Characterizing Unknown Events in MEG Data with Group Factor Analysis

Sami Remes[1], Arto Klami[2], and Samuel Kaski[1,2]

[1] Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University, Finland
[2] Helsinki Institute for Information Technology HIIT,
Department of Computer Science, University of Helsinki, Finland

Abstract. Many current neuroscientific experiments can be seen as data analysis problems with two or more data sources: brain activity and stimulus features or, as in this paper, activity of two brains. These setups have been analyzed with Canonical Correlation Analysis or its multiple-source probabilistic extension Group Factor Analysis, which capture statistical dependencies between the data sources in correlating components. We relax the assumption of global correlations and search for correlating signals related to discrete events. The assumption is that the sources correlate only during events with known timings, inferred from a stimulus stream for instance, but the type or nature of each event is not known. The unsupervised modelling of the events can then be viewed as a generalization of conditional averaging. We apply the model on two-person MEG measurements, in a demonstration task of identifying which of the two persons utters a word.

Keywords: Bayesian modelling, canonical correlation analysis, factor analysis, latent variable models, magnetoencephalography, MEG

1 Introduction

Recently, neuroscientific experiments have been performed with increasingly more complex, natural stimuli ranging from speech and music to movies, which poses challenges for conventional data analysis. A practical solution is to include more data, from multiple sources, into the modelling phase. A case where we would naturally be interested in finding some dependencies between multiple sources is that of studying common responses of multiple subjects, each interpreted as a separate data source, exposed to the same stimuli (Kauppi et al., 2010). Another natural multi-source setup treats the brain activity and a feature representation of the stimulus as two data sources, using the dependencies between them to characterize responses to complex natural stimuli for which no clearly defined trials, epochs or events can be specified (Ylipaavalniemi et al., 2009).

One of the most typical approaches to consolidate information from two data sources is canonical correlation analysis (CCA; Hotelling, 1936; for more recent reviews see e.g. Hardoon et al., 2004 and Klami et al., 2013). CCA finds linear combinations of the features of each data set such that the correlation between these combinations is maximized over the whole experiment. In recent years CCA has gained popularity in neuroscience, for all brain imaging techniques. Koskinen et al. (2012) used a Bayesian mixture of CCAs model to find dependencies between the brain activity measured by magnetoencephalography (MEG) and continuous speech. Campi et al. (2013) used a non-linear CCA variant for analyzing MEG data from an experiment in which subjects were given the same naturalistic stimulus consisting of tactile, auditory and visual blocks. Correa et al. (2010) applied CCA to investigate how EEG relates to fMRI measurements. Deleus and Van Hulle (2011) studied functional connectivity in fMRI using a variant of CCA, which generalized the method for more than two data sources. Another multi-view generalization called group factor analysis (GFA) was presented by Virtanen et al. (2012).

The CCA-based approaches assume that the correlations between the data sources last over the whole experiment. Often however, the observed correlations might be weak and temporally localized. For example, it is unreasonable to expect a strong correlation between the brain activity and naturalistic auditory stimulation during periods of low stimulus activity, or when the subject is not paying attention to the stimulation. Instead, we should expect that the degree of correlation varies during the experiment and, in particular, the correlations are stronger during some temporally localized periods of interest. In this work we propose a novel extension to the CCA-family that purposefully seeks to find such temporally localized correlations.

The proposed solution is based on the GFA model (Virtanen et al., 2012), extending it to support temporally localized correlations. We assume that the data sources are independent except during short temporal windows associated with events in the experiment. We assume that the timing of these events is known, at least approximately, but that there can be different types of events that trigger different kinds of correlations during the events. From this perspective the proposed model can be interpreted as a generalization of the classical MEG analysis, where the response to events is found by averaging over repetitions of the same event (Luck, 2005). Even in a very complex naturalistic stimulus some events can be extracted for example from the stimulus recordings, but it is often not obvious how to group them together for averaging. Our model learns such a grouping automatically by assigning with each event a specific type of correlation, and it is also robust for imprecise timing of the events. In a sense, it allows estimating the average responses even when we cannot say in advance to which phenomena the events correspond, while extending the definition of an interesting event to one that elicits a specific type of correlation between the data sources.

