ACADEMIC PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

NeuroImage

NeuroImage 0 (2003) 000–000

www.elsevier.com/locate/ynimg

Variational Bayesian inference for fMRI time series

Will Penny,* Stefan Kiebel, and Karl Friston

Wellcome Department of Imaging Neuroscience, University College, London WC1N 3BG, UK

Received 26 August 2002; accepted 18 December 2002

Abstract

We describe a Bayesian estimation and inference procedure for fMRI time series based on the use of General Linear Models with Autoregressive (AR) error processes. We make use of the Variational Bayesian (VB) framework which approximates the true posterior density with a factorised density. The fidelity of this approximation is verified via Gibbs sampling. The VB approach provides a natural extension to previous Bayesian analyses which have used Empirical Bayes. VB has the advantage of taking into account the variability of hyperparameter estimates with little additional computational effort. Further, VB allows for automatic selection of the order of the AR process. Results are shown on simulated data and on data from an event-related fMRI experiment.

1. Introduction

In neuroimaging, the estimation and inferences about evoked responses have, thus far, rested largely upon classical inference. In statistics, however, there are two main frameworks for making inferences, classical inference and Bayesian inference. For a comparison of the different frameworks see Barnett (1999) and Casella and Berger (1990). Strong advocates of Bayesian analysis consider it the only logical and self-consistent framework for probabilistic inference. The rationale behind such claims is laid down in classic texts such as Box and Tiao (1992) and Bernardo and Smith (2000). Adoption of a Bayesian inference framework has led to a multitude of advances in areas such as image processing (Blake and Isard, 1998), signal processing (O'Ruanaidh and Fitzgerald, 1996), machine learning (Jordan, 1999), and pattern recognition (Bishop, 1995). This is especially important as developments in these fields have a follow-on impact on neuroimaging methodology. The initial impact is already being felt (Friston et al., 2002b).

Both (classical) maximum likelihood and Bayesian analysis use the same model of how the data are caused, often a linear model. However, they differ in both estimation and inference. Bayesian analysis can be considered an extension of maximum likelihood that relies upon the specification of prior expectations about the parameters of the model, e.g., activations. In maximum-likelihood estimation, the parameters are chosen to maximize the likelihood of obtaining the observed data. In Bayesian analysis the objective is to compute the probability of the activation given the data, that is, the posterior density. Through Bayes rule this requires the specification of priors on the parameters or activations.

Inference in classical statistics proceeds by considering the null hypothesis that there is no activation. A statistic is then formed whose distribution under the null hypothesis can be used to reject that hypothesis if the data are sufficiently unlikely. For example a T statistic is a linear compound of parameter estimates divided by the standard error. The standard error in turn is based on the variance of the compounding likelihood density. This variance corresponds to a hyperparameter (a parameter of a probability density function of parameters).

In Bayesian inference the probability that the activation or contrast of parameters exceeds some specified threshold can be computed directly from the posterior density. This posterior density is parameterized by its own hyperparameters. In short, to make an inference of a classical or Bayesian sort both the parameters and the hyperparameters of a model must be estimated. In classical inference the hyper-

* Corresponding author.
E-mail address: {wpenny,karl}@fil.ion.ucl.ac.uk (W. Penny).

1053-8119/03/$ – see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S1053-8119(03)00071-5

parameters are Restricted Maximum Likelihood (ReML) estimates. These are simply the values of the hyperparameter that maximize the probability of the data.

Critically, the variability in hyperparameter estimates must enter into the inference. This variability is expressed through the degrees of freedom of classical statistics. For hierarchical linear Gaussian models with multiple hyperparameters we show (Kiebel et al., 2002a) how this variability can be taken into account using a Satterthwaite-type approximation based on ReML estimates of the hyperparameters.

In Bayesian inference this variability can be taken into account by forming the full posterior over the parameters and hyperparameters and then integrating out (i.e., averaging over) the hyperparameters. The ensuing marginal distribution is the posterior density of the parameters required for inference. In practice, however, this integration is often problematic. Either time-consuming sampling approaches are used or the variability is simply ignored. In the empirical Bayes framework (Carlin and Louis, 2000), for example, the variability in the hyperparameters is typically ignored leading to the "overconfidence problem" (Friston et al., 2002a).

