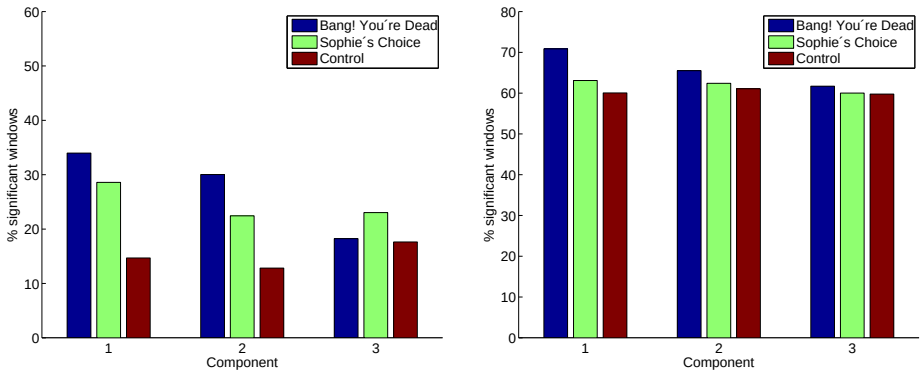
Figure B.12: $\lambda = 1000$, K $= 4$

APPENDIX C

Additional Results

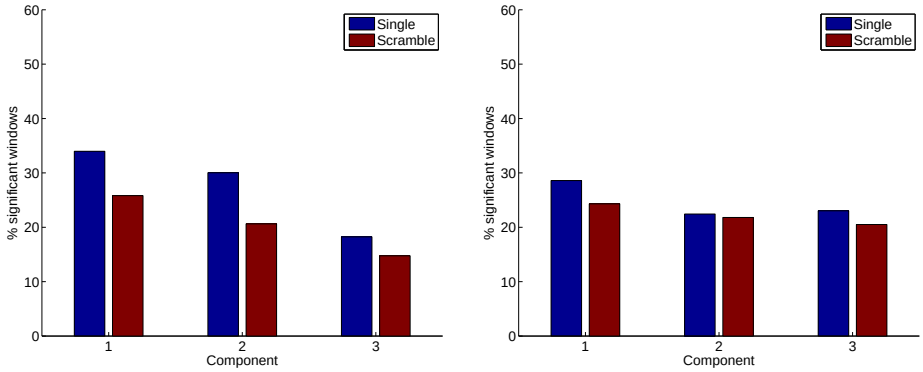
This appendix shows results omitted from chapter 4.

C.1 Single viewing significance



(a) Intra Subject **(b)** Inter Subject

Figure C.1: Single viewing significance controlled for multiple comparisons using FDR



(a) *Bang! You're Dead* **(b)** *Sophie's Choice*

Figure C.2: Comparison of significant correlations in the Single viewing IaSC between the original film and the scrambled version of *Bang! You're Dead* and *Sophie's Choice*. Controlled for multiple comparisons using FDR

C.2 Intra subject analysis

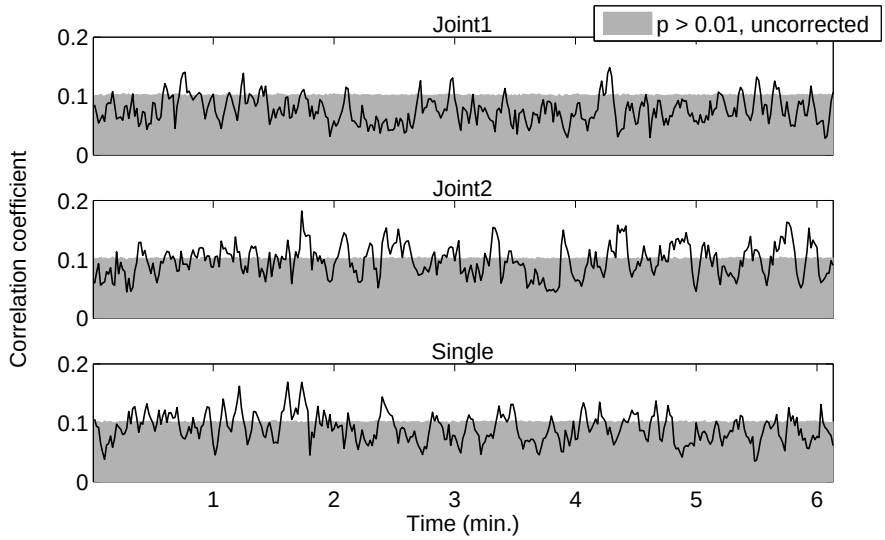


Figure C.3: Population IaSCs for the second CoCA component for the viewing of *Bang! You're Dead*.

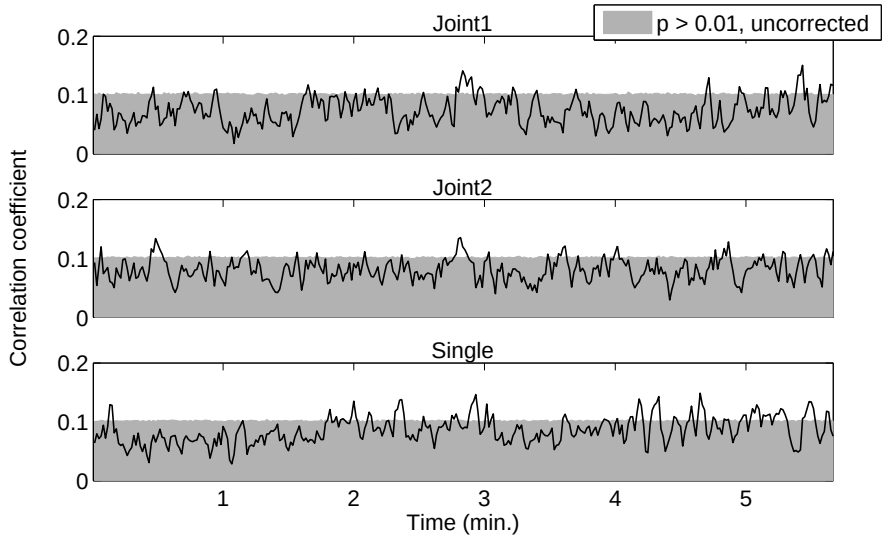


Figure C.4: Population IaSCs for the second CoCA component for the viewing of *Sophie's Choice*.

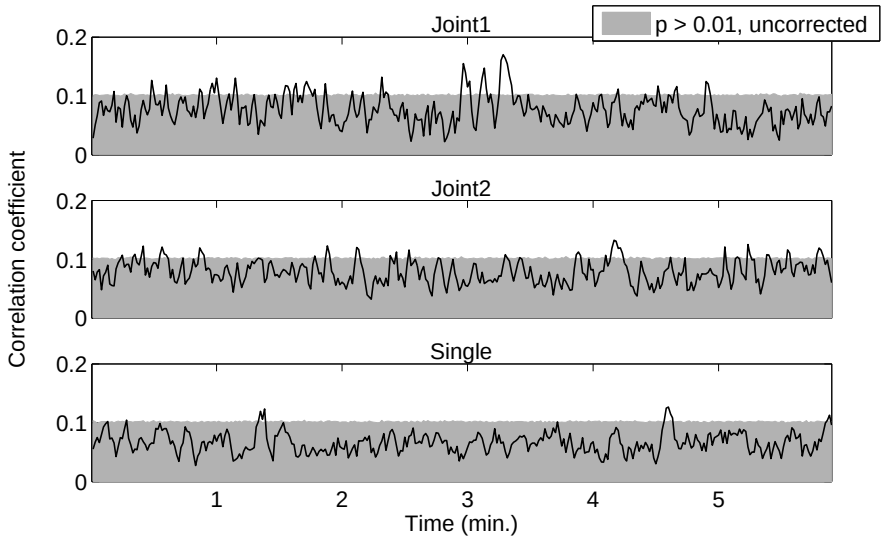


Figure C.5: Population IaSCs for the second CoCA component for the viewing of the control video.

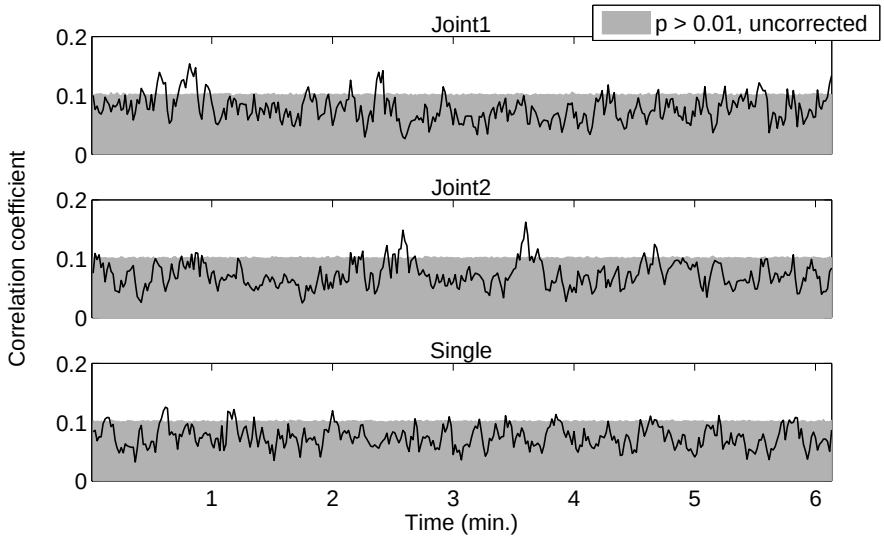


Figure C.6: Population IaSCs for the third CoCA component for the viewing of *Bang! You're Dead*.

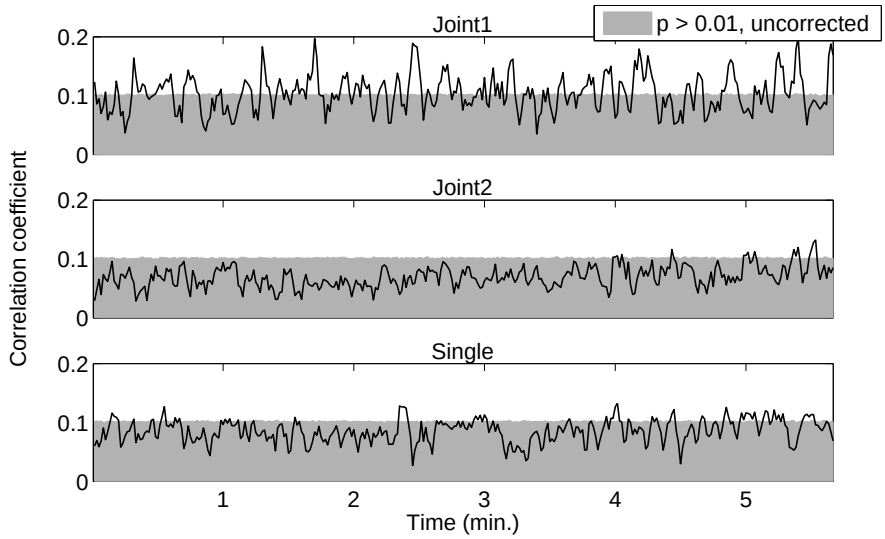


Figure C.7: Population IaSCs for the third CoCA component for the viewing of *Sophie's Choice*. Note that for joint 1 the component has been switched with what was estimated as being component 1, since this fitted the scalp maps in Dmochowski et al. [2012] better.

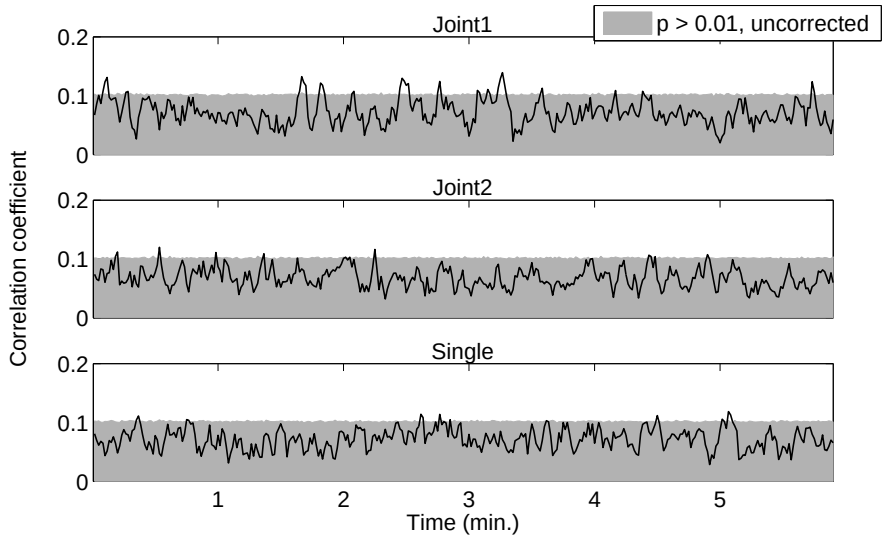


Figure C.8: Population IaSCs for the third CoCA component for the viewing of the control video.

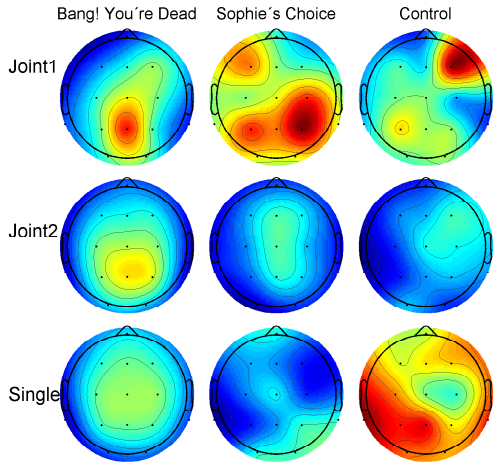


Figure C.9: Intra subject scalp projections for the second component recorded for all three films and the *joint 1*, *joint 2* and *single* viewing groups.

