Bayesian Linear Gaussian State Space Models for Biosignal Decomposition

Silvia Chiappa and David Barber [a]

IDIAP–RR 05-84

[a] IDIAP Research Institute, Rue du Simplon 4, CH-1920 Martigny, Switzerland

IDIAP Research Report 05-84

Bayesian 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.

1 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 (see for example [1]). Our criteria for the decomposition is that independent dynamical systems generate the sources which, under linear mixing, give rise to the observations. For any scalar source s^i and another source s^j and all times t, we seek a model of independent dynamics $p(s^i_{1:T}, s^j_{1:T}) = p(s^i_{1:T})p(s^j_{1:T})$. Furthermore, the aim is to find a matrix W that relates the sources to observations v_t through linear mixing $v_t = Ws_t$, where[1] $v_t = vert(v_t^1, \ldots, v_t^V)$, $s_t = vert(s_t^1, \ldots, s_t^C)$. This is a form of Independent Components Analysis (ICA) [2] although differs from the more standard assumption of independence at each time step, that is $p(s^i_{1:T}, s^j_{1:T}) = \prod_{t=1}^{T} p(s_t^i)p(s_t^j)$. Whilst there are many methods to deal with such temporal dependence (see [2]), in biosignal analysis it is important to have a method which encodes strong constraints such as desired frequencies of the sources. A closely related technique to ours is Nonlinear Dynamical Factor Analysis (NDFA) [3, 4]. Whilst being an attractive and powerful method, standard NDFA places no constraint that the observations are formed from mixing independent *scalar* dynamic sources, which makes interpretation of the resulting factors difficult. Furthermore, NDFA does not directly constrain the factors to contain particular frequencies so that in [4], in order to extract rhythmic activity, bias is incorporated by initializing the model with band-filtered principal components of the data. In addition, NDFA uses nonlinear state dynamics and mixing, which hampers inference and makes the incorporation of known constraints more complex. We therefore consider a simpler linear model which is nevertheless powerful, yet remains relatively interpretable and tractable. An important issue in decomposing signals into sources is the number of appropriate sources and also their complexity. To address this we use a Bayesian analysis of the Linear Gaussian State Space Model (LGSSM), as in [5], but constrained in order that independent dynamical processes can be identified and furthermore that scalar sources can be extracted from the signal.

2 Linear Gaussian state space models

In LGSSMs [6, 7, 8] the visible observation $v_t \in \mathcal{R}^V$ is linearly related to the hidden state vector $h_t \in \mathcal{R}^H$ by

$$v_t = Bh_t + \eta_t^v, \;\; \eta_t^v \sim \mathcal{N}(0, \Sigma_V) \; ,$$

where $\mathcal{N}(0, \Sigma_V)$ denotes a Gaussian distribution with zero mean and covariance Σ_V. The transition dynamics is also linear,

$$h_t = Ah_{t-1} + \eta_t^h, \; \eta_t^h \sim \mathcal{N}(0, \Sigma_H) \; .$$

Probabilistically, we express this as

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

with $p(v_t|h_t) = \mathcal{N}(Bh_t, \Sigma_V)$, $p(h_t|h_{t-1}) = \mathcal{N}(Ah_{t-1}, \Sigma_H)$ and $p(h_1) = \mathcal{N}(\mu, \Sigma)$. To make independent dynamical subsystems we use a block diagonal transition matrix $A = diag\left(A^1, \ldots, A^C\right)$, and state noise $\Sigma_H = diag\left(\Sigma_H^1, \ldots, \Sigma_H^C\right)$, where each block has dimension H_c. A one dimensional source s_t^c for each independent dynamical system is formed from $s_t^c = \mathbf{1}_c^{\mathsf{T}} h_t^c$, where $\mathbf{1}_c$ is a $H_c \times 1$ unit vector[2]. We can represent this as $s_t = Ph_t$, where $P = diag(\mathbf{1}_1^{\mathsf{T}}, \ldots, \mathbf{1}_C^{\mathsf{T}})$, $h_t = vert(h_t^1, \ldots, h_t^C)$ and h_t^c is a

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

[2] A more general approach would be to project to more than one dimension. This may then be interpreted as using a higher dimensional linear state space approximation of [3] which seeks independent subspace dynamics.

