Bayesian Correlated Component Analysis for inference of joint EEG activation

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

[1] Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Kongens Lyngby, Denmark, email: lkai@dtu.dk

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

***Abstract*—We propose a probabilistic generative multi-view model to test the representational universality of human information processing. The model is tested in simulated data and in a well-established benchmark EEG dataset.**

Keywords: Latent variable model, multi-view, EEG, variational inference, canonical correlation analysis

I. Introduction

We are interested in how the human brain solves computational problems such as decoding high level information from given natural stimulus such as e.g. watching a movie. Assuming that the movie brain interaction is jointly optimized we expect a certain amount of universality in the representations and processes used in the brains of subjects watching a given movie. Neuroscience based on movie stimulus has been pursued by Hasson et al. see e.g. [1]. They introduced a correlation approach between anatomically aligned brains. This is based on a rather strong assumption of universality, namely that both the extracted information (what) and representations (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 (here abbreviated CorrCA) was recently proposed by Dmochowski et al. [2]. Within the multivariate framework, a natural relaxation of the strong universality hypothesis, would be to let decoded content (what) be identical between subjects, while their representations, hence, the 'where' be more individual. Such an approach corresponds to analysis by the multivariate technique known as canonical correlation analysis (CCA) [3]. In CCA we search for individual stationary spatial networks with similar temporal activation among subjects. A model incorporating temporal structure and including both joint and individual signal components was developed by Lukic et al. [4]. A probabilistic approach to CCA also including the possibility of both joint and individual components was proposed by Klami et al. [5].

Here we will analyze a probabilistic model inspired by the work of Dmochowski et al. [2], however, with the possibility of learning the degree of universality from data. The latter is implemented in hierarchical Bayesian approach that allows variable degree of non-universality of representations (where) in individual subjects. We illustrate the performance and our approximate inference procedures in both simulation studies and in a benchmark EEG set. The specific contributions of this work are: 1) Formulation of a generative model and inference for Bayesian correlated component analysis (BCorrCA); 2) A principled scheme for inference of correlated components more than two simultaneous subjects; 3) Validation on simulated data and benchmark EEG data.

Authors with $(*)$ made equal contributions to this work

978-1-4799-4149-0/14/$31.00

II. Finding correlated components through eigenvalue decomposition

We first briefly review the two existing multivariate approaches both of which are derived for two view data sets. 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 of 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 $k \neq k'$ [5]. 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 [3]

$$\begin{aligned} \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{w}^{(1)} &= \rho_1\mathbf{w}^{(1)} \\ \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{w}^{(2)} &= \rho_1\mathbf{w}^{(2)}. \end{aligned} \tag{1}$$

where ρ is an eigenvalue of the system. Correlated component analysis (CorrCA) is a related approach with the additional constraint of a single set of weights that works for filtering both data sets. This stronger universality assumptions can also benefit from 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 [2]. In correlated component analysis the weights are thus estimated through a single eigenvalue decomposition [2],

$$\left(\mathbf{R}_{11} + \mathbf{R}_{22}\right)^{-1}\left(\mathbf{R}_{12} + \mathbf{R}_{21}\right)\mathbf{w} = \rho_2\mathbf{w}. \tag{2}$$

In an extended version of this work in preparation, we discuss the robustness of CorrCA to variability in subject weights.

