Bayesian Nonstationary Autoregressive Models for Biomedical Signal Analysis

Michael J. Cassidy* and William D. Penny

***Abstract*—We describe a variational Bayesian algorithm for the estimation of a multivariate autoregressive model with time-varying coefficients that adapt according to a linear dynamical system. The algorithm allows for time and frequency domain characterization of nonstationary multivariate signals and is especially suited to the analysis of event-related data. Results are presented on synthetic data and real electroencephalogram data recorded in event-related desynchronization and photic synchronization scenarios.**

***Index Terms*—Autoregressive modeling, Bayesian, Kalman smoother, variational Bayes.**

I. Introduction

We present an algorithm for modeling nonstationary multivariate time series and apply it to the analysis of biomedical signals. The model consists of a multivariate autoregressive (MAR) process with time-varying coefficients that adapt according to a linear dynamical system (LDS). While such a model is not new, the contribution of this paper is to present a fully Bayesian implementation which allows us to retain the full generality of the approach while deriving a practical algorithm.

The implementation has been made possible by the adoption of the "variational Bayesian (VB)" framework [1]. This paper represents a further step in the development of Bayesian signal processing algorithms where we have, so far, applied VB to stationary MAR models [2] and univariate Gaussian [3] and Non-Gaussian [4] AR models.

In Section II, we describe our basic time series model. Other authors have also developed algorithms for VB learning applied to linear dynamical systems, so we contrast our model with the one derived in [15], where a general class of conjugate-exponential models were considered with the linear state-space model emerging as a special case. One purpose of this paper is to try and bridge the gap between the specialized signal processing community and the experimental scientists who could find many useful applications of these new theoretical ideas. The nonstationary MAR model is of particular interest in this respect due to the strong physical interpretation of its model parameters.

Manuscript received April 3, 2001; revised May 28, 2002. The work of M. Cassidy was supported in part through an MRC nonclinical research training fellowship and in part by SmithKline Beecham. The work of W. Penny was supported by the Wellcome Trust. *Asterisk indicates corresponding author.*

*M. J. Cassidy is with the Sobell Department of Neurophysiology, Institute of Neurology, Queen Square, University College London, London, U.K. (e-mail: m.cassidy@ion.ucl.ac.uk).

W. D. Penny is with the Wellcome Department of Imaging Neuroscience, Institute of Neurology, Queen Square, University College London, London, U.K.

Publisher Item Identifier 10.1109/TBME.2002.803511.

In Section III, we describe the key principles behind the VB approach. In Section IV, we show how to apply VB to nonstationary MAR models and derive a set of update rules which are a generalization of Shumway and Stoffer's [5] expectation-maximization (EM) algorithm for time-series analysis. In Section V, we apply our algorithm to some synthetic data and also to electroencephalogram (EEG) data obtained from two different experimental scenarios. The first is a photic synchronization experiment, in which the EEG activity can be forced into synchrony with a strobe light flashing at frequencies close to the resting alpha state. The second example presents results from an event-related desynchronization (ERD) experiment. ERD describes the phenomenon where just before a voluntary movement takes place, the spectral power drops in the EEG recorded over the motor cortex in the alpha frequency band. These two scenarios provide data with well documented physiological changes that are suitable for study with our algorithm.

II. The Nonstationary MAR Model

In what follows, the notation $\mathcal{N}$, $\mathcal{G}$ and $\mathcal{W}$ refer to the Gaussian, Gamma, and Wishart probability densities which, for convenience, are defined in Appendix.

Our model is a linear dynamical system for T d-variate observations and a latent-space of dimension k. The state-space equations are

$$\begin{aligned} x_{t+1} &= Ax_t + w_t \\ y_t &= C_t x_t + v_t \end{aligned} \tag{1}$$

where $p(w) = \mathcal{N}(0, Q^{-1})$, $p(v) = \mathcal{N}(0, R^{-1})$, and Q and R are precision (inverse covariance) matrices. The variables x_t are state variables and y_t the observations. They are vectors of dimension $k \times 1$ and $d \times 1$, respectively.

A common way of writing the MAR(p) process expresses the term $C_t x_t$ as $\sum_{n=1}^{p} A_n y_{t-n}$, where A_n are the MAR coefficients at lag n (not to be confused with the state transition matrix A) and p is the "order" of the model. If one concatenates the coefficient matrices A_n in a row to form a $d \times (p \times d)$ matrix $\mathcal{A}$, one can then stack the columns of $\mathcal{A}^T$ to form the $k \times 1$ vector $x_t = \mathrm{vec}(\mathcal{A}^T)$, where vec denotes the stacking operation and $k = pd^2$. C_t is the appropriate lag matrix at time t, which embodies past values of the observations $y_{t-1}, \ldots, y_{t-p}$. We have

$$p(y_t|x_t) = \mathcal{N}(C_t x_t, R^{-1}) \tag{2}$$

$$p(x_t|x_{t-1}) = \mathcal{N}(Ax_{t-1}, Q^{-1}) \tag{3}$$

$$p(x_1) = \mathcal{N}(\mu_1, \Sigma_1) \tag{4}$$

which define the observation model, state transition model and initial state distribution. $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate Gaussian with mean μ and covariance Σ.

If we knew A, Q, and R then the hidden state variables (MAR coefficients) could be inferred from the preceeding time series using a Kalman filter [6]. The benefits of Kalman filters over least mean squares (LMS) and recursive least squares (RLS) algorithms are well known and are demonstrated, for example, in [7].

Now, in this paper we concern ourselves, not with real-time processing of signals, but with the retrospective analysis of data sets that have already been collected. For this reason, we can use both preceeding and succeeding time series samples for state estimation. This leads to the Kalman smoother [8].

In practice, however, we do not know A, Q, and R and so, in the engineering literature, these matrices are either set to arbitrary values or various *ad-hoc*/suboptimal procedures are used for their estimation. In the statistics literature, however, it has long been known that these matrices can be learned using an EM algorithm [5]. In the E-step, the hidden variables are estimated using a Kalman smoother and, in the M-step, the A, Q, and R matrices are updated. The E and M steps are iterated so as to maximize the likelihood of the data.

A problem with the EM approach, however, is that like all maximum-likelihood (ML) methods it is sensitive to overfitting. For nonstationary models of the type, we are considering the situation is particularly acute as we have many model parameters (MAR matrices at *each* time point, state transition and noise matrices etc.) and may often have few data points (i.e., short-time series). This has prevented a wider application of the EM algorithm in this context. In this paper, however, we show how overfitting can be prevented using a Bayesian approach; priors over parameters act as regularizers and so prevent overfitting (see e.g., [9] for a treatment of Bayesian methods in signal processing).

In Bayesian learning, one is generally interested in calculating the evidence of various models. To calculate the evidence one just multiplies the model likelihood by the priors, and then integrates over the parameters. In some cases, this integral can be solved analytically, but when the evidence integral is analytically intractable one has to resort to approximation techniques. The VB approach is one such technique and as was mentioned earlier, has been applied to the linear Gaussian state-space model in [15]. The model presented here differs from the model presented in [15] in a number of respects, which we now discuss.

The first point to note is that we have a constrained C_t matrix because the elements of C_t are determined by the previous values of the time series in the manner described above. In [15], this matrix is constant and unconstrained (i.e., $C_t = C$). The MAR model, therefore, shows a dependency between y_t and the previous value y_{t-1} which is not present in the standard state-space model. In this paper, we make the assumption that these variables are actually independent, an assumption that has been made in previous ML nonstationary MAR algorithms [7], [19]. During the review process, it has been brought to our attention that the estimation equations of the Kalman filter can also be designed for state-space models where the matrix C_t is a function of the previous observations [17]. In this paper, however, we retain the independence assumption and leave this possible extension to future work.

The fact that C is constant in [15] allows the authors to set the state noise covariance to the identity because arbitrary rescaling of this noise can be achieved through changes to A, C, and R. In our model, we are not free to absorb rescaling in C_t, so we must infer Q. In addition, the authors of [15] constrain R to be diagonal whereas we keep the observation noise covariance matrix unconstrained. These differences prevent an easy translation between the results of [15] and those presented here. Nevertheless, the reader is advised to refer to [15] and the companion paper [16] for more details on VB learning in linear Gaussian state-space models.