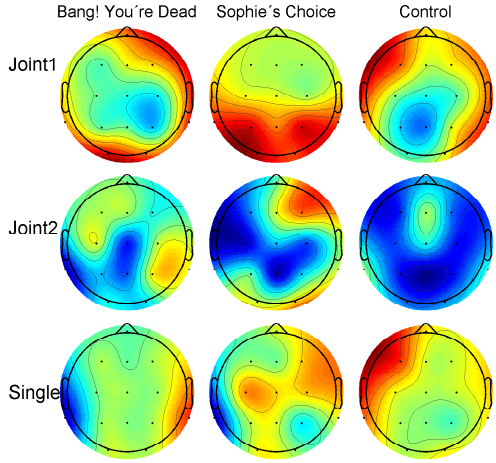
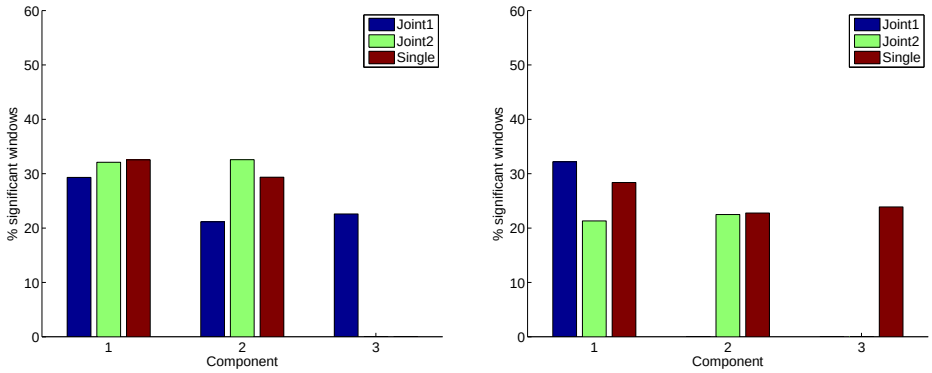
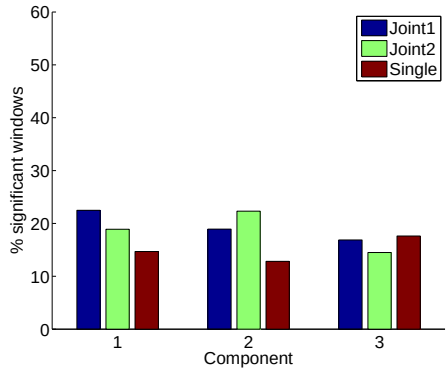


Figure C.10: Intra subject scalp projections for the third component recorded for all three films and the *joint 1*, *joint 2* and *single* viewing groups.



(a) *Bang! You're Dead*

(b) *Sophie's Choice*



(c) *Control*

Figure C.11: Comparison of significant correlations for the IaSCs for different groups of subjects. The level of significance have been controlled for multiple comparisons using FDR.

C.3 Inter subject analysis

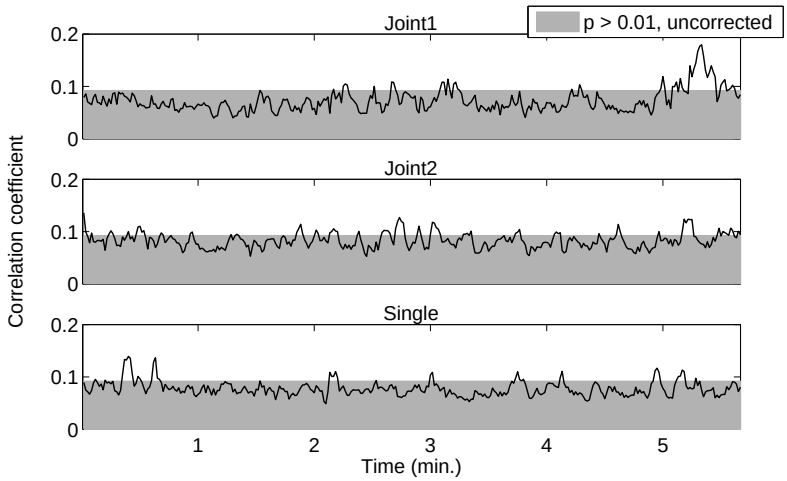


Figure C.12: Population ISCs for the first CoCA component for the viewing of *Sophie's Choice* in the first and second joint group as well as the single viewings.

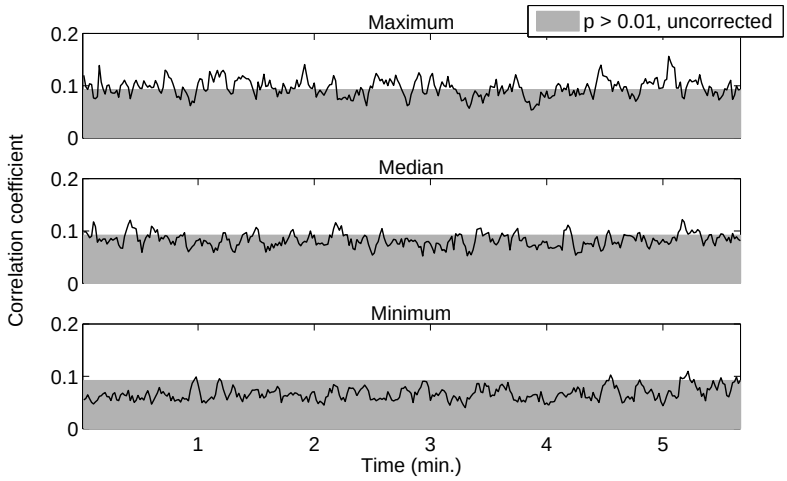
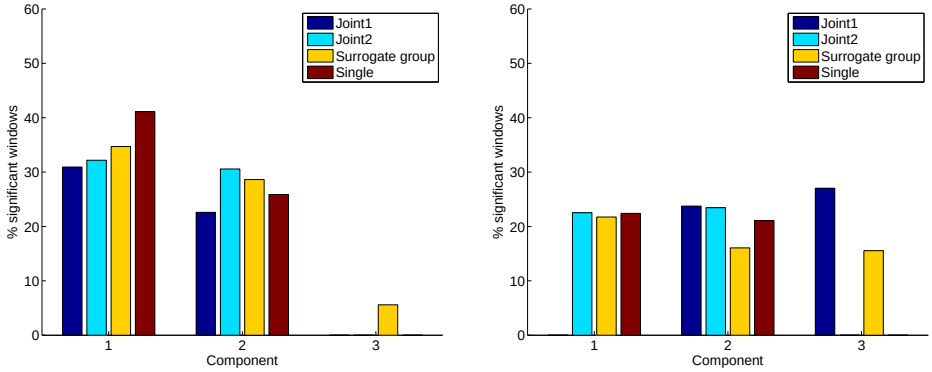
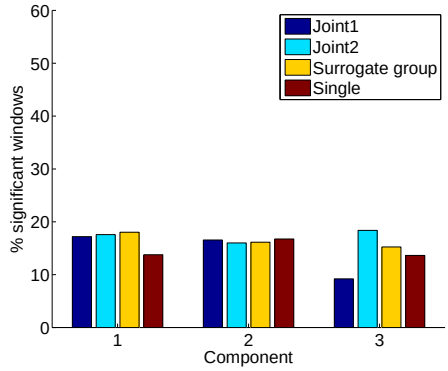


Figure C.13: Population ISCs for the first CoCA component for the viewing of *Sophie's Choice*. Each ISC is stemming from a surrogate group of eight subjects picked at random from *joint 1* and *joint 2*. The three ISCs seen in this figure attained the maximum, median and minimum number of significant windows.



(a) *Bang! You're Dead*

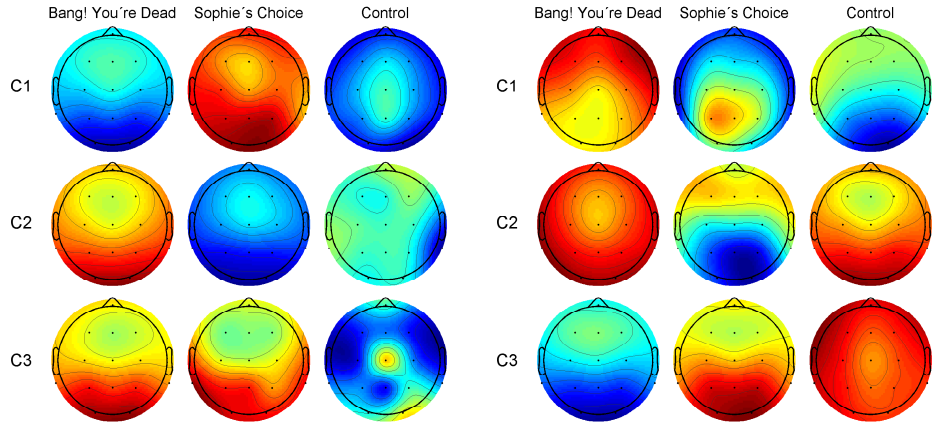
(b) *Sophie's Choice*



(c) *Control*

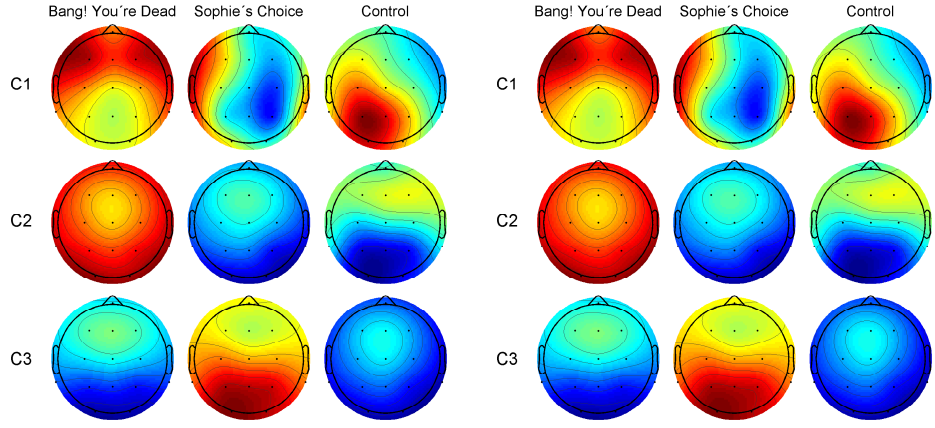
Figure C.14: Comparison of significant correlations for the ISCs for different groups of subjects. The level of significance have been controlled for multiple comparisons using FDR. For the surrogate group then mean number of significant windows across subject is shown.

C.4 Scalp projections from BCoCA



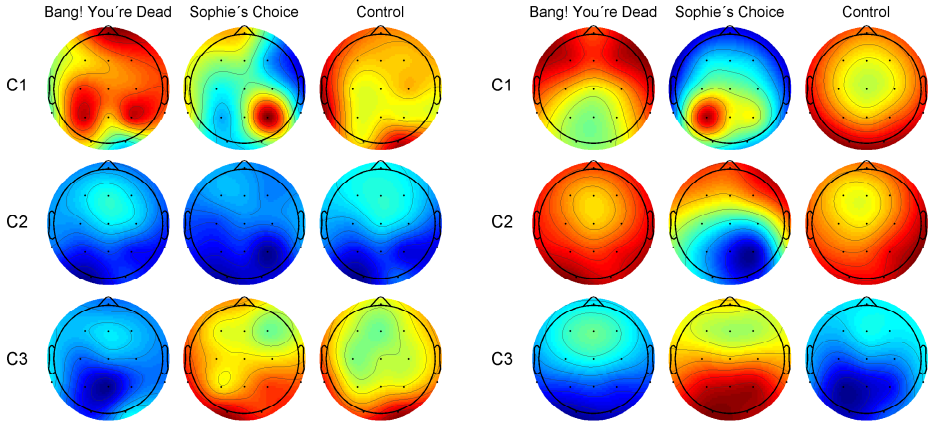
(a) Inter subject **(b)** Intra subject

Figure C.15: Single viewing



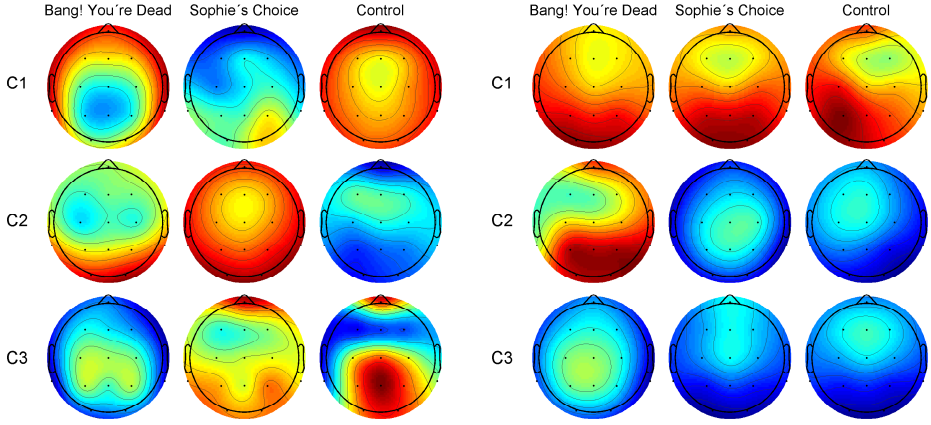
(a) Inter subject **(b)** Intra subject

Figure C.16: Joint viewing



(a) Inter subject **(b)** Intra subject

Figure C.17: Joint 1



(a) Inter subject **(b)** Intra subject

Figure C.18: Joint 2

APPENDIX D

Information Regarding Experimental Setup

This appendix contains relevant information regarding the recording of EEG for this thesis. It contains the questionnaires the subjects where presented with before and after the EEG recording. As most subjects were Danes, they were presented with a Danish version. Table D.2 show subject information regarding under which conditions the films were seen and how they were perceived, as well as biometric information. The experimental log (in danish) is also included.

