Variational Bayesian Mixture of Robust CCA Models

Jaakko Viinikanoja, Arto Klami, and Samuel Kaski

Aalto University School of Science and Technology
Department of Information and Computer Science
Helsinki Institute for Information Technology HIIT
http://www.cis.hut.fi/projects/mi/

Abstract. We study the problem of extracting statistical dependencies between multivariate signals, to be used for exploratory analysis of complicated natural phenomena. In particular, we develop generative models for extracting the dependencies, made possible by the probabilistic interpretation of canonical correlation analysis (CCA). We introduce a mixture of robust canonical correlation analyzers, using t-distribution to make the model robust to outliers and variational Bayesian inference for learning from noisy data. We demonstrate the improvements of the new model on artificial data, and further apply it for analyzing dependencies between MEG and measurements of autonomic nervous system to illustrate potential use scenarios.

Keywords: Bayesian data analysis, canonical correlation analysis, data fusion, latent variable models, robust models.

1 Introduction

Noisy estimates of human brain activity can be obtained with several measurement techniques, such as functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG). Given a controlled experiment, even relatively simple approaches can shed light on brain functions. For example, linear regression from brain activity to stimulus covariates can reveal which brain regions are related to the task at hand. For uncontrolled experiments, such as analysis of the brain functions in natural environments, the classical approaches, however, fall short since there are no simple covariates available and unsupervised analysis cannot separate the relevant variation from the rest.

A recent approach to tackling the problem is to consider correlations between the brain activity measurements and multivariate vectorial representations of the stimulus [6,17]. This allows using the stimulus still as a supervision signal, despite the representation not being condensed as simple covariates. Instead, we need to assume the stimulus representation contains noise just like the brain activity measurements do. The actual signal is separated from the noise by making one simple assumption: Statistical dependencies between the brain activity and the stimulus must be related to processing the stimulus, while independent variation in either signal should be seen as structured noise.

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 370–385, 2010.

Canonical correlation analysis (CCA) is a standard approach for finding correlations between two multivariate sources (see, *e.g.*, [9]), and is the method used in [17] for analysis of fMRI data under natural stimulation. CCA has, however, several limitations that make it suboptimal in practical applications. It assumes the signals are stationary, does not come with an easy way of estimating the number of correlated components, is not robust against outliers, and estimating the reliability of the components is difficult. Building on the probabilistic interpretation of CCA, we have earlier introduced a model that removes the stationarity assumption through mixture modeling and automatically learns the model complexity from data [10]. The earlier model, however, has practical limitations that prevent using it for neuroinformatics applications. In this article we improve the model further, by introducing more efficient and more easily interpretable inference procedure, and by making the model robust to outliers.

The model in [10] used posterior sampling for inference, and was formulated as a Dirichlet process mixture for estimating the model complexity, which makes the model suitable for small sample sizes but the inference becomes very inefficient for large data sets. In particular, the model does not have a fully conjugate prior, and hence less efficient sampling strategies need to be applied. Consequently, applying the model for analysis of MEG data would be beyond computationally feasible by some orders of magnitude. Furthermore, neuroscientific interpretation of the results of such a model would be difficult since the posterior sampler returns a set of results that need to be processed further for conclusive summaries. In this paper we solve both of these issues by switching to variational inference, which results in highly efficient optimization and also makes interpretation more straighforward since the approach is deterministic, while retaining the mixture capability and automatic relevance determination prior for inferring the number of correlating components.

The robustness to outliers is obtained by replacing the generative assumption of Gaussian noise by that of Student's t-distribution, modeled as a scale-mixture [11]. Similar representation was earlier used in the robust CCA variant of [2], but they only sought a maximum likelihood estimate for the model parameters instead of considering the full posterior distribution. Robust variational inference has earlier been presented only for simpler projection methods such as robust PCA [8,12] and robust factor analysis [5], using two different alternative approximations that have not been compared earlier. We show that there is no noticeable difference in accuracy or computational complexity between the two alternatives.

We illustrate the technical properties of the model using artificial data, showing the improvements in a set of experiments. We also demonstrate what kind of real analysis scenarios the model is useful for, by using it to learn dependencies between brain oscillations and autonomic nervous system (ANS) response under emotional sound stimuli. The emotions of the user are strongly visible in ANS but only vaguely in MEG, and correlations between these two pinpoint possible hypotheses on what part of the variation in the signals captured by MEG might be related to the emotions. In brief, the ANS measurements can be considered as noisy descriptions of the stimulus itself. The model is shown to find interpretable clusters that capture much stronger correlations than stationary analysis could. The model also outperforms the alternative of first clustering the data and then applying CCA separately for each of the clusters [7].

2 Bayesian CCA

The probabilistic CCA (PCCA) [3] is a generative probabilistic model for two multi-dimensional data sources $\mathbf{X}_1 = [\mathbf{x}_{11}, \ldots, \mathbf{x}_{1N}]$ and $\mathbf{X}_2 = [\mathbf{x}_{21}, \ldots, \mathbf{x}_{2N}]$. The model is written as

$$\begin{aligned} \mathbf{t}_n &\sim \mathcal{N}(\mathbf{t}_n|0, \mathbf{I}_D) \\ \mathbf{x}_{1n}|\mathbf{t}_n &\sim \mathcal{N}(\mathbf{x}_{1n}|\mathbf{W}_1\mathbf{t}_n + \boldsymbol{\mu}_1, \boldsymbol{\Psi}_1) \\ \mathbf{x}_{2n}|\mathbf{t}_n &\sim \mathcal{N}(\mathbf{x}_{2n}|\mathbf{W}_2\mathbf{t}_n + \boldsymbol{\mu}_2, \boldsymbol{\Psi}_2), \end{aligned} \tag{1}$$

where the $\boldsymbol{\Psi}$ denote the precision matrices of the normal distribution. The latent variables $\mathbf{t}$ encode the low-dimensional statistically dependent part while projection matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ specify how this dependency is manifested in each of the data sources. Bach and Jordan [3] established the connection between this probabilistic model and classical CCA by showing that the maximum-likelihood solution of the model coincides with the classical solution except for an arbitrary rotation in the latent space and projection matrices. PCCA as such does not solve any of the problems classical CCA has since it is merely an equivalent description, but the probabilistic formulation makes justified extensions possible. In the remainder of this section, we walk through the extensions and modifications required for creating a practically applicable dependency modeling tool for real-world signals.

