Bayesian Multivariate Autoregressive Models with Structured Priors

W.D. Penny and S.J. Roberts

Department of Engineering Science,
Oxford University.
E-mail: wpenny,sjrob@robots.ox.ac.uk

September 11, 2000

Abstract

We describe a Variational Bayesian (VB) learning algorithm for parameter estimation and model order selection in Multivariate Autoregressive (MAR) models. We explore the use of structured priors in which subsets of coefficients are grouped together and constrained to be of a similar magnitude. This allows MAR models to be more readily applied to high dimensional data and to data with greater temporal complexity. We also compare the VB model order selection criterion with the Minimum Description Length (MDL) approach. Results are presented on synthetic, physiological and electro-encephalogram (EEG) data.

1 Introduction

The Multivariate Autoregressive (MAR) process is used to model multiple time series data in such fields as geophysics [16], economics [10] and biomedicine [5]. It can also be seen as a parametric multivariate spectral estimation procedure and will provide parsimonious estimation of coherences and partial coherences [14] [8]. 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 of order d^2 parameters.

In this paper 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 number of 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.

Whilst 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 [3] and Independent Component Analysis [7]. We have also recently applied VB to univariate Autoregressive (AR) models [19] and univariate Non-Gaussian AR models [21].

In the first section of the paper we describe the multivariate autoregressive process as a probabilistic Bayesian model. We then describe the VB method for the estimation of MAR parameters and compare the VB model order selection criterion to the Minimum Description Length (MDL) approach. We then introduce the methodology behind the use of structured priors and present results on analysing synthetic, physiological and EEG data.

2 Multivariate Autoregressive Models

An MAR(p) model predicts the next value in a d-dimensional time series (ie. 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 $\boldsymbol{y}_n = [y_n(1), y_n(2), ..., y_n(d)]$ is the nth sample of a d-dimensional time series, each $\boldsymbol{A}(i)$ is a d-by-d matrix of coefficients and $\boldsymbol{e}_n = [e_n(1), e_n(2), ..., e_n(d)]$ is additive 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 $\boldsymbol{x}_n = [\boldsymbol{y}_{n-1}, \boldsymbol{y}_{n-2}, ..., \boldsymbol{y}_{n-p}]$ 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 $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{E}$ are $\boldsymbol{x}_n$, $\boldsymbol{y}_n$ and $\boldsymbol{e}_n$ respectively and there are $n = 1..N$ 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; see eg. Box and Tiao (p. 423) [4]. From these sources we know that given a data set $\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{Y}\}$ the likelihood of the data is given by

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

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

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

is the unnormalised error covariance matrix.

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

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

where $vec(\boldsymbol{W})$ denotes the columns of $\boldsymbol{W}$ being stacked on top of each other (for more on the *vec* notation, see [15]). To recover the matrix $\boldsymbol{W}$ we simply 'un-stack' 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

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

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

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

The noise precision matrix is taken to have the prior

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

which is the uninformative prior for multivariate linear regression; see Box and Tiao [4] (page 426) for a 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

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

The log-likelihood of the data is

$$\log p(\boldsymbol{D} \mid \boldsymbol{w}, \boldsymbol{\Lambda}) = -\frac{dN}{2}\log 2\pi + \frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}Tr(\boldsymbol{\Lambda}\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{w})) \tag{12}$$

2.1 Maximum Likelihood

The Maximum Likelihood (ML) solution (see eg. [25]) for the MAR coefficients is

$$\boldsymbol{W}_{ML} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{13}$$

The maximum likelihood noise covariance, S_{ML}, can be estimated as

$$\begin{aligned} \boldsymbol{S}_{ML} &= \frac{1}{N-k}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_{ML})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_{ML}) \\ &= \frac{1}{N-k}\boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{W}_{ML}) \end{aligned} \tag{14}$$

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] (page 321)

$$\boldsymbol{\Sigma}_{ML} = \boldsymbol{S}_{ML} \otimes (\boldsymbol{X}^T \boldsymbol{X})^{-1} \tag{15}$$

where $\otimes$ denotes the Kronecker product, which is defined in the Appendix.

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. In this paper we propose an alternative solution based on the Variational Bayesian framework which we now describe.

3 Variational Bayes

The 'evidence' or 'marginal likelihood' of a probabilstic model is the likelihood of the model after its parameters have been integrated out. The log evidence can be written as

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

Using Jensen's inequality we have

$$\log p(\boldsymbol{D}) \geq F(p) \tag{17}$$

where

$$F(p) = \int q(\boldsymbol{\theta}|\boldsymbol{D}) \log \frac{p(\boldsymbol{D},\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{D})} d\boldsymbol{\theta} \tag{18}$$

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 ie. if $q(\boldsymbol{\theta}|\boldsymbol{D}) = p(\boldsymbol{\theta}|\boldsymbol{D})$. The aim of Variational Bayesian (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

$$F(p) = L_{av} - KL(q||p) \tag{19}$$

where

$$L_{av} = \int q(\boldsymbol{\theta}|\boldsymbol{D}) \log p(\boldsymbol{D}|\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{20}$$

and

$$KL(q,p) = \int q(\boldsymbol{\theta}|\boldsymbol{D}) \log \frac{q(\boldsymbol{\theta}|\boldsymbol{D})}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{21}$$

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 $q(\theta|\boldsymbol{D})$. 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 penalizes more complex models.

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

$$q(\boldsymbol{\theta}|\boldsymbol{D}) = q(\boldsymbol{w}|\boldsymbol{D}) q(\alpha|\boldsymbol{D}) q(\boldsymbol{\Lambda}|\boldsymbol{D}) \tag{22}$$

This essentially arises because the prior (equation 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 [12] 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.