A. Priors

In this section, we describe the priors used in our model. Firstly, we place a Gaussian prior on the state transition matrix

$$p(\mathrm{vec}(A)|\alpha) = \mathcal{N}(\mathrm{vec}(I[k]), \alpha^{-1} I[k^2]). \tag{5}$$

Here, I represents the identity matrix and the Gaussian density $\mathcal{N}(\mu, \Sigma)$ is defined in Appendix (45). Therefore, in the absence of evidence to the contrary, the state transition is inferred to be equal to the identity matrix implying that there is no deterministic evolution of the state. The quantity α is the prior precision of the state-transition elements. We then place a Gamma (see (47)) (hyper) prior on the precision

$$p(\alpha) = \mathcal{G}(b_p, c_p). \tag{6}$$

We make this "uninformative" (i.e., broad) by using the settings

$$\begin{aligned} b_p &= 1000 \\ c_p &= 0.001. \end{aligned} \tag{7}$$

For the state noise and observation noise precisions, one can use

$$\begin{aligned} p(Q) &= \mathcal{W}_k(q_p, D_p) \\ p(R) &= \mathcal{W}_d(r_p, B_p) \end{aligned} \tag{8}$$

where $\mathcal{W}_N(a, B)$ is a Wishart density (see (49)). In this paper, however, we set $q_p = r_p = 0$ and $D_p = B_p = 0$ giving the improper priors

$$\begin{aligned} p(Q) &\propto |Q|^{-(k+1)/2} \\ p(R) &\propto |R|^{-(d+1)/2}. \end{aligned} \tag{9}$$

The overall joint density is

$$\begin{aligned} p(S, D, \theta) = &\prod_{t=1}^{T} p(y_t|x_t, R) \\ &\times \prod_{t=2}^{T} p(x_t|x_{t-1}) p(x_1) p(A|\alpha) p(\alpha) p(Q) p(R), \end{aligned} \tag{10}$$

where $D = \{y_t\}$ denotes the data, $S = \{x_t\}$ denotes the hidden variables and θ denotes the model parameters A, Q, R, and α. The overall probabilistic model is shown in Fig. 1, which indicates dependencies between the different variables.

III. Variational Bayes

The central quantity of interest in Bayesian learning is the posterior distribution $p(\theta, S|D)$. This implies estimation both

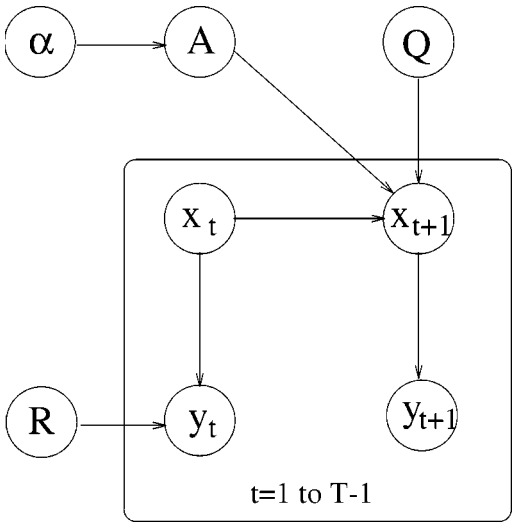

Fig. 1. Nonstationary autoregressive model. The autoregressive coefficients x_t are modeled as a linear dynamical system. The precision of the observations is given by R and the coefficients evolve deterministically via the transform A and stochastically via additive noise with precision Q. Each element of A has prior precision α.

of the parameters θ, the hidden variables S and the uncertainties associated with their estimation. In our nonstationary MAR model, the MAR coefficients x_t take the place of the hidden variables and the other variables (state noise and observation noise precision matrices etc.) are considered as parameters. The model parameters θ and hidden variables S can all be learned within the VB framework, a full tutorial on which is given in [1]. In what follows, we describe the key features.

Given a probabilistic model of the data, the log of the "evidence" or "marginal likelihood" can be written as

$$\begin{aligned}\log p(D) &= \int q(\theta,S|D)\log p(D)d\theta dS \\ &= \int q(\theta,S|D)\log\frac{p(D,\theta,S)}{p(\theta,S|D)}d\theta dS \\ &= \int q(\theta,S|D)\log\left[\frac{q(\theta,S|D)p(D,\theta,S)}{p(\theta,S|D)q(\theta,S|D)}\right]d\theta dS \\ &= F+KL. \end{aligned} \tag{11}$$

Here, $q(\theta,S|D)$ is to be considered, for the moment, as an arbitrary density. We have

$$F=\int q(\theta,S|D)\log\frac{p(D,\theta,S)}{q(\theta,S|D)}d\theta dS, \tag{12}$$

which is known (to physicists) as the negative variational free energy and

$$KL=\int q(\theta,S|D)\log\frac{q(\theta,S|D)}{p(\theta,S|D)}d\theta dS \tag{13}$$

is the KL-divergence [10] between the density $q(\theta,S|D)$ and the true posterior $p(\theta,S|D)$.

Equation (11) is the fundamental equation of the VB-framework. Importantly, because the KL-divergence is always positive [10], F provides a lower bound on the model evidence. Moreover, because the KL-divergence is zero when the two densities are the same, F will become equal to the model evidence when $q(\theta,S|D)$ is equal to the true posterior. For this reason, $q(\theta,S|D)$ can be viewed as an *approximate posterior*.

The aim of VB-learning is to maximize F and so make the approximate posterior as close as possible to the true posterior. To obtain a practical learninig algorithm we must also ensure that the integrals in F are tractable. One generic procedure for attaining this goal is to assume that the approximating density factorizes over groups of parameters (in physics this is known as the mean-field approximation). Thus, following [11], we consider:

$$q(\theta,S|D)=q(S|D)\prod_i q(\theta_i|D) \tag{14}$$

where θ_i is the ith group of parameters. The distributions which maximize F can then, via the calculus of variations, be shown to be

$$q(\theta_i|D)=\frac{\exp[I(\theta_i)]}{\int\exp[I(\theta_i)]d\theta_i} \tag{15}$$

where

$$I(\theta_i)=\int q(\theta^{\setminus i}|D)q(S|D)\log p(D,S|\theta)d\theta^{\setminus i}dS \tag{16}$$

and $\theta^{\setminus i}$ denotes all parameters *not* in the ith group. A similar expression exists for $q(S|D)$. Note that, importantly, this means we are able to determine the optimal analytic *form* of the component posteriors. This is to be contrasted with, for example, Laplace approximations where we have to arbitrarily fix the form of the component posteriors to be Gaussian [9].

The above principles lead to a set of coupled update rules for the *parameters* of the component posteriors, iterated application of which leads to the desired maximization. Further, by computing F for models of different order, we can perform model order selection (see e.g., [3]), although this is beyond the scope of the present paper. The free energy expression for our model is derived in Appendix. Updates for the parameters of the hidden variable posterior are analogous to the E-step in EM learning and the other updates are analogous to the M-Step.

IV. Update Rules

By plugging in the likelihood and priors for our nonstationary AR model (from Section II) into (15), the optimal components of the approximate posterior turn out to be

$$\begin{aligned} q(\mathrm{vec}(A)) &= \mathcal{N}(\mathrm{vec}(\bar{A}),A_c) \\ q(\alpha) &= \mathcal{G}(b,c) \\ q(Q) &= \mathcal{W}_k(q,D) \\ q(R) &= \mathcal{W}_d(r,B) \\ q(x_t) &= \mathcal{N}(\mu_t,\Sigma_t). \end{aligned} \tag{17}$$

Note that, for each component, the form of the approximate posterior is the same as the prior. In fact, this is no accident, as we chose the priors so as to achieve this (for a discussion of such "conjugate" priors, see [12]). In what follows, we show how the parameters of these distributions are updated. We also show that if we remove the priors we recover Shumway and Stoffer's MLEM algorithm [5].

A. E-Step

In this step, we update our distribution over hidden variables x_1 and x_t using a modified smoothing algorithm. As

emphasized in [15], the only difference between VB-Kalman smoothing and the standard Kalman smoothing algorithm is the use of expectations instead of point estimates (e.g., $\bar{A}$ instead of A). A lengthy derivation of the variational Kalman smoother relations is presented in [16], so are not re-presented here. Here we only note that the modified Kalman smoothing algorithm requires computation of expectation terms of the form $E\left[AXA^T\right]$, where X is some matrix. Although these expectations are relatively straightforward to evaluate, they are computationally intensive, and especially so for models with high MAR model order. We, therefore, use the approximation

$$E\left[AXA^T\right] \approx \bar{A}X\bar{A}^T \tag{18}$$

which gives qualitatively similar results and is much quicker to compute. As a result our E step is identical in form to the traditional Kalman smoother. For convenience, these recursions are given in Appendix. The computational complexity of the E-step is, therefore, comparable with that of the standard Kalman smoother.