Our main goal is in describing the computational model. Nevertheless, we still demonstrate it on a recent naturalistic MEG recording. We apply the model on a two-person MEG setup where the brain activity of two individuals interacting with each other are being recorded at the same time, with the instrumentation setup of Baess et al. (2012). The experimental setup is that of a word game; the two persons interact by alternatingly uttering isolated words that emerge to form a story, and we treat the MEG recordings of the individual subjects as two data sources. We consider the beginnings of the word utterances as events and obtain imprecise estimates of their timing by monitoring the microphone signal. The model then assumes that the MEG signals correlate during these events but not outside them, and we show that it automatically recognizes two different types of correlations between the sources. These types correspond to the obvious dichotomy of which subject utters the word, which provides us a ground truth for validating the model.

2 Model

2.1 Group Factor Analysis

Group factor analysis (GFA) is an extension of Bayesian CCA (Klami et al., 2013) to multiple data sources. Alternatively, it can be seen as an extension of factor analysis that treats the data sources similarly to how factor analysis treats individual variables. The model was introduced by Virtanen et al. (2012), and it is briefly summarized below.

Letting $\mathbf{y}_i^{(m)} \in \mathbb{R}^{D_m}$ denote the ith sample of the mth data set, the model is given by

$$\mathbf{y}_i^{(m)}|\mathbf{W}, \mathbf{z}_i, \tau_m \sim \mathcal{N}\left(\mathbf{W}^{(m)}\mathbf{z}_i, \tau_m^{-1}\mathbf{I}\right), \tag{1}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The data are modeled with unknown latent variables $\mathbf{z}_i \in \mathbb{R}^K$ corresponding to K factors, which are then mapped to the observation space with the linear projections $\mathbf{W}^{(m)}$ specific to each source. The latent variables are shared between all sets and can hence model correlations between the sources.

Typically the data sources have also non-trivial variation independent of the other sources, which cannot be modeled with the spherical noise model $\boldsymbol{\Sigma} = \tau_m^{-1}\mathbf{I}$. GFA models them with some of the K factors by controlling the values of $\mathbf{W}^{(m)}$ with a specific type of sparsity constraint; if some factor k is independent of the source m, we want $\mathbf{w}_k^{(m)}$ (the kth column of $\mathbf{W}^{(m)}$) to be zero. The desired structure is achieved with the automatic relevance determination (ARD; MacKay, 1995) prior

$$\mathbf{w}_k^{(m)}|\ \alpha_k^{(m)} \sim \mathcal{N}\left(\mathbf{0}, \left(\alpha_k^{(m)}\right)^{-1}\mathbf{I}\right), \tag{2}$$

where each of the α-parameters has independently (an almost non-informative) gamma prior $\alpha_k^{(m)} \sim \mathcal{G}(10^{-14}, 10^{-14})$. If for some m and k the parameter $\alpha_k^{(m)}$ becomes large, then the corresponding column of $\mathbf{W}^{(m)}$ is almost zero and thus the kth factor in the model is irrelevant in explaining the variation in dataset m.

To provide an alternative viewpoint, we denote $\boldsymbol{\alpha}^{(m)} = \left(\alpha_1^{(m)} \ldots \alpha_K^{(m)}\right)^T$. Then the matrix

$$\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}^{(1)} \ldots \boldsymbol{\alpha}^{(M)}\right)^T$$

acts as a kind of a loading matrix indicating which factors load on which datasets. This shows how GFA generalizes regular factor analysis to model relationships between the data sources instead of individual variables.

To further clarify the model, we provide a mapping from the mathematical notation above to the example application studied in this paper. We have $M = 2$ sources that correspond to MEG recordings of two subjects. Each sample is a time instance and the features are the MEG gradiometer sensors. The factors for which $\alpha_k^{(m)}$ is small for both m model correlations between the two brains, whereas the remaining factors model correlations between the gradiometer sensors of each subject. The model automatically learns, using the ARD prior, how many factors to use for both tasks.

2.2 Event-related Extension

As mentioned in Introduction, GFA and other CCA-based models assume that the correlations persist over all of the samples, that is, over the whole experiment. Next we introduce an extension of GFA that relaxes this assumption, and instead assumes that the data sources correlate only during pre-specified events. We assume that the timing of these events is known, but that there can be different types of events that induce different kinds of correlations between the sources. The goal is to learn linear mappings describing the nature of these different types of correlations, while simultaneously inferring for each event to which category it belongs.