The above procedure gives rise to update formulae for the weights, weight precision and noise precision matrix. These formulae, 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

$$I(\boldsymbol{w}) = \int\int q(\boldsymbol{\Lambda}|\boldsymbol{D})q(\alpha|\boldsymbol{D}) \log\left[p(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{\Lambda})p(\boldsymbol{w}|\alpha)\right] d\alpha d\boldsymbol{\Lambda} \tag{23}$$

then by substituting equations 11 and 22 into equation 18 and dropping those terms which are not a function of $\boldsymbol{w}$, it is possible to write

$$F(p) = -KL(q(\boldsymbol{w}|\boldsymbol{D}), \exp[\boldsymbol{I}(\boldsymbol{w})]) \tag{24}$$

The negative free energy is therefore maximised when

$$q(\boldsymbol{w}|\boldsymbol{D}) = \exp[I(\boldsymbol{w})] \tag{25}$$

Substituting in the likelihood and priors gives

$$\begin{aligned} I(\boldsymbol{w}) &= -\int q(\boldsymbol{\Lambda}|\boldsymbol{D})Tr(\boldsymbol{\Lambda}\boldsymbol{E_D}(\boldsymbol{w}))d\boldsymbol{\Lambda} - \int q(\alpha|\boldsymbol{D})\alpha E(\boldsymbol{w})d\alpha \\ &= -\frac{1}{2}Tr(\hat{\boldsymbol{\Lambda}}\boldsymbol{E_D}(\boldsymbol{w})) - \hat{\alpha}E(\boldsymbol{w}) \end{aligned} \tag{26}$$

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

$$\begin{aligned} \boldsymbol{\Lambda}_D &= \hat{\boldsymbol{\Lambda}} \otimes (\boldsymbol{X}^T\boldsymbol{X}) \\ \hat{\boldsymbol{\Sigma}} &= (\boldsymbol{\Lambda}_D + \hat{\alpha}\boldsymbol{I})^{-1} \\ \hat{\boldsymbol{w}} &= \hat{\boldsymbol{\Sigma}}\boldsymbol{\Lambda}_D\boldsymbol{w}_{ML} \end{aligned} \tag{27}$$

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 $\hat{\alpha} = 0$, ie. no prior on the weights, we recover the Maximum Likelihood MAR solution. The matrix $\hat{\boldsymbol{W}}$ is formed by 'un-stacking' the vector $\hat{\boldsymbol{w}}$ to form a $(p \times d)$-by-d matrix.

3.2 Updating the weight precisions

If we let

$$I(\alpha) = \int q(\boldsymbol{w}|\boldsymbol{D}) \log [p(\boldsymbol{w}|\alpha)p(\alpha)] \, d\boldsymbol{w} \tag{28}$$

then by substituting equations 11 and 22 into equation 18 and dropping those terms which are not a function of α, it is possible to write

$$F(p) = -KL(q(\alpha|\boldsymbol{D}), \exp[\boldsymbol{I}(\alpha)]) \tag{29}$$

The negative free energy is therefore maximised when

$$q(\alpha|\boldsymbol{D}) = \exp[I(\alpha)] \tag{30}$$

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

$$\begin{aligned} 1/b'_\alpha &= E(\hat{\boldsymbol{w}}) + \frac{1}{2} Tr(\hat{\boldsymbol{\Sigma}}) + \frac{1}{b_\alpha} \\ c'_\alpha &= \frac{k}{2} + c_\alpha \\ \hat{\alpha} &= b'_\alpha c'_\alpha \end{aligned} \tag{31}$$

These update equations are the same as for the univariate AR case [19].

3.3 Updating the noise precision matrix

If we let

$$I(\boldsymbol{\Lambda}) = \int q(\boldsymbol{w}|\boldsymbol{D}) \log p(\boldsymbol{D}|\boldsymbol{w}, \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}) d\boldsymbol{w} \tag{32}$$

then by substituting equations 11 and 22 into equation 18 and dropping those terms which are not a function of $\boldsymbol{\Lambda}$, it is possible to write

$$F(p) = -KL(q(\boldsymbol{\Lambda}|\boldsymbol{D}), \exp[\boldsymbol{I}(\boldsymbol{\Lambda})]) \tag{33}$$

The negative free energy is therefore maximised when

$$q(\boldsymbol{\Lambda}|\boldsymbol{D}) = \exp[I(\boldsymbol{\Lambda})] \tag{34}$$

By substituting the likelihood from equation 4 we get

$$\begin{aligned} I(\boldsymbol{\Lambda}) &= \int q(\boldsymbol{w}|\boldsymbol{D}) \left[\frac{N}{2} \log |\boldsymbol{\Lambda}| - Tr(\boldsymbol{\Lambda} \boldsymbol{E}_{\boldsymbol{D}}(\boldsymbol{W})) \right] d\boldsymbol{w} + \log p(\boldsymbol{\Lambda}) \\ &= \frac{N-d-1}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2} Tr \left(\boldsymbol{\Lambda} [\boldsymbol{E}_{\boldsymbol{D}}(\hat{\boldsymbol{W}}) + \boldsymbol{\Omega}] \right) \end{aligned} \tag{35}$$

where $\boldsymbol{\Omega}$ is a d-by-d matrix formed as follows. Firstly, we partition the weight covariance matrix $\hat{\boldsymbol{\Sigma}}$ into d-by-d submatrices, each of dimension $(p \times d)$-by-$(p \times d)$. The i,jth such submatrix, $\hat{\boldsymbol{\Sigma}}(i,j)$, contains all the covariances among the MAR coefficients associated with the ith and jth time series. The i,jth entry in $\mathbf{\Omega}$ is then given by

$$\mathbf{\Omega}_{i,j} = Tr(\hat{\mathbf{\Sigma}}(i,j)\boldsymbol{X}^T\boldsymbol{X}) \tag{36}$$