B. M-Step

We now present the results of the VB M step of our algorithm. For clarity, we include a detailed derivation of the update for R. We only present final results for the other updates, as the procedure for derivation is identical in each case.

1) Update for R: Equations (15) and (16) give the general procedure for updating parameters. Applied to the observation noise precision matrix R, they give

$$q(R|D) \propto \exp[I(R)] \tag{19}$$

where

$$\begin{aligned} I[R] =& \sum_{t=1}^{T} \int q(x_t) \left[\log p(y_t|x_t, R) + \log p(R)\right] dx_t \\ =& \sum_{t=1}^{T} \int q(x_t) \\ & \times \left[-\frac{1}{2}(y_t - C_t x_t)^T R (y_t -_t x_t) \right. \\ & \left. + \frac{(T-d-1)}{2} \log |R| \right] \\ =& \frac{(T-d-1)}{2} \log |R| \\ & - \frac{1}{2} \mathrm{tr} \left\{ \left[\sum_{t=1}^{T} (y_t - C_t \mu_t)(y_t - C_t \mu_t)^T \right. \right. \\ & \left. \left. + C_t \Sigma_t C_t^T \right] R \right\}. \end{aligned} \tag{20}$$

The required density is, therefore, in the form of a Wishart $q(R|D) = \mathcal{W}_d(r, B)$, where

$$\begin{aligned} r =& r_p + T \\ B =& B_p + \sum_{t=1}^{T} (y_t - C_t \mu_t)(y_t - C_t \mu_t)^T + C_t \Sigma_t C_t^T \\ \bar{R} =& rB^{-1}. \end{aligned} \tag{21}$$

The second term in B is the average observed covariance that is not explained by $C_t x_t$. This is identical to the VB update for the noise precision in a stationary MAR model (see [2, Eq. (28)]).

2) Update for A: The update for the state-transformation matrix A, where $q(\mathrm{vec}(A)) = \mathcal{N}(\mathrm{vec}(\bar{A}), A_c)$ is given by

$$\begin{aligned} A_c^{-1} =& \left(\sum_{t=2}^{T} M_{t-1} \right) \otimes \bar{Q} + \bar{\alpha} I[k^2] \\ \mathrm{vec}(\bar{A}) =& A_c \mathrm{vec} \left(\bar{Q} \sum_{t=2}^{T} M_{t,t-1} + \alpha I[k] \right) \end{aligned} \tag{22}$$

where M_t and $M_{t,t-1}$ are intermediate quantities obtained from the E-step and defined in (36). For $\bar{\alpha} = 0$, the Qs cancel, leaving

$$\bar{A}_{ML} = \left(\sum_{t=2}^{T} M_{t,t-1} \right) \left(\sum_{t=2}^{T} M_{t-1} \right)^{-1} \tag{23}$$

which is identical to the ML update (see [13, Eq. (18)]).

3) Update for Q: The update for the state precision matrix Q where $q(Q) = \mathcal{W}_k(q, D)$ is given by

$$\begin{aligned} q =& q_p + T - 1 \\ D =& D_p + \sum_{t=2}^{T} \left[M_t - \bar{A} M_{t-1,t} - M_{t,t-1} \bar{A}^T \right. \\ & \left. + \bar{A} M_{t-1} \bar{A}^T + f(M_{t-1}) \right] \\ \bar{Q} =& qD^{-1}. \end{aligned} \tag{24}$$

The quantity $f(M_{t-1})$ arises because the functional to be optimized (as a functional of Q) in this M step contains the term $\mathrm{tr}\left((M_{t-1} \otimes Q) A_c\right)$, which comes from integrating over A. One wants to be able to factorise out Q within the trace, which can be done by decomposing A_c as a sum of tensor products. If one writes

$$A_c = \sum_i s_i u_i u_i^T, \tag{25}$$

where s_i and u_i are eigenvalues and eigenvectors of A_c, one has

$$\begin{aligned} u_i =& \mathrm{vec}(U_i) \\ =& \mathrm{vec} \left(\sum_j c_{ij} \xi_j^i \otimes \eta_j^{iT} \right) \\ =& \sum_j c_{ij} \xi_j^i \otimes \eta_j^i, \end{aligned} \tag{26}$$

where U_i are $(k \times k)$ matrices formed from u_i and ξ, η are the vectors obtained by a singular value decomposition of U. By substituting for A_c in the trace term, one can show that

$$f(M_{t-1}) = \sum_{ijk} s_i c_{ij} c_{ik} \eta_j^{iT} M_{t-1} \eta_k^i \xi_k^i \xi_j^{iT}. \tag{27}$$

It is also possible to show that

$$\begin{aligned} \bar{A}_{ML} M_{t-1,t} = & \bar{A}_{ML} M_{t-1,t} + M_{t,t-1} \bar{A}_{ML}^T \\ & - \bar{A}_{ML} M_{t-1} \bar{A}_{ML}^T. \end{aligned} \tag{28}$$

We can remove the priors by setting $A_c = 0$ giving

$$\bar{Q}_{ML}^{-1} = \frac{1}{T-1} \sum_{t=2}^{T} \left(M_t - \bar{A}_{ML} M_{t-1,t} \right) \tag{29}$$

which is identical to the ML update (see [13, Eq. (20)]).

4) Update for α: The update for the precision of the state-transformation matrix α, where $q(\alpha) = \mathcal{G}(b, c)$, is given by

$$\begin{aligned} c &= c_p + \frac{k^2}{2} \\ \frac{1}{b} &= \frac{1}{b_p} + \frac{1}{2}\mathrm{vec}(\bar{A} - I[k])^T \mathrm{vec}(\bar{A} - I[k]) + \frac{\gamma}{2} \end{aligned} \quad (30)$$

where $\gamma = \mathrm{tr}(A_c)$. While we could derive VB updates for μ_1 and Σ_1, this would have little overall effect on the solution. These parameters are, therefore, currently fixed to their ML values.

C. Practical Issues

To initialize the algorithm we run the data set through one iteration of EM [5]. We then update the parameters of the approximate posterior (see beginning of Section IV), i.e., the parameters r and B using (21), $\bar{A}$ and A_c using (22), q and D using (24) and b and c using (30). We then update the estimates of the MAR coefficients using the modified Kalman smoothing algorithm described in Section IV-A. These pseudo E and M steps are then iterated. The progress of the learning algorithm is monitored by computing the negative free energy, F, (see the Appendix) and the algorithm is deemed to converge if F changes by less than 0.01%.

The computational complexity of the M-step is comparable with the ML algorithm except for the update for Q which requires singular value decomposition at each iteration. The impact on the algorithm's run-time is discussed in Section V.

While the nonstationary MAR model is defined in the time-domain, it is possible to convert to a frequency domain representation, as described in [18]. This allows for estimation, not only of the spectra for each channel but also of the cross-spectra. These cross-spectra can then be decomposed into phase and coherence components.

Finally, we note that the algorithm is well suited to dealing with multiple realizations of an observation sequence such as occur during event-related experimental paradigms. This is achieved by running the E-step once for each realization and averaging the estimates of x_t^t, Σ_t^t and other sufficient statistics as described in [13]. These multiple E-steps are interleaved with a single M-step. Thus, unlike conventional analysis of event-related data where averaging takes place in the observation space, in our model averaging takes place in the space of hidden variables. The results from applying this procedure to event-related data are given in Section V.

