Hierarchical bayesian modeling of inter-trial variability and variational Bayesian learning of common spatial patterns from multichannel EEG

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1109/ICASSP.2010.5495663">http://dx.doi.org/10.1109/ICASSP.2010.5495663</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://hdl.handle.net/1721.1/69677">http://hdl.handle.net/1721.1/69677</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.</td>
</tr>
</tbody>
</table>
Hierarchical Bayesian Modeling of Inter-Trial Variability and Variational Bayesian Learning of Common Spatial Patterns from Multichannel EEG

Wei Wu1,2,3, Zhe Chen1,3, Shangkai Gao2, Fellow, IEEE, and Emery N. Brown1,3, Fellow, IEEE

Abstract—In numerous neuroscience studies, multichannel EEG data are often recorded over multiple trial periods under the same experimental condition. To date, little effort is aimed to learn spatial patterns from EEG data to account for trial-to-trial variability. In this paper, a hierarchical Bayesian framework is introduced to model inter-trial source variability while extracting common spatial patterns under multiple experimental conditions in a supervised manner. We also present a variational Bayesian algorithm for model inference, by which the number of sources can be determined effectively via automatic relevance determination (ARD). The efficacy of the proposed learning algorithm is validated with both synthetic and real EEG data. Using two brain-computer interface (BCI) motor imagery data sets we show the proposed algorithm consistently outperforms the common spatial patterns (CSP) algorithm while attaining comparable performance with a recently proposed discriminative approach.

I. INTRODUCTION

We consider an EEG blind source separation problem when data are available for multiple experimental conditions. This type of problem is pervasive in neuroscience studies as many experimental designs involve contrasting brain states under multiple conditions. In general, a number of functionally distinct assemblies of brain activities underlie these experimental conditions. Each neural assembly may consist of synchronized activities at multiple brain regions, be they close or distant to each other. At the scalp surface each neural assembly assumes an associated spatial pattern that is a smeared version of the assembly’s activities in the brain, due to volume conduction effects. Collectively, the scalp EEG signals can be modeled as a linear and instantaneous mixture of the spatial patterns for different assemblies. Therefore, the signal processing challenge is to extract individual spatial patterns from multichannel EEG signals. In the sequel, in order to keep consistent with the terminology in the blind source separation literature we simply call each neural assembly a “source”.

A wide range of approaches have been proposed to identify the hidden sources (or brain patterns), among which two are noteworthy. One unsupervised blind source separation methodology that has proven to be successful in EEG signal processing is independent component analysis (ICA) [1]. In ICA, the temporal activities of the non-Gaussian sources are assumed to be mutually statistically independent. However, ICA attempts to separate the sources for each individual experimental condition without using the class label information. In addition, the statistical independence that is assumed for mathematical convenience is hard to justify or fully satisfy in practice. Another promising algorithm that has been introduced to extract spatial patterns from EEG data is the common spatial patterns (CSP) algorithm [2] (a.k.a. Fukunaga-Koontz transform (FKT) [3] in the pattern recognition field). In contrast to ICA, CSP works in a supervised fashion with an aim to maximize the variance ratio between the two given conditions. Notably, CSP has been successfully employed in brain-computer interfaces (BCIs) as a spatial filtering algorithm, as evidenced by recent international BCI competitions. Despite that it optimizes a discriminative criterion, our earlier work nonetheless provided a generative formulation of CSP [4]. In addition to providing insights, the generative formulation of CSP has the advantage of opening the possibility for further improvement. The main contribution of this paper is to propose a hierarchical extension of the generative model that can take into account inter-trial variability in the EEG data. Another contribution of this paper is that we derive a variational Bayesian learning algorithm for the inference of the hierarchical model, which enables us not only to obtain probabilistic estimates of the model parameters but also to automatically determine the model size (i.e., source number). We provide both simulated and real data experiments to validate the proposed model and algorithm.

II. METHODS

Notations: Let $k$ be the index for experimental conditions; $i = 1, \ldots, N_k$ be the index for trials; $j = 1, \ldots, J_k$ be the index for the sample points of each trial; $c = 1, \ldots, C$ be the index for channels; $m = 1, \ldots, M$ be the index for sources. $N(\mu, \Sigma)$ denotes the Gaussian distribution with mean $\mu$ and covariance $\Sigma$; $\text{Gam}(a, b)$ denotes the gamma distribution defined as $P(y|a, b) = \frac{1}{\Gamma(a)} b^a y^{a-1} e^{-by}$.