Table D.1: Information regarding which combinations of receiver experienced crosstalk. **Y** = Crosstalk experienced, n = no crosstalk. The crosstalk experienced using receiver pairs 2 and 9, 7 and 11, were only experienced using certain transmitters and could therefore be circumvented be using specific transmitters to certain receivers.

Receiver	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	—	—	—	—	—	—	—	—	—	—	—
2	**Y**	—	—	—	—	—	—	—	—	—	—
3	n	n	—	—	—	—	—	—	—	—	—
4	n	n	n	—	—	—	—	—	—	—	—
5	n	n	n	n	—	—	—	—	—	—	—
6	n	n	n	n	n	—	—	—	—	—	—
7	n	n	n	n	n	n	—	—	—	—	—
8	n	n	n	n	n	n	n	—	—	—	—
9	n	**Y**	n	n	n	n	n	n	—	—	—
10	n	n	n	n	**Y**	n	n	n	n	—	—
11	n	n	n	n	n	n	**Y**	n	n	n	—

Table D.2: Information regarding the subjects, under which condition they saw the movies, and how they percieved them. **S** = Scrambled scenes, **NS** = Non-scrambled scenes, **J1** = The first joint viewing, **J2** = The second joint viewing.

Subject no.	Condition	Order of movies	Age	Hours of sleep	German proficiency	Right handed	Seen the movies before	Understood the movies
16	S	1	20	7	3	Yes	No	Yes
10	NS	2	21	9	3	No	No	Yes
9	S	1	19	9	2	Yes	No	Yes
15	S	1	22	6	2	Yes	No	Yes
3	NS	2	20	7,5	2	Yes	No	Yes
6	NS	1	20	7,5	3	Yes	No	Yes
11	S	1	18	10	2	Yes	No	Yes
2	NS	5	20	9	3	Yes	No	Yes
4	NS	6	21	7	2	Yes	No	Yes
5	NS	1	21	9	2	Yes	No	Yes
14	S	1	20	9	2	Yes	No	Yes
12	S	1	21	8,5	2	No	No	Yes
13	S	2	20	8,5	1	Yes	No	Yes
7	NS	3	21	7	2	Yes	No	Yes
1	NS	2	19	8	1	Yes	No	Not Sophie's
8	S	2	19	8	1	Yes	No	Not Bang!
23	NS	4	24	7	2	Yes	No	Yes
22	S	1	25	8	2	Yes	No	Not Bang!
24	NS	3	25	8	2	Yes	No	Yes
21	S	1	25	5	2	Yes	No	Yes
17	S	1	21	7	1	Yes	No	Yes
19	S	1	21	7	2	No	No	Yes
18	NS	6	22	7,5	1	No	No	Yes
20	NS	6	23	8	2	Yes	No	Yes
33	J1	6	32	7	3	Yes	No	Yes
28	J1	6	25	8	2	Yes	No	Yes
30	J1	6	24	8	2	Yes	No	Yes
31	J1	6	26	8	1	Yes	No	Yes
27	J1	6	25	5,5	2	Yes	No	Yes
25	J1	6	21	8	2	Yes	No	Yes
29	J1	6	25	10	2	Yes	No	Maybe
26	J1	6	21	8	1	Yes	No	Yes
32	J1	6	25	7	2	Yes	No	Yes
34	J2	6	22	10	2	Yes	No	Yes
36	J2	6	20	9	1	Yes	Sophie's	Yes
35	J2	6	25	8,5	2	Yes	No	Yes
41	J2	6	25	8	2	Yes	No	Yes
37	J2	6	25	9	2	Yes	No	Yes
42	J2	6	22	6	3	Yes	No	Yes
40	J2	6	22	6,5	3	No	No	Yes
39	J2	6	24	8	2	Yes	Sophie's	Yes
38	J2	6	23	8	1	Yes	No	Yes

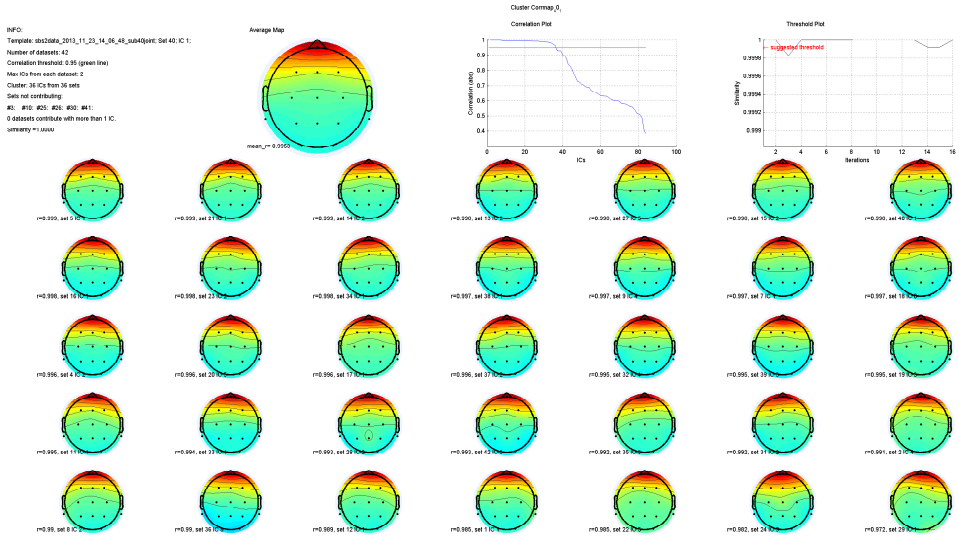


Figure D.3: CORRMAP correlation of ICs 1 - 35

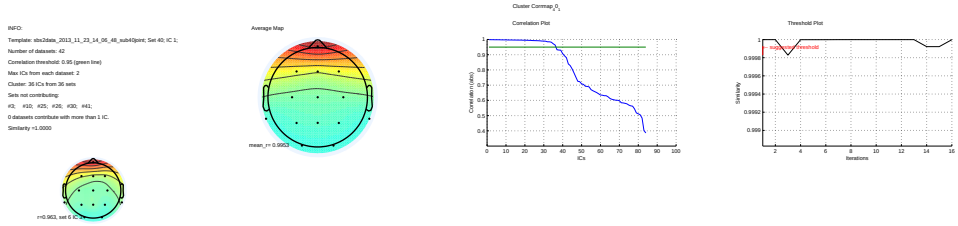


Figure D.4: CORRMAP correlation of IC 36

Questionnaire before EEG-measurement

Name: _____ **subject no.:** _____

Movie: _____

1. Are you right handed? Yes ☐ No ☐
2. Normal sight/corrected to normal vision Yes ☐ No ☐
3. Normal hearing: Yes ☐ No ☐
4. How many hours have you slept last night? _____
5. Age _____
6. Do you have a psychiatric record? Yes ☐ No ☐
7. Do you have a neurologic record? Yes ☐ No ☐
8. Have you ingested drugs or medication the last 24 hours? Yes ☐ No ☐
   a. If yes, which? _____
9. Level of German proficiency
   a. Fluent ☐
   b. Good understanding of the language ☐
   c. Basic understanding of the language ☐
   d. None ☐
10. Are you interested in participating in future experiments? If you answer ”Yes”, your sex, age and contact information will be saved for future use. Yes ☐ No ☐
11. Can we use pictures of the experimental setup, where you appear for our Master thesis, article or other things that regard this experiment? Yes ☐ No ☐

Mobiles and other electric equipment have to be removed before the experiment.

I hearby confirm, that I agree to participate in a experiment with EEG recordings during viewing films. I am informed that I participate voluntarily and that I can, at any time and without reasons, can redraw my consent to participate.

Date : _____ Signature : _____

Questionnaire after EEG-measurement

Name:		**Subject no.:**	

1. Had you seen the movies before
   a. The black/white movie Yes ☐ No ☐
   b. The movie in colour Yes ☐ No ☐
2. Did you understand the movie in german? (the one in colour) Yes ☐ No ☐

3. Which scenes made the strongest impression in the black/white movie?

_____

_____

_____

_____

4. Which scenes made the strongest impression in the movie in colour?

_____

_____

_____

_____

07.11.13

Forspøgslog

12.10. No. 1, Navn: Malene

- Andreas monterer Emocap.
- FP hæfter seddel om Scrambled / non-scrambled : N
- Skår med kørsel for valg af film rækkefølge: 2
- FP udfylder spørgeskema og seddel til udbetaling
- Andreas renser og udfylder elektrode hullerne med gel
- FP forklares at hun skal ind i et andet rum og se to film. Hun må ikke bevæge sig og skal ~~forsøge~~ forsøge at holde øjnene i ro indtil filmen er færdig

12.25 - Det viser sig svært at få godt signal i TP9 og TP10 pga pigernes hår og cappens dårlige tilslutning

12.30 - FP følges ind i forsøgslokalet og forklares hvordan forsøget kommer til at forløbe.

- Det forklares at hun skal ligge inden for skærmen.
- Andreas tænder kamera og film. Efter 45Hz tonen tager FP hovedtelefonerne i ørene og Andreas lukker en knap med knaphænderen umiddelbart til højre for FP's hoved.

13.15 - FP kommer ud hvorefter Simon går ind og sletter 8 EEG optagelsen samt kameraet.

- FP kommer med tilbage til forberedelsesrummet og udfylder spørgeskema omhvirvig forståelse af filmen.

- FPZ får monteret hat mens hun udfylder papir

14.55 - Hun følges ind i undersøgelsesrummet og forklares at hun skal sidde stille, kigge på skærmen uden at bevæge øjnene for meget.

- Video startes, optagelse af EEG startes

14.15.02 - FP kommer ind og siger at der ikke er lyd i filmen. Det er et problem med signal-delen som rettes og videoen startes igen.

15.20 - FP3 ankommer og får monteret hætten.

- Både tablet og målebånd ligger i eksperiment lokalet

- FP udfylder dokumenter og forklarer proceduren

16.35 - FP bliver desværre nødt til at vente i 10 min med hætten monteret fordi vi venter på den sidste

- Efter måling opdages det at EEG kablen er løbet tør for strøm.

15.55	-	FP4 ankommer og udfylder dokumenter
	-	Får monteret cap og forklaret proceduren; hvordan filmen forløber, og hvordan hun skal forholde sig
15.15	-	Cappen er monteret og forsøgspersonen venter
	-	
16:45	-	FP5 ankommer, starter med at udfylde dokumenter
	-	Cappen monteres og udfyldes med gel
	-	Monteret 10 min inden forsøg
18.00	-	Kommer tilbage og ~~udfylder~~ udfylder seddel
17.30	-	FP6 ankommer og udfylder dokumenter
	-	Får cap på
	-	Alt forløber efter planen
14:30	-	FP 12 ankommer, udfylder ark og monteret cap
	-	Stor cap og sender 2
	-	FP får forsøget forklaret
	-	FP bliver instrueret i ikke at falde i søvn, bygge rundt osv.
15:18	-	FP er færdig og udfylder skema
15:18	-	FP13 ankommer og udfylder dokumenter. Sidder i samme rum som FP12 i mens.
	-	Str. 54 og sender 02

16:19 - FP14 ~~et~~ ankommer og udfylder dokumenter
- Lille cap - sender 03
- FP bliver informeret om forsøget og hvordan hun skal forholde sig
16:32 - Bliver fulgt ind i rummet
17:00 - Kommer tilbage og udfylder spørgeskema

16:45 - FP15 ankommer, trækker, slår med tørning og udfylder dokumenter
17:40 - alt er forløbet som planlagt, FP går igen

- FP16, str. 58, sender 3
17:45 - får rutinen forklaret og fulgt ind i undersøgelsesrummet

13.11.13
17.36 - FP17 ankommer og udfylder dokumenter
- Sender 04 og cap str.
- FP får proceduren forklaret og monteret cap
- Den første elektrode op føres sig ~~~~ underligt. Først er fodbindelsen god, men da optagelsen skulle til at starte faldt kontaktansen til ~11
- Forsøget forløber ellers som planlagt

14.11.13
10.46 - FP18 ankommer og udfylder dokumenter. Hun sagde her en del og er muligvis en smule syg
- ~~Den første cap var defekt~~. Tror måske stikket var vådt
- Str. 58, sender 03

11.00	- FP19 anhommer og ~~udfyl~~ udfylder dokumenter
	- Får monteret cap og forklaret ~~p~~ procedueren.
	- Alt forløber som planlagt
12.10	- Kommer ud og udfylder dokumenter
11.25	- FP20 anhommer og udfylder dokumenter.
	- Tiden er desværre steredet lidt så hun sidder og venter sammen med den som skal ind inden og derefter den som kommer ud (FP18 & 19)
11.50	- FP får monteret cap
12.10	- Bliver vist ind i fødelokalet
	- Forsøg forløber som planlagt
15.11.13	
12.00	FP22 anhommer sammen med FP23 og får monteret cap samt får forklaret procedueren
	- Str. 58, sender 02
	- ~~FP23, Str. 54~~
9.00	FP21 (Trine), Str. 58, ~~sort~~ sender 02, Standard forløb
13.10	- FP23, Str. 54, sender 03, ventede siden kl. 12
13.30	FP24 anhommer og ~~f~~ udfylder papirerne.
	- Cap monteres og hun får proceduren forklaret
	- Forsøgt løber efter plamen, men der blev ikke noteret sender eller modtager

Forsøgslog

Fredag 8/8

8. november 2013 11:25

FS 7

Tid	
11.30	FS ankommer. Bliver informeret om forsøg og om ikke at lave for mange artefakter. Udfylder spørgeskemaer. Trækker NS
11.35	Cap str 54 monteres. Sender nr. 02 bruges. Lav konduktans ved elektrode i panden. Hej på resten af elektroder.
11.55	Forsøg påbegyndes.
12.35	Forsøg slut.
12.40	FS har afkrydset psykiatrisk journal. FS oplyser at hun ikke har fået medicin og at det er afsluttet.

