Bayesian multivariate autoregressive models with structured priors

W.D.Penny and S.J.Roberts

Abstract: A variational Bayesian (VB) learning algorithm for parameter estimation and model-order selection in multivariate autoregressive (MAR) models is described. The use of structured priors in which subsets of coefficients are grouped together and constrained to be of a similar magnitude is explored. This allows MAR models to be more readily applied to high-dimensional data and to data with greater temporal complexity. The VB model order selection criterion is compared with the minimum description length approach. Results are presented on synthetic and electroencephalogram data.

1 Introduction

The multivariate autoregressive (MAR) process is used to model multiple time series data in such fields as geophysics [1], economics [2] and biomedicine [3]. It can also be seen as a parametric multivariate spectral estimation procedure and will provide parsimonious estimation of coherences and partial coherences [4, 5]. One factor preventing its wider application, however, is the explosion in the size of the model as the number of time-series increases; MAR models for d time series have parameters of order d^2.

We show how the variational Bayesian (VB) framework can be applied to MAR models. By using 'structured priors' in which subsets of coefficients are constrained to be of a similar magnitude, the effective degrees of freedom in the model can be constrained. This allows MAR models to be more readily applied to high-dimensional data. Also, VB provides a model order selection criterion which can be used to select the appropriate number of time-lags.

While the Bayesian methodology has a long history the use of VB is relatively new; the key idea of VB is to find an approximation to the true posterior density which minimises the Kullback–Liebler (KL) divergence between these two densities. Notable recent applications are to principal component analysis [6] and independent component analysis [7]. We have also recently applied VB to univariate autoregressive (AR) models [8] and univariate non-Gaussian AR models [9].

2 Multivariate autoregressive models

An MAR(p) model predicts the next value in a d-dimensional time series (i.e. there are d time series), $\boldsymbol{y}_n$ as a linear combination of the p previous vector values of the time series

$$\boldsymbol{y}_n = \sum_{i=1}^{p} \boldsymbol{y}_{n-i} \boldsymbol{A}(i) + \boldsymbol{e}_n \tag{1}$$

where $\boldsymbol{y}_n = [y_n(1), y_n(2), \ldots, y_n(d)]$ is the nth sample of a d-dimensional time series, each $\boldsymbol{A}(i)$ is a d-by-d matrix of coefficients and $\boldsymbol{e}_n = [e_n(1), e_n(2), \ldots, e_n(d)]$ is Gaussian noise having zero mean and precision (inverse covariance) matrix $\boldsymbol{\Lambda}$. We have assumed that the data mean has been subtracted from the time series.

The model can be written in the standard form of a multivariate linear regression model as follows:

$$\boldsymbol{y}_n = \boldsymbol{x}_n \boldsymbol{W} + \boldsymbol{e}_n \tag{2}$$

where $\boldsymbol{x}_n = [\boldsymbol{y}_{n-1}, \boldsymbol{y}_{n-2}, \ldots, \boldsymbol{y}_{n-p}]$ are the p previous multivariate time series samples and $\boldsymbol{W}$ is a $(p \times d)$-by-d matrix of MAR coefficients (weights). There are therefore a total of $k = p \times d \times d$ MAR coefficients.

If the nth rows of $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{E}$ are $\boldsymbol{y}_n$, $\boldsymbol{x}_n$ and $\boldsymbol{e}_n$, respectively, and there are $n = 1, \ldots, N$ samples then we can write

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{E} \tag{3}$$

where $\boldsymbol{Y}$ is an N-by-d matrix, $\boldsymbol{X}$ is an N-by-$(p \times d)$ matrix and $\boldsymbol{E}$ is an N-by-d matrix. Writing the MAR model in this form allows us to make contact with the large body of statistical literature devoted to the multivariate linear regression model, e.g. Box and Tiao ([10], p. 423). From these sources we know that given a data set $\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{Y}\}$ the likelihood of the data is given by

$$p(\boldsymbol{D} \mid \boldsymbol{W}, \boldsymbol{\Lambda}) = (2\pi)^{-dN/2} |\boldsymbol{\Lambda}|^{N/2} \exp\left[-\frac{1}{2} Tr(\boldsymbol{\Lambda} \boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{W})) \right] \tag{4}$$

where $|\,|$ denotes the determinant, $Tr(\,)$ the trace and

$$\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{W}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}) \tag{5}$$

is the unnormalised error covariance matrix.

Before defining the priors in our Bayesian model we introduce the *vec* notation

$$\boldsymbol{w} = vec(\boldsymbol{W}) \tag{6}$$

IEE Proceedings online no. 20020149

DOI: 10.1049/ip-vis:20020149

Paper first received 22nd September 2000 and in revised form 28th September 2001

W.D. Penny is with the Wellcome Department of Cognitive Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK

S.J. Roberts is with the Department of Engineering Science, Oxford University, Parks Road, Oxford OX1 3PJ, UK

where $vec(\boldsymbol{W})$ denotes the columns of $\boldsymbol{W}$ being stacked on top of each other (for more on the *vec* notation, see [11]). To recover the matrix $\boldsymbol{W}$ we simply 'unstack' the columns from the vector $\boldsymbol{w}$. The transformation of a matrix into a vector in this way is a standard method for implicitly defining a probability density over a matrix.

In our Bayesian model we assume that the weights are drawn from a zero-mean Gaussian prior with an isotropic covariance having precision α

$$p(\boldsymbol{w}|\alpha) = \left(\frac{\alpha}{2\pi}\right)^{k/2} \exp(-\alpha E(\boldsymbol{w})) \tag{7}$$

where we define the 'weight error' as

$$E(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{8}$$

The weight precision is drawn from a Gamma prior

$$p(\alpha) = \mathrm{Ga}(\alpha; b_\alpha, c_\alpha) \tag{9}$$

The noise precision matrix is taken to have the prior

$$p(\boldsymbol{\Lambda}) = |\boldsymbol{\Lambda}|^{-(d+1)/2} \tag{10}$$

This density is an 'improper distribution' as it does not integrate to unity. The prior is, however, only assumed to have this form over a range in which the likelihood is appreciably nonzero. Outside this range the prior tails off to zero. This choice of prior is the 'uninformative prior' for multivariate linear regression: see Chapter 1 and page 426 of Box and Tiao [10] for a full discussion. We concatenate all the parameters of the model into the vector $\boldsymbol{\theta} = [\boldsymbol{w}, \alpha, \boldsymbol{\Lambda}]$. We also assume independence between the parameter groups so that the overall prior is

$$p(\boldsymbol{\theta}) = p(\boldsymbol{w}|\alpha)p(\alpha)p(\boldsymbol{\Lambda}) \tag{11}$$

2.1 Maximum likelihood

The maximum likelihood (ML) solution [12] for the MAR coefficients is

$$\boldsymbol{W}_{ML} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{12}$$

The maximum likelihood noise covariance $\boldsymbol{S}_{ML}$ can be estimated as

$$\begin{aligned}\boldsymbol{S}_{ML} &= \frac{1}{N-k}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_{ML})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_{ML}) \\ &= \frac{1}{N-k}\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{W}_{ML})\end{aligned} \tag{13}$$

where $k = p \times d \times d$. Again, we define $\boldsymbol{w}_{ML} = vec(\boldsymbol{W}_{ML})$. The ML parameter covariance matrix for $\boldsymbol{w}_{ML}$ is given by ([13], p. 321)

$$\boldsymbol{\Sigma}_{ML} = \boldsymbol{S}_{ML} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{14}$$

where $\otimes$ denotes the Kronecker product (see, e.g. p. 477 in Box and Tiao [10]).