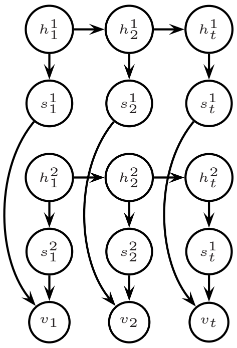


Figure 1: The variable h_t^c represents the vector dynamics of component c, which are projected by summation to form the dynamics of the scalar s_t^c (higher dimensional projections are an obvious extension). These sources are linearly mixed to form the visible observation vector v_t.

$H_c \times 1$ dimensional vector representing the state of dynamical system c. The graphical structure of this model is presented in Fig 1. As in standard ICA, we assume linear mixing of the sources to make the observation $v_t = W s_t + \eta_t^v$. Since the sources are formed from projection, the resulting emission matrix is constrained to be of the form[3]

$$B = WP\,,$$

where W is the $V \times C$ mixing matrix and P is a $C \times H$ projection, $H = \sum_c H_c$. Such a constrained form for B is needed to provide interpretable scalar sources[4].

Bayesian Linear Dynamical Systems

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

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

Here we take the ML-II ('evidence') framework, which involves maximising $p(v_{1:T}|\alpha,\beta)$ with respect to the hyperparameters α,β [3, 5, 9]. 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 each dynamical system to be close to a desired transition $\hat{A}$ (possibly zero) by using

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

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

[3] Unlike [5, 3] we cannot then assume $\Sigma_H \equiv I$ by parameter rescaling.

[4] Other projections would be equally valid.

Variational Bayes

We would like to optimize equation (1) with respect to α and β, but this is difficult due to the intractability of the integrals. Instead we consider the bound [3, 5, 9]

$$\log p(v_{1:T}|\alpha,\beta) \geq H_q(A,W,h_{1:T}) + \langle \log p(v_{1:T}, h_{1:T} A, W)\rangle_{q(A,W,h_{1:T})} \ , \tag{2}$$

where we dropped the explicit dependence on the hyperparameters on the right hand side[5]. 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 simplifying choices of the variational distribution q, we hope to achieve a tractable lower bound on the likelihood, which we may then optimize with respect to q, α, β. The key approximation in Variational Bayes is $q(A,W|h_{1:T}) \equiv q(A,W)$. This assumption allows other simplifications to follow, without further loss of generality. Since A and W separate in equation (2), optimally $q(A,W) = q(A)q(W)$ and hence

$$\log p(v_{1:T}|\theta) \geq -D(q(A),p(A)) - D(q(W),p(W)) + H_q(h_{1:T}) + \langle \log p(v_{1:T}, h_{1:T}|A,W)\rangle_{q(h_{1:T})q(A)q(W)} \ , \tag{3}$$

where $D(q(x),p(x))$ is the KL divergence $\langle \log q(x)/p(x)\rangle_{q(x)}$.

Determining $q(W)$

By examining equation (3), the contribution of $q(W)$ can be interpreted as the KL divergence between $q(W)$ and a Gaussian distribution in W (since $\log p(W|\beta) = -\frac{1}{2}\sum_{i,j}\beta_j W_{ij}^2 + const$). Hence, optimally, $q(W)$ is a Gaussian, for which we simply need to find the mean and covariance. For a quadratic form $x^T M x - 2x^{\mathsf{T}} m$ the covariance is M^{-1} and the mean is $M^{-1}m$. Hence the covariance $[\Sigma_W]_{ij,kl} \equiv \langle (W_{ij} - \langle W_{ij}\rangle)(W_{kl} - \langle W_{kl}\rangle)\rangle$ (averages wrt $q(W)$) is given by the inverse of the quadratic contribution

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

where $\tilde{h}_t = P h_t$. The mean is given by

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

where $q(h_{1:T})$ needed in the above is determined below.