V. Results

A. Simulations

Our first results compare the EM and VB algorithms for spectral estimation of a univariate time-series. As this is a univariate time series the MAR model reverts to an AR model. This data set has previously been investigated using a Kalman filter approach [19] and consists of a single sinusoid with a frequency f that is itself subject to a sinsuoidally varying phase modulation

$$y_t = 5\sin(2\pi f(t + 0.05\sin(2\pi f_{\mathrm{mod}} t))) + e_t. \quad (31)$$

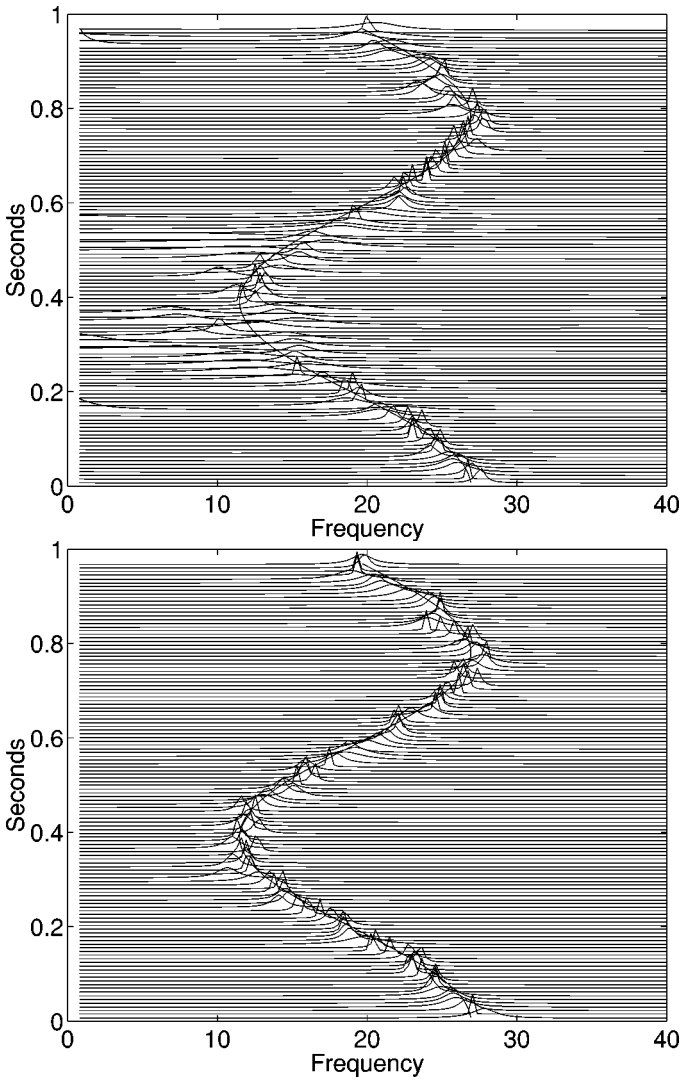

Fig. 2. A sinusoidally phase-modulated mode with low variance additive noise. The solid line shows the known instantaneous frequency underlying the data. Spectral estimation using a nonstationary AR(4) model inferred using EM (top) and VB (bottom). This shows the characteristic overfitting of the EM algorithm as opposed to the well-behaved VB procedure.

The instantaneous frequency underlying this signal (see e.g., [14, page 368]) is given by

$$f^i(t) = f + 0.1\pi f f_{\mathrm{mod}} \cos(2\pi f_{\mathrm{mod}} t). \quad (32)$$

We set $f = 19.2$ Hz and $f_{\mathrm{mod}} = 1.28$ Hz. We produced 1 second of data sampled at 128 Hz with the observation noise variance set to a small value ($\sigma_e^2 = 0.2$) and then applied the nonstationary MAR model with model order set to $p = 4$. The spectral estimates from EM and VB are shown in Fig. 2. Another data set was then generated, this time with a higher noise level ($\sigma_e^2 = 2$) and a higher order model was applied ($p = 8$). Fig. 3 shows the EM and VB spectra. For both noise levels, we see that EM is prone to overfitting especially when the model order is over-specified and the signal is noisy. In contrast, VB is robust to mis-specification of the model order and provides good spectral estimates even for noisy data. This is because the Bayesian priors act as regularizers. For both levels of noise, we see that the VB model tracks the true instantantaneous frequency underlying the data.

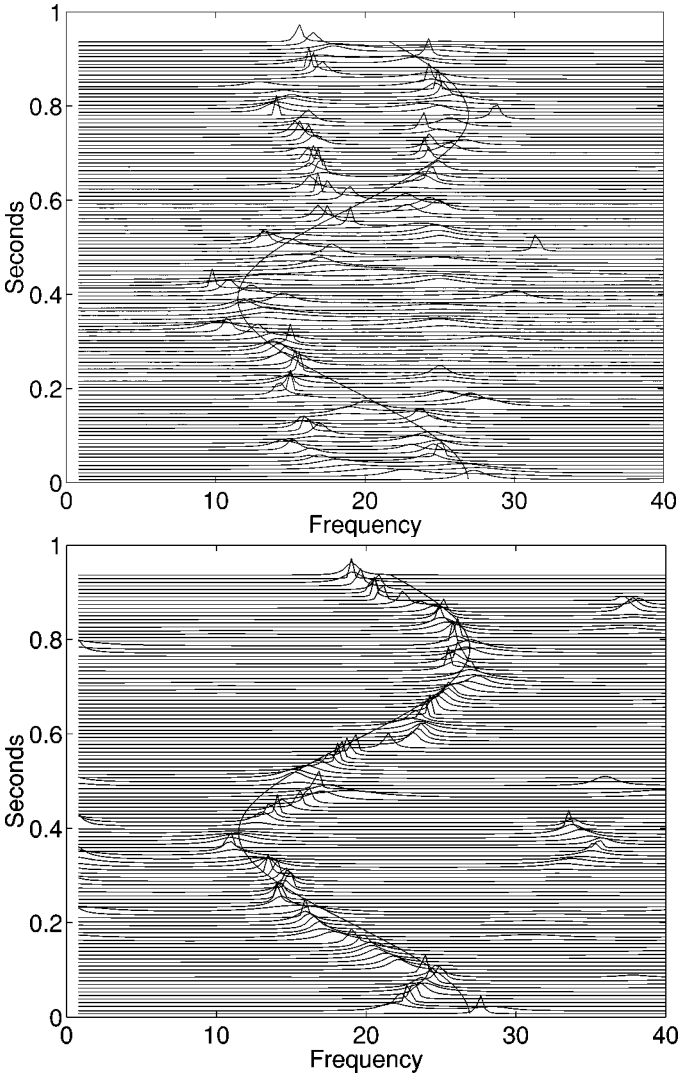

Fig. 3. A sinusoidally phase-modulated mode with high variance additive noise. The solid line shows the known instantaneous frequency underlying the data. Spectral estimation using a nonstationary AR(8) model inferred using EM (top) and VB (bottom). This shows the characteristic overfitting of the EM algorithm as opposed to the well-behaved VB procedure.

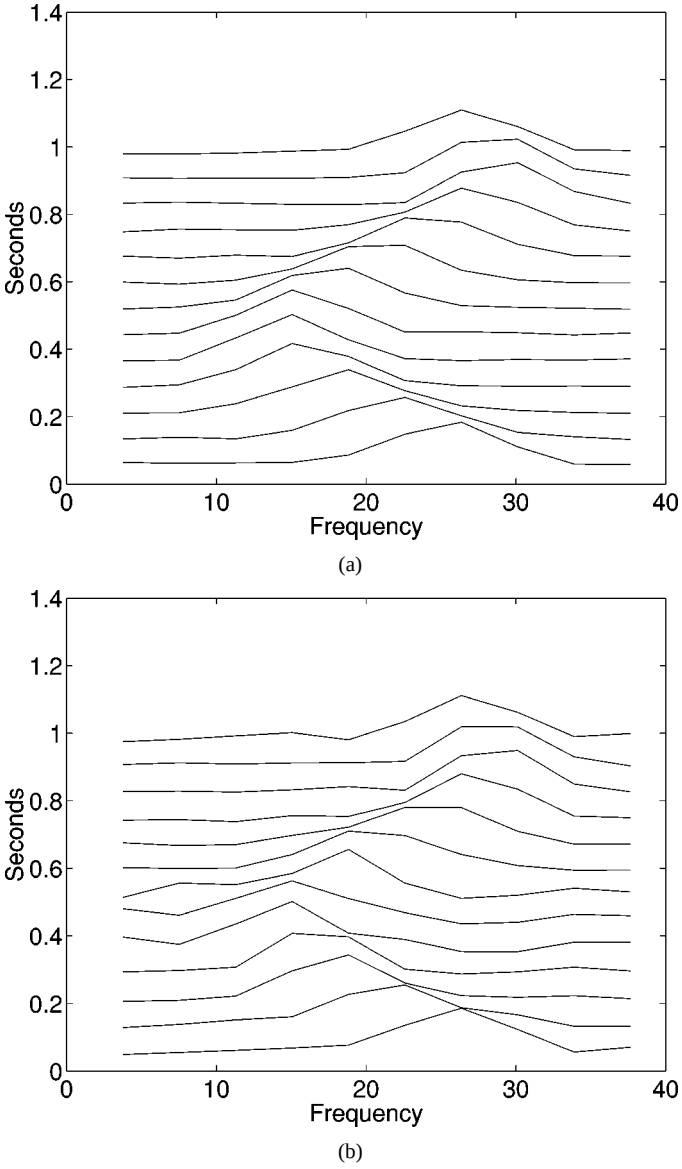

Fig. 4. Short-time fourier transform estimates of the spectrogram for (a) the low noise variance data in Fig. 2 and (b) the high noise variance data in Fig. 3. To achieve a reasonable temporal resolution the STFT trades off spectral resolution.