III. Probabilistic Correlated Component Analysis

Inspired by the probabilistic principal component analysis introduced by [6], an approach to CCA was presented in [7] using latent variables. They formulated a probabilistic generative model based on Gaussian distributed common sources, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, mixed to form two noisy observed datasets, $\mathbf{x}^{(m)} \sim \mathcal{N}\left(\mathbf{A}^{(m)}\mathbf{z}, \mathbf{\Phi}^{(m)}\right)$, for $m = \{1, 2\}$, with $\mathbf{\Phi}^{(m)}$ representing the covariance matrix for the observation noise of dataset m. $\mathbf{A}^{(m)} \in \mathbb{R}^{D\times K}$ signifies the mixing matrix[1], where each of the K columns represents the mixing of one source. To avoid discrete model selection in the related case of probabilistic PCA [8] introduced automatic relevance determination (ARD), with a Gaussian distribution for each column in the mixing matrix, $\mathbf{A}^{(m)} \sim \prod_k^K \mathcal{N}\left(\mathbf{A}_k^{(m)}|\mathbf{0}, \alpha_k^{-1}\right)$, which are regulated by the gamma distributed hyperparameter, $\boldsymbol{\alpha} \sim \prod_k^K \mathcal{G}a(\alpha_k|a_0, b_0)$.

These methods to probabilistic PCA and CCA has lead to different approaches to a Bayesian CCA e.g. [5], [10], [11] and group factor analysis (GFA) [12], the first practical multi-view generalization of Bayesian CCA. The novelty in these methods is mainly related to how they approximate the posterior distribution for the latent sources, using a full Bayesian treatment most employ variants of variational inference, an approach we also follow here.

Variational inference approximates the posterior distribution by a completely factorised variational distribution which is then optimized to match the posterior typically using the Kullback-Leibler divergence as a measure of the dissimilarity. The resulting algorithm consists of updating the lower bound with respect to each variable in turn such as expectation maximisation. In this setting the distributions from the exponential family often simplifies computation and provides for conjugate relationships between the prior and posterior distributions [13], [14].

The Bayesian correlated component analysis (BCorrCA) model proposed here introduces a relationship between the mixing matrices of the views, by including the common latent variable, $\mathbf{U}$, representing the mean mixing matrix across all datasets and the ARD variable λ which tunes how close the individual $\mathbf{A}^{(m)}$'s are to $\mathbf{U}$;

$$\mathbf{U} \sim \prod_k^K \mathcal{N}\left(\mathbf{v}_k|\mathbf{0}, \alpha_k^{-1}\right) \tag{3}$$

$$\mathbf{A} \sim \prod_m^M \prod_k^K \mathcal{N}\left(\mathbf{a}_k^{(m)}|\mathbf{u}_k, \lambda^{-1}\right), \quad \text{for } M \geq 2 \tag{4}$$

$$\lambda \sim \mathcal{G}a(a_0, b_0) \tag{5}$$

The attained updates are similar to the ones obtained in [10], with added regularisation of the mixing matrices, $\mathbf{A}$, by λ and $\mathbf{U}$. The result is an algorithm, which can implement both independent mixing matrices as in CCA (with a small λ), or completely aligned matrices as in CorrCA (with a large λ). Importantly it generalises in a straightforward manner to an arbitrary number of parallel data views.

[1] Authors use different letters for the mixing matrix. Most Bayesian models use the notation $\mathbf{W}$, probably stemming from [8], but as this letter is also used to define the demixing matrix, we choose here to use $\mathbf{A}$ as employed in [9].

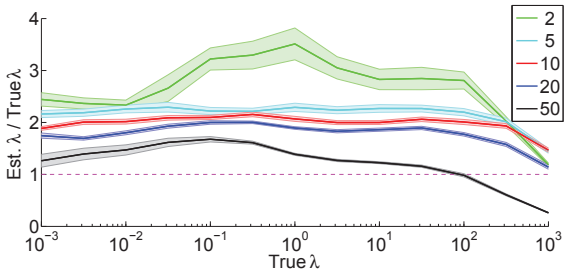


Fig. 1: Estimation of the similarity between the mixing matrices in simulated data with different number of views. The opaque area marks the standard error of the mean. The similarity is regulated through the parameter λ, and estimated using BCorrCA. The simulations were conducted with a single hidden source, 100 repetitions and a SNR of 3.