The first step is to make the inference more reliable by switching from the maximum likelihood solution to full Bayesian analysis, by complementing the likelihood with priors for the model parameters. We adopt the formulation of [10,16] for Bayesian CCA (BCCA)

$$\begin{aligned} \mathbf{w}_{ij}|\alpha_i &\sim \mathcal{N}(\mathbf{w}_{ij}|0, \mathrm{diag}(\alpha_{i1}, \ldots, \alpha_{iD})) \\ \alpha_{ij} &\sim \mathcal{G}(\alpha_{ij}|a_i, b_i) \\ \boldsymbol{\Psi}_i &\sim \mathcal{W}(\boldsymbol{\Psi}_i|\gamma_i, \boldsymbol{\Phi}_i) \\ \boldsymbol{\mu}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i|0, \beta_i\mathbf{I}), \end{aligned} \tag{2}$$

where $\mathcal{G}$ is the gamma distribution, $\mathcal{W}$ denotes the Wishart distribution, the subscript i is used to denote the data sources, and the distribution of data is given in (1). The rest of the symbols are hyper-priors of the model. The priors for the projection matrix row vectors $p(\mathbf{w}_{ij}|\alpha_i)$ and the precision prior $p(\alpha_{ij})$ implement the Automatic Relevance Determination (ARD) [14] which automatically controls the number of the components in the model by adjusting the precisions α_{ij} – the precisions for unnecessary components are driven to infinity, and hence the posterior peaks around the zero vector. Both [10,16] experimentally verified that the ARD mechanism detects the correct dimensionality of the latent space, which is the second necessary component for our practically usable model.

The Bayesian CCA model makes a strict assumption of Gaussian noise, which is problematic for many real life signals used as stimulus representations or brain activity measurements. This problem can be alleviated by replacing both the Gaussian noise and the Gaussian latent variables by Student's t-distribution (with ν degrees of freedom) that is more robust to outliers:

$$\begin{aligned}\mathbf{t}_n &\sim \mathcal{S}(\mathbf{t}_n|0, \mathbf{I}_D, \nu)\\ \mathbf{x}_{1n}|\mathbf{t}_n &\sim \mathcal{S}(\mathbf{x}_{1n}|\mathbf{W}_1\mathbf{t}_n + \boldsymbol{\mu}_1, \boldsymbol{\Psi}_1, \nu)\\ \mathbf{x}_{2n}|\mathbf{t}_n &\sim \mathcal{S}(\mathbf{x}_{2n}|\mathbf{W}_2\mathbf{t}_n + \boldsymbol{\mu}_2, \boldsymbol{\Psi}_2, \nu).\end{aligned}$$

For efficient inference, we exploit the latent infinite scale-mixture formulation of the t-distribution [11],

$$\mathcal{S}(\mathbf{t}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty du \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) \mathcal{G}\left(u|\nu/2, \nu/2\right).$$

Using this formulation, we can write the robust CCA model by adding an extra level of hierarchy

$$\begin{aligned}u_n &\sim \mathcal{G}\left(u_n|\nu/2, \nu/2\right)\\ \mathbf{t}_n|u_n &\sim \mathcal{N}(\mathbf{t}_n|0, u_n\mathbf{I}_D)\\ \mathbf{x}_{1n}|u_n, \mathbf{t}_n &\sim \mathcal{N}(\mathbf{x}_{1n}|\mathbf{W}_1\mathbf{t}_n + \boldsymbol{\mu}_1, u_n\boldsymbol{\Psi}_1)\\ \mathbf{x}_{2n}|u_n, \mathbf{t}_n &\sim \mathcal{N}(\mathbf{x}_{2n}|\mathbf{W}_2\mathbf{t}_n + \boldsymbol{\mu}_2, u_n\boldsymbol{\Psi}_2).\end{aligned}$$

This formulation has conjugate conditional distributions, which considerably simplifies inference. The above formulation for robust CCA has earlier been presented by [2], but they only considered the maximum likelihood estimate for the parameters. We couple the robust noise assumption with the priors for the Bayesian CCA model (2) to arrive at the novel model of Robust Bayesian CCA (RBCCA). It is a basic building block of our full model.

2.1 Mixture of Robust Bayesian CCAs

Next we turn our attention to removing the stationarity assumption, by replacing it with piecewise stationarity. In this work we follow our earlier model [10] and introduce a probabilistic mixture of robust Bayesian CCA models, letting each mixture cluster to model different kind of dependencies between the signals.

We formulate the probabilistic mixture by introducing an additional multinomial latent variable which generates the mixture assignment [13]. The robust mixture CCA model is therefore obtained by adding the latent variable $z_n \sim \text{Multinomial}(z_n|\boldsymbol{\pi})$, where $\boldsymbol{\pi}$ denotes the probabilities of the clusters (we use point estimates for $\boldsymbol{\pi}$, but the extension to Dirichlet prior would be straightforward), and conditioning all the rest of the latent variables and parameters on the value of z_n.

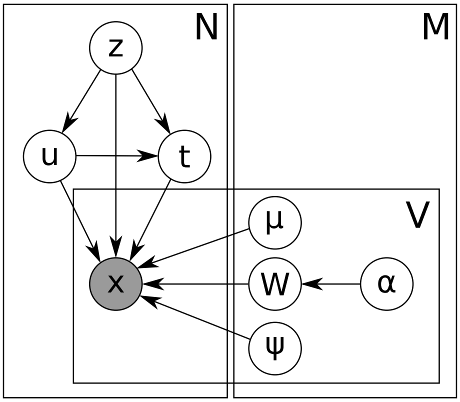

Fig. 1. Plate diagram of the mixture of robust CCA models. N denotes the samples, M the mixture clusters and V the two data sources. The hyper-priors of the variables are excluded for clarity.