In this paper we present the general approach, Variational Bayes (VB), that approximates the posterior density with an analytically tractable form based on the use of conjugate priors and the assumption of (a degree of) factorization in the posterior. This enables the posterior densities of the hyperparameters to be modeled and resolves the over-confidence problem. We introduce VB for functional neuroimaging time series and illustrate its application to the analysis of fMRI in the context of unknown hyperparameters governing serial correlations among the errors.

In section 2 we describe the time-series model. In section 3 we describe the Variational Bayes methodology and in section 4 show how it is applied to our model. This section makes extensive reference to mathematical derivations which are given in an appendix. In section 5 we present results on simulated data and on data from an event-related fMRI experiment.

2. Models of fMRI time series

A key issue in the analysis of fMRI time series is the concern that succesive samples are serially correlated. These correlations arise from neural, physiological, and physical sources including the pulsatile motion of the brain caused by cardiac cycles, local modulation of the static magnetic field by respiratory movement, and unmodeled neuronal activity. See Zarahn et al. (1997) and Woolrich et al. (2001) for a full discussion. Not all of this correlation can be removed by high-pass filtering as the required filter cutoffs would also remove much of the signal.

A standard approach to the analysis of fMRI time series employs voxel-wise General Linear Models (GLMs). The data at each voxel, Y, are explained with a set of effects that are incorporated into a design matrix, X. One then proceeds by fitting the model

$$Y = Xw + E \tag{1}$$

and making inferences based on the parameters, w. The voxel-wise GLM approach, pioneered in (Friston et al., 1995c) and developed in a Bayesian context (Friston et al., 2002b), allows one to produce functional maps of the human brain derived from single- or multiple-subject fMRI studies.

The serial correlation in the error time series, E, affects both the model fitting and the statistical inference. This is typically handled using a two-stage process where the correlation is estimated in the first stage and the parameters are estimated in the second stage. The diversity of ensuing approaches results from different characterisations of the serial correlation. These range from autoregressive (AR) processes (Friston et al., 1995b; Bullmore et al., 1996; Worsley et al., 2002), Autoregressive Moving Average (ARMA) processes (Locascio et al., 1997), AR plus white noise models (Purdon and Weisskoff, 1998), frequency domain models where the magnitude falls of as $1/f$ (Zarahn et al., 1997), or by multitapering (Woolrich et al., 2001). For a review of many of these approaches see Woolrich et al., (2001).

The two-stage process for handling the serial correlation can be extended to multiple iterations using ReML (Friston et al., 2002a) and this allows for both more accurate parameter estimation and statistical inference. While this is more computationally demanding (Worsley et al., 2002) and is subject to the law of diminishing returns (Bullmore et al., 1996; Woolrich et al., 2001) we nevertheless take such an iterative approach in this paper.

In this paper, we use the voxel-wise GLM approach in conjunction with AR error processes of arbitrary order. These are referred to as GLM-AR(p) models where p is the order of the AR process. The reason for this choice is that, of the many characterisations, AR processes are the most amenable to mathematical analysis. Further, as we will show, low-order AR processes are sufficient to characterise the serial correlation in fMRI time series (provided low-frequency drift terms are modeled as fixed effects).

Mathematically, the GLM-AR(p) model is given by

$$Y = Xw + E \tag{2}$$

$$E = \tilde{E}^T a^T + Z \tag{3}$$

where Y is a $[N \times 1]$ vector of fMRI time-series samples, X is the $[N \times k]$ design matrix (see Fig. 7 for an example of a design matrix), w is a $[k \times 1]$ vector of regression coef-

ficients, E is a $[N \times 1]$ vector of errors which are modeled as an AR process where a is a $[1 \times p]$ vector of AR coefficients, $\tilde{E}$ is a $[p \times N]$ matrix of "embedded" errors (see later) and Z is a $[N \times 1]$ vector of Independent and Identically Distributed (IID) Gaussian errors.

This same model can also be written in terms of the response at scan t

$$y_t = x_t w + e_t \tag{4}$$

$$e_t = \sum_{j=1}^{p} a_j e_{t-j} + z_t, \tag{5}$$

where y_t, x_t, e_t, and z_t are the tth rows of Y, X, E, and Z. For the design matrix in Fig. 7, for example, the row vector x_t corresponds to the tth row. The noise z_t is Gaussian with zero mean and precision (inverse variance) λ. The tth column of the embedded error matrix $\tilde{E}$ is

$$\tilde{E}_t = \begin{bmatrix} e_{t-1} \\ e_{t-2} \\ \cdots \\ e_{t-p} \end{bmatrix}. \tag{6}$$