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

$$\begin{aligned} \boldsymbol{B} &= \boldsymbol{E_D}(\hat{\boldsymbol{w}}) + \mathbf{\Omega} \\ a &= N \\ \hat{\mathbf{\Lambda}} &= a\boldsymbol{B}^{-1} \end{aligned} \tag{37}$$

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

$$KL_q(x) \equiv KL(q(x|\boldsymbol{D}), p(x)) \tag{38}$$

then the negative free energy can be computed, from equation 18, as

$$F(p) = L_{av} - KL_q(\boldsymbol{w}) - KL_q(\alpha) - KL_q(\mathbf{\Lambda}) \tag{39}$$

The average log-likelihood is given by

$$\begin{aligned} L_{av} &= \int q(\boldsymbol{\theta}|\boldsymbol{D}) \log p(\boldsymbol{D}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= -\frac{dN}{2}\log 2\pi + \frac{N}{2}\int q(\mathbf{\Lambda}|\boldsymbol{D}) \log|\mathbf{\Lambda}| d\mathbf{\Lambda} \\ &\quad -\frac{1}{2}\int q(\mathbf{\Lambda}|\boldsymbol{D}) q(\boldsymbol{w}|\boldsymbol{D}) Tr(\hat{\mathbf{\Lambda}}\boldsymbol{E_D}(\boldsymbol{w})) d\mathbf{\Lambda} d\boldsymbol{w} \\ &= -\frac{dN}{2}\log 2\pi + \frac{N}{2}L(a,\boldsymbol{B}) - \frac{1}{2}Tr(\hat{\mathbf{\Lambda}}[\boldsymbol{E_D}(\hat{\boldsymbol{w}}) + \mathbf{\Omega}]) \\ &= -\frac{dN}{2}\log 2\pi e + \frac{N}{2}L(a,\boldsymbol{B}) \end{aligned} \tag{40}$$

where $L(a, \boldsymbol{B})$ is defined in the Appendix (equation 63). By substituting the above expression into equation 39 and noting that many terms cancel, we get

$$F(p) = -\frac{N}{2}\log|\boldsymbol{B}| - KL_q(\boldsymbol{w}) - KL_q(\alpha) + \log\Gamma_d(N/2) \tag{41}$$

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 Variational Bayes (VB) framework maximises the negative free energy

$$F(p) = L_{av} - KL_q(\boldsymbol{\theta}) \tag{42}$$

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 acts as a penalty term which penalizes 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 [1, 6]

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

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

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

This is identical to the expression on p.12 of [16] referred to as Schwarz's Bayesian Criterion (SBC) (after dividing by $-Nd/2$ and dropping terms not dependent on model order). By comparsion with equation 41 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 [1]

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

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

5 Structured Priors

Instead of using the isotropic Gaussian prior of equation 8, where every coefficient has the same prior variance, we can split the coefficients up 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) [11], 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 $j = 1..G$ groups with k_j weights in the jth group and where

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

and $\boldsymbol{I}_j$ is a matrix with 1's 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

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

The other updates are exactly the same as for the global prior. The only other 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 above. 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 (i) weights involved in within-series prediction and (ii) weights involved in between-series predictions, ie. interactions. Because we can estimate the evidence for each model (ie. 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 above schemes allows us to test for many different types of interactions. For example, we could have separate groups for interaction and non-interaction at *each* time-lag; this gives rise to the 'Lag-Interaction Prior'. Similarly, we could group interactions between subsets of the time series together eg. for EEG data we could split the interactions into groups for within and between different cortical areas.

Finally, we note that k-means clustering procedures can be used to identify which weights should be grouped together. This is implemented by applying a MAR model with a global prior and then running k-means on the inferred coefficients. This provides a semi-automatic way of designing priors.

6 Other issues

We use uninformative priors (see eg. [17]) by setting $b = 10^3$ and $c = 10^{-3}$ for the Gamma prior on α (see 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 training if the proportional increase in $F(p)$ from one iteration to the next is less than 0.01 per cent.

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 above computation becomes a bottleneck in the VB algorithm. To overcome this we can just extract the first few 'significant' eigenvalues using, for example, the 'snapshot' method [24].

By analogy with the Bayesian Evidence framework [19] we can define a quantity γ, known as the effective number of 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 [16] where the coefficents 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, then by noting that the marginal distribution over the MAR coefficients is a multivariate t-distribution ([4], page 440), we can compute confidence intervals in a manner similar to that described in [16].

Finally, we note that the Power Spectral Density matrix can be computed from the MAR coefficients, as described in ([14], page 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 [16]

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

where

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

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

and the noise precision matrix is

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

We generated 50 data sets, each containing $N = 200$ values. Typically, 5 cycles of the VB update equations were required to reach convergence. Figure 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 illustrates an expected trend, which has been demonstrated more extensively for univariate AR models [19], namely, that as the correct model order increases (and the number of data points remains the same) the VB criterion is superior to BIC. Figure 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 $p = 4$.