FS 8

15.15	FS Ankommer. Bliver informeret. Udfylder skema. Trækker S
19.20	Får monteret Cap str 54. Har tykt hår, så til at starte med er alle konduktanser 0, men ekstra gel hjælp til normal konduktans. Har tendens til at miste konduktans ved pærer grundet hår der presser elektrode ud.
15.40	Forsøg startes. Glemte label, så sagde "forsøgsperson 8" ingen jeg gik ud af rummet.
16.05	Forsøg færdigt Havde svært ved at forstå handling i "Bang", ma spørgeskema viser at hun kunne stikke essensen af handling sammen.

FS 9

15.45 Fs Ankommer. Informeres. Skema, Trækker S.

15.50 For Cap str 54 på.

16.05 Forsøg startes

16.30 Forsøg slut. Fs faldt ud af at hun var god til tyste.
Skema viser hun har forstået det meste af handlingen.

Lørdag 9/11 Fs 10

12.55 Fs Ankommer

13.00 Fs trækker NS 2. Informeres. Udfylder skema.

Fs' hoved ligger mellem 54 og 58.
Endor med at få 54 på, men herdes øjer sliver hovet lidt i af Cappers øjehuller. Får sender 02.

Fs er en smule fugtig da hun har gået gennem regn.

13.20 Forsøg startes

APPENDIX E

Article in collaboration with Lucas Parra

Below follows a draft of an article in collaboration with Lucas Parra, co-author of Dmochowski et al. [2012]. The article has its focus on the derivation and testing of BCoCA. Content which is described in chapter 2 in this thesis.

Probabilistic Correlated Component Analysis Draft*

Andreas Trier Poulsen[1*], Simon Kamronn[1*],
Lucas Parra[2], and Lars Kai Hansen[1].

[1] Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Kongens Lyngby, Denmark

[2] Department of Biomedical Engineering,
City College of New York, City University of New York,
New York, NY, USA

December 17, 2013

We propose a probabilistic generative model for investigation of the universality of the representations used in human information processing. The model is tested in simulated data and in two well-established benchmark EEG data sets.

1 Introduction

We are interested in information processing in the human brain. In particular how the human brain solves computational problems such as decoding high level information from a movie. Assuming that the movie brain interaction is jointly optimized for this process we should expect a certain amount of optimality, hence universality, in the representations and processes used in the brains of subjects watching a movie.

Such an approach to neuroscience, said to be based on naturalistic stimuli, has been pursued by Hasson et al. [1, 2, 3]. They introduced a correlation approach between anatomically aligned brains. This is based in a rather strong assumption of universality, namely that both the extracted information (what) and the representation (where) are shared among subjects. To exploit the full spatio-temporal patterns of correlation and increase sensitivity, a multivariate version of this approach, so-called correlated component analysis was recently proposed by Dmochowski et al. [4].

Within the multivariate framework, a natural relaxation of the strong universality hypothesis, would be to investigate if the decoded content (what) was identical between subjects, but representations, hence, the 'where' individual. Such an approach corresponds to the

*This work is funded by Lundbeckfonden via CIMBI Center for Integrated Molecular Brain Imaging.
Authors with (∗) made equal contributions to this work

multivariate approach known as canonical correlation analysis (CCA) [5]. The CCA approach searches for individual stationary spatial networks with similar temporal activation among subjects and was generalized to account for both joint and individual signal components by Lukic et al. [6]. A probabilistic approach to CCA also including the possibility of both joint and individual components was proposed by Klami et al. [7].

Here we will analyze a probabilistic model which focuses on extracting joint components inspired by the work of Dmochowski et al., however, with the possibility of learning the degree of universality from data. The latter is implemented by hierarchical Bayesian approach that allows variable degree of non-universal representations (where) in individual subjects. We illustrate the performance and the approximate inference procedures invoked in both simulation studies and in electro-encephalographic (EEG) data. EEG was used to illustrate the model proposed by [4], while functional magnetic resonance imaging (fMRI) was used in [1, 2]. Dmochowski et al. [4] argue that voxel-wise correlations in blood oxygenation level dependent (BOLD) signals are unable to capture weak activity over distant regions, as well as that the poor temporal resolution of fMRI inhibits precise estimation synchronized information processing. However, the spatial resolution of EEG represents a drawback relative to fMRI, hence the test for similarity of spatial networks can only be answered at a limited spatial resolution.

2 Finding correlated components through eigenvalue decomposition

We first briefly review the two existing multivariate approaches.

Given two multivariate spatio-temporal datasets, $\mathbf{X}^{(1)} \in \mathbb{R}^{D_1 \times N}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{D_2 \times N}$, with $\{D_1, D_2\}$ defining the number of measured features and N the number time samples, CCA seeks to estimate weights, $\{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}\}$, which maximise the correlation between $\mathbf{y}_1 = \mathbf{X}^{(1)T}\mathbf{w}_k^{(1)}$ and $\mathbf{y}_2 = \mathbf{X}^{(2)T}\mathbf{w}_k^{(2)}$. At the same time CCA constrains the estimated weights with the condition that $\mathbf{X}^{(1)T}\mathbf{w}_k^{(1)}$ and $\mathbf{X}^{(1)T}\mathbf{w}_{k'}^{(1)}$ are uncorrelated for $k \neq k$ [7]. Introducing the sample covariance matrix, $\mathbf{R}_{ij} = \frac{1}{N}\mathbf{X}^{(i)}\mathbf{X}^{(j)T}$, CCA finds the weights analytically through eigenvalue decompositions

$$\begin{aligned} \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{w}^{(1)} &= \rho^2\mathbf{w}^{(1)} \\ \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{w}^{(2)} &= \rho^2\mathbf{w}^{(2)}. \end{aligned} \tag{1}$$

Correlated component analysis is a related approach, but differentiates itself by finding a single set of weights that works for filtering both datasets. This stronger universality assumptions is also motivated by its fewer degrees of freedom. Furthermore it does not require the somewhat artificial orthogonality between weights, which is less meaningful in, e.g., EEG where the weights are spatial networks [4]. In Correlated component analysis the weights are thus estimated through a single eigenvalue decomposition [4],

$$\left(\mathbf{R}_{11} + \mathbf{R}_{22}\right)^{-1}\left(\mathbf{R}_{12} + \mathbf{R}_{21}\right)\mathbf{w} = 2 \cdot \frac{\sigma_{12}}{\sigma_{11}}\mathbf{w}. \tag{2}$$

2.1 Robustness of correlated component analysis to assumption of universal patterns

While correlated component analysis is based on the assumption of identical weights, we here show that the approach is indeed robust to differences in the 'true weights'. In

particular we here show that even if the two sets of weights are orthogonal, correlated components can still be found using the method.

The observations are assumed to consist of a single true signal mixed into D dimensions by a vector and gaussian noise;

$$\mathbf{X}_1 = \mathbf{a}_1\mathbf{z} + \epsilon \ , \quad \mathbf{X}_2 = \mathbf{a}_2\mathbf{z} + \epsilon. \tag{3}$$

Given enough samples, the sample covariance matrices can be defined as

$$\mathbf{R}_{11} = P \cdot \mathbf{a}_1\mathbf{a}_1^T + \sigma^2\mathbf{I} \ , \quad \mathbf{R}_{12} = P \cdot \mathbf{a}_1\mathbf{a}_2^T, \tag{4}$$

where P signifies the power of $\mathbf{z}$ and σ^2 signifies the noise variance. For simplicity the the weight vectors are assumed to have unit length.

The two matrices in (2) can now be written as

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1} = \frac{1}{P}\left(\mathbf{a}_1\mathbf{a}_1^T + \mathbf{a}_2\mathbf{a}_2^T + \frac{2\sigma^2}{P}\mathbf{I}\right)^{-1} \quad \Leftrightarrow \tag{5}$$

$$= \frac{1}{P}\left([\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix} + \frac{2\sigma^2}{P}\mathbf{I}\right)^{-1} \tag{6}$$

$$\mathbf{R}_{12} + \mathbf{R}_{21} = P \cdot [\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix} \tag{7}$$

using block matrix notation. Using $\mathbf{a}_1^T\mathbf{a}_2 = 0$, $\|\mathbf{a}_1\|^2 = \|\mathbf{a}_2\|^2 = 1$ and the Woodbury identity, (6) can be expressed as;

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1} = \frac{1}{2\sigma^2}\left(\mathbf{I} - \frac{P}{2\sigma^2}[\mathbf{a}_1\ \mathbf{a}_2] \cdot \left(\mathbf{I} - \frac{P}{2\sigma^2}\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix}[\mathbf{a}_1\ \mathbf{a}_2]\right)^{-1}\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix}\right) \tag{8}$$

$$= \frac{1}{2\sigma^2}\left(\mathbf{I} - \frac{P}{2\sigma^2}[\mathbf{a}_1\ \mathbf{a}_2]\left(\mathbf{I} - \frac{P}{2\sigma^2}\begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix}\right) \tag{9}$$

$$= \frac{1}{2\sigma^2}\left(\mathbf{I} - \frac{P}{2\sigma^2 + P}[\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix}\right). \tag{10}$$

The matrix product of (6) and (7) then gives

$$(\mathbf{R}_{11} + \mathbf{R}_{22})^{-1}(\mathbf{R}_{12} + \mathbf{R}_{21}) = \frac{P}{2\sigma^2}\left(\mathbf{I} - \frac{P}{2\sigma^2 + P}[\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix}\right)[\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix} \tag{11}$$

$$= \frac{P}{2\sigma^2}\left(1 - \frac{P}{2\sigma^2 + P}\right)[\mathbf{a}_1\ \mathbf{a}_2]\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix} \tag{12}$$

$$= \frac{P}{2\sigma^2 + P}(\mathbf{a}_1\mathbf{a}_1^T + \mathbf{a}_2\mathbf{a}_2^T) \tag{13}$$

Using the simplifying assumptions made earlier an eigenvector for (13) can be seen to have the form $\alpha\mathbf{a}_1 + \beta\mathbf{a}_2$ since

$$\frac{P}{2\sigma^2 + P}(\mathbf{a}_1\mathbf{a}_1^T + \mathbf{a}_2\mathbf{a}_2^T)(\alpha\mathbf{a}_1 + \beta\mathbf{a}_2) = \frac{P}{2\sigma^2 + P}(\alpha\mathbf{a}_2 + \beta\mathbf{a}_1). \tag{14}$$

It can be seen that $\alpha\mathbf{a}_1 + \beta\mathbf{a}_2$ is an eigenvector when either $\alpha = \beta$ or $\alpha = -\beta$ with $\pm\frac{P}{2\sigma^2+P}$ as eigenvalues. This means that when the true mixing weights of two datasets are orthogonal correlated component analysis finds a common weight, consisting of the mean of the true weights.

3 Probabilistic Correlated Component Analysis

Inspired by the probabilistic principal component analysis introduced by [8], a probabilistic approach to CCA was presented in [9] using latent variables. They formulated a probabilistic generative model based on Gaussian distributed common sources, $\mathbf{z}$, mixed to form two noisy observed datasets

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{15}$$

$$\mathbf{x}^{(m)} \sim \mathcal{N}\left(\mathbf{A}^{(m)}\mathbf{z}, \mathbf{\Phi}^{(m)}\right), \quad \text{for } m = \{1, 2\}, \tag{16}$$

with $\mathbf{\Phi}^{(m)}$ representing the covariance matrix for the observation noise of dataset m. $\mathbf{A}$[1] signifies the mixing matrix, where each column represents the mixing of one source, which means that if one posses prior knowledge of the number of hidden sources the dimension of the estimated mixing matrix can be reduced to $\mathbf{A} \in \mathbb{R}^{D\times K}$. This is an advantage when $K < D$, but presents the problem of chosing the right value for K. To avoid discrete model selection [10] introduced a hierarchical prior over $\mathbf{A}$ using the automatic relevance determination (ARD) framework

$$\mathbf{A}^{(m)} \sim \prod_k^K \mathcal{N}\left(\mathbf{A}_k^{(m)} | \mathbf{0}, \alpha_k^{-1}\right) \tag{17}$$

$$\boldsymbol{\alpha} \sim \prod_k^K \mathcal{G}a(\alpha_k | a_0, b_0), \tag{18}$$

where $\mathbf{A}_k$ signifies the k'th row in $\mathbf{A}$ and $\boldsymbol{\alpha}_k$ is a gamma distributed hyper parameter controlling the precision of $\mathbf{A}_k$.

This approach to CCA has lead to Bayesian CCA [12, 13], a hierarchical Bayesian spatio-temporal model [14] and latest Group Factor Analysis (GFA) [15], the first practical multi-view generalization of Bayesian CCA, and its two-view extension Bayesian Inter-Battery Factor Analysis (BIBFA) [7]. The latest addition divided the sources into shared and view-specific sources enabling the simplification of $\mathbf{\Phi}^{(m)}$ to a diagonal matrix improving computing time for high dimensional data.

The above mentioned articles where not the first to introduce these concepts to probabilistic CCA, but instead had their focus on how to approximate the posterior distribution for

[1] Authors use different letters for the mixing matrix. Most Bayesian models use the notation $\mathbf{W}$, probably stemming from [10], but as this letter is also used to define the demixing matrix, we have in this article chosen to use $\mathbf{A}$ as [11].

the hidden sources. Instead of the commonly used maximum likelihood or maximum a posteriori solutions through expectation maximization, the articles focus on a full Bayesian treatment employing either Gibbs sampling or variational inference. Both approaches have their own advantages and drawbacks. Here we will focus on variational inference, and Gibbs sampling will not be discussed further.