For the low and high noise data sets, the EM algorithm required 33 and 34 iterations to converge, which took 8 s and 12 s. The VB algorithm required 35 and 51 iterations, which took 11 s and 40 s. More generally, the computer time scales with MAR model order.

The main difference in the estimated parameters from the EM and VB approaches is in the inferred values of the diagonal elements of the state transition matrix A. Over a number of runs, for both levels of noise variance, VB values were typically 0.7–0.8 whereas EM values were typically about 0.3–0.4. This shows that the VB approach places greater emphasis on previous samples resulting in a smoother tracking.

The VB results should also be compared to the original Kalman filtering approach [19]. Although this produced good spectral estimates it required manual setting of state noise precision, q, and observation noise precision, r, parameters (more precisely, the ratio of q to r was set to a value that made the "relative reduction in prediction error due to adaption" equal to 0.5, a hand-crafted arbitrary value). In contrast, the VB algorithm is fully automatic; q and r (and the state transformation matrix) are inferred from the data.

We also compare results with those obtained with a short-time Fourier transform (STFT) approach [14]. We used windows of length 32 samples, the data in each window being processed by a 32-tap Hanning filter. To obtain reasonable temporal resolution these windows were overlapped by 24 samples. This provided 13 spectral estimates for the 1-s data period. The resulting spectrograms are shown in Fig. 4. Changes to the window length, overlap and filter parameters yielded very similar spectral estimates. The spectrograms from the STFT are very broad-band and are not at all competitive with those from the nonstationary MAR model.

Our second example looks at a bivariate data set consisting of two stationary regimes with an abrupt transition between them. The first regime consists of two coherent 40-Hz modes and the second of a 40-Hz mode and a 10-Hz mode. The signals were sampled at 128 Hz and 100 samples were generated for each

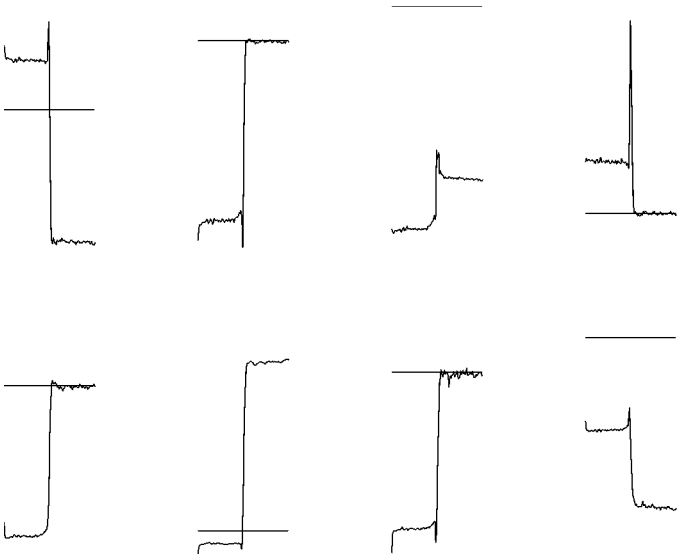

Fig. 5. Tracking of two-time series which, before the state transition, consist of two coherent 40-Hz modes and afterwards consist of a 10-Hz mode and a 40-Hz mode. There are 100 samples in each state. The left four plots show the evolution of the lag-1 coefficients and the right four show the lag-2 coefficients. The horizontal line indicates the zero level. After the state transition, the off-diagonal coefficients go to zero reflecting the "disconnection" of the two time series.

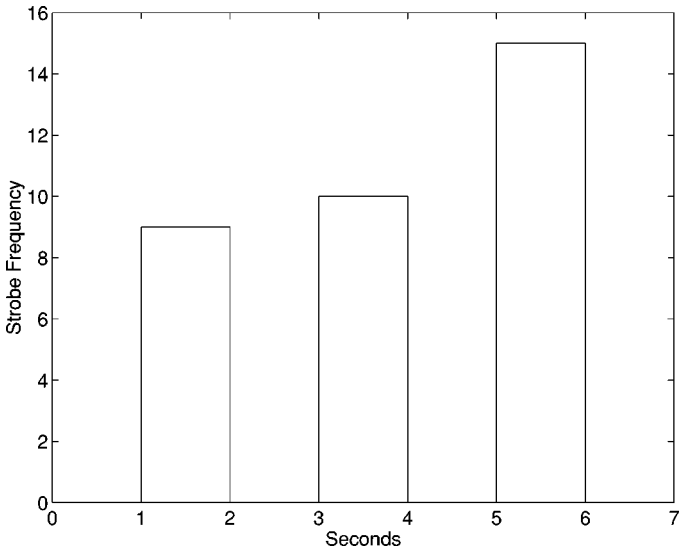

Fig. 6. Photic Stimulation Each trial of photic stimulation consisted of 1-s stimulus blocks interleaved with 1-s rest blocks.

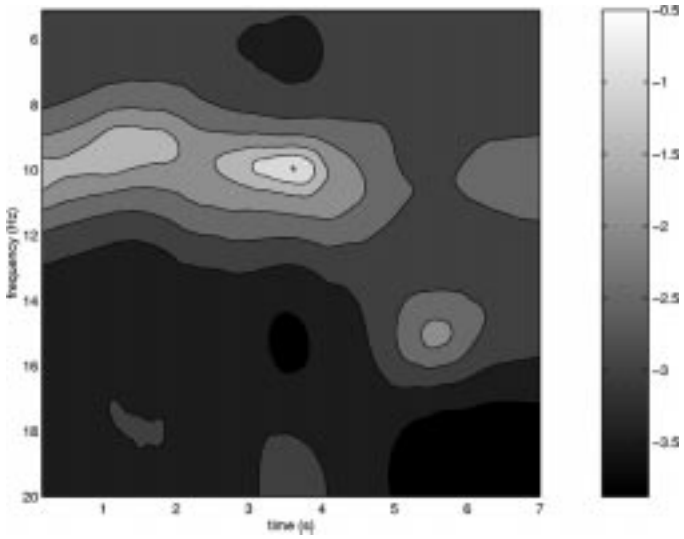

Fig. 7. The time varying log power spectrum of the EEG signal recorded from P3 (the P4 spectrum is very similar) during alternating periods of rest and strobe activity.

regime. We generated ten such time series and presented them to a nonstationary VB-MAR algorithm with model order set to $p = 2$. Fig. 5 shows the tracking of the MAR coefficients; a state-transition is noticeable half way through each coefficient time series. The change in magnitude of the off-diagonal coefficients reflects a disconnection of the time series in the second regime (i.e., they are no longer coherent).

B. Physiological Results

1) Photic Synchronization: The previous simulation examples demonstrate the strengths of our algorithm when compared to ML approaches. In this section, we shall investigate its performance on some real physiological data. The problem with this kind of testing is that an algorithm can produce a set of spectra but generally we do not know what the exact spectra should be (unlike the simulation examples). It is, therefore, important to try and find experimental scenarios in which the results are as predictable as possible. To this end, our first example applies the algorithm to some EEG data recorded during a photic synchronization experiment.

It is well known that one can cause EEG activity to synchronise with a flashing strobe. In this case, one has a certain amount of control over the spectral peaks that one expects to see, so properties of the algorithm (such as spectral resolution) can be tested within a true physiological context.