A. A Probabilistic Model for Learning CSPs

We first briefly describe a probability model [4] for learning CSPs from single-trial EEG under two experimental conditions. The probability model provides a generative view of CSP, which casts the solution as a maximum likelihood (ML) estimate from the probability model. The probability model is defined as

$$
\begin{align*}
\mathbf{x}_{ij}^{(k)} &= \mathbf{A} \mathbf{z}_{ij}^{(k)} + \mathbf{c}_{ij}^{(k)} \\
\mathbf{z}_{ij}^{(k)} &\sim N(0, \mathbf{A}^{-1}_k), \mathbf{c}_{ij}^{(k)} \sim N(0, \mathbf{\Psi}_k^{-1})
\end{align*}
$$

(1)

where $\mathbf{x}_{ij}^{(k)}$ is the zero-mean multichannel EEG signal at the $j$-th sample in the $i$-th trial for condition $k$, $\mathbf{z}_{ij}^{(k)}$ is the latent vector that contains the source activities, $\mathbf{c}_{ij}^{(k)}$ is the additive noise term, $\mathbf{A} \in \mathbb{R}^{C \times M}$ is the mixing matrix that contains spatial patterns, which is assumed to be identical for both conditions, $\mathbf{A}_1, \mathbf{A}_2, \mathbf{\Psi}_1, \mathbf{\Psi}_2$ are diagonal precision matrices, implying that the sources are uncorrelated, and that the observed variables are uncorrelated given the sources. The assumption of an identical mixing matrix for two conditions is useful if we hold the belief that the two conditions share certain spatial patterns as it enables us to use inter-class information to identify the mixing matrix. Another important consequence of the assumption is that it facilitates the uniqueness since the inherent rotational indeterminacy in the standard factor analysis model is eliminated up to scaling and...
permutation indeterminancies. The connection between model (1) and the CSP algorithm has been established in [4].

The inter-trial variability in EEG recordings is a common phenomenon. When multiple trials of EEG data are collected for each condition, as is typically the case for many EEG experiments, the assumption that the level of responses across trials is constant does not hold in general. Examples of trial-to-trial variations in human EEG recordings include the habituation effects [5], the P300 effects [6] and the event-related desynchronization/synchronization (ERD/ERS) effects [7]. However, model (1) is inadequate to capture inter-trial variations since the spatial covariance matrix of the data is fixed for each condition across trials. Below we provide a hierarchical extension of model (1) to address this issue in an effort to capture the inter-trial variability.

B. A Hierarchical Bayesian Model for Learning CSPs

The hierarchical modeling technique provides a powerful tool for characterizing both within-trial and inter-trial effects [8]. Specification and inference of a hierarchical model can be carried out under either classical or Bayesian framework. Here we follow a Bayesian framework as it provides a coherent and unified treatment for quantifying uncertainties. Moreover, as can be seen shortly, Bayesian framework allows us to employ the idea of automatic relevance determination (ARD) [9] to infer the source number.

Given multiple trials of EEG data recorded under two conditions, the hierarchical model can be constructed by modifying (1) as

\[
\begin{align*}
\mathbf{x}^{(ij)} &= \mathbf{A}\mathbf{z}^{(ij)} + \mathbf{e}^{(ij)} \quad (k = 1, 2) \\
\mathbf{z}^{(ij)} &\sim \mathcal{N}(\mathbf{0}, \mathbf{A}_k^{-1}) \\
\mathbf{A}_k &\sim \text{Gam}(\alpha, \beta) \\
\end{align*}
\]

Comparing models (2) and (1), we observe two major differences. First, both the data and the model parameters are now treated as random variables with assumed prior distributions. Second, the sources are no longer required to have the same variance components across trials. Effectively this allows the amplitude of each source to vary from trial to trial. Although the actual variability may not be restricted to the source amplitude, it is especially appropriate for modeling ERD/ERS that typically accompanies motor imagery since both phenomena reflect modulations in the power of ongoing EEG activity. In addition, there is a constraint that under the same condition, variance components are modeled by the same distribution a priori, which in our case is a gamma distribution. This constraint enforces our prior belief that every condition has an underlying mechanism of power modulation different from each other. The hyperparameters \(\{c_k, d_k\}\) in the prior distributions can be estimated in a principled way via empirical Bayes technique.