3.1 Variational inference

The joint probability distribution can be expressed as

$$p(\mathbf{X}) = p(\mathbf{H}, \mathbf{V})$$

with $\mathbf{H}$ and $\mathbf{V}$ being hidden and visible variables. Sometimes the joint distribution for a statistical model can get so complex that the true posterior distribution, $p(\mathbf{H}|\mathbf{V})$, becomes analytically intractable, in which case a suitable approximation, $q(\mathbf{H})$, can be a better option. In probabilistic variational inference the simplifying assumption is often that q is completely factorised as

$$q(\mathbf{H}) = \prod_i q_i(\mathbf{H}_i), \tag{19}$$

meaning that there are no conditional distributions in $q(\mathbf{H})$. This simplification is originally known in physics as *mean field theory* [16].

Variational inference uses the Kullback-Leibler (KL) divergence as a measure of the dissimilarity between the true distribution and its approximation, and seeks to minimise it. The KL divergence is defined as

$$\mathrm{KL}(q\|p) = \int q(\mathbf{H}) \ln \frac{q(\mathbf{H})}{p(\mathbf{H}|\mathbf{V})} d\mathbf{H}. \tag{20}$$

The evaluation of the KL divergence, as defined in (20) depends on the posterior distribution, but since this is assumed intractable the equation is not very useful in this form. Using the product rule (20) can be rearranged into an expression with distributions that are assumed analytically tractable;

$$\mathrm{KL}(q\|p) = \int q(\mathbf{H}) \ln \frac{q(\mathbf{H})}{p(\mathbf{H}, \mathbf{V})} d\mathbf{H} + \int q(\mathbf{H}) \ln p(\mathbf{V}) d\mathbf{H} \quad \Leftrightarrow \tag{21}$$

$$= \int q(\mathbf{H}) \ln \frac{q(\mathbf{H})}{p(\mathbf{H}, \mathbf{V})} d\mathbf{H} + \ln p(\mathbf{V}). \tag{22}$$

Defining the negative of the first term on the right hand side as $\mathcal{L}(q)$, a relationship between the true log likelihood and the approximation of the posterior distribution can be defined as

$$\ln p(\mathbf{V}) = \mathrm{KL}(q\|p) + \mathcal{L}(q). \tag{23}$$

Using Jensen's inequality it can be proven that the KL divergence is positive except when $q(\mathbf{H}) = p(\mathbf{H}|\mathbf{V})$, where it is zero. This means that $\mathcal{L}(q)$ cannot exceed the true log likelihood, and is therefore a lower bound for it. So when optimising $q(\mathbf{H})$ through minimisation of the KL divergence, one can instead do it through maximisation of $\mathcal{L}(q)$ [17].

Using the factorising assumption it can be proven that the lower bound can be maximised with respect to the approximated distribution for each variable, $q_j(\mathbf{H}_j)$, when

$$\ln q_j(\mathbf{H}_j) = \langle \ln p(\mathbf{H}, \mathbf{V}) \rangle_{\mathbf{H}/j} + C, \tag{24}$$

where $\langle\ \rangle_{\mathbf{H}/j}$ signifies the expectation with respect to all variables in $\mathbf{H}$, except $\mathbf{H}_j$. C is a constant term representing all terms of the expectation not dependent on the given variable, $\mathbf{H}_j$.

The resulting algorithm consists of updating the lower bound with respect to each variable in turn in a expectation maximisation like manner, where the order of updates can either be fixed or be chosen at random. The form of (23) makes exponential prior distributions a convenient choice with the benifit of having conjugate relationship between the prior and posterior distributions [16, 18].

3.2 The generative model

The Probabilistic or Bayesian correlated component analysis (BCoCA) model presented in this article is based on variational inference with assumptions of a exponential factorised posterior with conjugate priors, as described earlier in this section.

The priors for $\mathbf{z}$ and $\boldsymbol{\alpha}$ are defined as in (15,18) and the only change in the prior for $\mathbf{x}$ is the use of the precision matrix, $\boldsymbol{\Psi}$ instead of the covariance matrix, viz., the prior assigned to $\boldsymbol{\Psi}$ is a Wishart distribution;

$$\mathbf{x}^{(m)} \sim \mathcal{N}\left(\mathbf{A}^{(m)}{}_{\mathbf{Z}}, \boldsymbol{\Psi}^{(m)^{-1}}\right) \tag{25}$$

$$\boldsymbol{\Psi}^{(m)} \sim \mathcal{W}(\mathbf{S}_0, v_0). \tag{26}$$

The major differences lie in the prior for the weights, which have been expanded to include latent variable, $\mathbf{U}$, representing the mean weigth matrix across all datasets and the ARD variable λ which regularizes how close the $\mathbf{A}$'s lie to $\mathbf{U}$;

$$\mathbf{U} \sim \prod_k^K \mathcal{N}\left(\mathbf{u}_k | \mathbf{0}, \alpha_k^{-1}\right) \tag{27}$$

$$\mathbf{A}^{(m)} \sim \prod_k^K \mathcal{N}\left(\mathbf{a}_k^{(m)} | \mathbf{u}_k, \lambda^{-1}\right) \tag{28}$$

$$\lambda \sim \mathcal{G}a(a_0, b_0) \tag{29}$$

The joint probability is then given by

$$\begin{aligned} p(\mathbf{V}, \mathbf{H}) &= p(\mathbf{X}, \mathbf{Z}, \mathbf{A}, \mathbf{U}, \boldsymbol{\Psi}, \boldsymbol{\alpha}, \lambda) \\ &= p(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \boldsymbol{\Psi}) p(\mathbf{Z}) p(\Psi) p(\mathbf{A}|\mathbf{U}, \boldsymbol{\alpha}) p(\mathbf{U}|\lambda) p(\boldsymbol{\alpha}) p(\lambda), \end{aligned} \tag{30}$$

where $p(\mathbf{X}) = \prod_M^M p(\mathbf{X}^{(m)})$ and so forth.

Using variational inference results in the following approximated distributions;

$$q(\mathbf{Z}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{z}_n | \boldsymbol{\mu}_{\mathbf{z},n}, \Sigma_{\mathbf{z}}\right) \tag{31}$$

$$\Sigma_{\mathbf{z}}^{-1} = \sum_{m}^{M} \left\{ \left\langle \mathbf{A}^{(m)T} \Psi^{(m)} \mathbf{A}^{(m)} \right\rangle \right\} + \mathbf{I} \tag{32}$$

$$\boldsymbol{\mu}_{\mathbf{z},n} = \Sigma_{\mathbf{z}} \sum_{m}^{M} \left\langle \mathbf{A}^{(m)T} \right\rangle \left\langle \Psi^{(m)} \right\rangle \mathbf{x}_n^{(m)} \tag{33}$$

$$q(\boldsymbol{\Psi}) = \prod_{m}^{M} \mathcal{W}\left(\mathbf{S}_{\Psi}^{(m)}, v_{\Psi}\right) \tag{34}$$

$$\begin{aligned} \mathbf{S}_{\Psi}^{(m)^{-1}} = & \left\langle \mathbf{A}^{(m)} \sum_{n}^{N} \mathbf{z}_n \mathbf{z}_n^T \mathbf{A}^{(m)T} \right\rangle + \sum_{n}^{N} \mathbf{x}_n^{(m)} \mathbf{x}_n^{(m)T} \\ & - 2 \cdot \sum_{n}^{N} \mathbf{x}_n^{(m)} \left\langle \mathbf{z}_n^T \right\rangle \left\langle \mathbf{A}^{(m)T} \right\rangle + \mathbf{S}_0^{-1} \end{aligned} \tag{35}$$

$$v_{\Psi} = N + v_0 \tag{36}$$

$$q(\mathbf{A}^{(m)}) = \prod_{d=1}^{D} \mathcal{N}\left(\hat{\mathbf{a}}_d^{(m)} |\ \boldsymbol{\mu}_{\mathbf{a}_d}^{(m)}, \Sigma_{\mathbf{a}_d}^{(m)}\right) \tag{37}$$

$$\Sigma_{\mathbf{a}_d}^{(m)^{-1}} = \left\langle \psi_{dd}^{(m)} \right\rangle \sum_{n}^{N} \left\langle \mathbf{z}_n \mathbf{z}_n^T \right\rangle + \langle \lambda \rangle \mathbf{I} \tag{38}$$

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{a}_d}^{(m)} = & \Sigma_{\mathbf{a}_d}^{(m)} \left(\sum_{n}^{N} \langle \mathbf{z}_n \rangle \left\langle \psi_{(d,:)}^{(m)} \right\rangle \mathbf{x}_n^{(m)} + \langle \lambda \rangle \langle \mathbf{u}_d \rangle \right. \\ & \left. - \sum_{d' \neq d}^{D} \left\langle \psi_{dd'}^{(m)} \right\rangle \sum_{n}^{N} \left\langle \mathbf{z}_n \mathbf{z}_n^T \right\rangle \left\langle \mathbf{a}_{d'}^{(m)T} \right\rangle \right) \end{aligned} \tag{39}$$

$$q(\mathbf{U}) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{u}_k |\ \boldsymbol{\mu}_{\mathbf{u}_k}, \sigma_{\mathbf{u}_k}^2 \mathbf{I}\right) \tag{40}$$

$$\sigma_{\mathbf{u}_k}^{-2} = M \langle \lambda \rangle + \langle \alpha_k \rangle \tag{41}$$

$$\boldsymbol{\mu}_{\mathbf{u}_k} = \sigma_{\mathbf{u}_k}^2 \langle \lambda \rangle \sum_{m}^{M} \left\langle \mathbf{a}_k^{(m)} \right\rangle \tag{42}$$

$$q(\boldsymbol{\alpha}) = \prod_k^K \mathcal{G}a(\alpha_k | a_\alpha, b_{\alpha_k}) \tag{43}$$

$$a_\alpha = a_0 + \frac{D}{2} \tag{44}$$

$$b_{\alpha_k} = b_0 + \frac{\left\langle \mathbf{u}_k^T \mathbf{u}_k \right\rangle}{2} \tag{45}$$

$$q(\lambda) = \mathcal{G}a(\lambda | a_\lambda, b_\lambda) \tag{46}$$

$$a_\lambda = a_0 + \frac{MKD}{2} \tag{47}$$

$$b_\lambda = b_0 + \sum_k^K M \frac{\left\langle \mathbf{u}_k^T \mathbf{u}_k \right\rangle}{2} + \sum_m^M \left\{ \frac{\left\langle \mathbf{a}_k^{(m)T} \mathbf{a}_k^{(m)} \right\rangle}{2} - \left\langle \mathbf{a}_k^{(m)T} \right\rangle \langle \mathbf{u}_k \rangle \right\}. \tag{48}$$

where $\hat{\mathbf{a}}_d^{(1)}$ is a column vector corresponding to the d'th row of $\mathbf{A}$. Note that v_Ψ, a_α and a_λ are constants and can be defined before iterating over the other updates.

$\mathcal{L}(q)$ is often calculated to estimate the time of convergence, by setting a threshold for the relative change wrt. the previous iteration. It is usually derived as the sum of the expectations of each variable in $q(\mathbf{H})$ and $p(\mathbf{H}, \mathbf{V})$ wrt. $q(\mathbf{H})$ calculated independently. Inspired by [18] we have chosen to combine the expectations into one equation and let terms containing the same terms cancel each other out, where applicable. Since it is the change of the lower bound that is of interest, we also combined all constant terms into the common constant, C. This resulted in a simpler expression for $\mathcal{L}(q)$;

$$\begin{aligned}\mathcal{L}(q) = &\frac{1}{2} \sum_m^M \left\{ v_\Psi \ln |S_\Psi^{(m)}| + \sum_d^D \ln |\Sigma_{\hat{\mathbf{a}}_d}^{(m)}| \right\} - a_\lambda \ln b_\lambda \\ &+ \sum_k^K \left\{ -a_\alpha \ln b_{\alpha_k} + \frac{D}{2} \ln \sigma_{\mathbf{u}_k}^2 \right\} + \frac{1}{2} \ln |\Sigma_{\mathbf{z}}| \\ &- \frac{1}{2} \left(N \cdot \mathrm{Tr}(\Sigma_{\mathbf{z}}) + \sum_n^N \mu_{\mathbf{z},n}^T \mu_{\mathbf{z},n} \right) + C.\end{aligned} \tag{49}$$

This expression only calculates how the variables that are modified influence the lower bound. Therefore it cannot be used to directly compare with other models based on other priors. It can however be used for estimating a time of convergence and as measure to decide on the best result among multiple runs on the same data.

4 Performance on simulated data

4.1 Simulation Design

To measure the performance between BCoCA, correlated component analysis, and CCA, data is generated from the BCoCA model with a varying λ. This approach generates data from a pure correlated component analysis model, with equal true weights for all datasets, when $\lambda \gg 1$ and from a CCA model when $\lambda \ll 1$. From the model definition we get that

$$\mathbf{X}^{(m)} = \mathbf{A}_{\text{true}}^{(m)} \mathbf{Z} + \boldsymbol{\epsilon} \tag{50}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, where σ_ϵ^2 is varied to obtain the desired signal-to-noise ratio (SNR). $\mathbf{Z}$ is a $K \times N$ source matrix containing K time series. $\mathbf{A}$ is formulated as

$$\mathbf{A}_{\text{true}}^{(m)} = \mathbf{U} + \boldsymbol{\delta}^{(m)} \tag{51}$$

with $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{\alpha}^{-1})$ and $\boldsymbol{\delta}^{(m)} \sim \mathcal{N}(0, \lambda^{-1})$. The variance across views are hence only modelled by λ.

We have used up to four hidden sources, generated in the same manner as in [7], for comparability with their results. In this article we will mainly focus on the simple case of one hidden source corresponding to $K = 1$ in (50), meaning that the data is generated from one sinusoid and additive noise.