The full model including the mixture formulation for non-stationarity, ARD prior for choosing the number of correlated components within the clusters, and the t-distribution for handling outliers is represented in Figure 1. All of the models described above like RBCCA and mixture of Gaussian BCCAs are obtained as special cases of the full model, as are a number of other models. In particular, fixing $1/\alpha_{ij} = 0$ results in (robust) Gaussian mixture model [1], setting $\mathbf{\Psi}$ diagonal gives (robust) factor analysis [5], and further restricting it to be spherical leads to (robust) Bayesian PCA [8,12].

2.2 Variational Inference

For analysis, the above model formulation needs to be coupled with an inference algorithm. In particular, we need to learn the posterior distribution of the model parameters, and be able to make predictions for future data. Since the goal is to be able to apply the model for analysis of potentially very large data sets, we steer away from the computationally heavy earlier alternatives like the Gibbs sampling approach of [10] for mixture of BCCA inference, and instead choose to use the deterministic variational approximation. The resulting algorithm is computationally as efficient as finding the maximum likelihood or maximum a posteriori estimate through the EM algorithm, but maintains the advantage of full Bayesian analysis in capturing the uncertainty in the results. Next, we briefly summarize the variational Bayesian (VB) approach for inference, and only explain in more detail the choices specific for the novel parts of the model. For more extensive introduction to variational inference see, *e.g.*, [4].

The core of the inference process is in learning the posterior distribution $p(H|\mathbf{X}_1, \mathbf{X}_2, \Theta)$ of both the latent variables and the model parameters, denoted collectively as $H = \{\mathbf{Z}, \mathbf{U}, \mathbf{T}, \mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2\}$, given the observed data and model hyper-parameters Θ. Finding the true posterior is not feasible even for the basic CCA model, let alone the full robust mixture, because evaluating the marginal log-likelihood

$$\ln p(\mathbf{X}_1, \mathbf{X}_2|\Theta) = \ln\left(\int dH\, p(H, \mathbf{X}_1, \mathbf{X}_2|\Theta)\right)$$

is not tractable.

With variational Bayesian inference the problem is solved by approximating the true posterior with a variational distribution $q(H)$ from some limited class of distribution functions so that the inference remains tractable [4]. In practice, the class of distributions is limited by assuming that the full posterior factorizes into a number of (a priori) independent terms, and the optimal distribution in this class is chosen by minimizing the Kullback-Leibler divergence from the approximation to the true posterior $D_{\mathrm{KL}}(q(H)||p(H|\mathbf{X}_1, \mathbf{X}_2, \Theta))$. The full posterior is then found by an iterative EM-style algorithm. The resulting update formulas, based on the factorization described below, are given in the Appendix.

Following the variational Bayesian CCA by Wang [16], we consider a variational distribution for which the parameter part is fully factorised as

$$\prod_{i=1}^{2}\prod_{k=1}^{M} q(\mathbf{\Psi}_i^k)q(\boldsymbol{\mu}_i^k)q(\mathbf{W}_i^k)q(\alpha_i^k). \tag{3}$$

We then focus on the approximation used for the latent variables, naturally factorized over the data points as $\prod_{n=1}^{N} q(z_n, u_n, \mathbf{t}_n)$. It is clear that we need to consider separate terms for each cluster, $q(z_n)q(u_n, \mathbf{t}_n|z_n)$, but for the latter term two different tractable approximations are possible. For example [8,12] use the approximation $q(u_n)q(\mathbf{t}_n)$, assuming conditional independence between u_n and $\mathbf{t}_n$, whereas [5] chose $q(u_n)q(\mathbf{t}_n|u_n)$, not introducing any independence assumptions beyond those in the actual model. In our scenario both solutions are analytically tractable, conditioned on z_n.

To our knowledge, these two choices have not been compared before, nor the additional independence assumption justified. Since both are tractable and lead to implementations of comparable computational complexity, the relative accuracy of the two approximations is an interesting question for variational approximations of t-distribution models in general. Hence, we implemented both alternatives and empirically compare them in the experiments, showing that the difference in performance is negligible, making both alternatives valid. The formulas given in the Appendix assume the approximative $q(u_n)q(\mathbf{t}_n)$ factorisation.

Besides learning the posterior distribution of the latent variables and model parameters, we are naturally interested in making predictions for new data. Both inferring $\mathbf{x}_1$ given $\mathbf{x}_2$ (or vise versa) and inferring the latent variable $\mathbf{t}$ given $\mathbf{x}_1$ and/or $\mathbf{x}_2$ are useful for various application scenarios. Exact calculation of these distributions is again untractable due to the dependency on the hidden data posterior distributions. However, we can utilize the variational distributions to make the predictions tractable. For a new data point, $q(z, u, \mathbf{t})$ is chosen as the distribution which maximizes the variational lower bound of $\ln p(\mathbf{X}_{-i}|\Theta)$

where $-i$ denotes the observed data source(s). As an example, we consider the predictive density $p(\mathbf{x}_i|\mathbf{x}_{-i})$. Using the conditional independence of $\mathbf{x}_i$ and $\mathbf{x}_{-i}$ when $\mathbf{t}$ is known, and integrating out $\mathbf{t}$ results in

$$\mathbf{x}_i|z_k = 1, u, \boldsymbol{\Psi}_i, \boldsymbol{\mu}_i, \mathbf{W}_i \sim \mathcal{N}(\mathbf{x}_i|\mathbf{W}_i^k\boldsymbol{\mu}_{t_k} + \boldsymbol{\mu}_i^k, ((u\boldsymbol{\Psi}_i^k)^{-1} + \mathbf{W}_i^k\boldsymbol{\Sigma}_{t_k}(\mathbf{W}_i^k)^\top)^{-1}),$$

where the information from the observed data source is encapsulated in the parameters $\boldsymbol{\mu}_{t_k}$, $\boldsymbol{\Sigma}_{t_k}$ and the gamma distribution of u (these paremeters, however, are slightly different in comparison to the formulas in the Appendix as we observe only $\mathbf{X}_{-i}$). The above predictive density assumes the factored approximation for the latent variables; with the non-factored alternative the density is of the same form but $\boldsymbol{\Sigma}_{t_k}$ will explicitely depend on u.