EEG data was recorded from channels P3 and P4 (over the left and right parietal cortex), referenced to linked ears. Each trial of photic stimulation consisted of 1-s stimulus blocks interleaved with 1-s rest blocks (see Fig. 6). The frequencies of the stimulus blocks were 9, 10, and 15 Hz. EEG data was acquired throughout each trial at a sample rate of 80 Hz and a total of seven trials of data were analyzed using the multiple E-step approach described in Section IV-C.

Figs. 7 and 9 show a log-power spectrum and the coherence spectrum produced by our algorithm. One can see that it has successfully resolved the spectral peaks in the power and coherence at 9, 10, and 15 Hz, respectively. A harmonic peak in the coherence has also been detected at 18 Hz, between seconds 1 and 2. Note that in Fig. 7, the normal alpha activity at around 10 Hz is evident between periods of strobe activity (*e.g.*, at the beginning of the record). Also in Fig. 9, one should note that although stimulation synchronises both hemispheres, the resting alpha activity is not coherent between P3 and P4. It is for this reason that the peaks in Fig. 9 are better localized to the periods of photic stimulation. The algorithm is slow to detect the temporal changes and in this particular example, the discrete nature of the changes in stimulus would really be better captured by a hidden Markov model (the VB approach to the hidden Markov MAR model is considered in [20]). However, by comparison

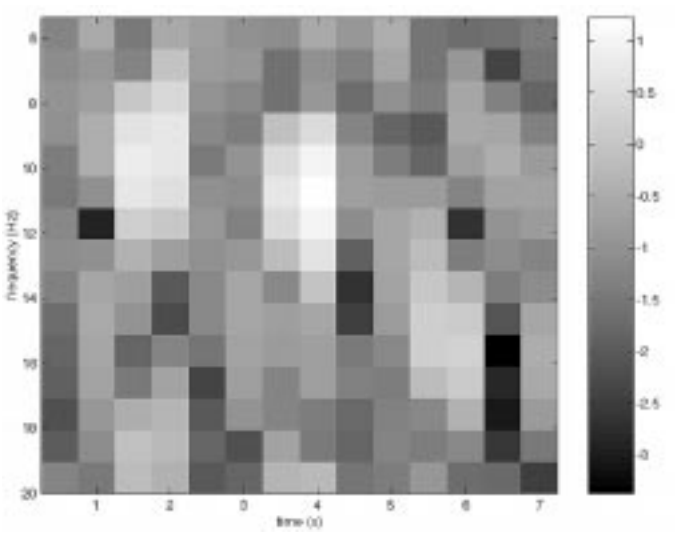

Fig. 8. The time varying log power spectrum of the EEG signal recorded from P3 during alternating periods of rest and strobe activity, calculated using the STFT.

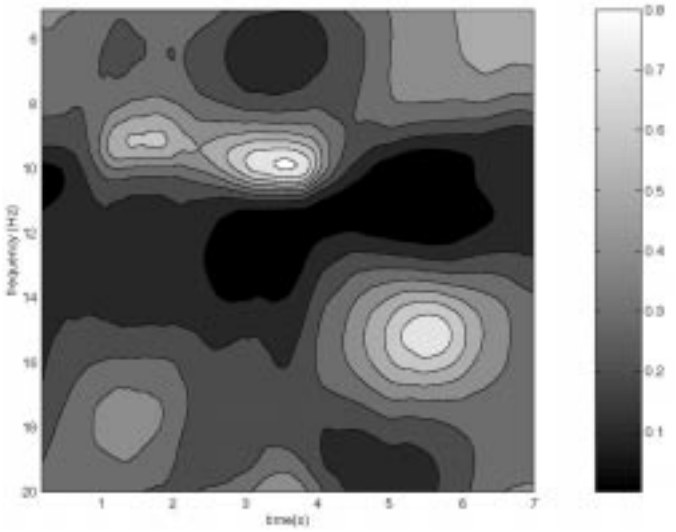

Fig. 9. The time varying coherence spectrum between the two EEG signals recorded from P3 and P4. The timing is as in the previous figure.

with other continuously changing nonstationary spectral estimation techniques, our algorithm performs well and manages to detect true spectral features of a physiological time series, even with a small data set.

It should be apparent that the small amount of data here prevents one from obtaining good results from a short-time fast Fourier analysis. Nevertheless, in Fig. 8 we present the best log-power spectrum that we could obtain with this method. We also note that we could not obtain any kind of time-varying coherence plot using this method. For wavelet analysis, another popular nonstationary spectral estimation paradigm, the estimation of coherence estimates is problematic [21] and in many experimental situations the coherence is of great interest due to its interpretation as a measure of functional coupling between different cortical areas or between cortex and muscle [22].

2) Event-Related Desynchronization: In our second example, ERD, the exact nature of the spectrum is not known with the same precision as the strobe scenario. However, many

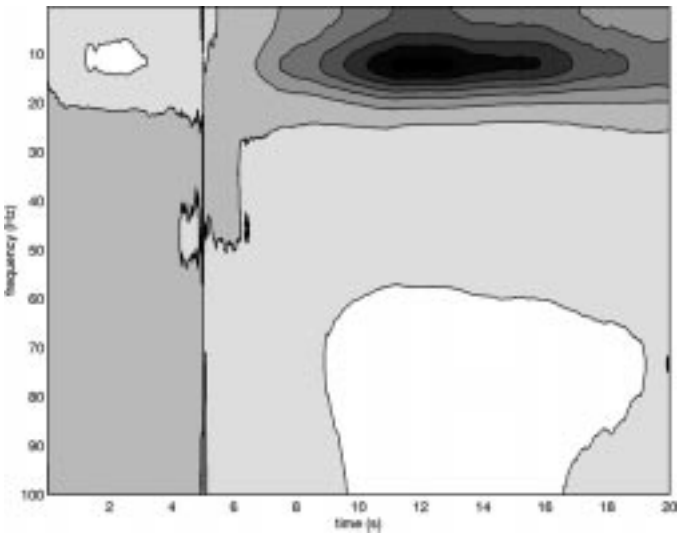

Fig. 10. Cusum showing ERD at low frequencies and event-related synchronization at high frequency over the right-hand area (C3) while movements were made with the left hand. The movement occurs at $t = 10$ s. White (black) areas denote an increase (decrease) in spectral power.

qualitative features of this phenomenon have been discovered by other researchers and provide us with a solid body of evidence for the analysis presented here. The point is that previously, researchers have had to average over large numbers of movement trials in order to discern the salient features. Here we demonstrate that with only a few trials, we are not only able to replicate the primary experimental findings of others but can also detect physiologically appropriate features that are insensitive to other Fourier-based averaging procedures.

ERD is a phenomenon that appears when one makes self-paced movements [23], [24]. In brief, the main effect is that one observes a drop in the EEG's spectral power in the alpha frequency band a few seconds before the the physical movement occurs. Then, after the movement terminates, the spectral power "rebounds" back to its resting level. This phenomenon, therefore, provides a nice example of an event-related spectral change amenable to study with our algorithm.

A right-handed subject was asked to make self-paced finger movements at a frequency of about 3/min. The arm, hand, and fingers were supported with the right forefinger held horizontal and unsupported. The self-paced movements involved a brief extension of the right finger followed by a return to the horizontal position, and EEG signals were recorded from C3, F3, Cz, Fz, C4, and F4, all referenced to the left ear. Eye movement artifacts and excess electromyogram (EMG) activity contaminating the EEG were removed using independent components analysis [25]. The cleaned signal taken from the contralateral hand area (C3) was then broken up into 20-s sections with the movement timed to $t = 10$ s. A trigger signal was recorded from an accelerometer placed on the distal phalanx and data was acquired simultaneously with the EEG recording. Twelve clean movements were concatenated and put through the VB algorithm (as described in Section IV-C). Spectra were derived and a cumulative sum (cusum) was computed in the time domain to reduce noise in the spectra and to bring out any genuine spectral changes. Fig. 10 shows a time-frequency plot (derived from the MAR coefficients) of this cusum. The baseline spectrum for

the cusum was taken as the average of the spectra over the first 5 s, which is the cause of the vertical line at $t = 5$ s. One can clearly see the ERD in the lower frequency bands, which begins about 3 s before the movement and then rebounds at about 5 s after the start of the movement. One can also see a smaller increase in power in the high-frequency band, centerd at 80 Hz, that follows the time course of the alpha desynchronization. This increase in power at around 80 Hz was much greater contralateral to the movement and, hence, is unlikely to be due to scalp EMG.