The problem with the ML approach, however, is model-overfitting. This can be overcome, to an extent, by combining ML with a model-order selection criterion such as the minimum description length (MDL). This is discussed in Section 4. We propose an alternative solution based on the variational Bayesian framework which we now describe.

3 Variational Bayesian framework

The 'evidence' or 'marginal likelihood' $p(\boldsymbol{D})$ of a probabilistic model is the likelihood of the model after its parameters have been integrated out. That is

$$p(\boldsymbol{D}) = \int p(\boldsymbol{D}, \boldsymbol{\theta}) d\boldsymbol{\theta} \tag{15}$$

The log of the evidence can be written as

$$\log p(\boldsymbol{D}) = \log \int p(\boldsymbol{D}, \boldsymbol{\theta}) d\boldsymbol{\theta} \tag{16}$$

This equality is unchanged by multiplying top and bottom by the same quantity, $q(\boldsymbol{\theta}|\boldsymbol{D})$ (which we shall soon see is the approximate posterior)

$$\log p(\boldsymbol{D}) = \log \int q(\boldsymbol{\theta}|\boldsymbol{D}) \frac{p(\boldsymbol{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{D})} d\boldsymbol{\theta} \tag{17}$$

Using Jensen's inequality (see, e.g. [6]) we have

$$\log p(\boldsymbol{D}) \geq F(p) \tag{18}$$

where

$$F(p) = \int q(\boldsymbol{\theta}|\boldsymbol{D}) \log \frac{p(\boldsymbol{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{D})} d\boldsymbol{\theta} \tag{19}$$

The quantity $F(p)$ is known (to physicists) as the negative variational free energy. This provides a lower bound on the log evidence, with equality if the approximating posterior is equal to the true posterior, i.e. if $q(\boldsymbol{\theta}|\boldsymbol{D}) = p(\boldsymbol{\theta}|\boldsymbol{D})$. The aim of VB learning or 'ensemble' learning is to maximise this lower bound. Using $p(\boldsymbol{D}, \boldsymbol{\theta}) = p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ we can write

$$F(p) = L_{av} - KL(q \,\|\, p) \tag{20}$$

where

$$L_{av} = \int q(\boldsymbol{\theta}|\boldsymbol{D}) \log p(\boldsymbol{D}|\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{21}$$

and

$$KL(q, p) = \int q(\boldsymbol{\theta}|\boldsymbol{D}) \log \frac{q(\boldsymbol{\theta}|\boldsymbol{D})}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{22}$$

The first term in $F(p)$ is the average log-likelihood of the data L_{av} where averaging takes place over the approximate posterior density $q(\boldsymbol{\theta}|\boldsymbol{D})$. The second term is the KL-divergence between the approximating posteriors and the priors. This increases with an increasing number of model parameters and so acts as a penalty term which penalises more complex models.

Maximising the negative free energy for a MAR model yields a posterior distribution which factorises as follows:

$$q(\boldsymbol{\theta}|\boldsymbol{D}) = q(\boldsymbol{w}|\boldsymbol{D})q(\alpha|\boldsymbol{D})q(\boldsymbol{\Lambda}|\boldsymbol{D}) \tag{23}$$

This essentially arises because the prior (11) is of the same form. We can then maximise F with respect to each of $q(\boldsymbol{w}|\boldsymbol{D})$, $q(\alpha|\boldsymbol{D})$ and $q(\boldsymbol{\Lambda}|\boldsymbol{D})$ separately. This is a generalisation of the procedure described by Mackay [14] who shows, in some detail, how $q(\boldsymbol{w}|\boldsymbol{D})$ and $q(\alpha|\boldsymbol{D})$ can be computed for a univariate linear regression model.

This procedure gives rise to update formulas for the weights, weight precision and noise precision matrix. These formulas, respectively, relate to parameters of the normal, Gamma and Wishart distributions which, for completeness, are defined in the Appendix.

3.1 Updating the weights

If we let

$$I(\boldsymbol{w}) = \int\int q(\boldsymbol{\Lambda}\,|\,\boldsymbol{D})q(\alpha\,|\,\boldsymbol{D})\log[p(\boldsymbol{D}\,|\,\boldsymbol{w},\boldsymbol{\Lambda})p(\boldsymbol{w}\,|\,\alpha)]d\alpha d\boldsymbol{\Lambda} \tag{24}$$

then by substituting (11) and (23) into (19) and dropping those terms which are not a function of $\boldsymbol{w}$, it is possible to write

$$F(p) = -KL(q(\boldsymbol{w}\,|\,\boldsymbol{D}), \exp[\boldsymbol{I}(\boldsymbol{w})]) \tag{25}$$

The negative free energy is therefore maximised when

$$q(\boldsymbol{w}\,|\,\boldsymbol{D}) = \exp[I(\boldsymbol{w})] \tag{26}$$

Substituting in the likelihood and priors gives

$$\begin{aligned} I(\boldsymbol{w}) &= -\int q(\boldsymbol{\Lambda}\,|\,\boldsymbol{D})Tr(\boldsymbol{\Lambda}\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{w}))d\boldsymbol{\Lambda} - \int q(\alpha\,|\,\boldsymbol{D})\alpha E(\boldsymbol{w})d\alpha \\ &= -\frac{1}{2}Tr(\hat{\boldsymbol{\Lambda}}\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{w})) - \hat{\alpha}E(\boldsymbol{w}) \end{aligned} \tag{27}$$

where $\hat{\alpha}$ and $\hat{\boldsymbol{\Lambda}}$ are the mean weight and noise precisions from the approximating densities (see following two Sections). The weight posterior is therefore a normal density $q(\boldsymbol{w}\,|\,\boldsymbol{D}) = \mathrm{N}(\boldsymbol{w};\, \hat{\boldsymbol{w}},\, \hat{\boldsymbol{\Sigma}})$ where

$$\begin{aligned} \boldsymbol{\Lambda}_D &= \hat{\boldsymbol{\Lambda}} \otimes (\boldsymbol{X}^T\boldsymbol{X}) \\ \hat{\boldsymbol{\Sigma}} &= (\boldsymbol{\Lambda}_D + \hat{\alpha}\boldsymbol{I})^{-1} \\ \hat{\boldsymbol{w}} &= \hat{\boldsymbol{\Sigma}}\boldsymbol{\Lambda}_D\boldsymbol{w}_{ML} \end{aligned} \tag{28}$$

where $\boldsymbol{I}$ denotes the identity matrix and $\boldsymbol{w}_{ML}$ denotes (the *vec* form of) the maximum likelihood MAR solution. Thus, the posterior precision matrix $\hat{\boldsymbol{\Sigma}}^{-1}$ takes the usual Bayesian form of being the sum of the data precision $\boldsymbol{\Lambda}_D$ plus the prior precision $\hat{\alpha}\boldsymbol{I}$. With $\hat{\alpha} = 0$, i.e. no prior on the weights, we recover the maximum likelihood MAR solution. The matrix $\hat{\boldsymbol{W}}$ is formed by 'unstacking' the vector $\hat{\boldsymbol{w}}$ to form a $(p \times d)$-by-d matrix.