The above expression does not yield a closed form expression for the distribution $p(\mathbf{x}_i|\mathbf{x}_{-i})$ but at least the two first moments are analytically tractable. The most important quantity is the conditional mean

$$\mathbb{E}[\mathbf{x}_i|\mathbf{x}_{-i}] = \sum_{k=1}^{M} q(z_k)\langle\mathbf{W}_i^k\rangle_{q(\{\mathbf{W}_i\})}\boldsymbol{\mu}_{t_k} + \langle\boldsymbol{\mu}_i^k\rangle_{q(\{\boldsymbol{\mu}_i\})}.$$

3 Model Validation

To validate that the model does what it promises, we performed two experiments using artificial data. First we show how replacing the Gaussian distribution with t-distribution considerably improves the accuracy of the model in presense of increasing amounts of outliers. At the same time we compare the two alternative variational factorizations for t-distributed latent variable models, showing that there is no noticeable difference in accuracy. Then we show how the model correctly captures non-stationarity with clusters and automatically extracts the correct number of correlated components.

In both of our artificial data experiments, we fix the hyperparameters to values corresponding to broad priors ($a_i = b_i = 0.1$, $\gamma_i = d_i + 1$, $\boldsymbol{\Phi}_i = 10^2\mathbf{I}$, $\beta_i = 1$) and consequently let the data determine the model parameters. The hyper-parameters $\boldsymbol{\pi}$ and ν:s are updated by maximizing the variational lower bound which leads to closed form and line-search update rules.

3.1 Robustness against Outliers

We start by showing the importance of robust modeling in presense of outliers. We first generate data (N=500) from the model (with single cluster), and then add a varying number of outlier data points drawn from the uniform distribution. We then compare the robust model with the two alternative variational approximations against the Gaussian model by measuring the variational lower bound and the mean error in predicting $\mathbf{x}_1$ from $\mathbf{x}_2$. Figure 2 shows how the performance is identical for the case of no outliers, meaning that there is no harm in using the robust variant, and that already for fairly modest ratios of outliers the robust variant is considerably better.

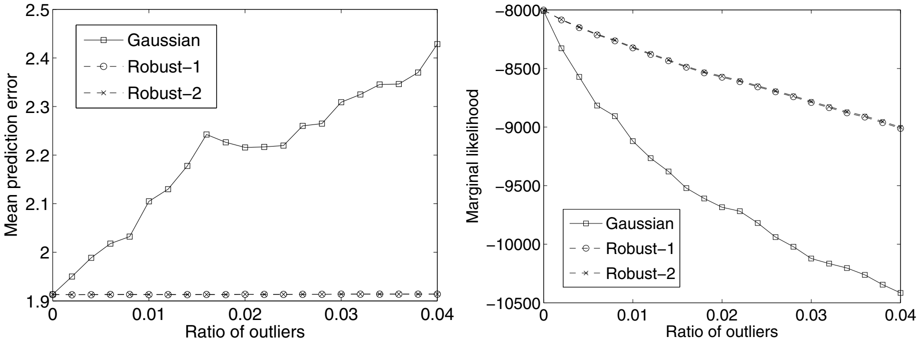

Fig. 2. Left: The mean prediction error for the robust CCA model stays essentially constant irrespective of the number of outliers in the data, whereas the Gaussian CCA model starts losing accuracy already for a low number of outliers. Even below 1% of samples being outliers the difference in accuracy is already noticeable. **Right:** The same effect is visible also in the variational lower bound. For both measures the curves for the two alternative approximations for the robust variant are completely overlapping, showing no difference.

The results also indicate that there does not seem to be any difference between the two variational approximations for the t-distribution. To further compare the alternatives, we construct an experiment designed to emphasize potential differences. We generate the data from the t-distribution with just $\nu = 2$ degrees of freedom, making the data very heavy-tailed (for high values of ν the t-distribution approaches Gaussian). The model is trained for $N = 10000$ data points, and 10 separate sets of 10000 data points are used for testing. We measure the error in predicting $\mathbf{x}_1$ given $\mathbf{x}_2$, both with the mean prediction error and quantiles of the error distribution to emphasize potential tail effects. The results, collected in Table 1, confirm that the accuracies are indeed comparable.

3.2 Model Selection

Next we show how the model comes with ready tools for choosing the model complexity. For any real analysis task both the number of clusters needed to correctly capture the non-stationary dependencies and the number of correlating components in each of the clusters are unknown. We show how the ARD prior for the projection matrices removes the need of explicitly specifying the number of components, by automatically ignoring unnecessary components, and how the marginal likelihood of the model reveals the correct number of clusters.

We created $M = 3$ clusters each consisting of 2000 points from the model

$$\begin{aligned}
\mathbf{t}_n &\sim \mathcal{N}(\mathbf{t}_n|0, \mathbf{I}) \\
\mathbf{x}_{1n}|\mathbf{t}_n &\sim \mathcal{N}(\mathbf{x}_{1n}|\mathbf{W}_1^k\mathbf{t}_n + \boldsymbol{\mu}_1^k, (\mathbf{L}_1^k\mathbf{L}_1^{k\top})^{-1}) \\
\mathbf{x}_{2n}|\mathbf{t}_n &\sim \mathcal{N}(\mathbf{x}_{2n}|\mathbf{W}_2^k\mathbf{t}_n + \boldsymbol{\mu}_2^k, (\mathbf{L}_2^k\mathbf{L}_2^{k\top})^{-1}),
\end{aligned}$$

Table 1. Quantitative analysis of the prediction errors of the two alternative variational approximations for t-distributions. The table shows the mean prediction error over 10000 independent test samples, averaged over 10 different realizations of the data set, as well as different quantiles of the distribution of the errors. The two factorizations are equal with respect to all of the measures.