IV. Performance on simulated data

Simulation Design

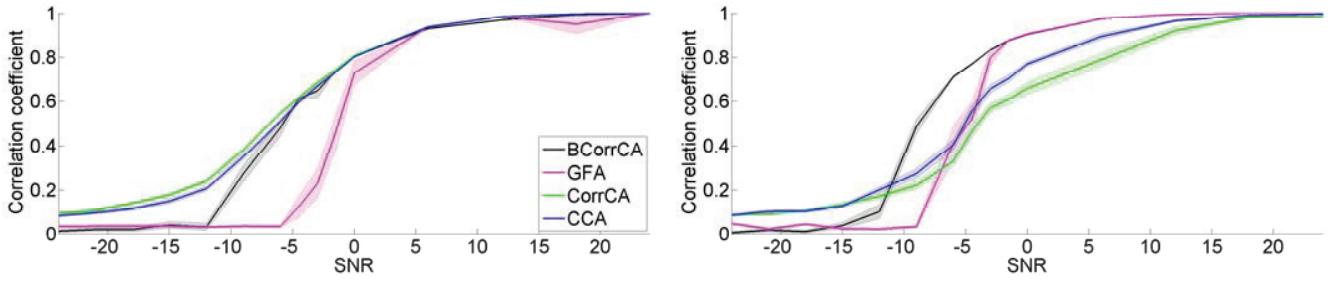
To validate and quantify the performance of BCorrCA, data was generated with a varying similarity between the mixing matrices by changing the 'true' λ. From the model definition we get that $\mathbf{X}^{(m)} = \mathbf{A}_{\text{true}}^{(m)}\mathbf{Z} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, 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. The mixing matrix is generated by $\mathbf{A}_{\text{true}}^{(m)} = \mathbf{U} + \boldsymbol{\delta}^{(m)}$, with $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{\alpha}^{-1})$ and $\boldsymbol{\delta}^{(m)} \sim \mathcal{N}(0, \lambda^{-1})$.

We have used up to four hidden sources, generated as in [5], for direct comparability with this work. Here we will mainly focus on the simple case of one hidden source corresponding to the data being generated from one sinusoid and additive noise, a more elaborate investigation pending a detailed report. For comparative analysis and benchmarking, the performances of BCorrCA, CorrCA, CCA, and GFA were estimated on the same data. For each combination of conditions either 20 or 100 datasets were randomly generated and the average correlation coefficient between the inferred source and the true source was chosen as the measure of performance.

CorrCA and CCA are designed to handle two views at a time. In case of multiple views we employed the scheme for combination proposed in [2], i.e., the views are concatenated in time so that all pairwise combinations are compared. This method has the disadvantage that the number of samples in the concatenated data scales as $M(M-1)$. The algorithms were tested at varying levels of SNR, number of views, M, and similarity between the true mixing matrices of each view. In each test the data had six dimensions and the number of observations was set to 500. When varying the number of views we used a total of 5,000 samples divided equally among the views.

Results

Figure 1 presents the results of simulations to test that BCorrCA can infer the correct level of view similarity (λ). Here the similarity between the true mixing matrices is varied by the 'true' λ parameter. It is seen that BCorrCA is in fact able to estimate this parameter's variation through the entire range. We find a small tendency to overestimate, which might stem from an interaction with the other ARD parameter, α, and an



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

Fig. 2: Performance of BCorrCA, GFA, CorrCA and CCA on simulated data measured by mean correlation coefficient and standard error of the mean with respect to the true source. In both subfigures the performance is tested under different levels of SNR, with either 2 or 5 views and 20 repetitions. The true weights are nearly equal in **(a)** with $\lambda = 10^3$ and dissimilar in **(b)** with $\lambda = 10^3$.

interaction with the number of views that is most pronounced when having only one hidden source.

The results from the tests on the simulated data can be seen in figure 2. We find that for high SNR levels all algorithms perform similar, but as the noise levels increase the latent models quickly drop towards zero correlation though BCorrCA do so less steeply and can perform at lower levels of SNR compared to GFA. This quick drop is caused by the Bayesian model's affinity to bias, hence to choose a 'zero-source' solution as the cost of a poor estimation gets too high. More aggressive priors could relieve this issue.