Determining $q(A)$

Since the dynamics are independent, optimally we have a factorised distribution $q(A) = \prod_c q(A^c)$, where $q(A^c)$ is Gaussian with covariance $[\Sigma_{A^c}]_{ij,kl} \equiv \left\langle \left(A_{ij}^c - \langle A_{ij}^c\rangle\right)\left(A_{kl}^c - \langle A_{kl}^c\rangle\right)\right\rangle$ (averages wrt $q(A^c)$) given by the inverse of the quadratic contribution. Momentarily dropping the dependence on the source c, the covariance for each source is

$$\left[\Sigma_A^{-1}\right]_{ij,kl} = \left[\Sigma_H^{-1}\right]_{i,k} \sum_{t=2}^{T} \left\langle h_{t-1}^j h_t^l \right\rangle + \alpha\delta_{i,k}\delta_{j,l} \ ,$$

and the mean is

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

where in the above all parameters and the variable h should be interpreted as pertaining to dynamic source c only (e.g. $h_{t-1}^{(c),j}$) and the averages are with respect to $q(h_{t-1}^c, h_t^c)$.

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

Determining $q(h_{1:T})$

Optimally $q(h_{1:T})$ is Gaussian since equation (3) is quadratic in $h_{1:T}$, being namely

$$-\frac{1}{2}\sum_{t=1}^{T}\left\langle (v_t - WPh_t)^{\mathsf{T}}\Sigma_V^{-1}(v_t - WPh_t)\right\rangle_{q(W)}$$
$$-\sum_{t=2}^{T}\frac{1}{2}\left\langle (h_t - Ah_{t-1})^{\mathsf{T}}\,\Sigma_H^{-1}\,(h_t - Ah_{t-1})\right\rangle_{q(A)} - \frac{1}{2}\,(h_1-\mu)^{\mathsf{T}}\,\Sigma^{-1}\,(h_1-\mu)\ .$$

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. This means that $q(h_{1:T})$ may be factored into $\prod_t q(h_t|h_{t-1})$. There are several mathematically equivalent ways to find the numerical form of the factorisation. In [5] novel smoother recursions are constructed. Here we take a simpler approach, motivated by the idea that when the covariances of $q(A)$ and $q(W)$ are zero the chain is exactly of the form of a standard Kalman Filter (KF), with emission matrix $B = WP$ and transition A. Our aim is to find an equivalent form for which the standard numerically stable Rauch-Tung-Striebel smoother recursions may be applied [6]. In order to do that, we rewrite

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

where $\langle B\rangle \equiv \langle W\rangle\, P$ and

$$[S_W]_{j,l} = \sum_{i,k=1}^{V}[\Sigma_W]_{ij,kl}\left[\Sigma_V^{-1}\right]_{i,k}, \qquad j,l \in 1,\ldots,H.$$

Similarly

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

where

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

To represent the above as a KF, with dynamics $\tilde{A}$, emission $\tilde{B}$ and observation $\tilde{v}$, we augment v_t and B as