Approximation	Mean	5% quantile	50% quantile	95% quantile
	Predict $\mathbf{x}_2 \vert \mathbf{x}_1$			
$q(u)q(\mathbf{t})$	8.9180	2.1766	6.2896	22.4195
$q(u)q(\mathbf{t} \vert u)$	8.9180	2.1766	6.2895	22.4177
	Predict $\mathbf{x}_1 \vert \mathbf{x}_2$			
$q(u)q(\mathbf{t})$	8.5027	2.0908	5.9635	21.4362
$q(u)q(\mathbf{t} \vert u)$	8.5028	2.0911	5.9636	21.4384

where the mean vector entries are drawn randomly so that the cluster centers are well seperated. The entries of the lower triangular matrices $\mathbf{L}_1$ and $\mathbf{L}_2$ are drawn from the uniform distribution between 0 and 0.5, with additional small positive entries added to the main diagonal, and the projection matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are generated as

$$\mathbf{W}_1^k = \sum_{j=1}^{d_k} \mathbf{w}_{1j}^k \mathbf{e}_j^\top$$

$$\mathbf{w}_{1j}^k \sim \mathcal{N}(\mathbf{w}_{1j}^k | \mathbf{1}, (10 * \mathbf{I})^{-1}),$$

where $d_k = \{3, 5, 7\}$ encodes the dimensionality of the latent space in each of the clusters. For $\mathbf{x}_2$ the procedure was the same.

Figure 3 shows two illustrations of the result, clearly demonstrating that the marginal likelihood grows until the correct number of clusters but does not improve further, indicating that coupling the likelihood with a reasonable prior on the number captures the correct complexity. The other sub-figure illustrates the projection matrix, revealing how only five components contain non-zero elements in the cluster that was created to have exactly five correlating components, even though the model was ran with maximal possible complexity (the number of dimensions that is here 50). Hence, the need for choosing the complexity is efficiently sidestepped. Note that the columns of the matrix have not been reordered for the illustration, but instead the approximation automatically learns the components roughly in the order of the magnitude.

4 MEG Analysis

To concretize the scenarios where searching for mutual dependencies is likely to be useful, we apply the model to analysis of brain response to natural stimulus. For natural stimuli the traditional approaches are not sufficient due to lack of repetition and control in the stimulus, and more data-driven approaches are needed. The primary purpose of the experiment is to illustrate potential uses for the model, and more detailed neuroscientific analysis is omitted.

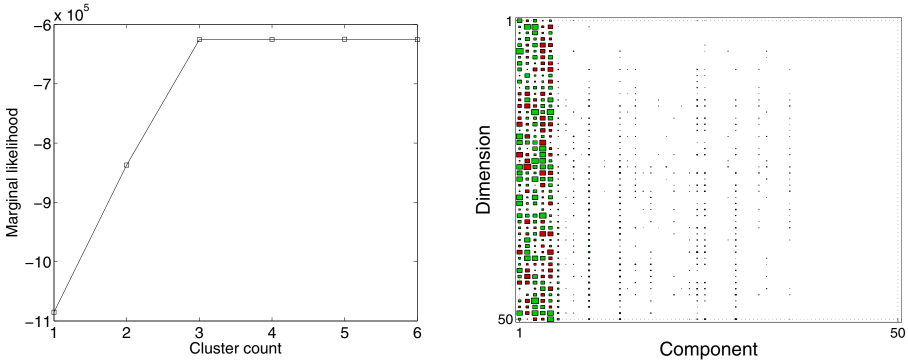

Fig. 3. Left: Marginal likelihood as a function of the number of clusters reveals that the data has three clusters. The likelihood remains constant for larger number of potential clusters because the model discovers only three clusters, leaving the rest empty. **Right:** Hinton plot of the projection matrix $\mathbf{W}_x^3$ of the three cluster model shows how the model correctly captures the correlating dimensions. The ARD prior automatically pushes the elements to zero for excess dimensions, keeping high values only for the five true components.

In particular, we demonstrate how statistical dependencies between brain activity measurements done with MEG and measurements of the autonomic nervous system (ANS) can be used to create hypotheses on where to look for emotional responses in MEG data. MEG measures the cortical brain activity while emotional stimuli mainly cause response in the deeper regions, and hence it is generally unknown to which degree emotional responses are visible in MEG data (see [15] for a analysis of a simple controlled experiment). Since ANS measurements are highly informative of emotional activity, correlations between the two sources provide a link between MEG and the emotions.

In this paper we present the results from the point of view of further validating the applicability of the model. In detail, we show how relaxing the stationarity assumption of the signal by mixture modeling reveals stronger correlations between the signals, and how the mixture components found by the model are interpretable and directly linked with the emotional stimuli labels not used in learning the model. These results complement the artificial experiments and show the model is directly applicable also for real scenarios, even for large sample sizes.

4.1 Data

We apply the model for joint analysis of brain oscillations and autonomic nervous system (ANS) response to emotionally loaded auditory stimuli [18]. Emotional sounds obtained from the International Affective Digitized Sounds (IADS-2) library with varying arousal and valence values were played while the brain activity of the test subjects was measured with MEG, and pupil diameter measured with iView X$^{\mathrm{TM}}$ MEG eye tracker was used as an example signal for the ANS activity. A total of 48 stimuli, each lasting 6 seconds, were played with 10 second shade-in and shade-out periods.

We extract dependencies between a single MEG channel, chosen based on preliminary analysis, and pupil diameter, as a demonstration of what the model can achieve. The approach directly generalizes to more MEG channels and would also be applicable to other measurements of ANS, such as galvanic skin response or heart-rate variability, or combinations of those.

After basic signal pre-processing consisting of resampling, de-trending, filtering and renormalisation, we apply a sliding rectangular window function to both one-dimensional signals. The stimuli-evoked responses are time-localised in the MEG signal, and the resulting window-based feature representation is a natural choice for such short time analysis. In the context of the CCA-type models, this representation encodes the time-amplitude information through the mean vector and frequency-amplitude information in the CCA projections. In other words, projecting the signal windows to the canonical scores corresponds to filtering.

The model hyperparameters are set in exactly the same way as for the artificial data except for the prior precisions which are smaller (with the order of magnitude estimated from the empirical covariance matrices), because the biomedical signals are known to be very noisy.

4.2 Results

Figure 4 (left) shows the marginal likelihood as a function of the number of clusters, showing that the data strongly supports more than one cluster and hence that the signal is clearly non-stationary. Solutions between two and five clusters are all sensible, whereas using more than five clusters does not improve the likelihood anymore. In fact, the excess clusters become empty during the learning process, and hence play no role.