We assume that outside the events the data sources are independent but that there can still be correlations between the different features within each data source. We model these with K factors

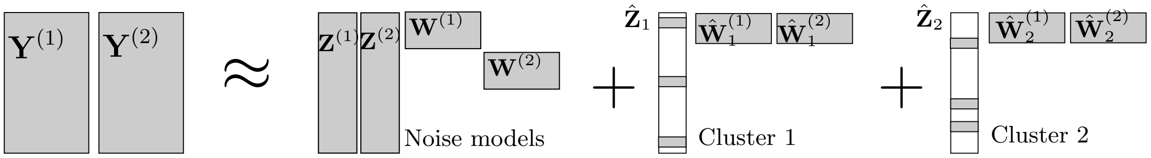

Fig. 1. Factorization of the data matrices in the proposed model, with gray color indicating non-zero elements. For the samples that do not belong to any event the two data sources are independent and modeled with the noise models. The samples within the events choose one of the event categories and use the factors specific to that category for modeling correlation during the event.

that use source-speficic latent variables $\mathbf{z}_i^{(m)}$. The correlations between the sources are modeled with E event categories, each capturing the correlations with $\hat{K}$ additional factors that use latent variables $\hat{\mathbf{z}}_i^e$ shared between the sources but specific to the event category. The resulting model is

$$\mathbf{y}_i^{(m)} \sim \mathcal{N}\left(\mathbf{W}^{(m)}\mathbf{z}_i^{(m)} + \sum_{e=1}^{E} \gamma_i^e \hat{\mathbf{W}}_e^{(m)} \hat{\mathbf{z}}_i^e, \tau_m^{-1}\mathbf{I}\right) \tag{3}$$

where γ_i^e is a binary variable indicating whether event e is active in sample i. For every sample that is outside the events we set $\gamma_i^e = 0$ for all event categories, whereas for the remaining samples we require that the set of consecutive samples during a specific event belong to exactly one event category. Finally, we set $\mathbf{z}_i^e = 0$ if $\gamma_i^e = 0$. Figure 1 illustrates the resulting factorization for a setup where we have $M = 2$ data sources and $E = 2$ event categories.

For controlling the complexity of the model we again use an ARD prior for the source-specific loading matrices $\mathbf{W}^{(m)}$; the model hence learns how many factors it needs for modeling the variation independent of other sources. The factors used for modeling the correlations during the events use the same prior as regular GFA, which allows learning which specific sources reflect the correlation.

3 Inference

For inference we adopt variational Bayes (VB). We write the factorized approximate posterior as

$$q(\Theta) = \prod_{m=1}^{M}\left(q(\boldsymbol{\alpha}^{(m)})q(\tau_m)\prod_{d=1}^{D_m} q(\mathbf{w}_d^{(m)})\prod_{i=1}^{N} q(\mathbf{z}_i^{(m)})\right)\prod_{e=1}^{E}\left(\prod_{i=1}^{N} q(\hat{\mathbf{z}}_i^e)\prod_{m=1}^{M} q(\hat{\boldsymbol{\alpha}}_e^{(m)})\prod_{d=1}^{D} q(\hat{\mathbf{w}}_d^e)\right), \tag{4}$$

where $D = \sum_{m=1}^{M} D_m$. With this factorization we will update the noise models (one for each data source) and the E shared event components separately, which allows for simpler updates following very closely those for Bayesian CCA by Klami et al. (2013).

Our extended model introduces a new variable $\boldsymbol{\gamma}$, and next we provide the necessary derivations for updating that parameter within the overall VB algorithm updating the whole model. For interpretability we do not provide a full posterior approximation for γ_i^e but instead use type-II maximum likelihood to provide binary decisions for each sample. To derive the objective function, we need to write the variational lower bound

$$\mathcal{L}(\boldsymbol{\gamma}) \propto \langle \log p(\mathbf{Y}|\Theta, \boldsymbol{\gamma})\rangle - \left\langle \frac{\log q(\hat{\mathbf{Z}}|\boldsymbol{\gamma})}{\log p(\hat{\mathbf{Z}}|\boldsymbol{\gamma})}\right\rangle \tag{5}$$