Within the Bayesian framework we also have to specify the prior distribution for the mixing matrix \(\mathbf{A}\). This is where the idea of ARD could come into play for model selection: Starting from the square (i.e., \(M = C\)) case, we assign each column of \(\mathbf{A}\) with an associated hyperparameter to inversely control its magnitude/relevance. The idea is that if the posterior distribution of the hyperparameter turns out to be concentrating upon large values, the corresponding column in the mixing matrix would essentially be switched off, with only the relevant columns remaining. In this fashion the number of the sources can be determined in conjunction with parameter estimation, which is in contrast with the classical model selection criteria such as AIC and BIC, where model selection and parameter estimation are treated at different stages. Specifically, in the current problem we place a hierarchical structure on the prior distribution of each column of the mixing matrix as

\[
\begin{align*}
a_{mn} &\sim \mathcal{N}(0, (\alpha^{(m)})^{-1}) \\
\alpha &\sim \text{Gam}(u, v)
\end{align*}
\]

where \(a_{mn}\) denotes the \(m\)-th column of \(\mathbf{A}\). \(I\) denotes the identity matrix, the hyperparameter \(\alpha^{(m)}\) denotes the precision of \(a_{mn}\). As opposed to the earlier hierarchical structure for modeling inter-trial variations, the hierarchical structure imposed here is primarily for the ease of implementing the ARD. The hierarchical structure also encourages the sparsity since the marginal distribution of \(a_{mn}\) by integrating out \(\alpha^{(m)}\) yields a Student-t distribution

\[
P(a_{mn}) \propto \left( u + \frac{a_{mn}^2}{2} \right)^{-\frac{u+m+1}{2}}
\]

Note that the Student-t distribution encompasses two well-known sparse priors as special cases: the Laplace prior is recovered when \(u=1\), whereas the Gaussian-Jeffreys prior is recovered in the limiting case when \(u \rightarrow 0\) and \(v \rightarrow 0\). In the current work we use the Gaussian-Jeffreys prior as it alleviates the need to determine further hyperparameters. In addition, we use the noninformative Jeffreys prior for \(\Psi_k\), i.e., \(\epsilon_k \sim 0\) and \(f_k \sim 0\), as mentioned earlier, we do not place a noninformative prior on \(\Lambda_k\) since we hold the belief that the priors for \(\Lambda_k\) and \(\Lambda_k^2\) should be different. The proposed hierarchical model is shown in Fig. 1 as a directed graph (using the tool of graphical models [10]).

C. The Variational Bayesian (VB) Algorithm for Model Inference

The essential goal of Bayesian inference is to compute the posterior distribution of the parameters and latent variables using Bayes’ rule. Exact Bayesian inference in many cases is intractable, thus various techniques for approximate inference have been proposed [10]. We adopt a variational approximation in our work as it permits fast computation and enables us to glean intuition as well as insight into the results. The VB inference is also invariant to re-parameterization of the model.

The VB approach can be viewed as a generalization of the expectation-maximization (EM) algorithm by maintaining posterior distributions over both parameters and latent variables. Specifically, VB maximizes a lower bound of the log marginal likelihood of the data over a restricted family of probability distributions

\[
\ln P(X) = \ln \int \frac{Q(\Theta)}{Q(\Theta)} d\Theta \geq \ln Q(\Theta) \ln \frac{P(X, \Theta)}{Q(\Theta)} d\Theta = \mathcal{L}(Q)
\]

where \(\Theta\) denotes collectively all the parameters and latent variables in the hierarchical model. The optimal \(Q\) that maximizes the lower bound \(\mathcal{L}\) is referred to as variational posterior. It can be shown that
the maximization of $\mathcal{L}$ is equivalent to minimizing the Kullback-Leibler (KL) divergence between the variational posterior and the true posterior. The basic idea of VB inference is to assume a certain structure for $Q$ in order to make $\mathcal{L}$ easy to compute. In this work we adopt the mean-field approximation, assuming that $Q$ can be factorized over $\Theta$. Due to space limit we present the update equations for variational posterior $Q^*$ below for model (2) by skipping step-by-step derivations. We also omit the formulæ for evaluating $\mathcal{L}$ as its only use is for checking the algorithmic convergence. The variational posteriors in each iteration are

$$Q^*(Z_k) = \prod_{i=1}^{C} \prod_{l=1}^{N} N(\mathbf{z}_{ik}, \Sigma_{ik}), Q^*(\Lambda) = \prod_{i=1}^{C} N(\mathbf{\Lambda}_i; \Sigma_{\Lambda}),$$