3.2 Updating weight precisions

If we let

$$I(\alpha) = \int q(\boldsymbol{w}\,|\,\boldsymbol{D})\log[p(\boldsymbol{w}\,|\,\alpha)p(\alpha)]d\boldsymbol{w} \tag{29}$$

then by substituting (11) and (23) into (19) and dropping those terms which are not a function of α, it is possible to write

$$F(p) = -KL(q(\alpha\,|\,\boldsymbol{D}), \exp[\boldsymbol{I}(\alpha)]) \tag{30}$$

The negative free energy is therefore maximised when

$$q(\alpha\,|\,\boldsymbol{D}) = \exp[\boldsymbol{I}(\alpha)] \tag{31}$$

By substituting the weight and weight precision priors we then see that the weight precision posterior is a Gamma density $q(\alpha\,|\,\boldsymbol{D}) = \mathrm{Ga}(\alpha;\, b'_\alpha,\, c'_\alpha)$ where

$$\begin{aligned} 1/b'_\alpha &= E(\hat{\boldsymbol{w}}) + \frac{1}{2}Tr(\hat{\boldsymbol{\Sigma}}) + \frac{1}{b_\alpha} \\ c'_\alpha &= \frac{k}{2} + c_\alpha \\ \hat{\alpha} &= b'_\alpha c'_\alpha \end{aligned} \tag{32}$$

These update equations are the same as for the univariate AR case [8].

3.3 Updating noise precision matrix

If we let

$$I(\boldsymbol{\Lambda}) = \int q(\boldsymbol{w}\,|\,\boldsymbol{D})\log[p(\boldsymbol{D}\,|\,\boldsymbol{w},\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})]d\boldsymbol{w} \tag{33}$$

then by substituting (11) and (23) into (19) and dropping those terms which are not a function of $\boldsymbol{\Lambda}$, it is possible to write

$$F(p) = -KL(q(\boldsymbol{\Lambda}\,|\,\boldsymbol{D}), \exp[\boldsymbol{I}(\boldsymbol{\Lambda})]) \tag{34}$$

The negative free energy is therefore maximised when

$$q(\boldsymbol{\Lambda}\,|\,\boldsymbol{D}) = \exp[I(\boldsymbol{\Lambda})] \tag{35}$$

By substituting the likelihood from (4) we get

$$\begin{aligned} I(\boldsymbol{\Lambda}) &= \int q(\boldsymbol{w}\,|\,\boldsymbol{D})\left[\frac{N}{2}\log|\boldsymbol{\Lambda}| - Tr(\boldsymbol{\Lambda}\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{W}))\right]d\boldsymbol{w} + \log p(\boldsymbol{\Lambda}) \\ &= \frac{N-d-1}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}Tr(\boldsymbol{\Lambda}[\boldsymbol{E}_{\boldsymbol{D}}(\hat{\boldsymbol{W}}) + \boldsymbol{\Omega}]) \end{aligned} \tag{36}$$

where

$$\boldsymbol{\Omega} = \sum_n (\boldsymbol{I}_d \otimes \boldsymbol{x}_n)\hat{\boldsymbol{\Sigma}}(\boldsymbol{I}_d \otimes \boldsymbol{x}_n)^T \tag{37}$$

The noise precision posterior is therefore a Wishart density $q(\boldsymbol{\Lambda}) = \mathrm{Wi}(\boldsymbol{\Lambda};\, a,\, \boldsymbol{B})$ where

$$\boldsymbol{B} = \boldsymbol{E}_{\boldsymbol{D}}(\hat{\boldsymbol{w}}) + \boldsymbol{\Omega} \qquad a = N \qquad \hat{\boldsymbol{\Lambda}} = a\boldsymbol{B}^{-1} \tag{38}$$

4 Model order selection

Because the negative free energy is an approximation to the model evidence we can use it for model order selection. If we write the KL-divergence between a generic approximate posterior and prior as

$$KL_q(x) \equiv KL(q(x\,|\,\boldsymbol{D}), p(x)) \tag{39}$$

then the negative free energy can be computed from (19) as

$$F(p) = L_{av} - KL_q(\boldsymbol{w}) - KL_q(\alpha) - KL_q(\boldsymbol{\Lambda}) \tag{40}$$

The average log-likelihood is given by

$$\begin{aligned} L_{av} &= \int q(\boldsymbol{\theta}\,|\,\boldsymbol{D})\log p(\boldsymbol{D}\,|\,\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= -\frac{dN}{2}\log 2\pi + \frac{N}{2}\int q(\boldsymbol{\Lambda}\,|\,\boldsymbol{D})\log|\boldsymbol{\Lambda}|d\boldsymbol{\Lambda} \\ &\quad -\frac{1}{2}\int q(\boldsymbol{\Lambda}\,|\,\boldsymbol{D})q(\boldsymbol{w}\,|\,\boldsymbol{D})Tr(\hat{\boldsymbol{\Lambda}}\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{w}))d\boldsymbol{\Lambda}d\boldsymbol{w} \\ &= -\frac{dN}{2}\log 2\pi + \frac{N}{2}L(a,\boldsymbol{B}) - \frac{1}{2}Tr(\hat{\boldsymbol{\Lambda}}[\boldsymbol{E}_{\boldsymbol{D}}(\hat{\boldsymbol{w}}) + \boldsymbol{\Omega}]) \\ &= -\frac{dN}{2}\log 2\pi e + \frac{N}{2}L(a,\boldsymbol{B}) \end{aligned} \tag{41}$$

where $L(a,\, \boldsymbol{B})$ is defined in (61). By substituting the expression into (40) and noting that many terms cancel, we get

$$F(p) = -\frac{N}{2}\log|\boldsymbol{B}| - KL_q(\boldsymbol{w}) - KL_q(\alpha) + \log\Gamma_d(N/2) \tag{42}$$

where Γ_d is the generalised gamma function (see the Section on the Wishart density in the Appendix). This last term is constant for a given N and d and so plays no part in model order selection.

4.1 Comparison with MDL

The VB framework maximises the negative free energy given in (20), where the first term is the average likelihood of the data and the second term is the KL-divergence between the approximating posteriors and the priors. This second term acts as a penalty which penalises more complex models.