as a function of $\boldsymbol{\gamma}$, omitting the constant terms. The relevant terms of the first part are

$$\langle \log p(\mathbf{Y}|\Theta, \boldsymbol{\gamma})\rangle \propto \sum_{i=1}^{N} \sum_{m=1}^{M} \left\langle -\frac{1}{2}\tau_m \left\| \hat{\mathbf{e}}_i^{(m)} - \sum_{e=1}^{E} \gamma_i^e \hat{\mathbf{W}}_e^{(m)} \hat{\mathbf{z}}_i^e \right\|^2 \right\rangle , \tag{6}$$

where we have defined $\hat{\mathbf{e}}_i^{(m)} = \mathbf{y}_i^{(m)} - \mathbf{W}^{(m)}\mathbf{z}_i^{(m)}$. Noting that $\gamma_i^k \gamma_i^l = \delta_{kl}$ and $(\gamma_i^e)^2 = \gamma_i^e$, (6) is a linear function of $\boldsymbol{\gamma}$. The second part of the lower bound comes from the prior

$$\left\langle \log p(\hat{\mathbf{Z}}|\boldsymbol{\gamma})\right\rangle \propto \sum_{i=1}^{N} \sum_{e=1}^{E} \left\langle -\frac{1}{2} \left\| \hat{\mathbf{z}}_i^e \right\|^2 | \gamma_i^e \right\rangle , \tag{7}$$

where the expectation depends on the corresponding γ_i^e; $\mathbf{z}$ is Gaussian when $\gamma = 1$, and a delta spike at zero when $\gamma = 0$. Since we assume that each event should be allocated to exactly one category, we couple all the samples within one event to have identical γ_i^e. The optimization problem is easy to solve due to linearity of (5) as a function of $\boldsymbol{\gamma}$. In the experimental section we demonstrate a scenario where we further assume that the event categories are equally common. This can be expressed as a linear constraint, which allows using generic integer linear programming solvers, such as `lp_solve` (Berkelaar et al., 2004) for finding the optimal values.

4 Experiments and results

We use data from an experiment where pairs of participants were simultaneously recorded at separate MEG devices and an audio-visual link was provided for communication between them (Baess et al., 2012). The pairs were instructed to play a word game where they took turns in saying one word at a time and the words were supposed to make up a sensible story. The length of the games ranged from 88 to 172 words between the different pairs, or approximately 5 minutes. The MEG data were preprocessed using the signal space separation method (Taulu et al., 2004), and the data were downsampled to 67 Hz from 1 000 Hz and high-pass filtered at 3 Hz.

We define an *event* here as a word spoken by either participant. Outside these events we assume no correlation between the brain activities. In total the events correspond to about 15% of the duration of the experiments, which implies that traditional methods seeking for global correlations should have difficult time capturing these temporally localized effects.

We run the model with $K = 25$, and two event categories with 4 components each. Thus most of the variation in the data will be captured by the subject-specific noise models and additional activity during the events then by one of the shared models. The amount of components was chosen based on a heuristic given by classical PCA: roughly 25 principal components account for about 90 percent of the variance. For each pair, we chose the best model by comparing the variational lower bounds of 50 random initializations.

We expect that the two event categories should correspond to the speaker identities for each individual word; these should obviously elicit different kind of correlations between the subjects. For validating whether the model learns these categories more accurately than simpler alternatives, we compare it with linear dimensionality reduction followed by k-means clustering of the samples within the events. We cluster the individual samples within the events into two clusters and then do majority voting for each event window. For the dimensionality reduction step we tried two alternatives, PCA that assumes the sources to be independent and GFA that attempts to model

Table 1. Clustering accuracy of words with respect to the speaker. For each pair the best method is written in boldface; see text for more details.

Pair	Number of events	Proposed model	GFA + k-means	PCA + k-means
A	102	**0.90**	0.55	0.57
B	104	**0.92**	0.51	0.51
C	172	**0.62**	0.55	0.59
D	116	**0.76**	0.61	0.59
E	88	**0.96**	0.50	0.60
F	170	**0.75**	0.57	0.51
G	160	**0.81**	0.58	0.52

global correlations between them. In effect, these comparison methods can be thought of as attempts to de-couple the elements of the proposed model into separate elements, and hence the comparison reveals the advantage of learning the joint model at once.