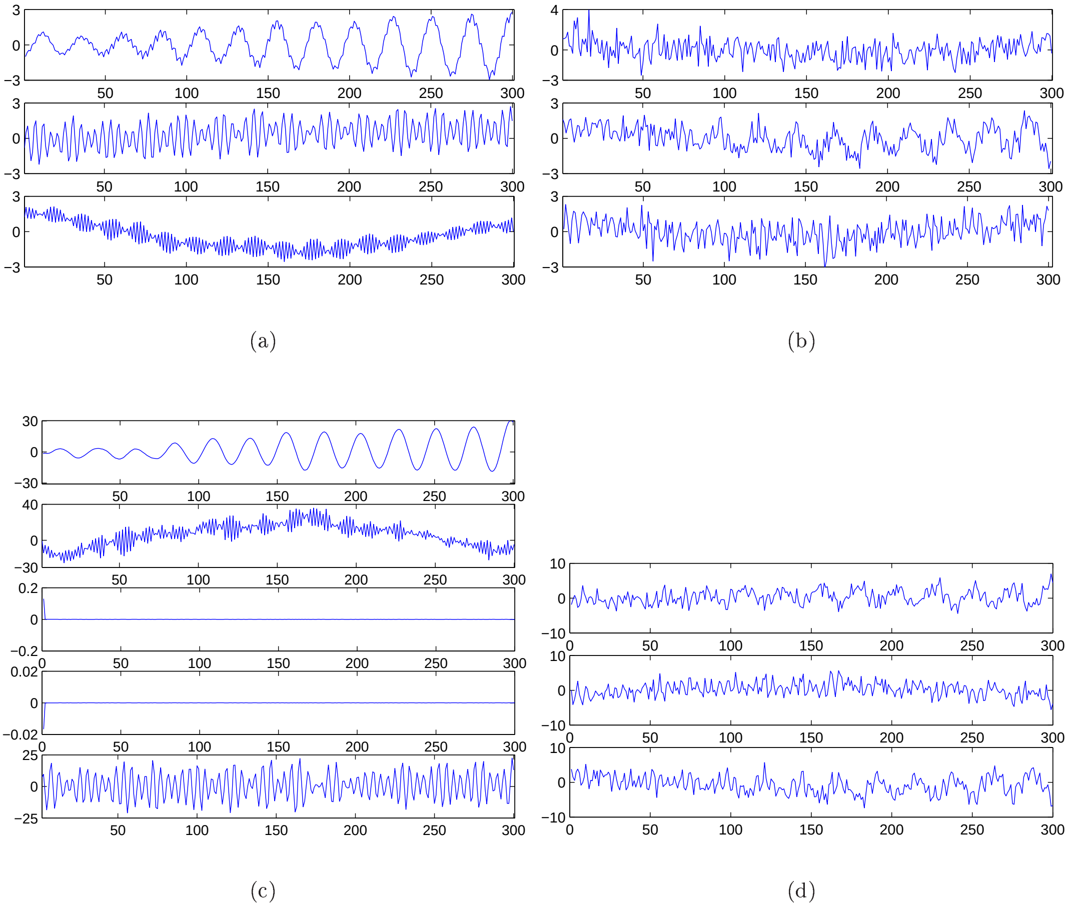
$$\tilde{v}_t = vert(v_t, \mathbf{0}, \mathbf{0}), \qquad \tilde{B} = vert(\langle B\rangle\, , U_A, U_W P),$$

where $\mathbf{0}$ is a $H \times 1$ zero vector and 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 KF is then completed by specifying $\tilde{\Sigma} \equiv \Sigma$, $\tilde{\mu} \equiv \mu$, $\tilde{A} \equiv \langle A\rangle$. Strictly speaking we need to make a slight adjustment and use 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 $\mathbf{0}$.

Finding the optimal Parameters

Differentiating equation (3) with respect to β_j and α_c we find that, optimally:

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



(a)

(b)

(c)

(d)

Figure 2: (a) Original (correlated) sources s. (b) Observations resulting from mixing the original sources, $v = Ws + \eta^v$, $\eta^v \sim \mathcal{N}(0, I)$. (c) Recovered sources using our method. (d) Independent sources found using FastICA.

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

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

The prior mean μ and covariance Σ are set to those of the distribution $q(h_1)$. Learning then proceeds by iterating the parameter update step followed by updating $q(h_{1:T}), q(A), q(W)$. The initial parameters are set randomly.