As the number of samples increases the parameter posterior becomes sharply peaked about the most probable values (which are also the maximum likelihood values) $\hat{\boldsymbol{\theta}}$. It can then be shown that in the large sample limit $N \to \infty$, $F(p)$ becomes equivalent to the Bayesian information criterion [15, 16]

$$BIC(p) = \log p(\boldsymbol{D} \mid \hat{\boldsymbol{\theta}}) - \frac{k}{2} \log N \tag{43}$$

which is itself equal to the negative of the minimum description length (MDL) i.e. $BIC(p) = -MDL(p)$. These popular model order selection criteria can therefore be seen as a limiting case of the VB framework. For the MAR model we have

$$BIC(p) = -\frac{N}{2} \log |\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{W}_{ML})| - \frac{k}{2} \log N \tag{44}$$

This is identical to the expression on p.12 of [1] referred to as Schwarz's Bayesian criterion (SBC) (after dividing by $-Nd/2$ and dropping terms not dependent on model order). By comparsion with (42) we see that the first term, the 'accuracy' term, is more or less the same as in VB; for BIC we have the determinant of the error covariance matrix, and for VB we have the determinant of the error covariance matrix averaged over the posterior weight density. The major difference is in the second term, which for VB is the sum of the weight and weight precision KL-divergences.

If we assume *a priori* that different model orders are equally likely (within some range) then a probability distribution over model order can be obtained from [15]:

$$q(p) = \frac{\exp(F(p))}{\sum_{p'} \exp(F(p'))} \tag{45}$$

This can also be applied to the BIC/MDL criterion.

5 Structured priors

Instead of using the isotropic Gaussian prior of (7), where every coefficient has the same prior variance, we can split the coefficients into groups and allow different groups to have different prior variances. This type of prior is known in the neural network field as automatic relevance determination (ARD) [17], so-called because by inspecting the inferred prior variances we can see which groups of parameters are relevant to the problem at hand.

For MAR models, as we shall see, different coefficients will lie naturally on different scales, and this property can be captured with 'structured priors'. For example, if the parameters are split into groups according to which time lag they are associated with (we call this the 'lag prior') then parameters associated with different lags will be of a different magnitude; generally, the magnitude will decrease with the lag until, for irrelevant lags, the magnitude should be zero. Structured priors are assumed to be of the form

$$p(\boldsymbol{w} \mid \{\alpha_j\}) = \prod_{j=1}^{G} \left(\frac{\alpha_j}{2\pi}\right)^{k_j/2} \exp(-\alpha_j E_j(\boldsymbol{w})) \tag{46}$$

where the weights have been split into $j = 1, \ldots, G$ groups with k_j weights in the jth group and where

$$E_j(\boldsymbol{w}) = \frac{1}{2} \boldsymbol{w}^T \boldsymbol{I}_j \boldsymbol{w} \tag{47}$$

and $\boldsymbol{I}_j$ is a matrix with ones along the diagonal that pick off coefficients in the jth group, and zeros elsewhere. Use of structured priors results in VB updates for the posterior weight covariance and weight precision as follows:

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= \left(\boldsymbol{\Lambda}_D + \sum_{j=1}^{G} \hat{\alpha}_j \boldsymbol{I}_j\right)^{-1} \\ 1/b'_\alpha(j) &= E_j(\hat{\boldsymbol{w}}) + \frac{1}{2} Tr(\boldsymbol{I}_j \hat{\boldsymbol{\Sigma}} \boldsymbol{I}_j) + \frac{1}{b_\alpha} \\ c'_\alpha(j) &= \frac{k_j}{2} + c_\alpha \\ \hat{\alpha}(j) &= b'_\alpha(j) c'_\alpha(j) \end{aligned} \tag{48}$$

The other updates are exactly the same as for the global prior. The only difference occurs when estimating $F(p)$; we replace $KL_q(\alpha)$ by $\sum_j KL_q(\alpha_j)$.

For MAR models we envisage many types of prior other than the lag prior described. This is because there are many natural and physically meaningful ways of grouping the MAR coefficients. For example, we could use an 'interaction prior' in which coefficients are split into two groups: weights involved in within-series prediction, and weights involved in between-series predictions, i.e. interactions. Because we can estimate the evidence for each model (i.e. a model using an interaction prior versus a model using a global prior) the data can tell us which model is more appropriate. This allows for statistical hypothesis testing in the usual manner.

Refinement of the schemes allows us to test for many different types of interactions. For example, we could have separate groups for interaction and noninteraction at each time lag; this gives rise to the 'lag-interaction prior'.

Similarly, we could group interactions between subsets of the time series together e.g. for EEG data we could split the interactions into groups for within and between different cortical areas.

6 Other issues

We use uninformative priors [18] by setting $b = 10^3$ and $c = 10^{-3}$ for the Gamma prior on α (see the Appendix for a definition of the Gamma density). The parameters of the VB posterior are initialised using the maximum likelihood solution; $\hat{\boldsymbol{w}} = \boldsymbol{w}_{ML}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{ML}$. The VB equations can then be applied iteratively until a consistent solution is reached. Convergence can be measured by evaluating the negative free energy. We calculate $F(p)$ at each iteration and terminate optimisation if the proportional increase in $F(p)$ from one iteration to the next is less than 0.01%.

Evaluation of the KL divergence between the weight posterior and prior requires computation of $\log |\boldsymbol{\Sigma}|$ which we implement using $\sum_{i=1}^{k} \log \lambda_i$ where λ_i are the eigenvalues of $\boldsymbol{\Sigma}$. This avoids the possibility of numerical problems when the number of weights k is large. When k is very large the computation becomes a bottleneck in the VB algorithm. To overcome this we extract just the first few 'significant' eigenvalues using, for example, the 'snapshot' method [19].

By analogy with the Bayesian evidence framework [8] we define a quantity γ, known as the effective degrees of freedom, as follows:

$$\gamma = k - \sum_{j=1}^{G} \alpha_j Tr(\boldsymbol{I}_j \boldsymbol{\Sigma} \boldsymbol{I}_j) \tag{49}$$

The rationale behind this is that parameter groups with posterior precision equal to their prior precision will have

$$\alpha_j Tr(\boldsymbol{I}_j \boldsymbol{\Sigma} \boldsymbol{I}_j) = k_j \tag{50}$$

That is, their values are not determined by the data. Consequently γ is the number of remaining data-determined parameters.

When evaluating the model-order criterion for a number of hypothesised model orders use could be made of the 'downdating' procedure described in [1] where the coefficients at order $p-1$ are computed from those estimated at model order p. This would save a good deal of computation, especially for large models.

If we were interested in making inferences about the values of the MAR coefficients themselves, by noting that the marginal distribution over the MAR coefficients is a multivariate t-distribution ([10], p. 440) we can compute confidence intervals in a manner similar to that described in [1].

Finally, we note that the power spectral density matrix can be computed from the MAR coefficients, as described in ([4], p. 408). From this matrix one can compute the power spectra, cross-spectra, coherences and partial coherences.