Event-related increases in high-frequency activity have hitherto only been reported in patients with subdural electrode grids [26]. These features could be sharpened up by including more movements in the average, as is done in more traditional approaches to event-related (de)synchronization. Physiologically, this result is important as it provides evidence that movement-related oscillatory neuronal activity is present in both high- and low-frequency bands.

VI. Discussion

We have proposed an algorithm for modeling nonstationary multivariate time series, conforming to a multivariate autoregressive process with time-varying coefficients that adapt according to a linear dynamical system. While such a model is not new, the contribution of this paper has been to present a Bayesian implementation which allows us retain the full generality of the model while deriving a practical algorithm.

By placing priors over model coefficients we have derived a VB algorithm which is a generalization of Shumway and Stoffer's [5]. EM algorithm for time-series analysis. Whereas the EM approach, being a ML algorithm, is prone to data overfitting, we have shown that the VB algorithm is robust.

An intermediate step between EM and VB can be implemented by placing priors over model coefficients and then estimating the maximum *a posteriori* (MAP) parameters. Effectively, the priors act as regularizers which prevent model overfitting. The MAP approach is computationally attractive but unfortunately requires the use of *ad-hoc* regularization parameters. In contrast, the VB algorithm provides an automatic method for finding the appropriate regularizers. For a discussion of these issues in the context of neural networks, see [27].

A further virtue of the VB approach is that it delivers posterior distributions over model parameters. These can then be used to provide predictive distributions for previously unseen data points. These predictive distributions are obtained by "integrating out" the model parameters and so capture all of the uncertainty of the modeling process. Predictive distributions of this sort have been obtained for VB Gaussian Mixture Models [11] and can also be obtained for nonstationary MAR models. In this paper, however, we have focussed on characterization rather than prediction and defer the use of predictive distributions to future papers.

While we have shown that the algorithm is insensitive to misspecification of model order, it is also possible to use our approximation of the model evidence as a model order selection criterion. In previous work, we have shown how this applies to the simpler cases of stationary MAR models (see [2]–[4]). The extension to the nonstationary case is straightforward, and will be demonstrated in following papers.

A further point we would like to emphasise is the suitability of the model for analysing event-related data sets. Instead of combining all the epochs together by taking a grand average, each epoch is treated as a separate observation sequence and all the sequences are presented to the VB algorithm. The algorithm then takes averages in the space of hidden variables (i.e., the MAR coefficients) rather than in the observed variables (i.e., the raw signals). This approach has been demonstrated here in two different EEG experiments.

From the experimental neuroscientist's point of view, the algorithm has a number of desirable properties. We have shown that it performs spectral estimation in scenarios where a short data set would prevent any sort of Fourier-based analyses. In addition, the MAR representation enables straightforward calculation of time-varying coherences and phases. These quantities are of interest to anyone wishing to study aspects of functional coupling. As we have noted, the VB framework helps one to find the most appropriate MAR model for a particular data set. Thus, the Bayesian framework combined with the MAR representation of time series provides a potentially powerful tool for the analysis of biomedical signals.

Appendix I
E-Step Kalman Smoother Recursions

Following [13], we write the expected value of x_t conditioned on all data up to time t as $x_t^t \equiv E[x_t|y_1^t]$. Similarly, the corresponding covariance is given by $\Sigma_t^t \equiv \mathrm{Var}[x_t|y_1^t]$.

A. Forward Recursions

This step implements the recursive computation of x_t^t and Σ_t^t from x_{t-1}^{t-1} and Σ_{t-1}^{t-1}

$$\begin{aligned} x_t^{t-1} &= \bar{A} x_{t-1}^{t-1} \\ \Sigma_t^{t-1} &= \bar{A}\Sigma_{t-1}^{t-1}\bar{A}^T + \bar{Q} \\ K_t &= \Sigma_t^{t-1} C_t^T \left(C_t \Sigma_t^{t-1} C_t^T + \bar{R}\right) \\ x_t^t &= x_t^{t-1} + K_t\left(y_t - C_t x_t^{t-1}\right) \\ \Sigma_t^t &= \Sigma_t^{t-1} - K_t C_t \Sigma_t^{t-1}. \end{aligned} \tag{33}$$

The procedure is initialized using $x_1^0 = \mu_1$ and $\Sigma_1^0 = \Sigma_1$ where the right-hand-side quantities are updated in the previous M-step [see (21), (22), and (24)].

B. Backward Recursions

The backward recursions compute x_t^t and Σ_t^t from x_{t-1}^{t-1} and Σ_{t-1}^{t-1}

$$\begin{aligned} J_{t-1} &= \Sigma_{t-1}^{t-1}\bar{A}^T\left(\Sigma_t^{t-1}\right)^{-1} \\ x_{t-1}^T &= x_{t-1}^{t-1} + J_{t-1}\left(x_t^T - \bar{A}x_{t-1}^{t-1}\right) \\ \Sigma_{t-1}^T &= \Sigma_{t-1}^{t-1} + J_{t-1}(\Sigma_t^T - \Sigma_t^{t-1})J_{t-1}^T. \end{aligned} \tag{34}$$

The procedure is initialized using $\Sigma_T^T = \Sigma_T$ and $x_T^T = x_T$ where the right-hand-side quantities are from the final forward recursion step. The forward and backward steps together allow us to compute x_t^T and Σ_t^T which are the first two moments of

x_t conditioned on the *whole* data set. We, therefore, have the update for $q(x_t) = \mathcal{N}(\mu_t, \Sigma_t)$ as

$$\begin{aligned}\mu_t \equiv& x_t^T \\ \Sigma_t \equiv& \Sigma_{t-1}^T.\end{aligned} \tag{35}$$

We also let

$$M_t \equiv \Sigma_t + \mu_t \mu_t^T \tag{36}$$

and derive $M_{t,t-1} \equiv \Sigma_{t,t-1}^T + \mu_t \mu_{t-1}^T$ where backward recursions are used for

$$\Sigma_{t-1,t-2}^T = \Sigma_{t-1}^{t-1} J_{t-2}^T + \left(\Sigma_{t,t-1}^T - \bar{A}\Sigma_{t-1}^{t-1}\right) J_{t-2}^T. \tag{37}$$

The last recursion is initialized using

$$\Sigma_{T,T-1}^T = (I - K_T C_t)\bar{A}\Sigma_{T-1}^{T-1}. \tag{38}$$

Appendix II
The Kronecker Product

If A is a $m \times m$ matrix and B is a $n \times n$ matrix, then the Kronecker product of A and B is the $(mn) \times (mn)$ matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1m}B \\ \dots & & \\ a_{m1}B & & a_{mm}B \end{bmatrix}. \tag{39}$$

For properties of the Kronecker product see, for example, in [12, p. 477].

Appendix III
Free Energy

The negative free energy (the lower bound on the model evidence) can be written as a sum of three terms. The first term is the average log-likelihood, where the expectation is taken with respect to the posterior density, and can be written as

$$L_{av} = \int\int q(x^\tau) q(\theta) \log p(D|\theta) dx^\tau d\theta. \tag{40}$$

The second term is the entropy of the hidden variables and is

$$H(x^\tau) = -\int q(x^\tau) \log q(x^\tau) dx^\tau, \tag{41}$$

while the final term penalises complex models and contains the Kullback–Leibler divergences between the posterior and prior distributions. This is

$$\begin{aligned}-\int q(\theta) \log\left(\frac{q(\theta)}{p(\theta)}\right) d\theta =& -KL(\alpha) - KL(A) - L_k(q, D) \\ & - L_d(r, B) + H(R) + H(Q)\end{aligned} \tag{42}$$

where the integral $L_N(a, B)$ is defined in (51) below. $KL(A)$, $KL(\alpha)$ and the entropy terms H are calculated from (46), (48), and (52), respectively. Many terms cancel in the sum, so we are left with the expression

$$\begin{aligned}F =& -\frac{dT}{2}\log 2\pi + \frac{k}{2}(\log 2\pi + T) - KL(\alpha) \\ & - KL(A) + \log Z_k(T-1, D) + \log Z_d(T, B) \\ & - \frac{1}{2}\sum_{t=1}^{T} \log|\Sigma_t|,\end{aligned} \tag{43}$$

where $Z_N(a, B)$ is also defined in Appendix.

