Bayesian Factorial Linear Gaussian State-Space Models for Biosignal Decomposition

Silvia Chiappa and David Barber

***Abstract*— We discuss a method to extract independent dynamical systems underlying a single or multiple channels of observation. In particular, we search for one dimensional subsignals to aid the interpretability of the decomposition. The method uses an approximate Bayesian analysis to determine automatically the number and appropriate complexity of the underlying dynamics, with a preference for the simplest solution. We apply this method to unfiltered EEG signals to discover low complexity sources with preferential spectral properties, demonstrating improved interpretability of the extracted sources over related methods.**

I. Introduction

Decomposing a multivariate time-series v_t^n, $t = 1, \ldots, T$, $n = 1, \ldots, V$ into a set of C simpler subsignals (sources) is a central goal in signal processing and is of particular interest in the analysis of biomedical signals. The goal of this paper is to introduce a model which can automatically determine the number of sources underlying the observations and in which we can bias the sources to be in certain frequency ranges. Furthermore, we are interested in taking into account the temporal structure of the time-series which can help in obtaining a good decomposition, especially when $C > V$. More specifically, our criterion for the decomposition is that independent dynamical systems generate the sources which, under linear noisy mixing, give rise to the observations. For any two scalar sources s_t^i and s_t^j and all times t, we seek a model of statistically independent dynamics $p(s_{1:T}^i, s_{1:T}^j) = p(s_{1:T}^i)p(s_{1:T}^j)$. Furthermore, the aim is to find a matrix W that relates the sources $s_t = vert(s_t^1, \ldots, s_t^C)$ to observations $v_t = vert(v_t^1, \ldots, v_t^V)$ through noisy mixing[1]. This is a form of Independent Components Analysis (ICA) [1] although it differs from the more standard assumption of independence at each time step, that is $p(s_{1:T}^i, s_{1:T}^j) = \prod_{t=1}^T p(s_t^i)p(s_t^j)$. We consider a Linear Gaussian State-Space Model (LGSSM), which is a powerful, yet interpretable and tractable, model. We constrain the LGSSM in order that independent dynamical processes can be identified and furthermore that scalar sources can be extracted from the signal. To determine the correct number of underlying processes and bias the solution towards a certain dynamics, we use a Variational Bayesian analysis which defines a prior distribution over the model parameters.

There are several existing decomposition methods which encode constraints such as desired frequencies of the independent sources (see for example [2], [3]). However, these methods do not automatically determine the correct number of underlying sources nor do they consider the dynamics of the signal in the model structure. A closely related technique to ours is (Non) Linear Dynamical Factor Analysis (NDFA) [4], [5]. Whilst being an attractive and powerful method, standard NDFA places no constraint that the observations are formed from mixing independent *scalar* sources, which makes interpretation of the resulting sources difficult. Furthermore, NDFA does not directly force the sources to contain particular frequencies but rather attempts to bias the discovered sources by careful initialization [5]. In addition, NDFA uses nonlinear state dynamics (and mixing), which hampers inference and makes the incorporation of known constraints more complex.

Inference in the Variational Bayesian LGSSM has previously been achieved using Belief Propagation, and differs from inference in the Kalman filtering/smoothing literature, for which highly efficient and stabilized procedures exist. A central contribution of this paper is to show how inference *can* be performed using the standard Kalman filtering/smoothing recursions by augmenting the original model.

[1] $vert(a, b, c)$ is the matrix formed by vertically stacking a, b and c.

II. Factorial Linear Gaussian State-Space Models

In LGSSMs [6], the hidden state vectors $h_{1:T}$ and the visible observations $v_{1:T}$ are linearly related by:

$$h_t = Ah_{t-1} + \eta_t^h, \;\; h_1 \sim \mathcal{N}(\mu, \Sigma), \;\; \eta_t^h \sim \mathcal{N}(\mathbf{0}_H, \Sigma_H)$$
$$v_t = Bh_t + \eta_t^v, \;\; \eta_t^v \sim \mathcal{N}(\mathbf{0}_V, \Sigma_V),$$