One of the main advantages of relaxing the stationarity assumption is that the regions of the data space showing strong dependency can be separated from the rest, to better capture the correlations. This should be manifested as some clusters having high correlations, while some other clusters learn to model the less dependent parts. We use this observation to construct a measure for comparing our model with the alternative solution of first clustering the data in the joint space and applying classical CCA for each of the clusters separately: For each model complexity we measure the difference between the highest correlations in the most and least dependent clusters. The comparison method also uses variational inference for learning the clusters, and the result is then turned into a hard clustering in order to compute the CCA. It hence follows the basic approach of [7], but the clustering model is replaced with a better one to compensate for the gain by our improved inference.

Figure 4 (right) shows how finding the clusters and the dependencies together improves compared to the alternative. The joint clustering is not able to separate the dependent parts from the independent ones, but instead merely divides the data into clusters of roughly equal size having correlations relatively close to

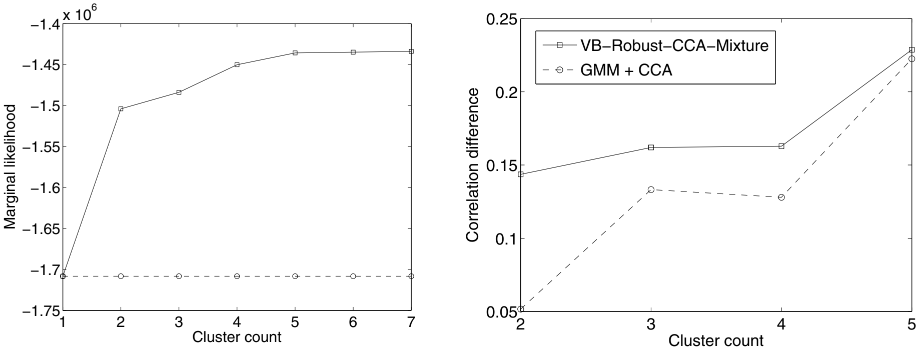

Fig. 4. Left: Marginal likelihood as a function of the number of clusters, showing how the data only supports at most five clusters. The baseline (dashed line) corresponds to the single-cluster solution, and the clear difference shows the improvement from relaxing the stationarity assumption. **Right:** Illustration of how the proposed model outperforms the alternative of first clustering the data and then applying classical CCA. The curves measure the ability of the model to separate different kinds of dependencies into different clusters, evaluated as the difference between the largest correlations in the most/least dependent cluster in the model.

each other. Increasing the number of clusters helps, since small enough clusters will start to capture the dependencies even when they were learned to model the joint distribution, but even then the joint clusters cannot be directly interpreted as capturing different kind of dependencies.

Next we take the three-cluster solution of our model for closer analysis in order to demonstrate that the clusters are interpretable. We do not proceed to analyze the projection vectors that would reveal the signal filters needed for full neuroscientific analysis, but instead idenfity the clusters based on the mean profiles of the pupil data (Figure 5; left) and show how the cluster identities are linked with the emotional stimuli labels (Figure 5; right). The labels were not used in learning the model, and hence this serves as an external validation. The mean vectors reveal that the largest cluster corresponds to no activity, while the other two clusters correspond to pupil dilation and contraction. The histogram of the stimuli labels in each of the clusters shows that the two smaller clusters are enriched with the positive and negative stimuli, respectively, proving that the model has learned not only a link between MEG and ANS, but that the link is indeed related to the underlying natural stimulus.

5 Discussion

Efficient and robust models for extracting statistical dependencies between multiple co-occurring data streams or signals are needed for exploratory analysis of complicated natural phenomena such as brain activity or cellular functions. We introduced a novel model that synthetises the latest advances in generative

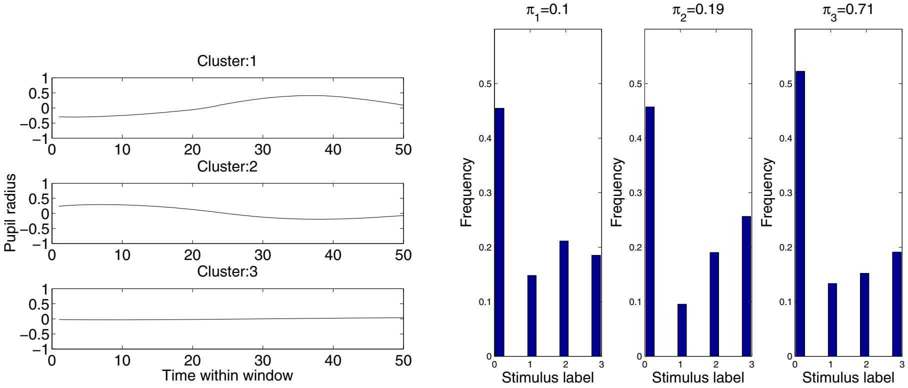

Fig. 5. Left: Mean of the pupil diameter in each of the clusters. **Right:** Distribution of the stimuli labels in each of the clusters. The stimuli enumeration from 0 to 3 refers to no stimulus, enomotionally neutral stimulus, positive stimulus and negative stimulus, respectively. Note how samples with positive stimulus labels are enriched in the first cluster and samples with negative stimulus label in the second cluster.

dependency modeling to create a practically applicable tool for large-scale analysis. The robust mixture of canonical correlation analyzers combines the mixture solution of [10] with the variational approximation of [16] and the robust CCA extension of [2] into a single model. The model is directly applicable to data sets of tens of thousands of observations, as demonstrated by the example application on the MEG data, and includes automatic solutions for model complexity selection. An open-source implementation of the model written in MATLAB is available at http://www.cis.hut.fi/projects/mi/software/vbcca/.

Furthermore, we studied alternative variational approximations for robust t-distribution models in general. Two different independence assumptions both lead to tractable approximations that have been used by earlier models [5,8,12]. We showed that the difference in the modeling accuracy between the two approximations is negligible, concluding that future variational approximations of scale-mixture models can choose either alternative based on the desired functional form for the predictive distributions, not needing to consider the modeling accuracy.