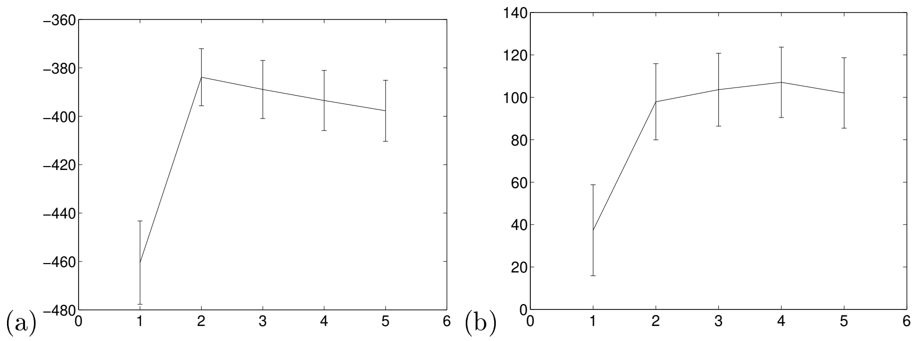


Figure 1: *A plot of the negative free energy, $F(p)$, versus model order p for data generated from a MAR(p) model with (a) $p = 2$ and (b) $p = 4$. We plot the average value calculated over 50 trials with error bars at plus and minus 1 standard deviation.*

7.2 Structured Priors

We generated 50 data points from the MAR(2) model described above 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 Figure 2 in the form of a 'Hinton diagram' [2]. 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 3 seconds worth of data from 5 sinusoidally varying time series, each sampled at 128Hz. The time series were generated independently, thus there is no interaction in the model so all the off-diagonal MAR coefficients should be zero. Figure 3 shows that use of the interaction prior clearly reduces the magnitude of the spurious coefficients. 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 equation 45).

Inferences of this sort become harder to make when we have fewer data points. For example, if we repeat the above procedure, but with 2 seconds of data and then only 1 second of data the above 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.

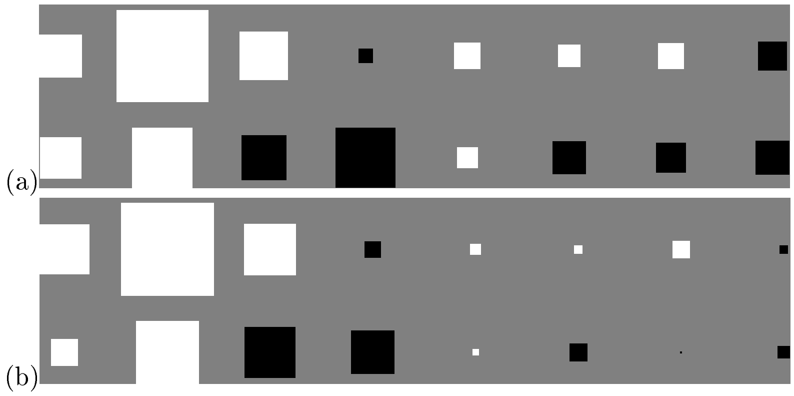


Figure 2: **Lag-Prior example**. *Estimation of MAR(4) model coefficients applied to $d = 2$ time series with (a) a Global-prior and (b) a Lag-prior. White indicates positive, black negative and the area is proportional to the magnitude. At each time lag there is a d-by-d matrix of coefficients; we therefore have 4 2-by-2 matrices with the time-lag increasing from left-to-right. This model was applied to data generated from an MAR(2) process. Thus, the true values of the 8 coefficients on the right of each figure are zero. Use of the Lag-prior reduces the magnitude of these irrelevant coefficients.*

As a third example, we generated a further six-dimensional data set in which the first 3 time-series were noisy versions of the same 10Hz sinewave and the next 3 were noisy versions of a 12.5Hz sinewave. Applying a MAR(1) model to this data with a global prior produced a model with evidence -463. An MAR(1) model with the interaction prior resulted in a model with evidence -468. For this example the global prior is therefore better than the interaction prior. Neither priors are particularly good, however, as they both neglect the block-diagonal structure. Application of a k-means clustering procedure (with $k = 2$) to the inferred coefficients from the first model clearly identified this structure. This allowed us to set up a prior semi-automatically and train a further MAR(1) model with this 'auto-prior'. This resulted in a model with evidence -449, much better than the others. The inferred coefficients are shown in Figure 4.

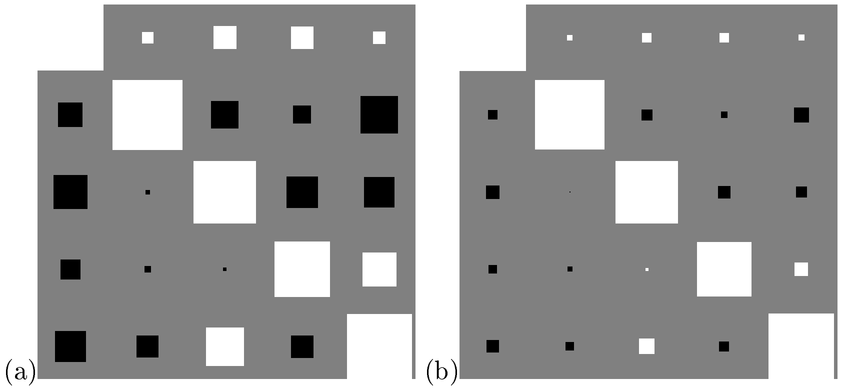


Figure 3: **Interaction-Prior example**. *Estimation of the matrix of MAR coefficients at lag 1, when the model is applied to data consisting of 5 independent, sinsusoidally varying time series with (a) a Global prior and (b) an Interaction prior. White indicates positive, black negative and the area is proportional to the magnitude. Use of the interaction prior reduces the magnitude of the off-diagonal coefficients (which, ideally, should be zero).*