The order of clusters being unidentifiable, we label the clusters solely for the purpose of measuring the accuracy by the label with most matches within a cluster. For the proposed model the mean accuracy over 7 pairs of participants is 0.82 (Table 1); three pairs get over 0.9 accuracy but for all pairs it is significantly better than chance ($p < 0.005$, all pairs, random permutation test). The comparison methods solve the task considerably worse, irrespective of the dimensionality reduction technique. The improvements are due to our method learning the background noise simultaneously with the events and allowing the temporal profile to vary between each individual event.

5 Conclusions

CCA-based methods have been used to study problems in neuroscience where data is collected from multiple sources. In this work, we presented a new CCA model that looks for temporally local correlations during pre-defined events, and thus in a way generalizes the usual event-based analysis prevalent in the MEG community. We used variational Bayesian methods for computing approximate posterior distribution of the model, and demonstrated the model on a two-person MEG data set. Our results show that the proposed model is able to separate the two event types in an unsupervised manner, and greatly outperformed straight-forward application of k-means clustering. Neuroscientific interpretation of our results is on-going.

Acknowledgments

This work was financially supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, grant no 251170; grant no 266969; BrIAn project of LASTU research program, no 135198). We thank Prof. A. Hyvärinen and Prof. R. Hari, the leaders of the BrIAn project, for fruitful collaboration. The dual-MEG data set was collected in Brain2Brain project funded by the European Research Council (Advanced Grant #232946 to Riitta Hari, Brain Research Unit, O.V. Lounasmaa Laboratory, Aalto University). We thank P. Baess, R. Hari, T. Himberg, L. Hirvenkari, V. Jousmäki, A. Mandel, J. Mäkelä, J. Nurminen, L. Parkkonen, and A. Zhdanov for the possibility to use the anonymized dual-MEG data, A. Mandel for preprocessing of the data set, and L. Hirvenkari for the stimulus timing. Computational resources were provided by Aalto Science-IT project.

Bibliography

Pamela Baess, Andrey Zhdanov, Anne Mandel, Lauri Parkkonen, Lotta Hirvenkari, Jyrki P Mäkelä, Veikko Jousmäki, and Riitta Hari. MEG dual scanning: a procedure to study real-time auditory interaction between two persons. *Frontiers in Human Neuroscience*, 6, 2012.

Michel Berkelaar, Kjell Eikland, and Peter Notebaert. lp_solve 5.5, open source (mixed-integer) linear programming system. Software, May 1 2004. URL http://lpsolve.sourceforge.net/5.5/.

Cristina Campi, Lauri Parkkonen, Riitta Hari, and Aapo Hyvärinen. Non-linear canonical correlation for joint analysis of MEG signals from two subjects. *Frontiers in neuroscience*, 7, 2013.

Nicolle M Correa, Tom Eichele, Tülay Adalı, Yi-Ou Li, and Vince D Calhoun. Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI. *Neuroimage*, 50(4):1438–1445, 2010.

Filip Deleus and Marc M Van Hulle. Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis. *Journal of Neuroscience methods*, 197(1): 143–157, 2011.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

Jukka-Pekka Kauppi, Iiro P Jääskeläinen, Mikko Sams, and Jussi Tohka. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. *Frontiers in neuroinformatics*, 4, 2010.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013.

Miika Koskinen, Jaakko Viinikanoja, Mikko Kurimo, Arto Klami, Samuel Kaski, and Riitta Hari. Identifying fragments of natural speech from the listener's MEG signals. *Human brain mapping*, 2012.

Steven J Luck. *An introduction to the event-related potential technique (cognitive neuroscience)*. A Bradford Book, 2005.

David JC MacKay. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

Samu Taulu, Matti Kajola, and Juha Simola. Suppression of interference and artifacts by the signal space separation method. *Brain Topography*, 16:269–275, 2004.

Seppo Virtanen, Arto Klami, Suleiman A Khan, and Samuel Kaski. Bayesian group factor analysis. In Neil Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR W&CP*, pages 1269–1277. JMLR, 2012.

Jarkko Ylipaavalniemi, Eerika Savia, Sanna Malinen, Riitta Hari, Ricardo Vigário, and Samuel Kaski. Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. *NeuroImage*, 48(1):176 – 185, 2009.