We also define the embedding matrices D and $\tilde{X}$ whose tth columns are given by

$$d_t = \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \cdots \\ y_{t-p} \end{bmatrix} \tag{7}$$

and

$$\tilde{X}_t = \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \cdots \\ x_{t-p} \end{bmatrix}, \tag{8}$$

where d_t is $[p \times 1]$ and $\tilde{X}_t$ is $[p \times k]$. Note that D is $[p \times N]$ and $\tilde{X}$ is $[p \times kN]$ and because $e_t = y_t - x_t w$ we have $\tilde{E}_t = d_t - \tilde{X}_t w$. To apply the above equations to a time series we simply ignore the first p values of y_t. This will have little effect on the ensuing inferences. In what follows the notation $\mathsf{N}(\mu, \Sigma)$ refers to the multivariate Normal distribution with mean μ and covariance Σ. The notation $\mathsf{Ga}(b,c)$ refers to the Gamma probability distribution with parameters b and c.

2.1. Likelihood

Equations (4) and (5) can be used to express the log-likelihood of the fMRI time series as

$$\log p(Y|w, a, \lambda) = \frac{-\lambda}{2} \sum_t ((y_t - a d_t) - (x_t - a\tilde{X}_t)w)^2 + \frac{N-p}{2} \log \frac{\lambda}{2\pi}. \tag{9}$$

This may equivalently be written as

$$\log p(Y|w, a, \lambda) = \frac{-\lambda}{2} \sum_t ((y_t - x_t w) - a(d_t - \tilde{X}_t w))^2 + \frac{N-p}{2} \log \frac{\lambda}{2\pi}. \tag{10}$$

We present the two versions because in the first, the regression coefficients w are more easily isolated, and in the second, the AR coefficients a are. This will simplify the math later.

2.2. Priors

In this paper we use vague priors on the model parameters

$$p(w|\alpha) = \mathsf{N}(0, \alpha^{-1} I) \tag{11}$$

$$p(a|\beta) = \mathsf{N}(0, \beta^{-1} I)$$

$$p(\lambda) = \mathsf{Ga}(b_0, c_0), \tag{12}$$

where $\alpha = 10^{-6}$, $\beta = 10^{-3}$, $b_0 = 1000$, $c_0 = 0.001$, and N and Ga refer to the Normal and Gamma densities defined in Appendix A. The value α is larger than β because the regression coefficients are typically larger than the autoregressive coefficients. The particular value used for β can, in principle, affect the model order selection process. This is discussed further in sections 4.2 and 5.2.

We choose vague priors because the focus of this paper is on modeling the error process. Future work will allow for spatial priors and for priors allowing information to be aggregated over voxels and subjects. For example, for a random effects analysis (Yandell, 1997) of data from multiple subjects a hierarchical prior such as

$$p(w|\alpha) = \mathsf{N}(w_{pop}, \alpha^{-1} I) \tag{13}$$

would be more appropriate, where w_{pop} are the population regression coefficients and α is the between-subject precision. Alternatively, one might wish to use a shrinkage prior based on the variability of the regression coefficients over voxels (see, e.g., Friston et al., 2002a). In this case

$$p(w|\alpha) = \mathsf{N}(0, \alpha^{-1} I), \tag{14}$$

where α is the precision of the regression coefficients over voxels.

Given a Gaussian likelihood function with IID errors the conjugate prior for the noise precision is a Gamma density. This is the prior we use in this paper and its mathematical form is described in the appendix. All of the parameters of our model are collectively written as θ. That is, $\theta = \{w, a, \lambda\}$. The prior over the parameters is

$$p(\theta) = p(w|\alpha)p(a|\beta)p(\lambda). \quad (15)$$

The log-likelihood in Eqs. (9) and (10) is also written as log $p(Y|\theta)$.

The posterior distribution, $p(w|Y)$, can now be computed by combining the prior and likelihood using Bayes' rule. For the model we have described, however, there is no analytic form for $p(w|Y)$. A common solution is to resort to sampling methods (Kiebel et al., 2002b). In this paper, however, we make use of the Variational Bayesian framework in which the true posterior density is approximated with a factorised density. In the numerical examples in this paper the accuracy of this approximation will be verified using Gibbs sampling.

3. Variational Bayes

The central quantity of interest in Bayesian learning is the posterior distribution $p(\theta|Y)$. This implies estimation of both the parameters θ and the uncertainties associated with their estimation. This can be achieved with the VB framework, a full tutorial on which is given in Lappalainen and Miskin (2002). In what follows we describe the key features.