7.3 Cognitive-EEG Data

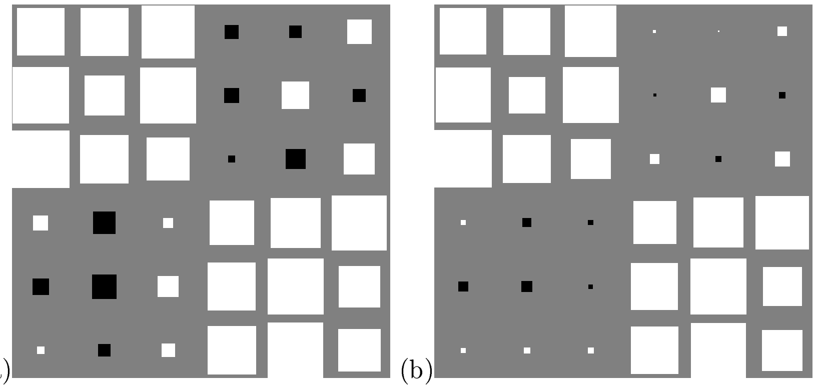
We analysed six-channel EEG data recorded whilst a subject performed different cognitive tasks. The data is derived from a large database collected from four subjects performing five different tasks [1], although in this paper we focus on just two of the tasks; (i) a baseline task and (ii) a maths task. The EEG records for a single task performance are shown in Figure 5. To ensure signal stationarity we analyse 1-second sub-sections of data. The following results were derived from a single subject (subject 2) performing the above two tasks, unless otherwise stated.

Figure 6 shows that for a global prior the optimal model order, averaged over 1-second sections of data, is two. With the interaction prior, this optimum increases to four. For the model with global prior the effective degrees of freedom (see equation 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 10 1-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 2 weight groups are hypothesized 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-

[1] This data is available from http:\\www.colostate.cs.edu\ ~anderson.



(a) (b)

Figure 4: **Auto-Prior example**. *Estimation of the matrix of MAR coefficients at lag 1, when the model is applied to data consisting of two groups of interacting sinsusoids with (a) an Interaction prior and (b) an 'Auto-prior'. The Auto-prior was derived from a k-means clustering procedure, and clearly identified the block-diagonal structure - placing block-diagonal coefficients in one group and off-block-diagonal in the other. White indicates positive, black negative and the area is proportional to the magnitude. Use of the auto-prior reduces the magnitude of the off-block-diagonal coefficients (which, ideally, should be zero). Results with the global prior were very similar to the Interaction prior.*

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.

We also looked at a 'hemispheric prior' where the coefficients were placed in three groups (i) one for all within-series terms (ii) one for all left hemisphere interactions, (iii) one for all right hemishere interactions and (iv) one for the cross-hemisphere interactions. The motivation for this prior being that the interactions between hemispheres may be of a different scale to those within each. This was not, however, shown to be the case as, for all sections of data from all tasks, the interaction prior had higher evidence.

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, 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 [8]. 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 (see Figure 5). This is picked up in the MAR models by the inference of highly non-diagonal noise covariance matrices. Finally, although the spatio-temporal 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

Table 1: **Evidence of MAR models with different priors: Cognitive Data.** *The evidence, as estimated by the negative free energy F is given for (i) an interaction-prior, which splits coefficients into two groups; one for interactions and one for within-series terms, (ii) an interaction prior at each time lag, (iii) a lag-prior,(iv) a global prior and (v) a '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 1-second sections of data; for each section we subtracted the mean value of F so that results from all sections could be meaningfully pooled.*

Prior	Evidence
Interaction	31 $\pm$ 2
Lag-Interaction	26 $\pm$ 2
Lag	-6 $\pm$ 4
Global	-23 $\pm$ 3
Random	-28 $\pm$ 3

Table 2: **Classification of EEG records using MAR features.** *Error rates (%) for using MAR features to discriminate between EEG data recorded during the baseline task versus the maths task, for using just the diagonal, off-diagonal or all of the MAR coefficients.*

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

which were identified using a stepwise forwards selection procedure [9]. Error rates were computed using a 5-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.

7.4 Physiological Data

We analysed a portion of the multi-channel physiological data set described in [23]. This is shown in Figure 8 and consists of three channels; heart rate, chest volume and blood-oxygen concentration.

Using a global prior, the optimal model order was seen to be six. This gave rise to a model with evidence 219. Application of the lag, lag-interaction and interaction priors gave evidences of 205, 198 and 247. The optimal model order for the interaction prior was at $p = 7$, again showing that the interaction prior allows a (marginally) greater extraction of temporal features.

We then applied the k-means clustering procedure to the MAR coefficients, with $k = 3$. This identified three main clusters (i) large positive weights, (ii)

Table 3: **Evidence of MAR models with different priors: Sleep data.** *The evidence, as estimated by the negative free energy F is given for (i) an interaction-prior, which splits coefficients into two groups; one for interactions and one for within-series terms, (ii) an interaction prior at each time lag, (iii) a lag-prior and (iv) a global prior. These were all calculated for a time-lag of $p = 4$ and were averaged over ten 1-second sections of data; for each section we subtracted the mean value of F so that results from all sections could be meaningfully pooled.*

Prior	Evidence
Interaction	8.0 $\pm$ 5.0
Lag	6.2 $\pm$ 4.0
Lag-Interaction	-1.6 $\pm$ 3.4
Global	-12.8 $\pm$ 5.0

large negative weights and (iii) small weights. We then created an 'Auto-prior' (although, it was somewhat hand-crafted) by placing the large weights in one group and the small weights in the other. Figure 9 shows the groupings. The resulting model had an evidence of 253, the highest of any model. This shows that clustering procedures are a promising method for designing suitable priors (see section 8 for more on this).

7.5 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 focussed on ten 1-second sections of data recorded whilst the subject was awake or in sleep-stage 4.

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 4, whereas use of the interaction prior increased this to 7. 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.