Appendix IV
Densities and Divergences

The KL-divergence between densities $q(x)$ and $p(x)$ is

$$KL(q, p) = \int q(x) \log \frac{q(x)}{p(x)} dx. \tag{44}$$

A. Normal Density

The multivariate Normal density is given by

$$\mathcal{N}(\mu, \Sigma) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{-d/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \tag{45}$$

The KL divergence for Normal densities $q(x) = \mathcal{N}(\mu_q, \Sigma_q)$ and $p(x) = \mathcal{N}(\mu_p, \Sigma_p)$ is

$$\begin{aligned}KL(q, p) =& 0.5 \log \frac{|\Sigma_p|}{|\Sigma_q|} + 0.5 \text{ trace } \left(\Sigma_p^{-1}\Sigma_q\right) \\ & + 0.5(\mu_q - \mu_p)^T \Sigma_p^{-1}(\mu_q - \mu_p) - \frac{d}{2}\end{aligned} \tag{46}$$

where $|\Sigma_p|$ denotes the determinant of the matrix Σ_p.

B. Gamma Density

The Gamma density is given by

$$\mathcal{G}(b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right). \tag{47}$$

For Gamma densities $q(x) = \mathcal{G}(b_q, c_q)$ and $p(x) = \mathcal{G}(b_p, c_p)$, the KL-divergence is

$$\begin{aligned}KL(q, p) =& (c_q - 1)\Psi(c_q) - \log b_q - c_q - \log\Gamma(c_q) \\ & + \log\Gamma(c_p) + c_p \log b_p \\ & - (c_p - 1)(\Psi(c_q) + \log b_q) + \frac{b_q c_q}{b_p}\end{aligned} \tag{48}$$

where $\Psi()$ is the digamma function [28].

C. Wishart Density

The Wishart distribution is given by ([29, page 85])

$$\mathcal{W}_d(a, B) = \frac{1}{Z_d(a, B)} |\Lambda|^{(a-d-1)/2} \exp\left[-\frac{1}{2}Tr(B\Lambda)\right] \tag{49}$$

where

$$Z_d(a, B) = 2^{ad/2} |B|^{-a/2} \Gamma_d\left(\frac{a}{2}\right) \tag{50}$$

and $\Gamma_d()$ is the generalized gamma function defined on page 62 of [29].

The entropy and KL-divergence of a Wishart can be defined in terms of the integral

$$L_d(a, B) = \int \mathcal{W}(a, B) \log|\Lambda| d\Lambda. \tag{51}$$

The entropy of $q(\Lambda) = \mathcal{W}(q, Q)$ is then given by

$$H(q) = -\left(\frac{q-d-1}{2}\right) L_d(q, Q) + \frac{qd}{2} + \log Z_d(q, Q). \tag{52}$$

The KL-Divergence between densities $q(\Lambda) = \mathcal{W}_d(q, Q)$ and $p(\Lambda) = \mathcal{W}_d(p, P)$ is given by

$$KL(q,p) = \left(\frac{q-d-1}{2}\right) L_d(q,Q) - \frac{qd}{2} + \frac{q}{2} Tr(PQ^{-1}) + \log \frac{Z_d(p,P)}{Z_d(q,Q)} - \left(\frac{p-d-1}{2}\right) L_d(p,P). \quad (53)$$

ACKNOWLEDGMENT

The authors would like to thank S. Roberts, K. Friston, S. Kiebel and especially P. Brown for their advice and support during the preparation of this manuscript.

REFERENCES

[1] H. Lappalainen and J. W. Miskin, "Ensemble learning," in *Advances in Independent Component Analysis*, M. Girolami, Ed. Berlin, Germany: Springer-Verlag, 2000.
[2] W. D. Penny and S. J. Roberts, "Bayesian multivariate autoregressive models with structured priors," Inst. Elect. Eng. Trans. Med., Image .Signal Processing, 2002, to be published.
[3] ——, "Bayesian methods for autoregressive models," presented at the IEEE Int. Workshop Neural Networks for Signal Processing, Sydney, Australia, 2000.
[4] ——, "Variational Bayes for nongaussian autoregressive models," presented at the IEEE Int. Workshop Neural Networks for Signal Processing, Sydney, Australia, 2000.
[5] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.
[6] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[7] M. Arnold, W. H. R. Miltner, H. Witte, R. Bauer, and C. Braun, "Adaptive AR modeling of nonstationary time series by means of Kalman filtering," *IEEE Trans. Biomed. Eng.*, vol. 45, pp. 553–562, May 1998.
[8] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural Computation*, vol. 11, no. 2, pp. 305–346, 1999.
[9] J. J. K. O'Ruaniaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. Berlin, Germany: Springer-Verlag, 1996.
[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*: John Wiley, 1991.
[11] H. Attias *et al.*, "A variational Bayesian framework for graphical models," in *NIPS 12*, T. Leen *et al.*, Eds. Cambridge, MA: MIT Press, 2000.
[12] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. New York: Wiley, 1992.
[13] Z. Ghahramani and G. E. Hinton. (1996) Parameter Estimation for Linear Dynamical Systems. Univ. Toronto, Dept. Comput. Sci, Toronto, ON, Canada. [Online] Available: http://www.gatsby.ucl.ac.uk/zoubin/papers.html. Tech. Rep. CRG-TR-96-2
[14] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1991.
[15] Z. Ghahramani and M. J. Beal, "Propagation algorithms for variational Bayesian learning," in *NIPS 13*, T. Leen, Ed. Cambridge, MA: MIT Press, 2001.
[16] M. J. Beal and Z. Ghahramani, *The Variational Kalman Smoother*. London, U.K.: Gatsby Computat. Neurosci. Unit, 2001.
[17] A. N. Shiryaev, *Probability*. Berlin, Germany: Springer-Verlag, 1984.
[18] S. L. Marple, *Digital Spectral Analysis With Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
[19] D. W. Skagen, "Estimation of running frequency spectra using a Kalman filter algorithm," *J. Biomed. Eng.*, vol. 10, pp. 275–279, May 1988.
[20] M. J. Cassidy and P. Brown, "Hidden Markov based autoregressive analysis of stationary and nonstationary electrophysiological signals for functional coupling studies," *J. Neurosci. Meth.*, vol. 116, pp. 35–53, 2002.
[21] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bull. Amer. Meteorological Soc.*, vol. 79, pp. 61–78, 1998.
[22] P. Brown, "Cortical drives to human muscle: The piper and related rhythms," *Progress in Neurobiology*, vol. 60, pp. 97–108, 2000.
[23] R. Salmelin and R. Hari, "Spatiotemporal characteristics of sensorimotor neuromagnetic rhythms related to thumb movement," *Neuroscience*, vol. 60, pp. 537–550, 1994.
[24] C. Toro, G. Deuschl, R. Thatcher, S. Sato, C. Kufta, and M. Hallett, "Event-related desynchronization and movement-related cortical potentials on the ECoG and EEG," *Electroencephalogr. Clin. Neurophysiol.*, vol. 93, pp. 380–389, 1994.
[25] T.-P. Jung, S. Makeig, A. J. Bell, and T. J. Sejnowski, "Independent components analysis of electroencephalographic data," *NIPS*, vol. 8, pp. 145–151, 1996.
[26] N. E. Crone, D. L. Miglioretti, B. Gordon, and R. P. Lesser, "Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis II. Event related synchronization in the gamma-band," *Brain*, vol. 121, pp. 2301–2315, 1998.
[27] D. J. C. Mackay, "Bayesian interpolation," *Neural Computat.*, vol. 4, pp. 415–447, 1992.
[28] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. V. P. Flannery, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 1992.
[29] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: Wiley, 1982.

Michael J. Cassidy received the Ph.D. degree in mathematics from Cambridge University, Cambridge, U.K., in 1998.

He is currently working in the Sobell Department of Neurophysiology at University College London, London, U.K., and is interested in the analysis of movement-related physiological data.

William D. Penny received the Ph.D. degree in electrical engineering from Imperial College, London, U.K., in 1993.

He has since worked as a Postdoctoral Researcher at University College London, London, U.K., Imperial College, and Oxford University, Oxford, U.K., and is now working as a brain-imaging statistician at University College London, London, U.K.