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

$$\begin{aligned}\log p(Y) &= \int q(\theta|Y)\ \log p(Y)d\theta \\ &= \int q(\theta|Y)\log\frac{p(Y,\ \theta)}{p(\theta|Y)}d\theta \\ &= \int q(\theta|Y)\log\left[\frac{q(\theta|Y)p(Y,\ \theta)}{p(\theta|Y)q(\theta|Y)}\right]d\theta \\ &= F + KL. \end{aligned} \quad (16)$$

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

$$F = \int q(\theta|Y)\log\frac{p(Y,\ \theta)}{q(\theta|Y)}\,d\theta, \quad (17)$$

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

$$KL = \int q(\theta|Y)\log\frac{q(\theta|Y)}{p(\theta|Y)}\,d\theta \quad (18)$$

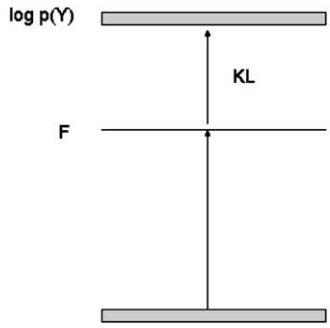

Fig. 1. The quantity F provides a lower bound on the log-evidence of the model with equality when the approximate posterior equals the true posterior.

is the KL divergence (Cover and Thomas, 1991) between the density $q(\theta|Y)$ and the true posterior $p(\theta|Y)$.

Equation (16) is the fundamental equation of the VB framework. Importantly, because the KL divergence is always positive (Cover and Thomas, 1991), F provides a lower bound on the model evidence. Moreover, because the KL divergence is zero when the two densities are the same, F will become equal to the model evidence when $q(\theta|Y)$ is equal to the true posterior. This is shown schematically in Fig. 1. For this reason $q(\theta|Y)$ can be viewed as an *approximate posterior*.

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

$$q(\theta|Y) = \prod_i q(\theta_i|Y), \quad (19)$$

where θ_i is the ith group of parameters. The distributions which maximise F can then, via the calculus of variations, be shown to be (Lappalainen and Miskin, 2000)

$$q(\theta_i|Y) = \frac{\exp[I(\theta_i)]}{\int \exp[I(\theta_i)]d\theta_i}, \quad (20)$$

where

$$I(\theta_i) = \int q(\theta^{\setminus i}|Y)\ \log p(Y,\ \theta)d\theta^{\setminus i} \quad (21)$$

and $\theta^{\setminus i}$ denotes all parameters *not* in the ith group. Note that, importantly, this means we are able to determine the optimal analytic *form* of the component posteriors [using eq. (20)]. This is to be contrasted with Laplace approximations where we have to arbitrarily fix the form of the component posteriors to be Gaussian (O'Ruaniaidh and Fitzgerald, 1996).

The above principles lead to a set of coupled update rules

```
Initialise;
While (ΔF > tol);
   Update Sufficient Statistics (SS) for regression coefficient
      distribution, {ŵ, Σ̂}, using equation 62
   Update SS for autoregressive coefficient
      distribution, {m, V}, using equation 49
   Update SS for noise precision
      distribution, {b_λ, c_λ}, using equation 75
   Calculate F using equation 28
   Let ΔF = (F^New − F^Old)/F^New
End
```

Fig. 2. Pseudo-code for VB algorithm. Update rules for the sufficient statistics of the distributions $q(w|Y)$, $q(a|Y)$, and $q(\lambda|Y)$ are applied until the relative increase in the objective function F is less than a specified tolerance, *tol.*

for the *parameters* of the component posteriors, iterated application of which leads to the desired maximisation. Further, by computing F for models of different order, we can perform model order selection (see, e.g., Roberts and Penny, 2002). The Bayesian Information Criterion (BIC) model order criterion has been shown to be a special case of the VB criterion (F), recovered in the limit of a large number of data points (Attias, 2000).