Figure 2 also shows how performance improves as the number of available views increases. We see that CCA and CorrCA actually perform worse, when the number of views increase and their mixing matrices are non-equal. With identical 'true' matrices the algorithms perform the same as BCorrCA. The increased performance stems from having more instances of the signal and then be able to average out noise. For the latent variable models two things are evident: BCorrCA outperforms GFA at low SNR and increasing the number of views is beneficial, as it increases the correlation with the true source even though the number of observations do not increase.

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 BCorrCA.

V. Validation on benchmark EEG from an audio-visual selective attention task

For the purpose of evaluating the algorithm on real EEG, we use a well-documented dataset from the auditory-visual attention shift study by the Swartz Center [15].

Experimental design

The paradigm was designed to test the effect of ageing on neural responses of the auditory and visual stimuli. The auditory stimuli were 100-ms duration 550-Hz (target) and 500-Hz (non-target) sine-wave tones and the visual stimuli were light-blue (target) and dark-blue (non-target) filled squares presented for 100 ms on a light-grey background. During the experiment the subject was notified to either attend to visual or auditory stimuli and press a button when detecting a relevant target.

Pre-processing

The continuous EEG was recorded from 33 channels of the international 10-20 system at 250 Hz from 49 subjects (5 discarded for missing events). The EEG was re-referenced to the left mastoid and segmented into response-locked epochs with 200 ms pre- and 900 ms post-response time. Epochs with values higher than $70\mu V$ or exceeding 5 standard deviations were automatically rejected to remove eye artefacts and electrical drift. The remaining epochs were bandpass filtered using a windowed sinc-filter with passband frequencies 0.1-40 Hz and baseline-corrected in relation to the 200 ms pre-response interval. The "infomax" ICA algorithm [16] was then used on the concatenated epochs to find the de-mixing matrix and isolate independent components containing eye artefacts. The "fully" automatic algorithm ADJUST [17] was used to identify noisy components which were then conservatively inspected manually and removed from the EEG. The EEG was then normalised with respect to its total power.

Finding correlated components

Because BCorrCA is sensitive to temporal misalignment the views used in this test was chosen based on the correlation between the averaged ERP of channel P3 for view 8 and the rest of the views. The six views with highest correlation (including self) were chosen and cropped to contain the same number of epochs (n=394). To test the reliability of the algorithm 100 epochs were randomly drawn from each view and concatenated to create 6 views of "continuous" EEG to be analysed in BCorrCA. The procedure was repeated 100 times in order to compute mean and standard error of the mean.

Results

The variance of and difference between view-specific filters represents an estimate of the similarity between these. The same applies to the time series components but the algorithm only computes a common component for all the views, however, by constructing the backward model filters, $\mathbf{W}$, the view-specific components can be found by $\mathbf{z}^{(m)} = \mathbf{w}^{(m)T}\mathbf{X}^{(m)}$. In table I we see that the average within-view variance for both filters and time series are significantly larger than the average between-view variance which suggests a high amount of universality in the neural representations. The same is illustrated in figure 3 by the relatively small standard deviation (mesh) compared to the average filter. Figure 4 shows in the small error a high level of reliability for both methods but the neural response is represented much stronger in BCorrCA.

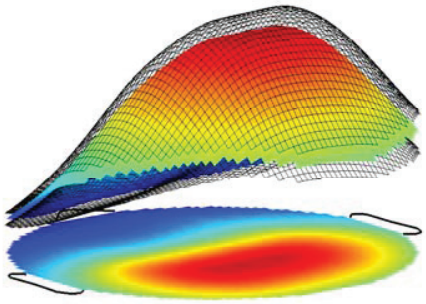


Fig. 3: Grand average filter projection and standard error (mesh). The filter is illustrated both as a 2- and 3-dimensional structure to show the between-view variation. Positive weights are located in the parieto-occipital region