4.1.1 Measure of performance

The correlation coefficient between the inferred sources and the true source was chosen as the measure of performance. Since the latent models infer a common $\mathbf{Z}$ for all datasets, the mean of the view specific $\mathbf{y}_1$ and $\mathbf{y}_2$ was used as the inferred sources for CCA and correlated component analysis. This improved their performance by approximately 10 - 20% compared to using only $\mathbf{y}_1$. For each condition 20 datasets were randomly generated from the distributions described in 4.1 and each algorithm was tested on the same data. The mean and standard deviation (std) for the 20 datasets were calculated and used to compare the performance between the four algorithms. In the tests with four hidden sources all correlation combinations between the inferred sources and the true ones were calculated, where each inferred source were only allowed to correlate with one true source and vice versa. The combination with the highest mean correlation were then chosen.

4.1.2 Testing conditions

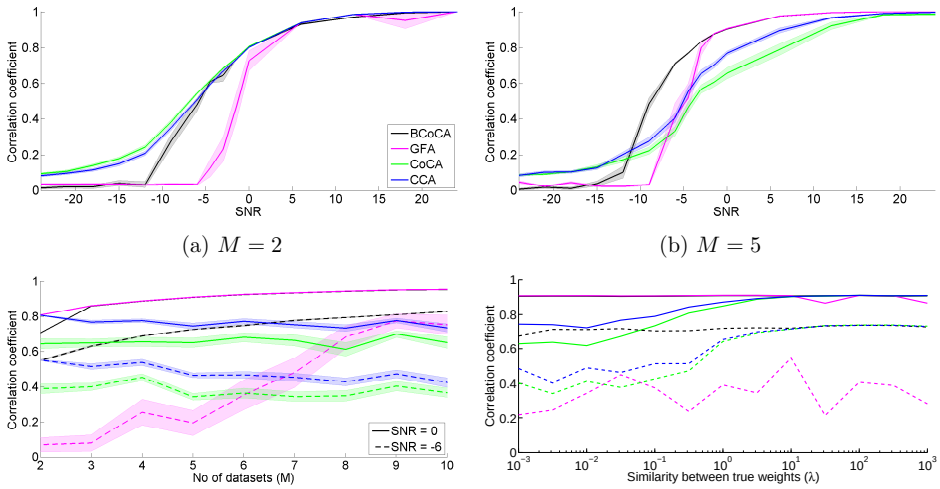
The algorithms were tested at varying levels of SNR, number of datasets and similarity between the true weights of each dataset. In each test the number of observations was set to 500, except when varying the number of datasets. The test was conducted on a total of 5.000 samples spread out equally among the datasets, so that each contained 2.500 samples for $M = 2$ and 500 samples for $M = 10$. All the conditions were tested with one and four hidden sources.

4.1.3 Correlated component analysis and CCA on multiple datasets

correlated component analysis and CCA can only compare two datasets at a time. In case of multiple dataset comparison this thesis will follow the same method as in [4], where the datasets are concatenated sample-wise into

$$\begin{aligned} \bar{\mathbf{X}}_1 &= [\mathbf{X}^{(1)}, \mathbf{X}^{(1)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}], \\ \bar{\mathbf{X}}_2 &= [\mathbf{X}^{(2)}, \mathbf{X}^{(3)}, \mathbf{X}^{(4)}, \mathbf{X}^{(3)}, \mathbf{X}^{(4)}, \mathbf{X}^{(4)}] \end{aligned} \tag{52}$$

so that all combinations of datasets will be compared. As both correlated component analysis and CCA use eigenvalues decomposition on the sample covariance matrices, using $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ corresponds to using the average pair-wise sample covariance matrices. This way the eigenvalue decomposition has to be calculated only once. However using this method the number of samples in $\bar{\mathbf{X}}_1$ scales by $M(M-1)/2$, with (52) showing the case of concatenating with four datasets.



(a) $M = 2$ (b) $M = 5$

(c) Varying M (d) $M = 5$

Figure 1: Performance of BCoCA, GFA, CoCA and CCA on simulated data measured by mean correlation coefficient and standard error of the mean with respect to the true source(s). In each subfigure the data is varied in one or two variables: In **(a)** and **(b)** the performance is tested under different levels of SNR and 2 or 5 datasets. **(c)** shows the performance with varying number of datasets and two levels of SNR. In **(d)** the similarity between the true weights are varied by the λ parameter and shows the correlation with two levels of SNR. In the last subfigure the std. was left out to avoid cluttering the graph. $\lambda = 10^{-3}$ for all but **(a)**, where $\lambda = 10^{3}$.

4.1.4 Testing conditions

The algorithms were tested at varying levels of SNR, number of datasets, M, and similarity between the true weights of each dataset. In each test the dataset had six dimensions and the number of observations was set to 500, except when varying the number of datasets. This test was conducted on a total of 5.000 samples spread out equally among the datasets, so that each contained 2.500 samples for $M = 2$ and 500 samples for $M = 10$. All the conditions were tested with one and four hidden sources.

4.2 Results

4.2.1 Performance in simulated data

The results from the tests on the simulated data, with one hidden source mixed into the observed datasets, can be seen in figure 1. Figures 1a and 1b show the performance on increasing values of SNR and 2 or 5 datasets. It can be seen that for high levels of SNR the algorithms perform the same, but as the noise levels increase the latent models, quickly drop towards zero correlation, though BCoCA do so less steeply and can perform at lower levels of SNR compared to GFA. This quick drop is due to the models choosing the zero-source solution as the cost of a poor estimation gets too high. BCoCA comes closer to zero

as this algorithm seemingly choose a source of constant zeros, as opposed to what appears to be low amplitude noise. Better initialisation and basing the prior hyperparameters on the observed data might improve the performance with varying levels of noise.

Figure 1c shows how performance improves as the number of available datasets increases. Here BCoCA and GFA are the only of algorithms tested, which directly generalises to more than two datasets. It can be seen that CCA and correlated component analysis actually performs worse, when the number of datasets increase. This is only true for datasets with non-similar weights. With equal true weights the algorithms perform the same as BCoCA. When considering that CCA and correlated component analysis deal with increasing datasets by concatenating them into two datasets, the importance of having equal weights makes sense as this then corresponds to actually just having two datasets. The increased performance must then stem from having more instances of the signal and then be able to average the noise out. For the two latent models two things are evident. That BCoCA again outperforms GFA at low levels of SNR and that increasing the number of datasets increases the correlation even though the number of observations do not increase. Some of this effect could stem from averaging out the random noise, when calculating the inferred source as the mean of the sources of estimated on each datasets.

Increasing the number of hidden sources to four decreased the mean correlation but did not change the relative performance between the algorithms, except for GFA, which had a performance closer to that of BCoCA.

All test was run at different levels of similarity between the true weights of each dataset by varying λ. Figure 1d shows the case of two datasets and one hidden source. As expected the effect can only be seen on correlated component analysis and BCoCA, but correlated component analysis handles the datasets with different sources better than initially anticipated. An explanation to this is discussed in 2.1.

5 Performance on EEG

In this section the performance of BCoCA will be evaluated on EEG from two separate experiments.

5.1 Auditory Stimulated Data

Using a cohort of 5 subjects, the auditory data was created by speaking two words to a person who then responded whether they were synonyms. The dataset was then separated into synonyms and non-synonyms by independent component analysis.

The data was bandpass filtered to 0.5 Hz – 200 Hz, re-sampled to 200 Hz and divided into epochs with the latency of the second word as zero. To reduce noise from eye movement the independent component with most activity in the eye region was used as a template to find similar components, using the function CORRMAP in EEGLAB [19], which were extracted from all sets. It has previously been shown that alpha band de-synchronisation is linked with tasks that require the subjects attention [20] so the band power of the alpha band (7–15Hz) was used as test data. To remove outliers the epochs where ordered with respect to latency of response time and only epochs 21–160 where used.

Intra-subject correlation

Intra-subject correlation (IaSC) is tested by using BCoCA on the five datasets in each condition respectively. The resulting filters are then used to find the components with maximum mutual correlation from each of the datasets and the coincidence in neural activity is measured by computing the correlation coefficient on an intra subject basis. The correlation is computed in a window equal to the sample-length of one epoch and the step size is 25% of this. The population IaSC is the average of all the individual IaSC. A two-tailed t-test shows that 91% of the windows are significantly correlated.

Inter-subject correlation

The inter-subject correlation (ISC) is found by pooling all the datasets and using BCoCA. The components from each dataset is correlated with all of the others and then averaged to get the population correlation. In figure 2 the average over all the epochs of the combined component clearly show the decrease of alpha activity after mention of the first word and almost immediate increase after the response. The variation in response time is quite significant and is probably due to difference in level of familiarity with the words. A two-tailed t-test shows that 90% of the windows are significantly correlated.

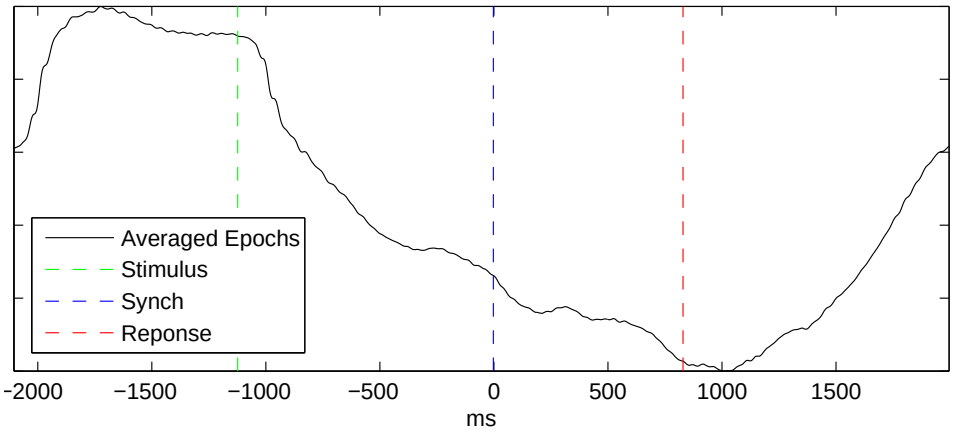


Figure 2: Averaged epoch of the first component from the inter-subject paradigm

5.2 Face-evoked response

A widely accepted theory of face recognition is the multi-component model of face-processing [21] in which the brain derives details about a person from physical aspects. These are used to create structural model that is passed on to other processes that are responsible for recognition, identification, expression analysis, etc. [22] conducted an experiment in which subjects where exposed to a series of images of faces or scrambled faces. The hypothesis from earlier [23] was that a negative peak around 170 ms(N170) post stimulus in the posterior region is greater when the subject is shown a face.

Paradigm and pre-processing

Based on Phase 1 in the study by [22] a subject was over two trials presented with 86 images of faces and 86 images of scrambled faces. The data was bandpass filtered(2-200 Hz), downsampled to 200 Hz and epoched using SPM12. The dataset is available online from the SPM website [24].

ERP Analysis

The epoched data was for each trial concatenated and tested using BCoCA, correlated component analysis and CCA. For the latter two the filters corresponding to the maximally correlated components where used. Epochs from both conditions were processed at the same time, yielding a single component that was then divided into epochs corresponding to the raw data. In figure 3a the averaged epochs for the two conditions is illustrated for the raw EEG and from this it is not possible to the events of interest. In 3b the averaged epoch for the first components for each condition shows that all the algorithms have extracted a coherent signal and that the results are very similar. In figure 3c the difference signal of the averaged epochs shows as expected that the negative peak at N170 is greater for the face condition. The algorithms all locate the time of this occurrence as around 190 ms which corresponds well with the literature [22]. To localize the neurons that are responsible for the face processing, the average of the epochs for the face condition at 170 ms was subtracted from the average scrambled condition. Projecting the channels onto a 2D scalp map as in figure 4d the illustration clearly shows that the posterior regions in the occipital lobe contribute more negatively in the face condition. Projecting the weights from BCoCA the illustration in figure 4a depicts the correlated neural activity. The result shows that the signals in the posterior region are highly correlated. The BCoCA algorithm has thus effectively extracted the component from the datasets that exactly depicts the neural activity of interest.

6 Discussion and Conclusion

Research in social neuroscience has during the previous decade shifted from being inherently single person studies of people observing others towards two-way interaction. The isolated paradigms of standard cognitive science only incorporates information-flow from the environment to the observer, but this approach is inadequate in the paradigm of *embodied cognition*. Interaction and emotional engagement between people are dynamic processes that couples them in a unit that is not readily separable [25]. These inter-personal interactions can be crucial to understanding the mechanisms of social cognition and so far hyperscanning is the only method to tap into inter-brain process [26]. To this purpose EEG is becoming an increasingly popular modality due to its high temporal resolution and recent advances in mobile equipment [27]. EEG data is corrupted by signals from external and internal sources. Internal such as intrinsic activity from the *default mode network*, a part of the brain that is active in the absence of externally oriented cognitive tasks, and external from muscle and eye artefacts. In an experimental paradigm using hyperscanning it is thus advantageous to extract data that is correlated across multiple datasets because the uncorrelated data, in this case noise, is ideally filtered out [4].

A Bayesian version of correlated component analysis gives new approaches to the extraction of these shared signals. Tests with artificial data showed that having more than two

datasets improved the extraction of the shared signals, even though the total amount of observations were the same. This shows promises of better extraction of data from experiments where individual subjects have been exposed to the same stimulus as by [4], but also higher efficiency, when dealing with larger groups and don't have to calculate the average of all pariwise correlations.