where $\mathcal{N}$ denotes a Gaussian distribution. The notation $\mathbf{0}_D$ stands for a $D \times 1$ zero vector. Probabilistically:

$$p(v_{1:T}, h_{1:T}) = p(v_1|h_1)p(h_1)\prod_{t=2}^{T} p(v_t|h_t)p(h_t|h_{t-1}),$$

with $p(v_t|h_t) = \mathcal{N}(Bh_t, \Sigma_V)$ and $p(h_t|h_{t-1}) = \mathcal{N}(Ah_{t-1}, \Sigma_H)$. To make independent dynamical subsystems we use block diagonal transition and state noise matrices A, Σ_H and Σ, where each block c has dimension H_c. A one dimensional source s_t^c for each independent dynamical subsystem is formed from $s_t^c = \mathbf{1}_c^\mathsf{T} h_t^c$, where $\mathbf{1}_c$ is a unit vector and h_t^c is the state of the dynamical system c. Combining the sources, we can write $s_t = Ph_t$, where $P = diag(\mathbf{1}_1^\mathsf{T}, \ldots, \mathbf{1}_C^\mathsf{T})$, $h_t = vert(h_t^1, \ldots, h_t^C)$. The resulting emission matrix is constrained to be of the form $B = WP$, where W is the $V \times C$ mixing matrix and P is a $C \times H$ projection, with $H = \sum_c H_c$. Such a constrained form for B is required to provide interpretable scalar sources.

Bayesian Factorial Linear Gaussian State-Space Models

In our Bayesian treatment of learning we define the priors $p(A|\alpha)$ and $p(W|\beta)$, where α and β are hyperparameters. We

do not define any prior for Σ_H, Σ_V, μ and Σ, which will formally be considered as hyperparameters[2]. The total set of hyperparameters is $\Theta = \{\alpha, \beta, \Sigma_H, \Sigma_V, \mu, \Sigma\}$. Therefore:

$$p(v_{1:T}|\Theta) = \int_{A,W} p(v_{1:T}|A, W, \Theta)p(A|\alpha)p(W|\beta)dAdW\,. \quad (1)$$

Here we take the ML-II ('evidence') framework, which involves maximizing $p(v_{1:T}|\Theta)$ with respect to Θ [4], [7]. Ideally, the number of sources effectively contributing to the observed signal should be small. This suggests the prior:

$$p(W|\beta) = \prod_{j=1}^{C} \left(\frac{\beta_j}{2\pi}\right)^{V/2} e^{-\frac{\beta_j}{2}\sum_{i=1}^{V} W_{ij}^2}\,.$$

We can bias A to be close to a desired transition $\hat{A}$ (possibly zero) by using:

$$p(A^c|\alpha_c) = \left(\frac{\alpha_c}{2\pi}\right)^{H_c^2/2} e^{-\frac{\alpha_c}{2}\sum_{i,j=1}^{H_c}\left(A_{ij}^c - \hat{A}_{ij}^c\right)^2}$$

for each component c, so that $p(A|\alpha) = \prod_c p(A^c|\alpha_c)$[3].

Variational Bayes

Optimizing Eq. (1) with respect to Θ is difficult due to the intractability of the integrals. Instead we consider the lower bound [4], [7]:

$$\begin{aligned}\mathcal{L} = \log p(v_{1:T}|\Theta) \geq & H_q(A, W, h_{1:T}) \quad (2)\\ & + \langle \log p(v_{1:T}, h_{1:T}, A, W)\rangle_{q(A,W,h_{1:T})}\,,\end{aligned}$$

where we dropped the explicit dependence on Θ on the rhs[4]. The notation $H_d(x)$ signifies the entropy of the distribution $d(x)$, and $\langle\cdot\rangle_{d(x)}$ denotes the expectation operator. For certain variational distributions q, we hope to achieve a tractable bound, which we may then optimize with respect to q and Θ. The key approximation in Variational Bayes (VB) is $q(A, W|h_{1:T}) \equiv q(A, W)$. Since A and W separate in the rhs of Eq. (2), optimally $q(A, W) = q(A)q(W)$, hence:

$$\begin{aligned}\mathcal{L} \geq & - D(q(A), p(A)) - D(q(W), p(W)) + H_q(h_{1:T})\\ & + \langle \log p(v_{1:T}, h_{1:T}|A, W)\rangle_{q(h_{1:T})q(A)q(W)} \equiv \mathcal{F}\,.\end{aligned}$$

$D(q(x), p(x))$ is the KL divergence $\langle \log q(x)/p(x)\rangle_{q(x)}$. The VB procedure iteratively performs co-ordinate wise ascent of $\mathcal{F}$ with respect to $q(W), q(A), q(h_{1:T})$ and Θ.