7 Results

7.1 Model order selection

To illustrate the VB-MAR approach we apply it to the MAR(2) process described by Neumaier and Schneider [1]

$$\boldsymbol{y}_n = \sum_{i=1}^{p} \boldsymbol{y}_{n-i} \boldsymbol{a}(i) + \boldsymbol{e}_n \tag{51}$$

where

$$\boldsymbol{A}(1) = \begin{pmatrix} 0.40 & 0.30 \\ 1.20 & 0.70 \end{pmatrix} \tag{52}$$

$$\boldsymbol{A}(2) = \begin{pmatrix} 0.35 & -0.40 \\ -0.30 & -0.50 \end{pmatrix} \tag{53}$$

and the noise precision matrix is

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1.00 & 0.50 \\ 0.50 & 1.50 \end{pmatrix}^{-1} \tag{54}$$

We generated 50 data sets, each containing $N = 200$ values. Typically, five cycles of the VB update equations were required to reach convergence. Fig. 1*a* shows a plot of the VB model order criterion as applied to this data, showing a clear peak at the correct model order $p = 2$. Over the 50 data sets, the VB criterion chose the correct model order 100% of the time, as did the BIC criterion. We then generated 50 data sets from MAR(3), MAR(4) and MAR(5) models (coefficients not shown) each containing 200 data samples. Over the 50 data sets the VB criterion chose the correct model order 96, 84 and 8% of the time, with BIC getting it right 86, 76 and 2% of the time, respectively. This shows that the VB criterion is superior to BIC, a trend which has been demonstrated more extensively for univariate AR models [8]. Fig. 1*b* shows a plot of the VB model-order criterion for the MAR(4) models, showing a clear peak at the correct model order $p = 4$. The reason why it is harder to select the correct model order as we increase p is that we have held the number of data points constant; the ratio of data points to coefficients has therefore reduced. This has the effect of making the model order curves less peaky at higher p.

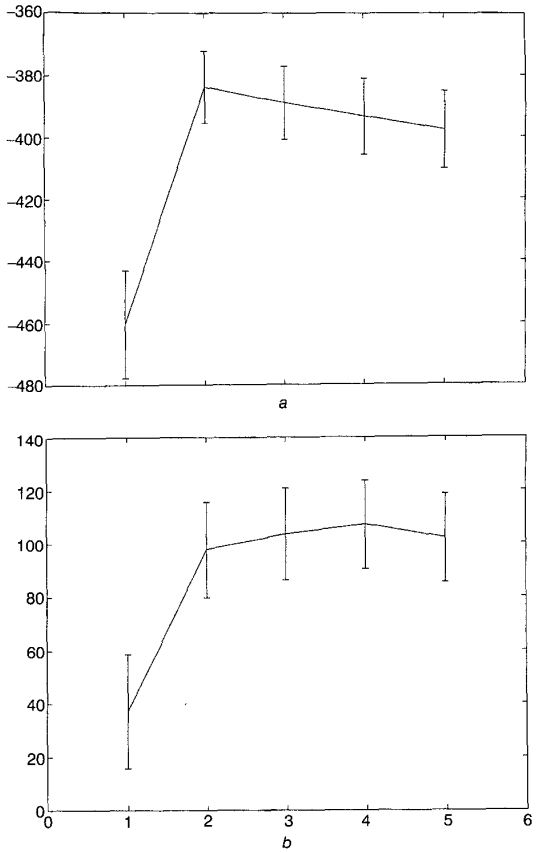

Fig. 1 *Plot of negative free energy F(p) against model order p on data generated from a MAR(p) model*

Average value calculated over 50 trials with error bars at ±1 standard deviation

a $p = 2$
b $p = 4$

7.2 Structured priors

We generated 50 data points from the MAR(2) model described and then applied a MAR(4) model, firstly using a global prior and secondly using a lag prior. The resulting coefficient estimates are shown in Fig. 2 in the form of a 'Hinton diagram' [20]. Use of the lag-prior reduces the magnitude of coefficients at lags 3 and 4 (the true value of these coefficients is zero). This effect becomes more dramatic as the number of time series d is increased, because each weight group contains d^2 parameters.

In the second example, the model was applied to three seconds worth of data from five sinusoidally varying time series, each sampled at 128 Hz. The time series were generated independently, thus there is no interaction in the model so all the off-diagonal MAR coefficients should be zero. Fig. 3 shows that use of the interaction prior clearly reduces the magnitude of the spurious coefficients.

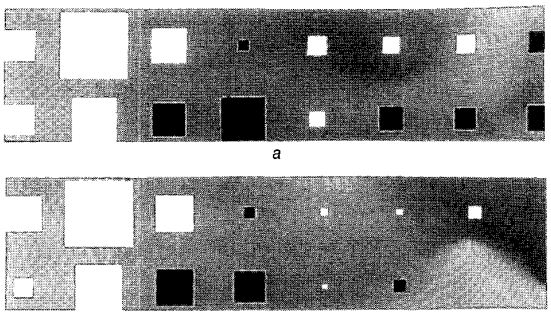

Fig. 2 *Lag-prior example: estimation of MAR(4) model coefficients applied to $d=2$ time series*

White indicates positive, black negative, and area is proportional to magnitude. At each time lag there is a d-by-d matrix of coefficients; we therefore have four 2-by-2 matrices with time-lag increasing from left to right. This model was applied to data generated from MAR(2) process. True values of the eight coefficients on the right of each Figure are zero. Use of the lag-prior reduces the magnitude of these irrelevant coefficients

a Global prior
b Lag prior

The inferred standard deviation of weights in each group $\sqrt{1/\alpha_j}$ was 0.27 for the within-series group and 0.03 for the interaction group. Moreover, the model using the interaction prior has higher evidence (as indicated by the negative free energy): -1940 as opposed to -1949. The hypothesis that the interaction weights are of a different magnitude to the non-interaction weights is therefore accepted with probability 0.9999 ($p=\exp(-1940)/[\exp(-1940)+\exp(-1949)]$; from (45)).

Inferences of this sort become harder to make when we have fewer data points. For example, if we repeat the procedure, but with two seconds of data and then only one second of data the hypothesis is accepted with probabilities of 0.95 and 0.89, respectively. Finally, for half a second of data we have $p=0.006$ and the hypothesis is rejected.

7.3 Cognitive-EEG data

We analysed six-channel EEG data recorded while a subject performed different cognitive tasks. The data is derived from a large database collected from four subjects performing five different tasks (this data is available from http://www.colostate.cs.edu/~anderson), although in this paper we focus on just two of the tasks: (i) a baseline task and (ii) a maths task (for which subjects were given non-trivial multiplication problems such as 49×78). The EEG records for a single task performance are shown in Fig. 4. To ensure signal stationarity we analyse one-second subsections of data. The following results were derived from a single subject (subject 2) performing the two tasks, unless otherwise stated.