The direct estimation of the shared response enables new methods for analysis in experiments with simultaneous stimulation of groups of subjects. Instead of analysing the pairwise correlations of subject respones, it is now possible to look at the correlation between each subject and the response shared by the entire group. In its present form BCoCA estimates the similarity between the responses of a group of subjects through ARD-parameter, λ, but a further expansion with a, λ_m, for each subject would give a indicator whether one or two subjects stand out from the responses of a larger group. Finally, does the latent model structure enable the direct estimation of the forward model used to visualise the scalp maps with the latent variable, $\mathbf{U}$, representing the shared forward model, which again might prove useful with larger groups of subjects.

The analysis of the auditory evoked EEG dataset showed a decrease of alpha activity after the task was given to the subjects which corresponds well with the expected deactivation of the default mode network. Significant inter-subject correlation of 90% of the first component on average across all subjects shows that the method has managed to extract a component that is highly correlated for all the datasets.

From the face-evoked dataset we saw the ability to extract the component from two datasets that would otherwise require manual inspection. The spatial filter from the algorithm correctly selected the posterior region as contributing to the component and the anterior as reducing.

The BCoCA algorithm presented in this communication still has room for further improvement. The tests on artificial data shoved that both latent models chose to turn of al their components, when the SNR got low, which might not always be the best solution. This, and the performance in general, could be improved by investigating the manner of priors and the initialisation of variables. The common choice is setting the hyperparameters to low values [10, 7], as was done in this article, but improvement could be found in conducting a pre-evaluation of data to find suitable values. Some latent models use other algorithms with a lower cost to find suitable values for initialisation [12, 14]. In the present form of BCoCA the precision matrix for the gaussian noise is modelled using the Wishart distribution, as this is the common choice as the conjugate distribution. Tests has shown that this variable has a very high influence on changes in performance and the evaluation of the lower bound. A focus for future work could therefore lie in investigating using other distributions for modelling the noise.

References

[1] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach, "Intersubject synchronization of cortical activity during natural vision," *science*, vol. 303, no. 5664, pp. 1634–1640, 2004.

[2] Uri Hasson, Orit Furman, Dav Clark, Yadin Dudai, and Lila Davachi, "Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding.," *Neuron*, vol. 57, no. 3, pp. 452–62, Feb. 2008.

[3] Uri Hasson, Rafael Malach, and David J Heeger, "Reliability of cortical activity during natural stimulation.," *Trends in cognitive sciences*, vol. 14, no. 1, pp. 40–8, Jan. 2010.

[4] Jacek P Dmochowski, Paul Sajda, Joao Dias, and Lucas C Parra, "Correlated components of ongoing EEG point to emotionally laden attention - a possible marker of engagement?," *Frontiers in human neuroscience*, vol. 6, no. May, pp. 112, Jan. 2012.

[5] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[6] a.S. Lukic, M.N. Wernick, L.K. Hansen, J. Anderson, and S.C. Strother, "A spatially robust ICA algorithm for multiple fMRI data sets," in *Proceedings IEEE International Symposium on Biomedical Imaging*. 2002, pp. 839–842, IEEE.

[7] Arto Klami, Seppo Virtanen, and Samuel Kaski, "Bayesian canonical correlation analysis," *Journal of Machine Learning Research*, vol. 14, pp. 965–1003, 2013.

[8] Michael E Tipping and Christopher M Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[9] Francis R Bach and Michael I Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

[10] Christopher M Bishop, "Variational principal components," 1999.

[11] Lucas C Parra, Clay D Spence, Adam D Gerson, and Paul Sajda, "Recipes for the linear analysis of EEG.," *NeuroImage*, vol. 28, no. 2, pp. 326–341, Nov. 2005.

[12] Chong Wang, "Variational bayesian approach to canonical correlation analysis," *Neural Networks, IEEE Transactions on*, vol. 18, no. 3, pp. 905–910, 2007.

[13] Arto Klami and Samuel Kaski, "Local dependent components," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 425–432.

[14] Wei Wu, Zhe Chen, Shangkai Gao, and Emery N Brown, "A hierarchical bayesian approach for learning sparse spatio-temporal decompositions of multichannel eeg," *Neuroimage*, vol. 56, no. 4, pp. 1929–1945, 2011.

[15] Seppo Virtanen, Arto Klami, Suleiman A Khan, and Samuel Kaski, "Bayesian group factor analysis," *arXiv preprint arXiv:1110.3204*, 2011.

[16] Christopher M Bishop et al., *Pattern recognition and machine learning*, vol. 1, Springer New York, 2006.

[17] John Winn, "Variational message passing and its applications," *Unpublished doctoral dissertation, Cambridge University*, 2003.

[18] Kevin P Murphy, *Machine learning: a probabilistic perspective*, The MIT Press, 2012.

[19] Arnaud Delorme and Scott Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis.," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[20] W Klimesch, M Doppelmayr, H Russegger, T Pachinger, and J Schwaiger, "Induced alpha band power changes in the human EEG and attention," *Neuroscience Letters*, vol. 244, no. 2, pp. 73–76, Mar. 1998.

[21] Vicki Bruce and Young Andy, "Understanding face recognition," *British journal of psychology*, vol. 77, no. 3, 1986.

[22] R.N. Henson, Y. Goshen-Gottstein, T. Ganel, L.J. Otten, A. Quayle, and M.D. Rugg, "Electrophysiological and Haemodynamic Correlates of Face Perception, Recognition and Priming," *Cerebral Cortex*, vol. 13, no. 7, pp. 793–805, July 2003.

[23] Shlomo Bentin, Truett Allison, and Aina Puce, "Electrophysiological studies of face perception in humans," *Journal of cognitive . . .*, vol. 8, no. 6, pp. 551–565, Nov. 1996.

[24] R.N. Henson, "Multimodal face-evoked dataset," .

[25] Leonhard Schilbach, Bert Timmermans, Vasudevi Reddy, Alan Costall, Gary Bente, Tobias Schlicht, and Kai Vogeley, "Toward a second-person neuroscience.," *The Behavioral and brain sciences*, vol. 36, no. 4, pp. 393–414, Aug. 2013.

[26] Ivana Konvalinka and Andreas Roepstorff, "The two-brain approach: how can mutually interacting brains teach us something about social interaction?," *Frontiers in human neuroscience*, vol. 6, no. July, pp. 215, Jan. 2012.

[27] Arkadiusz Stopczynski, Carsten Stahlhut, Jakob Eg Larsen, Michael Kai Petersen, and Lars Kai Hansen, "The Smartphone Brain Scanner: A Mobile Real-time Neuroimaging System," pp. 1–17, Apr. 2013.

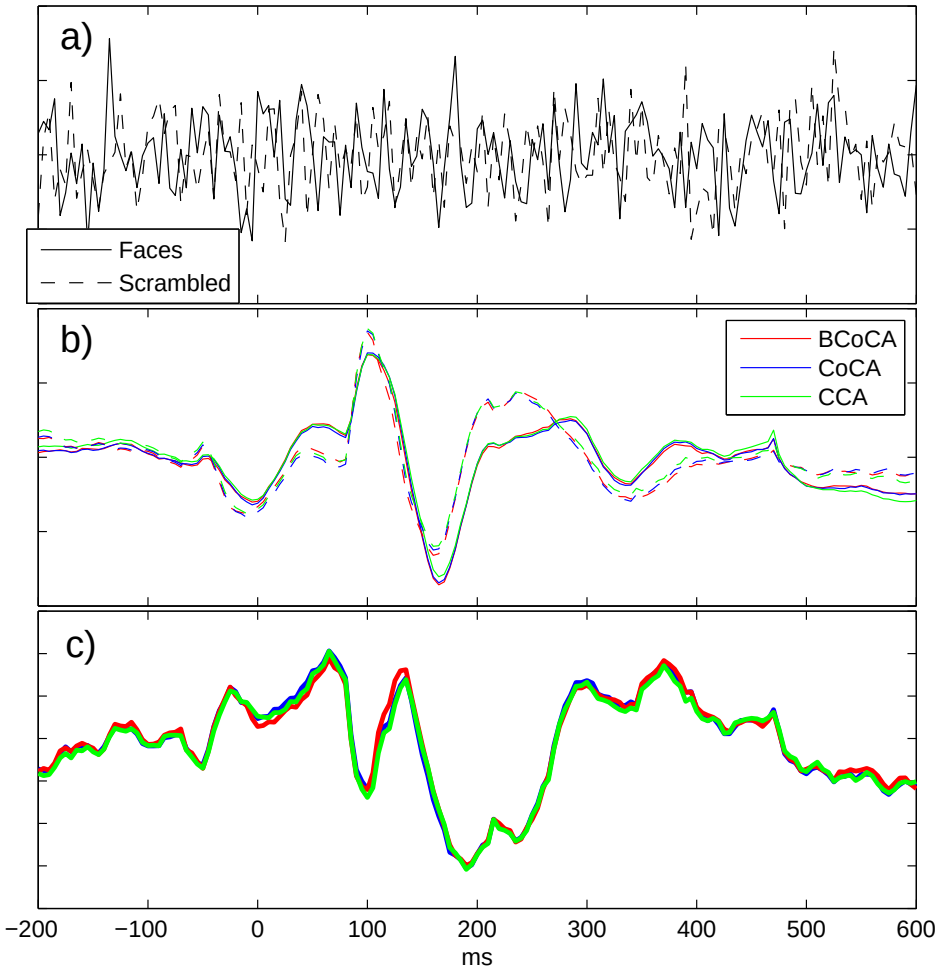
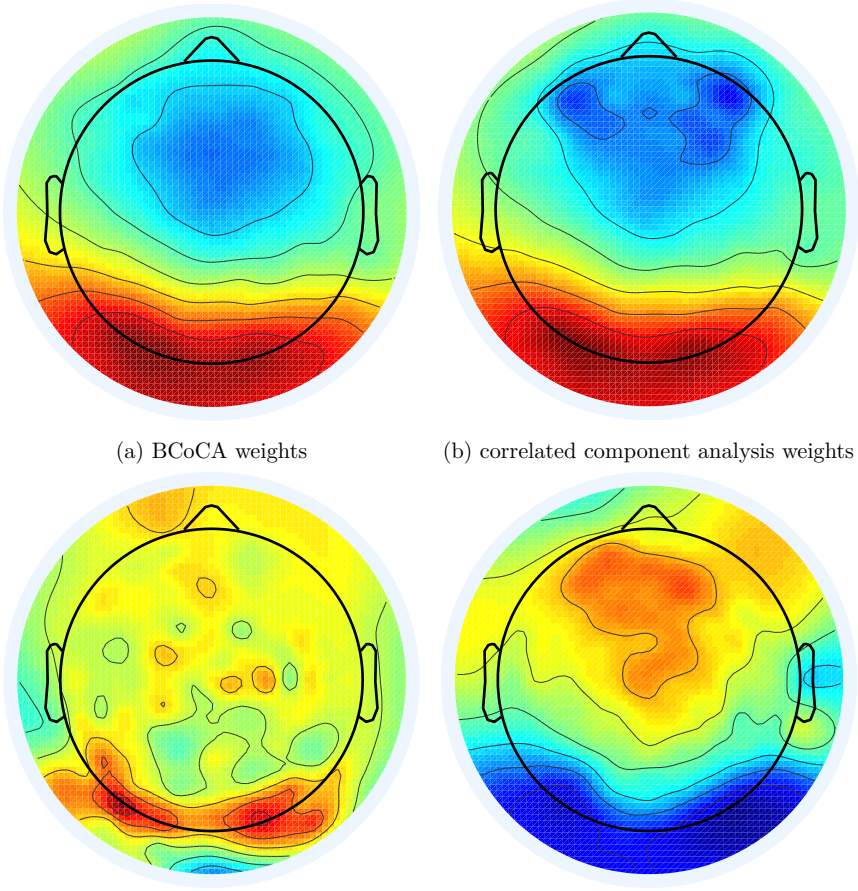


Figure 3: **a)** Averaged epochs across all channels of the raw EEG for faces and scrambled. **b)** Averaged epochs in component space found by BCoCA, correlated component analysis and CCA for face and scrambled condition **c)** Difference between average epoch for face and scrambled condition



(a) BCoCA weights

(b) correlated component analysis weights

(c) CCA

(d) Difference

Figure 4: (**a**) Scalp projections of weights from the BCoCA algorithm. (**b**) Scalp projections of weights from the correlated component analysis algorithm using the forward model [11] (**c**) Scalp projections of the average of the two spatial filters from CCA (**d**) Scalp projections of the average difference between epochs of faces and scrambled images at 170 ms. The blue colour in the posterior regions depicts a negative value.

Bibliography

Alberts, B., D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter (2010). *Essential Cell Biology*. Third. Garland Science.

Attias, H. (2000). "A variational Bayesian framework for graphical models". In: *Advances in neural information processing systems*.

Babiloni, F. and L. Astolfi (2012). "Social neuroscience and hyperscanning techniques: Past, present and future." In: *Neuroscience and biobehavioral reviews*.

Babiloni, F., F. Cincotti, D. Mattia, M. Mattiocco, F. De Vico Fallani, A. Tocci, L. Bianchi, M. G. Marciani, and L. Astolfi (2006). "Hypermethods for EEG hyperscanning." In: *IEEE Engineering in Medicine and Biology Society* 1, pp. 3666–9.

Bach, F. R. and M. I. Jordan (2005). *A Probabilistic Interpretation of Canonical Correlation Analysis*. Tech. rep. University of California, pp. 1–11.

Bell, A. J. and T. J. Sejnowski (1995). "An information-maximization approach to blind separation and blind deconvolution." In: *Neural computation* 7.6, pp. 1129–59.

Bentin, S., T. Allison, A. Puce, E. Perez, and G. McCarthy (1996). "Electrophysiological Studies of Face Perception in Humans." In: *Journal of cognitive neuroscience* 8.6, pp. 551–565.

Berger, J. and K. Milkman (2010). "Social transmission, emotion, and the virality of online content". In: *Wharton Research Paper*, pp. 1–52.

Bishop, C. M. (2006). "Pattern Recognition and Machine Learning". In: *Journal of Electronic Imaging* 16.4, p. 049901. arXiv: `0-387-31073-8`.