Determining $q(W)$

By examining $\mathcal{F}$, optimally, $q(W)$ is a Gaussian. The covariance $[\Sigma_W]_{ij,kl} \equiv \left\langle \left(W_{ij} - \langle W_{ij}\rangle\right)\left(W_{kl} - \langle W_{kl}\rangle\right)\right\rangle$ (averages wrt $q(W)$) is given by the inverse of:

$$\left[\Sigma_W^{-1}\right]_{ij,kl} = \left[\Sigma_V^{-1}\right]_{ik} \sum_t \left\langle \tilde{h}_t^j \tilde{h}_t^l \right\rangle_{q(h_t)} + \beta_j \delta_{ik}\delta_{jl}\,,$$

where $\tilde{h}_t = Ph_t$ and δ_{ij} is the Kronecker delta function. The mean is given by:

$$\langle W_{ij}\rangle = \sum_{k,l,n,t} [\Sigma_W]_{ij,kl} \left[\Sigma_V^{-1}\right]_{k,n} \left\langle \tilde{h}_t^l \right\rangle_{q(h_t)} v_t^n\,.$$

Determining $q(A)$

Optimally we have a factorized distribution $q(A) = \prod_c q(A^c)$, where $q(A^c)$ is Gaussian with inverse covariance given by (dropping the dependence on c):

$$\left[\Sigma_A^{-1}\right]_{ij,kl} = \left[\Sigma_H^{-1}\right]_{ik} \sum_{t=2}^{T} \left\langle h_{t-1}^j h_{t-1}^l \right\rangle_{q(h_{t-1})} + \alpha\delta_{ik}\delta_{jl}\,.$$

The mean is:

$$\langle A_{ij}\rangle = \sum_{k,l} [\Sigma_A]_{ij,kl} \left(\alpha\hat{A}_{kl} + \sum_n \left[\Sigma_H^{-1}\right]_{kn} \sum_{t=2}^{T} \left\langle h_{t-1}^l h_t^n \right\rangle_{q(h_{t-1:t})}\right).$$

Inference on $q(h_{1:T})$

Optimally $q(h_{1:T})$ is Gaussian since its log is quadratic in $h_{1:T}$, being namely[5]:

$$\begin{aligned}& -\frac{1}{2}\sum_{t=1}^{T} \left\langle (v_t - WPh_t)^{\mathsf{T}} \Sigma_V^{-1} (v_t - WPh_t)\right\rangle_{q(W)} \quad (3)\\ & -\frac{1}{2}\sum_{t=2}^{T} \left\langle (h_t - Ah_{t-1})^{\mathsf{T}} \Sigma_H^{-1} (h_t - Ah_{t-1})\right\rangle_{q(A)}\,.\end{aligned}$$

We can carry out the averages over A and W since $q(A)$ and $q(W)$ are Gaussian and the above is quadratic in the parameters A and W. In order to compute the required statistics $\langle h_t\rangle_{q(h_t)}$ and $\left\langle h_{t-1}h_t^{\mathsf{T}}\right\rangle_{q(h_{t-1:t})}$, our aim is to represent Eq. (3) as the $\log \tilde{q}(h_{1:T}|\tilde{v}_{1:T})$ of a LGSSM with some suitable parameters. To do that we use a mean + fluctuation decomposition:

$$\begin{aligned}& \left\langle (v_t - Bh_t)^{\mathsf{T}} \Sigma_V^{-1} (v_t - Bh_t)\right\rangle_{q(W)}\\ & = (v_t - \langle B\rangle h_t)^{\mathsf{T}} \Sigma_V^{-1} (v_t - \langle B\rangle h_t) + h_t^{\mathsf{T}} P^{\mathsf{T}} S_W P h_t\,,\end{aligned}$$

where $\langle B\rangle \equiv \langle W\rangle P$ and the fluctuation is by determined by:

$$[S_W]_{jl} = \sum_{i,k=1}^{V} [\Sigma_W]_{ij,kl} \left[\Sigma_V^{-1}\right]_{ik}, \quad j, l \in 1, \ldots, C. \quad (4)$$