Fig. 5 shows that for a global prior the optimal model order averaged over one-second sections of data is two. With the interaction prior, this optimum increases to four. For the model with the global prior the effective degrees of freedom (see (49)) was $\gamma=68\pm 1$ for the $p=2$ model and $\gamma=125\pm 1$ for the $p=4$ model. For the model with the interaction prior and $p=4$ we had $\gamma=56\pm 4$. Thus by using a structured prior we can constrain the effective degrees of freedom in the model and so extract information from longer time lags. These results were obtained by averaging over ten one-second blocks of EEG data from the baseline task.

We then looked at the evidence of models with a fixed model order but using different priors. The results are given in Table 1. They show that the interaction prior is most suitable for this data which implies that the coefficients naturally split into groups of two different magnitudes; one for the within-series coefficients and another for the interactions. This result also held when we repeated the analysis over the same subject performing all five different cognitive tasks.

One reason for the success of this prior is its simplicity; only two weight groups are hypothesised whereas the other priors have more groups. This is important as the negative free energy involves a penalty term which is the sum of the KL-divergences of the precisions over all the groups (see Section 5), which naturally becomes larger with more groups. But to show that simplicity was not the sole reason for the success of this prior we created a two-group prior where weights were randomly assigned to each group (see 'Random' in Table 1). This prior performed considerably worse than all the rest, indicating the importance of the diagonal structure.

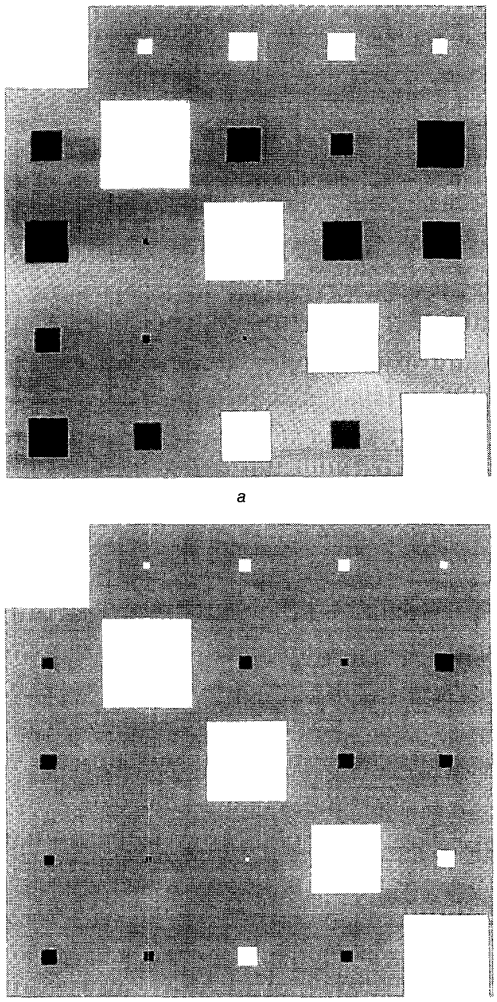

Fig. 3 *Interaction-prior example: estimation of matrix of MAR coefficients at lag 1, when model is applied to data consisting of five independent, sinusoidally varying time series*

White indicates positive, black negative, and area is proportional to magnitude. Use of the interaction prior reduces the magnitude of the off-diagonal coefficients (which, ideally, should be zero)

a Global prior
b Interaction prior

Table 1: Evidence of MAR models with different priors: cognitive data

Prior	Evidence
Interaction	31 ± 2
Lag-interaction	26 ± 2
Lag	-6 ± 4
Global	-23 ± 3
Random	-28 ± 3

Evidence, as estimated by the negative free energy F is given for (i) interaction-prior, which splits coefficients into two groups; one for interactions and one for within-series terms, (ii) interaction prior at each time lag, (iii) lag-prior, (iv) global prior and (v) 'random' prior where coefficients are randomly split into two groups. These were all calculated for a time lag of $p=2$ and were averaged over ten one-second sections of data; for each section we subtracted the mean value of F so that results from all sections could be meaningfully pooled

The success of the interaction prior, which typically resulted in the within-series weights being 15 to 20 times bigger than the between-series weights (e.g. Fig. 6), means that there is greater temporal information within each channel than between channels. This results in small values for the partial coherences between channels, a finding which is confirmed by results in [5]. This is not to be confused with the purely spatial interaction between channels, however, as indicated by the high degree of (instantaneous) correlation between the time series (Fig. 4). This is picked up in the MAR models by the inference of highly non-diagonal noise covariance matrices. Finally, although the spatiotemporal interactions are small, they are not insignificant. To emphasise this we present results on using MAR models to discriminate between the EEG recorded during the baseline and the maths tasks for three of the subjects. These results were computed using linear discriminant analysis applied to subsets of coefficients which were identified using a stepwise forwards selection procedure [21]. Error rates were computed using a five-fold cross-validation process. The results are shown in Table 2. They show that the off-diagonal coefficients are discriminative in themselves and that when combined with the diagonal terms we can get an overall improvement in discrimination accuracy.

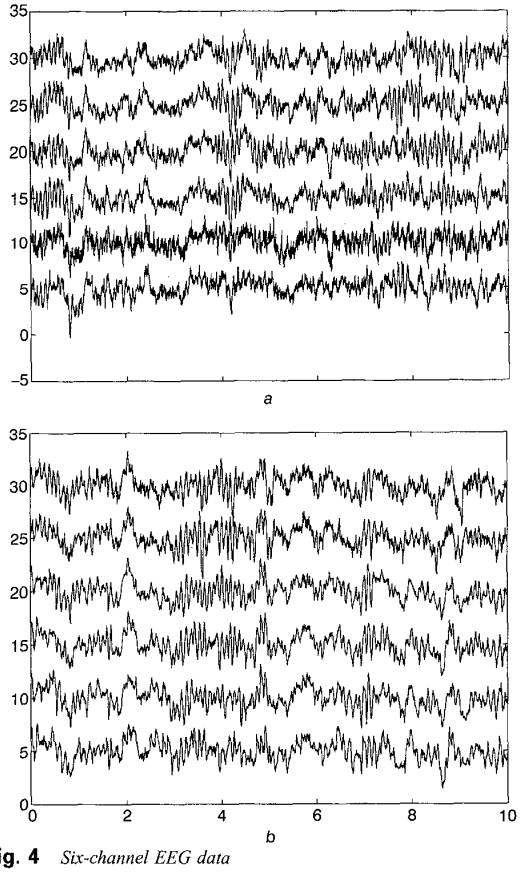

Fig. 4 *Six-channel EEG data*

Channels are recorded from one electrode on the right and one on the left of central (C), parietal (P) and occipital (O) cortex in the following order (from bottom to top in each part of the Figure): C-left, C-right, P-left, P-right, O-left and O-right. x-axis shows elapsed seconds