4. Variational Bayes for GLM-AR

We assume the following factorised form for the approximate posterior

$$q(\theta|Y) = q(w|Y)q(a|Y)q(\lambda|Y). \tag{22}$$

By plugging in the likelihood and priors for our GLM-AR model (from section 2) into Eq. (20), the approximate posteriors turn out to be

$$\begin{aligned} q(w|Y) &= \mathsf{N}(\hat{w},\hat{\Sigma}) \\ q(a|Y) &= \mathsf{N}(m,\ V) \\ q(\lambda|Y) &= \mathsf{Ga}(b_\lambda,\ c_\lambda). \end{aligned} \tag{23}$$

Note that, for each component, the form of the approximate posterior is the same as the prior. In fact, this is no accident, as we chose the priors to achieve this (for a discussion of such "conjugate" priors, see Box and Tiao, 1992). In the appendix we show how the parameters of these distributions are updated to maximise F [see Eq. (17)]. Parameter estimation in VB consists of iterative application of these update rules as shown by the pseudo-code in Fig. 2.

4.1. Initialisation

The distribution for the regression coefficients, $q(w|Y)$, is initialised by ignoring the autocorrelation in the errors. This is set using the well-known Ordinary Least Squares (OLS) solution

$$\begin{aligned} \hat{w} &= (X^TX)^{-1}X^TY \\ \hat{\Sigma} &= \sigma_e^2(X^TX)^{-1}, \end{aligned} \tag{24}$$

where

$$\sigma_e^2 = \frac{1}{N-k}\sum_t (y_t - x_t\hat{w})^2. \tag{25}$$

If we now assume the regression coefficients to be correct, the distribution for the AR coefficients can be set using the Maximum Likelihood (ML) solution [from inspection of Eq. (3)]

$$\begin{aligned} m &= (\tilde{E}\tilde{E}^T)^{-1}\tilde{E}E \\ V &= \sigma_z^2(\tilde{E}\tilde{E}^T)^{-1}, \end{aligned} \tag{26}$$

where

$$\sigma_z^2 = \frac{1}{N-p}\sum_t (e_t - \tilde{E}_t^T m^T)^2, \tag{27}$$

which uses $\tilde{E}_t = d_t - \hat{x}_t\hat{w}$, i.e., the value of $\hat{w}$ estimated in (24). Equation (24) constitutes the OLS update for the regression coefficients and Eq. (26) the OLS update for the AR coefficients.

4.2. Negative free energy

The negative free energy is used both to monitor convergence during parameter estimation and as a criterion for selecting the optimal AR order. As shown in the appendix, it can be computed as

$$F(p) = L_{av} - KL(\lambda) - KL(w) - KL(a), \tag{28}$$

where, for a generic parameter θ_i, $KL(\theta_i)$ denotes the KL divergence between the approximate posterior $q(\theta_i|Y)$ and the prior $p(\theta_i)$. Expressions for the KL divergences for the various densities are given in (Roberts and Penny, 2002). These KL terms should not be confused with the KL divergence in Eq. (18) which is between the approximate posterior and the true posterior.

The first term is given by

$$L_{av} = \frac{N-p}{2}\log\tilde{\lambda} - \frac{\bar{\lambda}}{2}\tilde{G} - \frac{N-p}{2}\log 2\pi \tag{29}$$

and

$$\log\tilde{\lambda} = \psi(c_\lambda) + \log b_\lambda \tag{30}$$

and where $\tilde{G}$ is computed from Eq. (77). Note that we write $F(p)$ here to emphasise the dependence of F on AR model order p. This dependence arises because F, being a lower bound on the model evidence, can be used as a model order selection criterion (see section 3). It consists of two terms: the average likelihood constitutes an accuracy term and the KL divergences constitute penalties for model complexity. The penalty term arises because $KL(a)$ increases with p. Just how much it increases, in part, depends on the value of β (the prior precision of the AR coefficients). In section 5.2, however, we provide a numerical example showing that this dependence is very weak.

When $F(p)$ is used for model order selection it is important that all models be given the same number of data points. For this reason the terms $(N - p)$ in Eqs. (29) and (75) should be replaced by $(N - p_{max})$, where p_{max} is the maximum putative model order.

The presence of the penalty terms in the objective function ($F(p)$) also prevents model overfitting, even for very large AR model orders. This comprises the VB solution to the over-confidence problem (Friston et al., 2002b).

One can imagine an alternative scheme for estimating the optimal AR order; fit a GLM model using OLS and then fit AR models to the residuals using a criterion such as BIC to choose the optimal order (Neumaier and Schneider, 2000). While this approach will give some indication of the true AR model order it is based on OLS estimators which, for any particular data sample, may contain a large error (that is, large in comparison to the VB estimate). Furthermore, in previous research on autoregressive modeling we have established that the VB selection criterion is superior to BIC (Roberts and Penny, 2002). To our knowledge, the VB scheme we have described is the only way for finding the optimal AR order from data sets with activations. A viable alternative is to focus on null data sets as in (Woolrich et al., 2001).