Similarly:

$$\begin{aligned}& \left\langle (h_t - Ah_{t-1})^{\mathsf{T}} \Sigma_H^{-1} (h_t - Ah_{t-1})\right\rangle_{q(A)}\\ & = (h_t - \langle A\rangle h_{t-1})^{\mathsf{T}} \Sigma_H^{-1} (h_t - \langle A\rangle h_{t-1}) + h_{t-1}^{\mathsf{T}} S_A h_{t-1}\,,\end{aligned}$$

$$[S_A]_{jl} = \sum_{i,k=1}^{H} [\Sigma_A]_{ij,kl} \left[\Sigma_H^{-1}\right]_{ik}, \quad j, l \in 1, \ldots, H.$$

To represent Eq. (3) as a LGSSM $\tilde{q}(h_{1:T}|\tilde{v}_{1:T})$, we augment v_t and B as:

$$\tilde{v}_t = vert(v_t, \mathbf{0}_H, \mathbf{0}_C), \quad \tilde{B} = vert(\langle B\rangle, U_A, U_W P),$$

[2]A Bayesian treatment of Σ_H, Σ_V, μ and Σ is straightforward using conjugate priors (see [7], [8]) but is not taken here for space restrictions and since we have little preference for constraining these parameters.

[3]For dimensional reasons, we can also assume a Gaussian prior on the columns of W with exponent $-\frac{1}{2}\beta_j W_j^{\mathsf{T}} \Sigma_V^{-1} W_j$. This simplifies the statistics of $q(W)$ and Eq. (4). The same holds for A. This is also convenient when we assume a prior for Σ_H and Σ_V, since it ensures conjugacy [7], [8].

[4]Strictly we should write throughout $q(\cdot|v_{1:T})$. We omit the dependence on $v_{1:T}$ for notational convenience.

[5]For simplicity, we ignore the contribution from h_1 and a constant term.

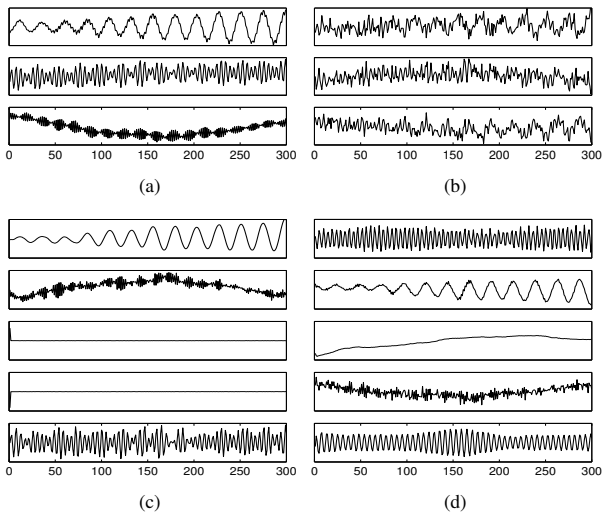

Fig. 1. (a) Original sources s_t. (b) Observations resulting from mixing the original sources, $v_t = Ws_t + \eta_t^v$, $\eta_t^v \sim \mathcal{N}(\mathbf{0}_V, I)$. (c) Recovered sources using the Bayesian FLGSSM. (d) Sources found with MAP FLGSSM. The retained sources have been rescaled to aid visualization.