2.1 Demonstration

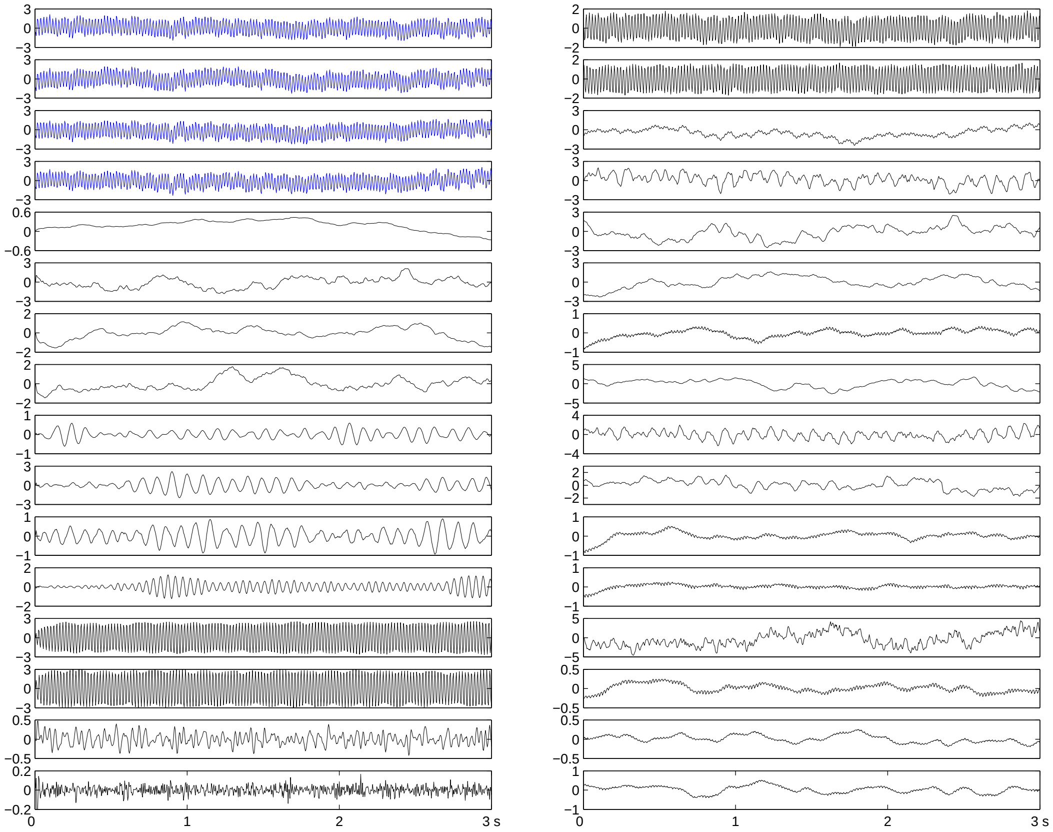
In a proof of concept experiment, we used a LGSSM to generate 3 sources with random 5×5 transition matrices A^c, $h_1 \sim N(0, I)$ and $\Sigma_H = I$. The sources were mixed into three observations $v_t = Ws_t + \eta_t^v$, for W chosen with elements from a zero mean unit variance Gaussian distribution, and $\Sigma_V = I$. We then trained a different LGSSM with 5 sources and 7 dimensions for each dynamic component c. To bias the model to find the simplest sources, we used $\hat{A}^c \equiv \mathbf{0}$ for all sources. In Fig 2a and Fig 2b we see the original sources and the noisy observations respectively. The observation noise is so high that a good estimation of the sources is possible only by taking the dynamics into account. In Fig 2c we see the estimated sources from our method after 400 iterations. Two of the 5 sources have been removed, the remaining three are a reasonable estimation of the original sources. The FastICA [2] result is given in Fig 2d. In fairness, FastICA cannot deal with noise and also seeks independent components, whereas in this example the sources are slightly correlated. Nevertheless, this example demonstrates that, whilst a standard method such as FastICA indeed produces independent components, this may not be a satisfactory result, since there is no search for simplicity of the underlying dynamical system, nor indeed may independence at each time point be a desirable criterion.