Bishop, C. (1999). "Variational principal components". In: *9th International Conference on Artificial Neural Networks: ICANN '99* 1999, pp. 509–514.

Blankertz, B., G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio (2007). "The non-invasive Berlin Brain-Computer Interface: fast acquisition of effective performance in untrained subjects." In: *NeuroImage* 37.2, pp. 539–50.

Böckler, A. and N. Sebanz (2012). "A co-actor's focus of attention affects stimulus processing and task performance: an ERP study." In: *Social neuroscience* 7.6, pp. 565–77.

Bruce, V. and Y. Andy (1986). "Understanding face recognition". In: *British journal of psychology* 77.3.

Bufalari, I., T. Aprile, A. Avenanti, F. Di Russo, and S. M. Aglioti (2007). "Empathy for pain and touch in the human somatosensory cortex." In: *Cerebral Cortex* 17.11, pp. 2553–61.

Carr, J. (1997). *Microwave & Wireless Communications Technology*. Elsevier Science.

Decety, J. and P. L. Jackson (2006). "A Social-Neuroscience Perspective on Empathy". In: *Current Directions in Psychological Science* 15.2, pp. 54–58.

Delic, E. and B. Ziady (2008). *DETERMINATION OF BIOSENSOR CONTACT QUALITY*.

Delorme, A. and S. Makeig (2004). "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." In: *Journal of neuroscience methods* 134.1, pp. 9–21.

Dmochowski, J. P., P. Sajda, J. Dias, and L. C. Parra (2012). "Correlated components of ongoing EEG point to emotionally laden attention - a possible marker of engagement?" In: *Frontiers in human neuroscience* 6.May, p. 112.

Dumas, G. (2011). "Towards a two-body neuroscience." In: *Communicative & integrative biology* 4.3, pp. 349–52.

Dumas, G., J. Nadel, R. Soussignan, J. Martinerie, and L. Garnero (2010). "Inter-brain synchronization during social interaction." In: *PloS one* 5.8, e12166.

Duun-Henriksen, J., T. W. Kjaer, R. E. Madsen, L. S. Remvig, C. E. Thomsen, and H. B. Sorensen (2012). "Correlation between intra-and extracranial background EEG". In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, pp. 5198–5201.

Emotiv (2012). *Epoc specifications*.

Enticott, P. G., P. J. Johnston, S. E. Herring, K. E. Hoy, and P. B. Fitzgerald (2008). "Mirror neuron activation is associated with facial emotion processing." In: *Neuropsychologia* 46.11, pp. 2851–4.

Fisher, R. A. et al. (1949). "The design of experiments." In: *The design of experiments*. 5th ed.

Goldin, P. R., C. a. C. Hutcherson, K. N. Ochsner, G. H. Glover, J. D. E. Gabrieli, and J. J. Gross (2005). "The neural bases of amusement and sadness: a comparison of block contrast and subject-specific emotion intensity regression approaches." In: *NeuroImage* 27.1, pp. 26–36.

Grill-Spector, K. (2003). "The neural basis of object perception". In: *Current Opinion in Neurobiology* 13.2, pp. 159–166.

Groppe, D. M., T. P. Urbach, and M. Kutas (2011). "Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review". In: *Psychophysiology* 48.12, pp. 1711–1725.

Hansen, L. K. and C. E. Rasmussen (1994). "Pruning from Adaptive Regularization". In: *Neural Computation* 6.6, pp. 1223–1232.

Hansen, L., A. Arvidsson, and F. Nielsen (2011). "Good friends, bad news-affect and virality in twitter". In: *Communications in Computer and Information Science* 185, pp. 34–43.

Hardoon, D. R., S. Szedmak, and J. Shawe-Taylor (2004). "Canonical correlation analysis: an overview with application to learning methods." In: *Neural computation* 16.12, pp. 2639–64.

Hasson, U., A. a. Ghazanfar, B. Galantucci, S. Garrod, and C. Keysers (2012). "Brain-to-brain coupling: a mechanism for creating and sharing a social world." In: *Trends in cognitive sciences* 16.2, pp. 114–21.

Hasson, U., Y. Nir, I. Levy, G. Fuhrmann, and R. Malach (2004). "Intersubject synchronization of cortical activity during natural vision." In: *Science (New York, N.Y.)* 303.5664, pp. 1634–40.

Hein, G. and T. Singer (2008). "I feel how you feel but not always: the empathic brain and its modulation." In: *Current opinion in neurobiology* 18.2, pp. 153–8.

Henson, R. "Multimodal face-evoked dataset". In:

Henson, R., Y. Goshen-Gottstein, T. Ganel, L. Otten, A. Quayle, and M. Rugg (2003). "Electrophysiological and Haemodynamic Correlates of Face Perception, Recognition and Priming". In: *Cerebral Cortex* 13.7, pp. 793–805.

Hotelling, H. (1936). "Relations between two sets of variates". In: *Biometrika* 28.3, pp. 321–377.

Jordan, M. I. (1999). "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 233.37, pp. 183–233.

Jung, T.-p., S. Makeig, and C. Humphries (2000). "Removing electroencephalographic artifacts by blind source separation". In: *Psychophysiology* 37.

Kassam, K. S., A. R. Markey, V. L. Cherkassky, G. Loewenstein, and M. A. Just (2013). "Identifying Emotions on the Basis of Neural Activation." In: *PloS one* 8.6, e66032.

Klami, A. (2013). "Bayesian Canonical Correlation Analysis". In: *Journal ofMachine Learning Research* 14, pp. 965–1003.

Klami, A. and S. Kaski (2007). "Local dependent components". In: *Proceedings of the 24th international conference on Machine learning - ICML '07*, pp. 425–432.

Klimesch, W., M. Doppelmayr, H. Russegger, T. Pachinger, and J. Schwaiger (1998). "Induced alpha band power changes in the human EEG and attention". In: *Neuroscience Letters* 244.2, pp. 73–76.

Konvalinka, I. and A. Roepstorff (2012). "The two-brain approach: how can mutually interacting brains teach us something about social interaction?" In: *Frontiers in human neuroscience* 6.July, p. 215.

Kourtis, D., N. Sebanz, and G. Knoblich (2013). "Predictive representation of other people's actions in joint action planning: an EEG study." In: *Social neuroscience* 8.1, pp. 31–42.

Lachat, F., L. Hugueville, J.-D. Lemaréchal, L. Conty, and N. George (2012). "Oscillatory Brain Correlates of Live Joint Attention: A Dual-EEG Study." In: *Frontiers in human neuroscience* 6.June, p. 156.

Luminet, O., P. Bouts, F. Delie, A. S. R. Manstead, and B. Rimé (2000). "Social sharing of emotion following exposure to a negatively valenced situation". In: *Cognition & Emotion* 14.5, pp. 661–688.

MacKay, D. (1996). "Bayesian methods for backpropagation networks". In: *Models of neural networks III*.

Maddock, R. J., A. S. Garrett, and M. H. Buonocore (2003). "Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task." In: *Human brain mapping* 18.1, pp. 30–41.

Makeig, S. and A. Bell (1996). "Independent component analysis of electroencephalographic data". In: *Advances in neural information processing systems* 8.

Mammone, N., F. La Foresta, and F. C. Morabito (2012). "Automatic Artifact Rejection From Multichannel Scalp EEG by Wavelet ICA". In: *IEEE Sensors Journal* 12.3, pp. 533–542.

Manly, B. F. (2007). *Randomization, bootstrap and Monte Carlo methods in biology*. Vol. 70. 3rd ed. CRC Press.

Minka, T., J. Winn, J. Guiver, and D. Knowles (2013). *Infer.NET 2.5*.

Minka, T. (2005). "Divergence measures and message passing". In: *Microsoft Research, Cambridge*.

Molgedey, L. and H. Schuster (1994). "Separation of a mixture of independent signals using time delayed correlations". In: *Physical Review Letters* 72.23, pp. 3634–3637.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. The MIT Press.

Nagel, J. H. (2000). "Biopotential Amplifiers". In: *The Biomedical Engineering Handbook: Second Edition*. CRC Press.

Nummenmaa, L., E. Glerean, M. Viinikainen, I. P. Jääskeläinen, R. Hari, and M. Sams (2012). "Emotions promote social interaction by synchronizing brain activity across individuals." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.24, pp. 9599–604.

Nunez, P. (1974). "The brain wave equation: a model for the EEG". In: *Mathematical Biosciences* 291, pp. 279–297.

Parra, L. C., C. D. Spence, A. D. Gerson, and P. Sajda (2005). "Recipes for the linear analysis of EEG." In: *NeuroImage* 28.2, pp. 326–341.

Pearson, K. (1896). "Mathematical Contributions to the Theory of Evolution.–On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs". In: *Proceedings of the Royal Society of London* 60.359-367, pp. 489–498.

Petersen, K. B. and M. S. Pedersen (2006). *The matrix cookbook*.

Pineda, J. a. (2005). "The functional significance of mu rhythms: translating "seeing" and "hearing" into "doing"." In: *Brain research. Brain research reviews* 50.1, pp. 57–68.

Poreisz, C., K. Boros, A. Antal, and W. Paulus (2007). "Safety aspects of transcranial direct current stimulation concerning healthy subjects and patients." In: *Brain research bulletin* 72.4-6, pp. 208–14.

Preston, S. D. and F. B. M. de Waal (2002). "Empathy: Its ultimate and proximate bases." In: *The Behavioral and brain sciences* 25.1, pp. 1–20, 1–20.

Purves, D. (2004). *Neuroscience*. Third. Sinauer.

Raz, G., Y. Jacob, T. Gonen, Y. Winetraub, T. Flash, E. Soreq, and T. Hendler (2013). "Cry for her or cry with her: context-dependent dissociation of two modes of cinematic empathy reflected in network cohesion dynamics." In: *Social cognitive and affective neuroscience*, pp. 1–9.

Raz, G., Y. Winetraub, Y. Jacob, S. Kinreich, A. Maron-Katz, G. Shaham, I. Podlipsky, G. Gilam, E. Soreq, and T. Hendler (2012). "Portraying emotions at their unfolding: a multilayered approach for probing dynamics of neural networks." In: *NeuroImage* 60.2, pp. 1448–61.

Rimé, B. (2009). "Emotion Elicits the Social Sharing of Emotion: Theory and Empirical Review". In: *Emotion Review* 1.1, pp. 60–85.

Rimé, B., C. Finkenauer, O. Luminet, E. Zech, and P. Philippot (1998). "Social Sharing of Emotion: New Evidence and New Questions". In: *European Review of Social Psychology* 9.1, pp. 145–189.

Rizzolatti, G. and M. Fabbri-Destro (2008). "The mirror system and its role in social cognition." In: *Current opinion in neurobiology* 18.2, pp. 179–84.

Saab, M. (2008). "Basic Concepts of Surface Electroencephalography and Signal Processing as Applied to the Practice of Biofeedback". In: 36.4, pp. 128–133.

Sartori, L., A. Cavallo, G. Bucchioni, and U. Castiello (2012). "From simulation to reciprocity: the case of complementary actions." In: *Social neuroscience* 7.2, pp. 146–58.

Schilbach, L., B. Timmermans, V. Reddy, A. Costall, G. Bente, T. Schlicht, and K. Vogeley (2013). "Toward a second-person neuroscience." In: *The Behavioral and brain sciences* 36.4, pp. 393–414.

Shteynberg, G., J. B. Hirsh, A. D. Galinsky, and A. P. Knight (2013). "Shared Attention Increases Mood Infusion." In: *Journal of experimental psychology. General* 142.2.

Singer, T. (2012). "The past, present and future of social neuroscience: a European perspective." In: *NeuroImage* 61.2, pp. 437–49.

Stopczynski, A., C. Stahlhut, J. E. Larsen, M. K. Petersen, and L. K. Hansen (2013). *The Smartphone Brain Scanner: A Mobile Real-time Neuroimaging System*. Tech. rep. DTU Compute, pp. 1–17. arXiv: 1304.0357.

Tipping, M. E. and C. M. Bishop (1999). "Probabilistic Principal Component Analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 611–622.

Tognoli, E., J. Lagarde, G. C. DeGuzman, and J. a. S. Kelso (2007). "The phi complex as a neuromarker of human social coordination." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.19, pp. 8190–5.

Viola, F. C., J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener (2009). "Semi-automatic identification of independent components representing EEG artifact." In: *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* 120.5, pp. 868–77.

Virtanen, S., A. Klami, S. A. Khan, and S. Kaski (2011). "Bayesian group factor analysis". In: *arXiv preprint arXiv:1110.3204*.

Wang, C. (2007). "Variational Bayesian approach to canonical correlation analysis." In: *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 18.3, pp. 905–10.

Webster, J. (1984). "Reducing motion artifacts and interference in biopotential recording". In: *Biomedical Engineering, IEEE Transactions on* 2.12, pp. 22–23.

Wicker, B., C. Keysers, J. Plailly, J. P. Royet, V. Gallese, and G. Rizzolatti (2003). "Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust." In: *Neuron* 40.3, pp. 655–64.

Widmann, A. and E. Schröger (2012). "Filter effects and filter artifacts in the analysis of electrophysiological data." In: *Frontiers in psychology* 3.July, p. 233.

Winn, J. M. (2004). "Variational Message Passing and its Applications". PhD thesis.

Winn, J. and C. M. Bishop (2005). "Variational message passing". In: *Journal of Machine Learning Research* 6, pp. 661–694.

Wu, W., Z. Chen, S. Gao, and E. N. Brown (2011). "A hierarchical Bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG." In: *NeuroImage* 56.4, pp. 1929–45.

Yun, K., K. Watanabe, and S. Shimojo (2012). "Interpersonal body and neural synchronization as a marker of implicit social interaction." In: *Scientific reports* 2, p. 959.