where U_A is the Cholesky decomposition of S_A, so that $U_A^\mathsf{T} U_A = S_A$. Similarly, U_W is the Cholesky decomposition of S_W. The equivalent LGSSM is then completed by specifying $\tilde{A} \equiv \langle A \rangle$, $\tilde{\Sigma}_H \equiv \Sigma_H$, $\tilde{\Sigma}_V \equiv diag(\Sigma_V, I, I)$, $\tilde{\mu} \equiv \mu$, $\tilde{\Sigma} \equiv \Sigma$[6]. In this way *any* standard inference routines in the literature may be applied to compute $q(h_t) = \tilde{q}(h_t|\tilde{v}_{1:T})$, including those specifically addressed at improving numerical stability [9]. In the experiments, we used the standard predictor-corrector filtering and Rauch-Tung-Striebel smoothing [9]. A minor modification to the standard predictor-corrector filtering routine may be applied for computational efficiency (see [8] for details). This method is considerably simpler and more general than the procedure given in [7], which is based on Belief Propagation and do not correspond to any of the standard forms in the Kalman filtering/smoothing literature.

Finding the Optimal Θ

Differentiating $\mathcal{F}$ with respect to Θ we find that, optimally:

$$\beta_j = \frac{V}{\sum_i \left\langle W_{ij}^2 \right\rangle_{q(W)}}, \qquad \alpha_c = \frac{H_c^2}{\sum_{i,j} \left\langle [A^c - \hat{A}^c]_{ij}^2 \right\rangle_{q(A^c)}},$$

$$\Sigma_V = \frac{1}{T} \sum_{t=1}^{T} \left\langle (v_t - WPh_t)(v_t - WPh_t)^\mathsf{T} \right\rangle_{q(W)q(h_t)},$$

$$\Sigma_H^c = \frac{1}{T-1} \sum_{t=2}^{T} \left\langle \left(h_t^c - A^c h_{t-1}^c\right)\left(h_t^c - A^c h_{t-1}^c\right)^\mathsf{T} \right\rangle_{q(A^c)q(h_{t-1:t}^c)},$$

$$\Sigma = \left\langle (h_1 - \mu)(h_1 - \mu)^\mathsf{T} \right\rangle_{q(h_1)}, \qquad \mu = \langle h_1 \rangle_{q(h_1)} .$$

A. Demonstration

In a proof of concept experiment, we used a FLGSSM to generate 3 sources s_t^c with random 5×5 transition matrices A^c, $\mu = \mathbf{0}_H$ and $\Sigma \equiv \Sigma_H \equiv I$, see Fig. 1a. The sources were mixed into three observations $v_t = Ws_t + \eta_t^v$, for W chosen with elements from a zero mean unit variance Gaussian distribution, and $\Sigma_V = I$ (Fig. 1b). We then trained a Bayesian FLGSSM with 5 sources and 7×7 transition matrices A^c. To bias the model to find the simplest sources, we used zero matrices $\hat{A}^c$ for all sources. In Fig. 1c we plot the estimated sources from our method after convergence. Two of the 5 sources have been removed, and the remaining three are a reasonable estimation of the original sources. Another possible approach for introducing prior knowledge is to use a Maximum a Posteriori (MAP) procedure by adding a prior term to the original log-likelihood $\log p(v_{1:T}|A, W, \Theta) + \log p(A|\alpha) + \log p(W|\beta)$. However, it is not clear how to reliably find the *hyperparameters* α and β in this case. One solution is to estimate them by optimizing the new objective function jointly with respect to the parameters and hyperparameters (this is the so-called joint map estimation – see for example [10]). The complexity of this approach is similar to the unaugmented Bayesian LGSSM, although in this case solving a Sylvester equation is required for updating the parameters. A typical result of using this joint MAP approach on the artificial data is presented in Fig. 1d. The joint MAP does not estimate the hyperparameters well, so that an incorrect number of sources is found, and the sources are not as well estimated as in the Bayesian procedure.