For the sleep-stage 4 data, the optimal model orders decreased; to 3 for the global prior and to 6 for the interaction prior. This illustrates a known trend for sleep EEG data; that optimal model order decreases with increasing sleep stage [18].

8 Discussion

In this paper we have shown how to apply the Variational Bayesian (VB) 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 [19], 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. Whilst 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.

Finally, we have presented some evidence to show that clustering procedures can be used to semi-automatically design appropriate model priors. In future work we intend to fully automate this procedure by using a Gaussian Mixture Prior. In this model each term in equation 47 is replaced by a univariate Gaussian Mixture model. This means that weights are soft-assigned to groups and the assignments, means, variances and number of groups can be inferred using (almost) the same VB method as described in [20]. This will form a subroutine of the present VB-MAR algorithm.

A The Kronecker Product

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

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & ... & a_{1m}\boldsymbol{B} \\ ... & & \\ a_{m1}\boldsymbol{B} & & a_{mm}\boldsymbol{B} \end{bmatrix} \tag{55}$$

For properties of the Kronecker product see, for example, page 477 in Box and Tiao [4].

B Densities and divergences

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

$$KL(q,p) = \int q(x) \log \frac{q(x)}{p(x)} dx \tag{56}$$

B.1 Normal density

The multivariate Normal density is given by

$$\mathsf{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{57}$$

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

$$KL(q,p) = 0.5 \log \frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} + 0.5 Trace(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_q) + 0.5(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) - \frac{d}{2} \tag{58}$$

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

B.2 Gamma density

The Gamma density is given by

$$\mathsf{Ga}(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right) \tag{59}$$

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

$$\begin{aligned} KL(q,p) &= (c_q - 1)\Psi(c_q) - \log b_q - c_q - \log \Gamma(c_q) \qquad (60) \\ &+ \log \Gamma(c_p) + c_p \log b_p - (c_p - 1)(\Psi(c_q) + \log b_q) + \frac{b_q c_q}{b_p} \end{aligned}$$

where $\Psi()$ is the digamma function [22].

B.3 Wishart density

The Wishart distribution is given by ([15], page 85)

$$\mathsf{Wi}(\boldsymbol{\Lambda}; a, \boldsymbol{B}) = \frac{1}{Z(a, \boldsymbol{B})} |\boldsymbol{\Lambda}|^{(a-d-1)/2} \exp\left[-\frac{1}{2} Tr(\boldsymbol{B}\boldsymbol{\Lambda})\right] \tag{61}$$

where

$$Z(a, \boldsymbol{B}) = 2^{ad/2} |\boldsymbol{B}|^{-a/2} \Gamma_d(a/2) \tag{62}$$

and $\Gamma_d()$ is the generalised gamma function defined on page 62 of [15].

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

$$L(a, \boldsymbol{B}) = \int \mathsf{Wi}(\boldsymbol{\Lambda}; a, \boldsymbol{B}) \log |\boldsymbol{\Lambda}| d\boldsymbol{\Lambda} \tag{63}$$

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

$$H(q) = -\left(\frac{q-d-1}{2}\right) L(q, \boldsymbol{Q}) + \frac{qd}{2} + \log Z(q, \boldsymbol{Q}) \tag{64}$$

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

$$\begin{aligned} KL(q,p) &= \left(\frac{q-d-1}{2}\right) L(q, \boldsymbol{Q}) - \left(\frac{p-d-1}{2}\right) L(p, \boldsymbol{P}) \qquad (65) \\ &\quad -\frac{qd}{2} + \frac{q}{2} Tr(\boldsymbol{P}\boldsymbol{Q}^{-1}) + \log \frac{Z(p, \boldsymbol{P})}{Z(q, \boldsymbol{Q})} \end{aligned}$$

References

[1] H. Attias. A Variational Bayesian Framework for Graphical Models. In T. Leen et al, editor, *NIPS 12*, Cambridge, MA, 2000. MIT Press.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[3] C.M. Bishop. Variational Principal Components. In *International Conference on Artificial Neural Networks*, pages 509–514, 1999.

[4] G.E.P Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis.* John Wiley, 1992.

[5] S.L. Bressler, M. Ding, and W. Yang. Investigation of cooperative cortical dynamics by multivariate autoregressive modeling of event-related local field potentials. *Neurocomputing*, 26-27:625–631, 1999.

[6] D.M. Chickering and D. Heckerman. Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. Technical Report MSR-TR-96-08, Microsoft Research, 1996.

[7] R. Choudrey, W.D. Penny, and S.J. Roberts. An Ensemble Learning approach to Independent Component Analysis. In *IEEE International Workshop on Neural Networks for Signal Processing, Sydney, Australia*, 2000.

[8] M. Kaminski, K. Blinowska, and W. Szelenberger. Topographic Analysis of coherence and propagation of EEG activity during sleep and wakefulness. *Electroencephalography and Clinical Neurophysiology*, 102:216–227, 1997.

[9] D.G. Kleinbaum, L.L. Kupper, and K.E. Muller. *Applied Regression Analysis and Other Multivariable Methods.* PWS-Kent, Boston, 1988.

[10] H. Lutkepohl. *Introduction to Multiple Time Series Analysis.* Springer Verlag, 1993.

[11] D.J.C. Mackay. Bayesian non-linear modelling for the energy prediction competition. *ASHRAE Transactions*, 100:1053–1062, 1994.

[12] D.J.C. Mackay. Ensemble Learning and Evidence Maximization. Technical report, Cavendish Laboratory, University of Cambridge, 1995.

[13] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley, 1997.

[14] S.L. Marple. *Digital spectral analysis with applications.* Prentice-Hall, 1987.