Acknowledgments

The authors belong to the Adaptive Informatics Research Centre, a CoE of the Academy of Finland. The project was supported by the aivoAALTO project of the Aalto University and by the Academy of Finland decision number 133818, and was in part supported by the PASCAL2 EU NoE. We thank Siina Pamilo, Miika Koskinen, and Riitta Hari (Brain Research Unit, Low Temperature Laboratory, Aalto University School of Science and Technology) for providing the data used in the neuroscientific experiment.

Appendix: Variational EM-Update Equations

The factorization $q(u_n|z_n)q(\mathbf{t}_n|z_n)$ results in the variational distributions of the following form. The parameters of the approximation are implicitly defined as the symbols for which the right hand sides are conditioned on:

$$
\begin{aligned}
q(\{\boldsymbol{\Psi}_i\}) &= \prod_{k=1}^{M} \mathcal{W}(\boldsymbol{\Psi}_i^k|\widetilde{\gamma}_i^k, \widetilde{\boldsymbol{\Phi}}_i^k)) \\
q(\{\boldsymbol{\mu}_i\}) &= \prod_{k=1}^{M} \mathcal{N}(\boldsymbol{\mu}_i^k|\boldsymbol{\mu}_{\mu_i^k}, \boldsymbol{\Sigma}_{\mu_i^k}) \\
q(\{\mathbf{W}_i\}) &= \prod_{k=1}^{M} \prod_{j=1}^{d_i} \mathcal{N}(\mathbf{W}_{ij}^k|\boldsymbol{\mu}_{W_{ij}^k}, \boldsymbol{\Sigma}_{W_{ij}^k}) \\
q(\{\alpha_i\}) &= \prod_{k=1}^{M} \prod_{j=1}^{d} \mathcal{G}(\alpha_{ij}^k|a_{ij}^k, b_{ij}^k) \\
q(\mathbf{t}_n|z_{nk}=1) &= \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_{t_{nk}}, \boldsymbol{\Sigma}_{t_{nk}}^{-1}) \\
q(u_n|z_{nk}=1) &= \mathcal{G}(u_n|\alpha_{q_{nk}}, \beta_{q_{nk}}) \\
q(z_n) &= \text{Multinomial}(z_n|r_n).
\end{aligned}
$$

The update rules needed for learning the parameters of the approximation are then given by the following formulas, where $\langle A\rangle_{q(\cdot)}$ denotes the expectation of A with respect to $q(\cdot)$. In addition, we denote the dimensionality of $\mathbf{x}_{in}$ with d_i and the latent space dimensionality with D:

$$
\begin{aligned}
\boldsymbol{\mu}_{t_{nk}} &= \boldsymbol{\Sigma}_{t_k}\left(\sum_{i=1}^{2}\langle(\mathbf{W}_i^k)^\top\boldsymbol{\Psi}_i^k(\mathbf{x}_{in}-\boldsymbol{\mu}_{ik})\rangle_{q(\{\boldsymbol{\mu}_i\})q(\{\boldsymbol{\Psi}_i\})q(\{\mathbf{W}_i\})}\right) \\
\boldsymbol{\Sigma}_{t_{nk}}^{-1} &= \langle u_n\rangle_{q(u_n|z_{nk}=1)}\left(\sum_{i=1}^{2}\langle(\mathbf{W}_i^k)^\top\boldsymbol{\Psi}_i^k\mathbf{W}_i^k\rangle_{q(\{\boldsymbol{\Psi}_i\})q(\{\mathbf{W}_i\})}+\mathbf{I}_D\right) \\
&= \langle u_n\rangle_{q(u_n|z_{nk}=1)}\boldsymbol{\Sigma}_{t_k}
\end{aligned}
$$

$$
\begin{aligned}
\alpha_{q_{nk}} &= \frac{\nu_k+\sum_{i=1}^{2}d_i+D}{2} \\
\beta_{q_{nk}} &= \nu_k/2+\langle\frac{1}{2}\mathbf{t}_n^\top\mathbf{t}_n\rangle_{q(\mathbf{t}_n|z_{nk}=1)} \\
&\quad+\sum_{i=1}^{2}\langle\frac{1}{2}(\mathbf{x}_{in}-\mathbf{W}_i^k\mathbf{t}_n-\boldsymbol{\mu}_i^k)^\top\boldsymbol{\Psi}_i^k(\mathbf{x}_{in}-\mathbf{W}_i^k\mathbf{t}_n-\boldsymbol{\mu}_i^k)\rangle_{q(*)} \\
&\quad\text{where } q(*)=q(\mathbf{t}_n|z_{nk}=1)q(\{\boldsymbol{\mu}_i\})q(\{\boldsymbol{\Psi}_i\})q(\{\mathbf{W}_i\}))
\end{aligned}
$$