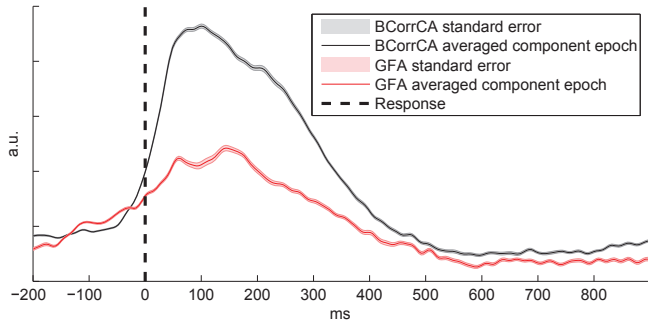


Fig. 4: Average ERP of component from BCorrCA and GFA. Mean and standard error is computed from 100 repetitions of 100 randomly selected epochs from each view

TABLE I: Variance of time series components and filters

	Within-view	Between-view
Filters	0.046	0.018
Time series	0.095	0.010

VI. Discussion and Conclusion

During the past decade research in social neuroscience has shifted from being inherently single person studies of people observing others towards multi-way interaction between multiple persons [18], which calls for methods that are able to adapt to the level of universality in neural representations across brains.

The probabilistic implementation of correlated component analysis presented here provides a new approach to the extraction of shared representations and information. Tests on artificial data showed that multiple views improve the extraction of shared signals, even when the total amount of observations were kept the same. Direct inference of shared response in the face of intersubject variability of representation enables new methods for analysis in experiments with simultaneous stimulation of groups of subjects. Instead of analysing pairwise correlations of subject response the methods proposed here allow us to infer correlated attention and other joint activity in large cohorts. Our analysis of the auditory-visual attention shift EEG dataset showed the expected response post-motor potential following the key press and the response is markedly stronger using BCorrCA compared to GFA. In Table I we list the total variation within and between subjects (views). Note, the within subject variability (i.e., variability across electrode weights) is larger than the between subject variability. The estimated response time course is highly universal for the subjects entering the analysis.

Acknowledgements

This work is funded by Lundbeckfonden via CIMBI and by the Danish Strategic Research Council via the Neuro24/7 project.

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] 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.

[3] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[4] Ana S. Lukic, Miles N. Wernick, Lars Kai Hansen, Jon Anderson, and Stephen 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.

[5] Arto Klami, Seppo Virtanen, and Samuel Kaski, "Bayesian canonical correlation analysis," *Journal of Machine Learning Research*, vol. 14, pp. 965–1003, 2013.

[6] 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.

[7] Francis R Bach and Michael I Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

[8] Christopher M Bishop, "Variational principal components," 1999.

[9] 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.

[10] Chong Wang, "Variational bayesian approach to canonical correlation analysis," *Neural Networks, IEEE Transactions on*, vol. 18, no. 3, pp. 905–910, 2007.

[11] 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.

[12] Seppo Virtanen, Arto Klami, Suleiman A Khan, and Samuel Kaski, "Bayesian group factor analysis," *arXiv preprint arXiv:1110.3204*, 2011.

[13] Christopher M Bishop et al., *Pattern recognition and machine learning*, vol. 1, Springer New York, 2006.

[14] Kevin P Murphy, *Machine learning: a probabilistic perspective*, The MIT Press, 2012.

[15] R Ceponiene, M Westerfield, M Torki, and J Townsend, "Modality-specificity of sensory aging in vision and audition: evidence from event-related potentials.," *Brain research*, vol. 1215, pp. 53–68, June 2008.

[16] Anthon J. Bell and Terrence J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution.," *Neural computation*, vol. 7, no. 6, pp. 1129–59, Nov. 1995.

[17] Andrea Mognon, Jorge Jovicich, Lorenzo Bruzzone, and Marco Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features.," *Psychophysiology*, vol. 48, pp. 229–240, July 2010.

[18] 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.