[15] R.J. Muirhead. *Aspects of Multivariate Statistical Theory.* John Wiley, 1982.

[16] A. Neumaier and T. Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *Submitted to ACM Trans. Math. Softw.*, 2000.

[17] J.J.K. O'Ruaniaidh and W.J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing.* Springer, 1996.

[18] J. Pardey, S. Roberts, and L. Tarassenko. A Review of Parametric Modelling Techniques for EEG Analysis. *Med. Eng. Phys.*, 18(1):2–11, 1996.

[19] W.D. Penny and S.J. Roberts. Bayesian Methods for Autoregressive Models. In *IEEE International Workshop on Neural Networks for Signal Processing, Sydney, Australia*, 2000.

[20] W.D. Penny and S.J. Roberts. Variational Bayes for 1-dimensional Mixture Models. Technical Report PARG-2000-01, Department of Engineering Science, Oxford University, 2000.

[21] W.D. Penny and S.J. Roberts. Variational Bayes for Non-Gaussian Autoregressive Models. In *IEEE International Workshop on Neural Networks for Signal Processing, Sydney, Australia*, 2000.

[22] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C.* Cambridge, 1992.

[23] D.R. Rigney, A. L. Goldberger, W.C. Ocasio, Y. Ichimaru, G.B. Moody, and R.G. Mark. Multi-Channel Physiological Data: Description and Analysis (Data Set B). In A.S. Weigend and N.A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 105–129. Addison-Wesley, 1994.

[24] L. Sirovich. Turbulence and the dynamics of coherent structures. *Quarterly Applied Mathematics*, 45(3):561–590, 1987.

[25] S. Weisberg. *Applied Linear Regression.* John Wiley, 1980.

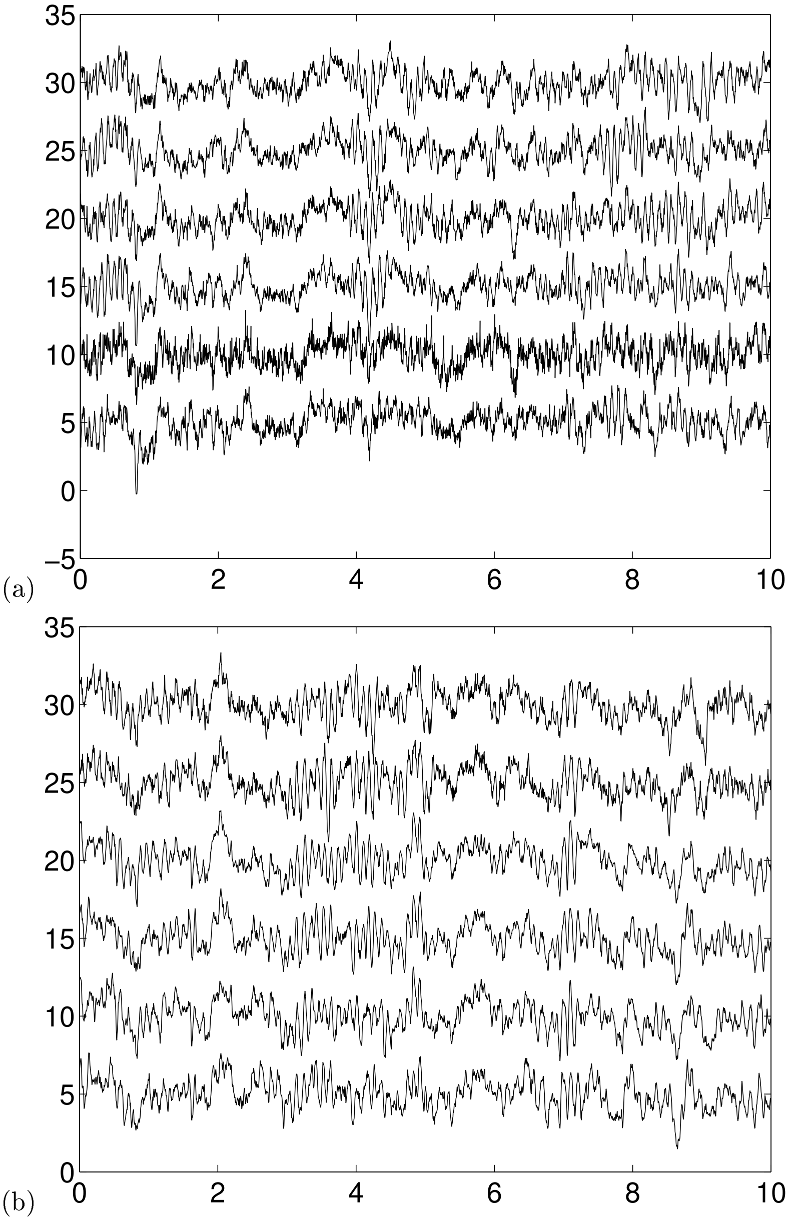


Figure 5: **EEG Data** *Six-channel EEG recorded during (a) a baseline task and (b) a maths task. The channels are recorded from one electrode on the right and one on the left of Central (C), Parietal (P) and Occipital (O) cortex in the following order (going up the page); C-left, C-right, P-left, P-right, O-left and O-right. The x-axis shows elapsed seconds.*