$$
\begin{aligned}
\ln \rho_{nk} &= \ln \boldsymbol{\pi}_k - D_{\mathrm{KL}}(q(u_n|z_{nk}=1)||p(u_n|z_{nk}=1)) \\
&- \langle D_{\mathrm{KL}}(q(\mathbf{t}_n|z_{nk}=1)||p(\mathbf{t}_n|z_{nk}=1,u_n))\rangle_{q(u_n|z_{nk}=1)} \\
&+ \sum_{i=1}^{2} \langle \ln p(\mathbf{x}_{in}|z_{nk}=1,u_n,\mathbf{t}_n)\rangle_{q(\{\boldsymbol{\mu}_i\})q(\{\boldsymbol{\Psi}_i\})q(\{\mathbf{W}_i\})q(u_n|z_{nk}=1)q(\mathbf{t}_n|z_{nk}=1)} \\
r_{nk} &= \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nk}} \\
\widetilde{\gamma}_i &= \gamma_i + \sum_{n=1}^{N} q(z_{nk}) \\
(\widetilde{\boldsymbol{\Phi}}_i^k)^{-1} &= (\boldsymbol{\Phi}_i)^{-1} + \sum_{n=1}^{N} q(z_{nk})\langle((\mathbf{x}_{in} - \mathbf{W}_i^k\mathbf{t}_n - \boldsymbol{\mu}_i^k)\dots \\
&\times (\mathbf{x}_{in} - \mathbf{W}_i^k\mathbf{t}_n - \boldsymbol{\mu}_i^k)^\top u_n)\rangle_{q(u_n,\mathbf{t}_n|z_{nk}=1)q(\{\mathbf{W}_i\})q(\{\boldsymbol{\mu}_i\})} \\
\boldsymbol{\Sigma}_{\boldsymbol{\mu}_i^k}^{-1} &= \beta_i\mathbf{I} + \sum_{n=1}^{N} q(z_{nk})\langle u_n\boldsymbol{\Psi}_i^k\rangle_{q(u_n|z_{nk}=1)q(\{\boldsymbol{\Psi}_i\})} \\
\boldsymbol{\mu}_{\mu_i^k} &= \boldsymbol{\Sigma}_{\mu_i^k}(\sum_{n=1}^{N} q(z_{nk})\langle u_n\boldsymbol{\Psi}_i^k(\mathbf{x}_{in} - \mathbf{W}_i^k\mathbf{t}_n)\rangle_{q(u_n,\mathbf{t}_n|z_{nk}=1)q(\{\mathbf{W}_i\})q(\{\boldsymbol{\Psi}_i\})}\big) \\
\boldsymbol{\Sigma}_{W_{ij}^k}^{-1} &= \langle \mathrm{diag}(\alpha_i^k)\rangle_{q(\{\alpha_i\})} + \sum_{n=1}^{N} q(z_{nk})\langle u_n\mathbf{t}_n\mathbf{t}_n^\top(\boldsymbol{\Psi}_i^k)_{(j,j)}\rangle_{q(u_n,\mathbf{t}_n|z_{nk}=1)q(\{\boldsymbol{\Psi}_i\})} \\
\boldsymbol{\mu}_{W_{ij}^k} &= \boldsymbol{\Sigma}_{W_{ij}^k}(\sum_{n=1}^{N} q(z_{nk})\langle t_n(\boldsymbol{\Psi}_i^k)_{(j,:)}u_n(\mathbf{x}_{in} - \boldsymbol{\mu}_i^k)\rangle_{q(u_n,\mathbf{t}_n|z_{nk}=1)q(\{\boldsymbol{\mu}_i\})q(\{\boldsymbol{\Psi}_i\})} \\
&- \sum_{n=1}^{N}\sum_{l\neq j}^{d_i} q(z_{nk})\langle u_n\mathbf{t}_n\mathbf{t}_n^\top(\boldsymbol{\Psi}_i^k)_{(j,l)}\mathbf{W}_{il}^k\rangle_{q(u_n,\mathbf{t}_n|z_{nk}=1)q(\{\boldsymbol{\mu}_i\})q(\{\boldsymbol{\Psi}_i\})}\big) \\
a_{ij}^k &= a_i + d_i/2 \\
b_{ij}^k &= b_i + \langle||\mathbf{W}_{i_{(:,j)}}^k||^2\rangle_{q(\{\mathbf{W}_i\}))}/2
\end{aligned}
$$

References

1. Archambeau, C., Verleysen, M.: Robust Bayesian clustering. Neural Networks 20, 129–138 (2007)
2. Archambeau, C., Delannay, N., Verleysen, M.: Robust probabilistic projections. In: Cohen, W., Moore, A. (eds.) Proceedings of ICML 2006, the 23rd International Conference on Machine Learning, pp. 33–40. ACM, New York (2006)
3. Bach, F.R., Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis. Tech. Rep. 688, Department of Statistics, University of California, Berkeley (2005)
4. Beal, M.: Variational algorithms for approximate Bayesian inference. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, UK (2003)

5. Chatzis, S., Kosmopoulos, D., Varvarigou, T.: Signal modeling and classification using a robust latent space model based on t distributions. IEEE Transactions on Signal Processing 56(3), 949–963 (2008)
6. Correa, N., Li, Y.O., Adali, T., Calhoun, V.D.: Canonical correlation analysis for feature-based fusion of biomedical imaging modalities to detect associative networks in schizophrenia. Special issue on fMRI analysis for Human brain mapping. IEEE J. Selected Topics in Signal Processing 2(6), 998–1007 (2008)
7. Fern, X., Brodley, C.E., Friedl, M.A.: Correlation clustering for learning mixtures of canonical correlation models. In: Kargupta, H., Kamath, C., Srivastava, J., Goodman, A. (eds.) Proceedings of the Fifth SIAM International Conference on Data Mining, pp. 439–448 (2005)
8. Gao, J.: Robust L1 principal component analysis and its Bayesian variational inference. Neural Computation 20(2), 555–572 (2008)
9. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
10. Klami, A., Kaski, S.: Local dependent components. In: Ghahramani, Z. (ed.) Proceedings of ICML 2007, the 24th International Conference on Machine Learning, pp. 425–432. Omnipress (2007)
11. Liu, C., Rubin, D.: ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica 5, 19–39 (1995)
12. Luttinen, J., Ilin, A., Karhunen, J.: Bayesian robust PCA for incomplete data. In: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation, pp. 66–73 (2009)
13. McLachlan, G.J., Peel, D.: Finite Mixture Models. John Wiley & Sons, New York (2000)
14. Neal, R.: Bayesian learning for neural networks. Lecture Notes in Statistics, vol. 118. Springer, New York (1996)
15. Onoda, K., Okamoto, Y., Shishida, K., Hashizume, A., Ueda, K., Yamashita, H., Yamawaki, S.: Anticipation of affective images and event-related desynchronization (ERD) of alpha activity: An MEG study. Brain Research 1151, 134–141 (2007)
16. Wang, C.: Variational Bayesian approach to canonical correlation analysis. IEEE Transactions on Neural Networks 18, 905–910 (2007)
17. Ylipaavalniemi, J., Savia, E., Malinen, S., Hari, R., Vigário, R., Kaski, S.: Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. NeuroImage 48, 176–185 (2009)
18. Yokosawa, K., Pamilo, S., Hirvenkari, L., Ramkumar, P., Pihko, E., Hari, R.: Activation of auditory cortex by anticipating and hearing emotional sounds: a magnetoencephalographic study. In: 16th Annual Meeting of the Organization for Human Brain Mapping (June 2010)