B. Application to EEG Analysis

In Fig. 2a we plot three seconds of EEG data recorded from 4 channels (located in the right hemisphere) while a person is performing imagined movement of the right hand. As is typical in EEG, each channel shows drift terms below 1 Hz which correspond to artifacts of the instrumentation, together with the presence of 50 Hz mains contamination. These effects mask the rhythmical activity related to the mental task, mainly centered at 10 and 20 Hz, which we want to extract. Standard ICA methods such as FastICA do not find satisfactory sources based on raw 'noisy' data, and preprocessing with band-pass filters is usually required. Additionally, in EEG research, flexibility in the number of recovered sources is important, since there may be many independent oscillators of interest underlying the observations and we would like some way to automatically determine their effective number. To preferentially find sources at particular frequencies, we specified a block diagonal matrix $\hat{A}^c$ with each block being a rotation at the desired frequency ω: $\begin{pmatrix} \cos(2\pi\omega/N) & -\sin(2\pi\omega/N) \\ \sin(2\pi\omega/N) & \cos(2\pi\omega/N) \end{pmatrix}$, where N is the number of samples per second. In order to extract the dominant drifts below 1 Hz, the mains contaminations and the oscillations related to the mental task, we defined the following 16 groups of frequencies ω: [0.5], [0.5], [0.5], [0.5]; [10,11], [10,11], [10,11], [10,11]; [20,21], [20,21], [20,21], [20,21]; [50], [50], [50], [50]. Hence, the total hidden dimension of the FLGSSM is $H = 48$. The temporal evolution of the sources obtained after training the Bayesian FLGSSM is shown in Fig. 2b (grouped by frequency range). This method removed 4 unnecessary sources from the mixing matrix W, that is one [10,11] Hz and three [20,21] Hz sources. We can see that the

[6]Strictly, we need a time-dependent emission $\tilde{B}_t = \tilde{B}$, for $t = 1, \ldots, T-1$. For time T, $\tilde{B}_T$ has the Cholesky factor U_A replaced by a zero matrix.

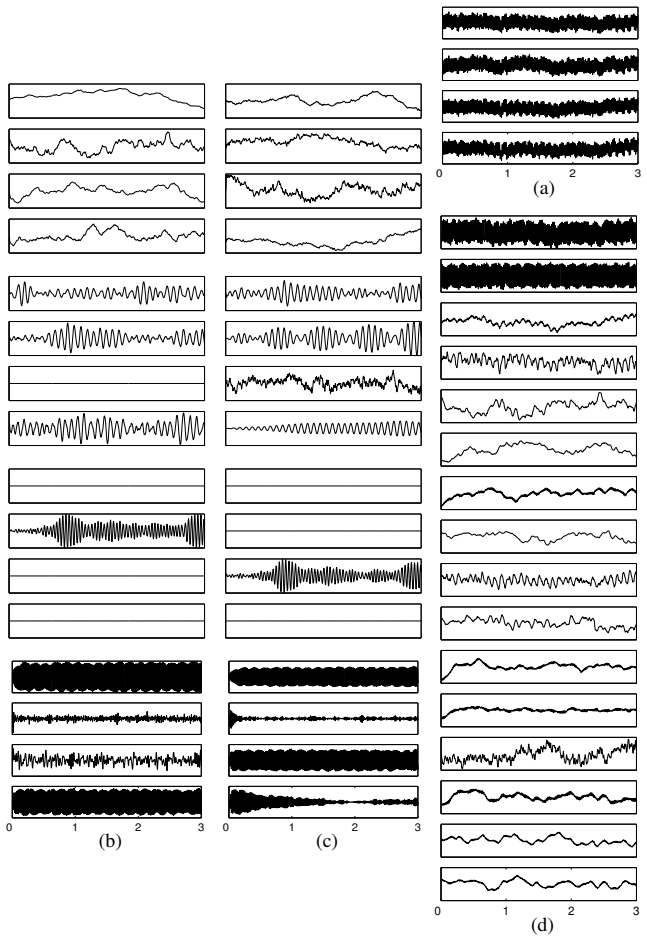

Fig. 2. (a) Three seconds of unfiltered EEG data recorded from 4 electrodes. (b) The 16 sources s_t estimated by the Bayesian FLGSSM. (b) Sources estimated by the MAP FLGSSM. (c) The 16 factors estimated by NDFA. The retained sources have been rescaled to aid visualization.

first 4 sources (counting from the top down) contain dominant low frequency drift, source 5, 6 and 8 contain [10,11] Hz, while source 10 contains [20,21] Hz centered activity. Out of the 4 sources initialized to 50 Hz, only 2 retained 50 Hz activity, while the A^c of the other two have changed to model other frequencies present in the EEG.