2.2 Application to EEG analysis

In Fig 3a (blue), we show three seconds of EEG data recorded from 4 channels (located in the right hemisphere) while a subject is performing imagined movement of his right hand. As is typical in EEG, each channel shows low frequency drift terms, together with the presence of 50 Hz mains contamination, which masks the information related to the mental task, mainly centered at 10 and 20 Hz. 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. However, even with pre-filtering, the number of components is usually restricted in ICA to be equal to the number of channels. In EEG this is potentially too restrictive since there may be many independent oscillators of interest underlying the observations and we would like some way to automatically determine the effective number of such oscillators. We used our method with 16 sources and, 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. The frequencies for the 16 sources were [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] Hz respectively. After training, the Bayesian approach removed 4 unnecessary sources from the mixing matrix W, that is one [10,11] Hz and three [20,21] Hz sources. The temporal evolution of the 12 retained sources is presented in Fig 3a (black). We can see that effectively the first 4 sources contain dominant low frequency drift, the following 3 contain [10,11] Hz, while the 8th 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 last two sources have changed in order to model other frequencies present in the signals. In order to asses the advantage of using prior frequencies for extracting task-related information and the potential limitations of using a linear model, we have compared our method with NDFA [3]. We extracted 16 factors using a NDFA model in which both MLPs had one hidden layer of 30 neurons. The other parameters were set to the default values. In Fig 3b 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 our method, there are 2 main 50 Hz sources (first two factors), even if 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. From this example we can conclude that, while the two methods give similar results, the prior specification of independent dynamical processes at particular frequencies has helped our model to better isolate the activity of interest into a smaller number of sources, and, among these sources, to separate the contribution of oscillators at different frequencies, that is 10 Hz and 20 Hz oscillators.

3 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. This procedure is closely related to others previously proposed in the literature, but has the particular property that the sources are themselves projections from higher dimensional independent linear dynamical



(a)

(b)

Figure 3: (a) The top four (blue) signals are the original unfiltered EEG channel data. The remaining 12 subfigures are the sources s estimated by our method. (b) The 16 factors estimated by NDFA after convergence (800 iterations).

systems. A particular advantage of the linear dynamics approach is the tractability of inference, and we demonstrated how this can be achieved reliably by conversion to a standard Kalman Smoother form. 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. The method is able then to automatically extract signals, for example, biased towards particular frequencies. One disadvantage of the current model is that some signals or artifacts in for example EEG may be so complex that they are difficult to model with a stationary state space model. One possibility would be to include additional components which are assumed temporally independent (as in the standard ICA), along the lines of those used in [10]. We think that the development of this and related methods may provide useful tools particularly in multichannel signal analysis.

References

[1] A. Delorme and S. Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, (134):9–21, 2004.

[2] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley, 2001.

[3] Harri Valpola and Juha Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[4] 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, ICA 2001*, pages 457–462, 2001.

[5] M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* Phd thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[6] Y. Bar-Shalom and Xiao-Rong Li. *Estimation and Tracking : Principles, Techniques and Software.* Artech House, Norwood, MA, 1998.

[7] M. West and J. Harrison. *Bayesian forecasting and dynamic models.* Springer, 1999.

[8] Rasmus Kongsgaard Olsson and Lars Kai Hansen. A harmonic excitation state-space approach to blind separation of speech. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *NIPS*, pages 993–1000. MIT Press, Cambridge, MA, 2005.

[9] D.J.C. MacKay. Ensemble learning and evidence maximisation. Unpublished manuscipt : www.variational-bayes.org, 1995.

[10] S. Chiappa and D. Barber. EEG Classification using Generative Independent Component Analysis. *Neurocomputing*, 2006. To appear.