a Baseline task
b Maths task

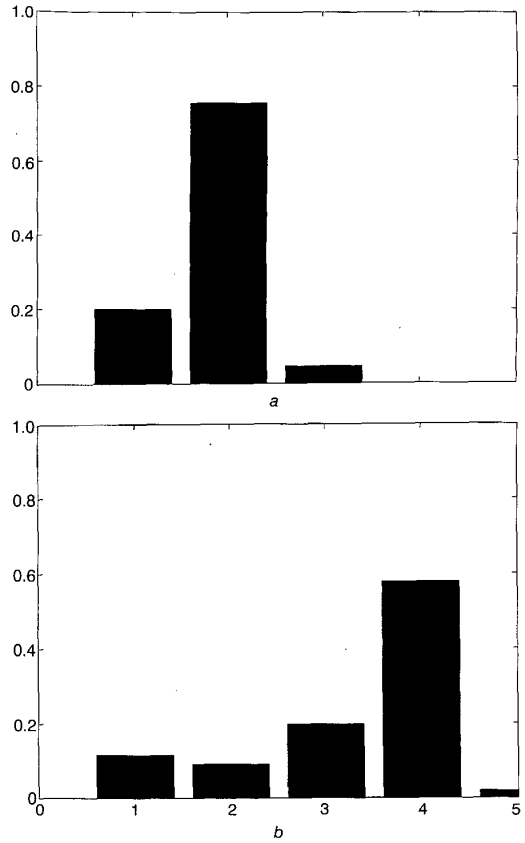

Fig. 5 *EEG model-order selection: plot of P(p), probability of model order p*

Probabilities calculated from (45) and averaged over ten one-second sections of data. By disregarding spurious spatial interactions the MAR model using the interaction prior is able to extract more temporal information

a Global prior
b Interaction prior

7.4 Sleep-EEG data

We analysed data from a sleep-EEG database where the EEG was recorded from six-channels over left and right central, parietal and occipital cortex. We focused on ten one-second sections of data recorded while the subject was awake or in sleep-stage 4.

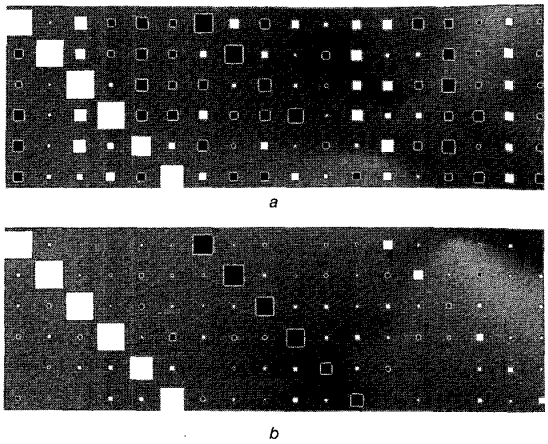

Fig. 6 *Hinton diagram of MAR(3) model fitted to EEG data recorded during maths task*

At each of three time lags (left to right) we have a 6 × 6 coefficient matrix. The interaction prior reduces the magnitude of the off-diagonal terms

a Global prior
b Interaction prior

Table 2: Classification of EEG records using MAR features

	Error rates, %		
Data	Diagonal	Off-diagonal	All
Subject 1	7.9 + 5.3	3.6 + 2.5	4.3 ± 3.0
Subject 2	2.1 ± 3.1	0.7 ± 1.6	0.0 ± 0.0
Subject 3	9.3 ± 3.2	25.0 ± 6.7	8.6 ± 5.4

Error rates (%) for using MAR features to discriminate between EEG data recorded during the baseline task against maths task, for using just the diagonal, off-diagonal or all of MAR coefficients

For the awake data we fitted MAR(4) models with different priors. The evidence of the resulting models is given in Table 3 which shows that the lag and the interaction priors are best supported (the difference between them is not significant). For the global prior, the optimal model order was four, whereas use of the interaction prior increased this to seven. The effective degrees of freedom for the global prior models was $\gamma = 117 \pm 5$ for $p = 4$ and $\gamma = 181 \pm 12$ for $p = 7$. For the interaction prior we had $\gamma = 145 \pm 8$ for $p = 7$. Thus by reducing the effective degrees of freedom using a structured prior we can extract more temporal information.

Table 3: Evidence of MAR models with different priors: sleep data

Prior	Evidence
Interaction	8.0 ± 5.0
Lag	6.2 ± 4.0
Lag-interaction	−1.6 ± 3.4
Global	−12.8 ± 5.0

Evidence, as estimated by negative free energy F is given for (i) interaction-prior, which splits coefficients into two groups: one for interactions and one for within-series terms, (ii) interaction prior at each time lag, (iii) lag-prior and (iv) global prior. These were all calculated for a time lag of $p = 4$ and were averaged over ten one-second sections of data; for each section we subtracted the mean value of F so that results from all sections could be meaningfully pooled

For the sleep-stage 4 data the optimal model orders decreased to three for the global prior and to six for the interaction prior. This illustrates a known trend for sleep EEG data: that optimal model order decreases with increasing sleep stage [23].

8 Discussion

We have shown how to apply the variational Bayesian framework to MAR models. By using 'structured priors' in which subsets of coefficients are constrained to be of a similar magnitude, the effective number of degrees of freedom in the model can be constrained. This allows MAR models to be more readily applied to high-dimensional data, or to data with greater temporal complexity.

Also, VB provides a model-order selection criterion which can be used to select the appropriate number of time lags. Our experiments have shown this to be superior to the MDL criterion. In earlier work [8] we have also shown this to be the case for univariate AR models.

We have also shown that the negative variational free energy (which approximates the model evidence) can be used to choose an appropriate prior. While choosing your prior after seeing the data may at first seem nonsensical, in fact, this is not the case. The range of priors available, in effect, enriches the MAR model class, and the choice of appropriate prior reduces to a model selection problem; not of model order as is usually the case but of model structure.