The MAP FLGSSM approach is presented in Fig. 2c. We can see that none of the [10,11] Hz sources has been removed, even if contribution of source 8 to the observations is relatively small. One A^c biased at [10,11] Hz includes other frequencies in addition to 10 Hz (source 7). As in the Bayesian case, only one [20,21] Hz component is retained. There are two dominant 50 Hz components, however none of 50 Hz sources has been removed. In conclusion, the Bayesian FLGSSM seems better able to remove unnecessary components and gives cleaner sources at the desired frequencies.

To asses the advantage of using prior frequencies for extracting task-related information and the potential limitations of using a linear model, we compared our method with NDFA [4]. We extracted 16 factors using a NDFA model in which both MLPs had one hidden layer of 30 neurons. In Fig. 2d we show the temporal evolution of the resulting factors. The first 10 factors from the top give the strongest contribution to the observations. In agreement with the Bayesian FLGSSM, there are 2 main 50 Hz sources (first two factors), although a small 50 Hz activity is present also in other factors, namely 7, 11, 12 and 14. The slow drift has not been isolated and is present in almost all factors. The information related to hand movement, namely [10,20] Hz activity, is spread over factors 3, 4, 9, 10 and 13, which however contain also other frequencies. The prior specification of independent dynamical processes at particular frequencies has therefore helped the Bayesian FLGSSM to better isolate the activity of interest into a smaller number of sources and, among these sources, to separate the contribution of oscillators at 10 Hz and 20 Hz.

III. Conclusion

We presented a method to identify independent dynamical sources in noisy temporal data, based on a Bayesian procedure which automatically biases the solution to finding a small number of sources with preferential dynamics. This procedure is closely related to others previously proposed in the literature, but has the property that the sources are themselves projections from higher dimensional independent linear dynamical systems. Here we concentrated on the projection to a single dimension since this aids interpretability of the signals, being of particular importance for applications in biomedical signal analysis. An advantage of our linear dynamics approach is tractability of inference, and we demonstrated how the statistics of the hidden variables in the Bayesian LGSSM can be estimated by using any Kalman filtering/smoothing routine. The method is able then to automatically extract signals, for example, biased towards particular frequencies.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley and Sons, 2001.
[2] K. Knuth, "Bayesian source separation and localization," in *SPIE'98: Bayesian Inference for Inverse Problems*, 1998, pp. 147–158.
[3] J. Särelä and H. Valpola, "Denoising source separation," *Journal of Machine Learning Research*, vol. 6, pp. 233–272, 2005.
[4] H. Valpola and J. Karhunen, "An unsupervised ensemble learning method for nonlinear dynamic state-space models," *Neural Computation*, vol. 14, pp. 2647–2692, 2002.
[5] J. Särelä, H. Valpola, R. Vigário, and E. Oja, "Dynamical Factor Analysis of Rhythmic Magnetoencephalographic Activity," in *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, 2001, pp. 457–462.
[6] Y. Bar-Shalom and X. Li, *Estimation and Tracking: Principles, Techniques and Software*. Artech House, 1998.
[7] M. Beal, "Variational Algorithms for Approximate Bayesian Inference," Ph.D. Thesis, Gatsby Computational Neuroscience Unit, University College London, UK, 2003.
[8] S. Chiappa, "Analysis and classification of EEG signals using probabilistic models for brain computer interfaces," Ph.D. Thesis, EPFL, Lausanne, Switzerland, 2006.
[9] M. Grewal and A. Andrews, *Kalman Filtering: Theory and Practice Using MATLAB*. John Wiley and Sons, Inc., 2001.
[10] S. Saquib, C. Bouman, and K. Sauer, "ML parameter estimation for Markov random fields with applicationsto Bayesian tomography," *IEEE Transactions on Image Processing*, vol. 7, pp. 1029–1044, 1998.