where $\mathbf{z}_{ik}$ denotes the transpose of the $c$-th row of $\mathbf{Z}$. The parameters are given by

$$\Sigma_{\Lambda} = \text{diag}(\langle \mathbf{\Lambda} \rangle),~\Sigma_{\Lambda}^{-1} = \left[ \sum_{k=1}^{N} \sum_{i=1}^{J} \sum_{j=1}^{N} (\mathbf{\Lambda}_k)^{ij}(\mathbf{\Lambda}_k)^{ij\top} \right]^{-1},$$

where $\langle \cdot \rangle$ denotes the expectation over the random variables inside the angle bracket. An ideal way of computing the empirical Bayes estimates of the hyperparameters $\{c_k, d_k\}$ is by maximizing the marginal likelihood of the data. However, since the marginal likelihood is difficult to evaluate for the same reason why exact Bayesian inference is intractable, we instead resort to maximizing the lower bound of $\mathcal{L}$, hoping that this would increase the marginal likelihood as well, or even if it does not, the bound would at least become tighter. Maximizing $\mathcal{L}$ w.r.t. the hyperparameters $\{c_k, d_k\}$ leads to the following update equations

$$-F(c_k) + F(c_k^{\text{new}}) = \ln \frac{d_k}{d_k^{\text{new}}}, \quad \frac{d_k^{\text{new}}}{d_k} = \frac{\gamma(c_k)}{\gamma(c_k^{\text{new}})}$$

where $F$ denotes the digamma function. $\{c_k, d_k\}$ can then be obtained by solving the above fixed-point equations iteratively.

### III. EXPERIMENTAL RESULTS

#### A. Results on Synthetic Data

First, we compare the performance of the VB algorithm for the hierarchical model (HVB), the standard CSP algorithm, and the EM algorithm for maximum likelihood estimation (as described in [4]) on the reconstruction accuracy of the mixing matrix via Monte Carlo simulations. In each Monte Carlo run, two sets of 10 mutually uncorrelated sources are generated, with each set corresponding to one class. Each source comprises 1,000 data points that are independently and identically Gaussian distributed with zero mean. The standard deviations of the 10 sources from one class are in descending order from 10 to 1, while for the other class the standard deviations are ascending from 1 to 10. A $20 \times 10$ mixing matrix is also randomly generated, with each entry zero-mean Gaussian distributed. Additive white Gaussian noise is simulated, resulting in observations with varying SNR from 0 to 20 dB.

The noisy mixture signals are then presented to all three algorithms. Since CSP specifically addresses the square mixing ($20 \times 20$) case, we select the 10 columns that correspond to the 5 largest and 5 smallest eigenvalues to form the estimated mixing matrix $\mathbf{A}$. For EM, we test its sensitivity to the mis-specification of the source number by choosing $M = 12$ (the EM estimate will converge to m.l.e. for the true source number case when $M = 10$ is assumed), which exceeds the true source number by only 2. The estimated mixing matrix is then formed similarly to the CSP case. For HVB, the 10 columns in the posterior mean of $\mathbf{A}$ that correspond to the 10 smallest $c_k$ are selected to form $\mathbf{A}$.

The Amari index is defined as a measure of the closeness of $\hat{\mathbf{A}}$ and the true mixing matrix $\mathbf{A}$ [4], which is invariant to permutation and scaling of the columns of $\mathbf{A}$. The mean/SD of Amari indices (averaged over 50 runs) are plotted in Fig. 2. With no surprise, the index increases with increasing SNR for all methods. However, HVB outperforms both CSP and EM when the source number is mis-specified under the same SNR.

To illustrate the capability of HVB in identifying the true source number, the algorithms’ results of one Monte Carlo run for 20dB SNR are illustrated in Fig. 3 using the Hinton diagrams. Note that CSP fails to prune the redundant columns in the mixing matrix, whereas HVB can recover the true source dimensionality.

#### B. Results on Real EEG Data

Next, we demonstrate the prediction performance of the hierarchical Bayesian learning framework on two multi-trial BCI data sets. Both EEG data sets were recorded during motor imagery experiments in which subjects were instructed to perform imaginary movement in each trial. The first data set (left/right-hand imagery), which is described in detail in [11], consists of 7 subjects, with 120 trials of data per class for each subject. The second data set (right-hand/foot imagery), which is data set IVa from BCI Competition.