9 References

1 NEUMAIER, A., and SCHNEIDER, T.: 'Estimation of parameters and eigenmodes of multivariate autoregressive models', *ACM Trans. Math. Softw.*, 2001, **27**, pp. 27–57
2 LUTKEPOHL, H.: 'Introduction to multiple time series analysis' (Springer Verlag, 1993)
3 BRESSLER, S.L., DING, M., and YANG, W.: 'Investigation of cooperative cortical dynamics by multivariate autoregressive modeling of event-related local field potentials', *Neurocomputing*, 1999, **26–27**, pp. 625–631
4 MARPLE, S.L.: 'Digital spectral analysis with applications' (Prentice-Hall, 1987)
5 KAMINSKI, M., BLINOWSKA, K., and SZELENBERGER, W.: 'Topographic analysis of coherence and propagation of EEG activity during sleep and wakefulness', *Electroencephalogr. Clinical Neurophysiol.*, 1997, **102**, pp. 216–227
6 BISHOP, C.M.: 'Variational principal components'. Proceedings of international conference on *Artificial neural networks*, 1999, pp. 509–514
7 CHOUDREY, R., PENNY, W.D., and ROBERTS, S.J.: 'An ensemble learning approach to independent component analysis'. Presented at the IEEE international workshop on *Neural networks for signal processing*, Sydney, Australia, 2000
8 PENNY, W.D., and ROBERTS, S.J.: 'Bayesian methods for autoregressive models'. Presented at the international workshop on *Neural networks for signal processing*, Sydney, Australia, 2000
9 PENNY, W.D., and ROBERTS, S.J.: 'Variational Bayes for non-Gaussian autoregressive models'. Presented at the IEEE international workshop on *Neural networks for signal processing*, Sydney, Australia, 2000
10 BOX, G.E.P., and TIAO, G.C.: 'Bayesian inference in statistical analysis' (Wiley, 1992)
11 MUIRHEAD, R.J.: Aspects of multivariate statistical theory' (Wiley, 1982)
12 WEISBERG, S.: 'Applied linear regression' (Wiley, 1980)
13 MAGNUS, J.R., and NEUDECKER, H.: 'Matrix differential calculus with applications in statistics and econometrics' (Wiley, 1997)
14 MACKAY, D.J.C.: 'Ensemble learning and evidence maximisation'. Technical report, Cavendish Laboratory, University of Cambridge, 1995
15 ATTIAS, H.: 'A variational Bayesian framework for graphical models' *in* LEEN, T., *et al.* (Eds.): 'Proceedings of NIPS 12' (MIT Press, Cambridge, MA, 2000)
16 CHICKERING, D.M., and HECKERMAN, D.: 'Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables'. Technical report MSR-TR-96-08, Microsoft Research, 1996

17 MACKAY, D.J.C.: 'Bayesian nonlinear modelling for the energy prediction competition', *ASHRAE Trans.*, 1994, **100**, pp. 1053–1062
18 O'RUANIAIDH, J.J.K., and FITZGERALD, W.J.: 'Numerical Bayesian methods applied to signal processing' (Springer, 1996)
19 SIROVICH, L.: 'Turbulence and the dynamics of coherent structures', *Q. Appl. Math.*, 1987, **45**, (3), pp. 561–590
20 BISHOP, C.M.: 'Neural networks for pattern recognition' (Oxford University Press, Oxford, 1995)
21 KLEINBAUM, D.G., KUPPER, L.L., and MULLER, K.E.: 'Applied regression analysis and other multivariable methods' (PWS-Kent, Boston, 1988)
22 RIGNEY, D.R., GOLDBERGER, A.L., OCASIO, W.C., ICHIMARU, Y., MOODY, G.B., and MARK, R.G.: 'Multichannel physiological data: description and analysis (data set B)' *in* WEIGEND, A.S., and GERSHENFELD, N.A. (Eds.): 'Time series prediction: forecasting the future and understanding the past' (Addison-Wesley, 1994), pp. 105–129
23 PARDEY, J., ROBERTS, S., and TARASSENKO, L.: 'A review of parametric modelling techniques for EEG analysis', *Med. Eng. Phys.*, 1996, **18**, (1), pp. 2–11
24 PENNY, W.D., and ROBERTS, S.J.: 'Variational Bayes for 1-dimensional mixture models'. Technical report PARG-2000-01, Department of Engineering Science, Oxford University, 2000
25 PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T., and FLANNERY, B.V.P.: 'Numerical recipes in C' (Cambridge, 1992)

10 Densities and divergences

10.1 Normal density

The multivariate normal density is given by

$$\mathrm{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \times \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \quad (55)$$

The KL divergence for normal densities $q(\boldsymbol{x}) = \mathrm{N}(\boldsymbol{x};\ \boldsymbol{\mu}_q,\ \boldsymbol{\Sigma}_q)$ and $p(\boldsymbol{x}) = \mathrm{N}(x;\ \boldsymbol{\mu}_p,\ \boldsymbol{\Sigma}_p)$ is

$$KL(q,p) = 0.5\log\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} + 0.5Tr(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) + 0.5(\boldsymbol{\mu}_q-\boldsymbol{\mu}_p)^T\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_q-\boldsymbol{\mu}_p) - \frac{d}{2} \quad (56)$$

where $|\boldsymbol{\Sigma}_p|$ denotes the determinant of the matrix $\boldsymbol{\Sigma}_p$.

10.2 Gamma density

The Gamma density is given by

$$\mathrm{Ga}(x;b,c) = \frac{1}{\Gamma(c)}\frac{x^{c-1}}{b^c}\exp\left(\frac{-x}{b}\right) \quad (57)$$

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

$$KL(q,p) = (c_q-1)\Psi(c_q) - \log b_q - c_q - \log\Gamma(c_q) + \log\Gamma(c_p) + c_p\log b_p - (c_p-1)(\Psi(c_q)+\log b_q) + \frac{b_q c_q}{b_p} \quad (58)$$

where $\Psi(\ )$ is the digamma function [25].

10.3 Wishart density

The Wishart distribution is given by [11, p. 85]

$$\mathrm{Wi}(\boldsymbol{\Lambda};a,\boldsymbol{B}) = \frac{1}{Z(a,\boldsymbol{B})}|\boldsymbol{\Lambda}|^{(a-d-1)/2}\exp\left[-\frac{1}{2}Tr(\boldsymbol{B}\boldsymbol{\Lambda})\right] \quad (59)$$

where

$$Z(a,\boldsymbol{B}) = 2^{ad/2}|\boldsymbol{B}|^{-a/2}\Gamma_d(a/2) \quad (60)$$

and $\Gamma_d(\ )$ is the generalised gamma function defined in ([11], p. 62).

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

$$L(a,\boldsymbol{B}) = \int \mathrm{Wi}(\boldsymbol{\Lambda};a,\boldsymbol{B})\log|\boldsymbol{\Lambda}|d\boldsymbol{\Lambda} \quad (61)$$

The entropy of $q(\boldsymbol{\Lambda}) = \mathrm{Wi}(\boldsymbol{\Lambda};\ q,\ \boldsymbol{Q})$ is then given by

$$H(q) = -\left(\frac{q-d-1}{2}\right)L(q,\boldsymbol{Q}) + \frac{qd}{2} + \log Z(q,\boldsymbol{Q}) \quad (62)$$

The KL divergence between densities $q(\boldsymbol{\Lambda}) = \mathrm{Wi}(\boldsymbol{\Lambda};\ q,\ \boldsymbol{Q})$ and $p(\boldsymbol{\Lambda}) = \mathrm{Wi}(\boldsymbol{\Lambda};\ p,\ \boldsymbol{P})$ is given by

$$KL(q,p) = \left(\frac{q-d-1}{2}\right)L(q,\boldsymbol{Q}) - \left(\frac{p-d-1}{2}\right)L(p,\boldsymbol{P}) - \frac{qd}{2} + \frac{q}{2}Tr(\boldsymbol{P}\boldsymbol{Q}^{-1}) + \log\frac{Z(p,\boldsymbol{P})}{Z(q,\boldsymbol{Q})} \quad (63)$$