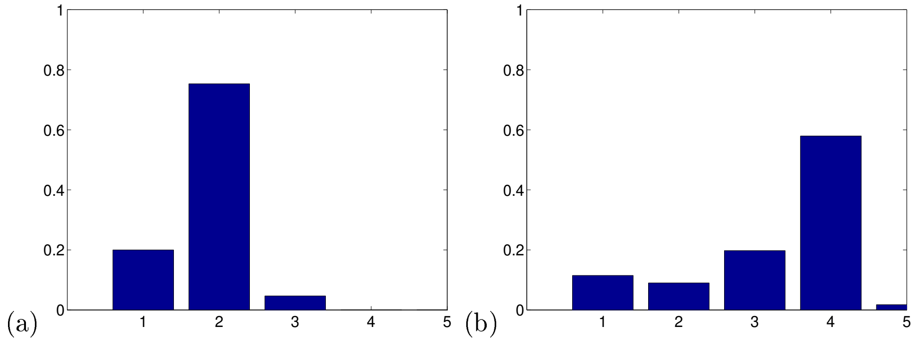


Figure 6: **EEG Model Order Selection** *A plot of $P(p)$, the probability of model order p versus number of time-lags p for (a) a global prior and (b) an interaction prior. These probabilities were calculated from equation 45 and averaged over ten 1-second sections of data. By disregarding spurious spatial interactions the MAR model using the interaction prior is able to extract more temporal information.*

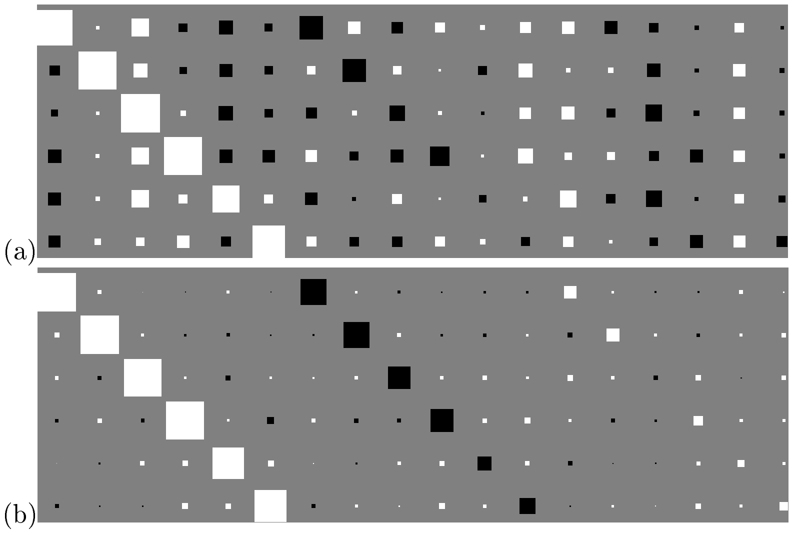


Figure 7: **Hinton diagram of an MAR(3) model fitted to EEG data recorded during a maths task** *for (a) a global prior and (b) an interaction prior. At each of the 3 time-lags (going left to right across the page), we have a 6×6 coefficient matrix. The interaction prior reduces the magnitude of the off-diagonal terms.*

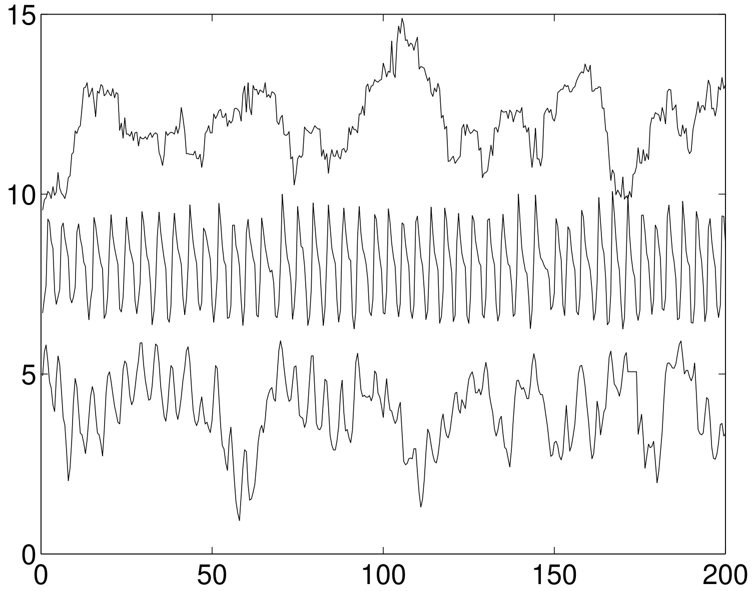


Figure 8: **Physiological Data** *The figure shows the heart-rate, respiration and blood-oxygen concentraion (going up the page) recorded over a 200-second period.*

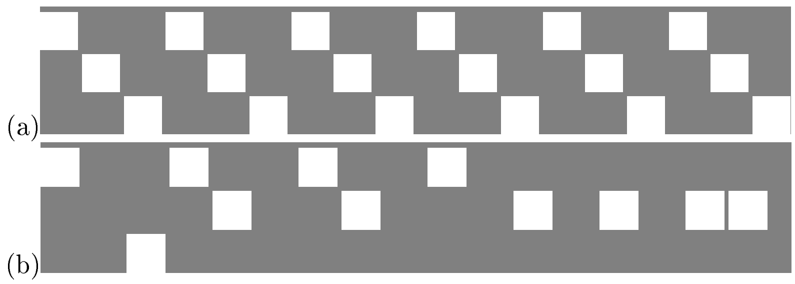


Figure 9: **Comparison of prior structure** *For each of the above figures, at each of the six time-lags (going left to right across the page) we have a* 3×3 indicator *matrix. The white squares* indicate *coefficients belonging to the first group (they are not the actual values of the coefficients). All others belong to the second group. Figure (a) shows the groupings for an Interaction prior and Figure (b) for an Auto-prior. The auto prior is for a MAR(6) model derived from the physiological data. This prior leads to an MAR model with a higher evidence than one with an interaction prior.*