---

Fig. 2. Comparison of Amari indices between the HVB, CSP and EM algorithms at different signal-to-noise ratios (SNR).

Fig. 3. An illustration where ARD can effectively determine the source number. (a) Hinton diagram of the true mixing matrix $\mathbf{A}$; (b, c) Hinton diagrams of $\mathbf{A}$ estimated from CSP and HVB, respectively. Each entry in the matrices is represented by the size of the area of a white (positive value) or black (negative value) square.
III, consists of 5 subjects, with 280 trials per each.

We compare HVB with CSP on the first data set. For each subject, the complete 240 trials are split into a training set of 160 trials and a test set of 80 trials, with an equal number of trials for each class. All learning procedures are performed on the training set, while the test set is reserved only for the evaluation of classification accuracies. To avoid the potential bias that may be introduced by ad-hoc tuning, we apply identical preprocessing settings, i.e., band-pass filtering (8-30Hz), time windowing (a 3.5s imagination period), and channel selection (all 32 channels), to the raw EEG data before presenting them to the two algorithms. Fisher linear discriminant analysis is then employed as the classifier for both algorithms. As the input to the classifier, the feature vectors for each trial is constructed to be the log-variance of the sources. The conventional way of selecting n sources for each class based on the variance ratio is adopted for both algorithms. For each subject, the optimal n is determined using 10-fold cross-validation on the training set. The classification accuracies of the two algorithms for the 7 subjects are listed in Table I. It can be seen that HVB achieves superior performance than CSP for 4 of 7 subjects, while achieving the same prediction performance for the other 3 subjects. By examining the results, for all the subjects we confirm that HVB can indeed extract physiologically meaningful spatial patterns that CSP sometimes fails to identify (data not shown due to space limit).

When testing the publicly available BCI competition data set, we compare the performance of HVB, CSP, and a recently proposed discriminative algorithm named logistic regression with spectral regularization (LRSR) [12]. We use the same settings as those provided on [13] to preprocess the raw data for each algorithm. For LRSR, 10-fold cross-validation is employed to estimate the optimal regularization parameter for each subject. The log-matrix heuristic provided on [13] is adopted to ensure LRSR to obtain its best performance. The classification accuracies for the three algorithms on the test sets are listed in Table I. As shown, HVB and LRSR have comparable prediction performances, which are both consistently better than CSP’s on 4 subjects. However, it should be noted that LRSR is a discriminative algorithm designed specifically for classification instead of modeling the data, whereas our generative model offers a better data interpretation.

IV. DISCUSSION AND FUTURE WORK

The hierarchical modeling framework proposed in this paper is an extension of our earlier work presented in [4]. The CSP algorithm deals with multi-trial EEG data by simply ignoring trial-to-trial variations and assuming a constant spatial covariance matrix for each experimental condition, whereas in our new framework the inter-trial variations are taken into account in that the amplitude is allowed to vary across trials. Although the possible factors of variability in reality are by no means restricted to amplitude, it is nonetheless appropriate for modeling ERD/ERS as both phenomena reflect modulations in the power of ongoing EEG activity. A similar treatment for event-related potentials (ERPs) can be found in [14]. At this point, it is worth clarifying the distinctions between the modeling framework in this paper and the one in [14]: Because ERP is a phase-locked response in EEG, it is typically modeled as a fixed effect, i.e., its waveform is assumed to be constant across trials. By contrast, induced responses such as ERD/ERS reflect changes in ongoing EEG activity that are not phase-locked to any external stimuli, hence they are modeled as random effects.

In the experimental section of this paper, for real data analysis we have chosen BCI classification as an example to demonstrate the effectiveness of the proposed algorithm. Such consideration is due to the fact that the classification accuracy in BCI provides an objective measure to compare the performance between different algorithms. However, the applicability of our algorithm is not confined to BCI studies. Indeed, the algorithm readily applies to other neuroscience studies where spatial patterns are of interest given multi-trial multichannel EEG data from multiple experimental conditions. Extension of the current framework to cases where there are more than two conditions is straightforward by expanding the number of sub-models in model (2).

Finally, our work serves as a basis for exploring various extensions. First, we can integrate the idea of replacing Gaussian distribution by Student-t distribution as in [4] into the current framework, which can make the model more robust against the data outliers. Second, similar models can be developed to learn common spatial patterns of ERP from EEG data. Again, it would be very interesting to compare its prediction performance with the discriminative framework in [15] on ERP data sets, such as P300.

REFERENCES