5. Results

5.1. Synthetic data I

We generated data from a known GLM-AR model

$$y_t = x_t w + e_t \tag{31}$$

$$e_t = a e_{t-1} + z_t, \tag{32}$$

where $x_t = 1$ for all t, $w = 2.7$, $a = 0.3$, and $1/\lambda = \mathrm{Var}(z) = \sigma^2 = 4$. We generated $N = 128$ samples. Now, given any particular values of w, a, λ it is possible to compute the exact posterior distribution up to a normalisation factor, as

$$p(w, a, \lambda|Y) \propto p(Y|w, a, \lambda)\, p(w|\alpha)\, p(a|\beta)\, p(\lambda). \tag{33}$$

If we evaluate the above quantity over a grid of values w, a, λ we can then normalise it so it sums to one and so make plots of the exact posterior density.

Fig. 3 compares the exact and approximate posterior joint densities for w, a. In the true posterior it is clear that there is a dependence between w and a (the width of the density over w depends on a) and that the approximate posterior used in VB ignores this dependence. Fig. 4 compares the exact and approximate posterior marginal densities for w, a and σ^2 showing good agreement. This example epitomises the VB approach, showing that accurate estimation of the marginal distributions is possible without detailed modeling of the joint distributions.

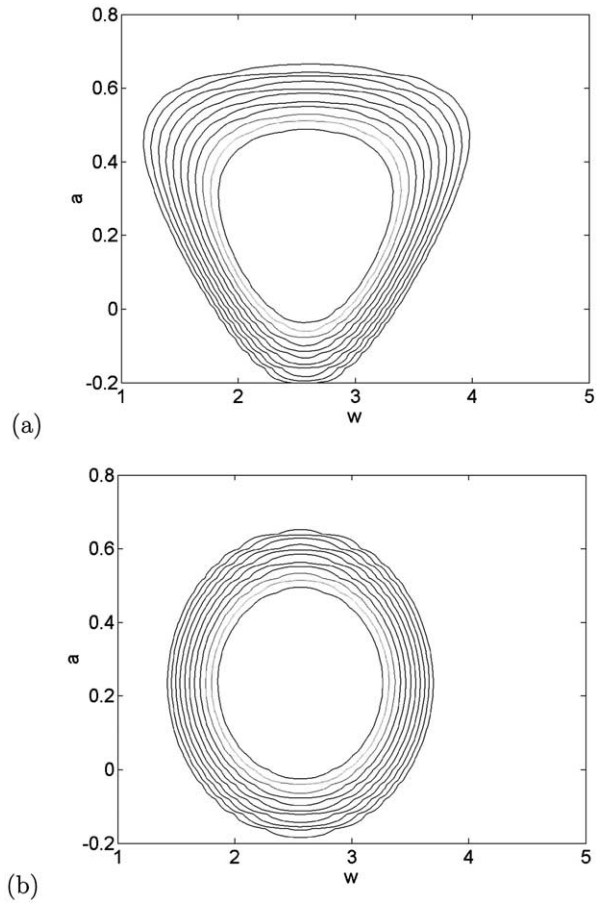

Fig. 3. The figures show contour lines of constant probability density from (a) the exact posterior $p(a, w|Y)$ and (b) the approximate posterior used in the VB algorithm, $q(a, w|Y)$ for the example in section 5.1. This clearly shows the effect of the factorisation, $q(a, w|Y) = q(a|Y)q(w|Y)$.

5.2. Synthetic data II

We generated data from a larger GLM-AR model having two regression coefficients and three autoregressive coefficients. While it is possible, in principle, to plot the exact posteriors using the method described previously, this would require a prohibitive amount of computer time. We therefore validated the VB algorithm by comparing it with Gibbs sampling (Kiebel et al., 2002b).

We used the model

$$y_t = x_t w + e_t \tag{34}$$

$$e_t = \sum_{j=1}^{p} a_j e_{t-j} + z_t, \tag{35}$$

where x_t is a two-element row vector, the first element flipping between a '−1' and '1' with a period of 40 scans (i.e., 20 −1's followed by 20 1's) and the second element being '1' for all t. The two corresponding entries in w reflect the size of the activation, $w_1 = 2$, and the mean signal level, $w_2 = 3$. We used an AR(3) model for the errors with

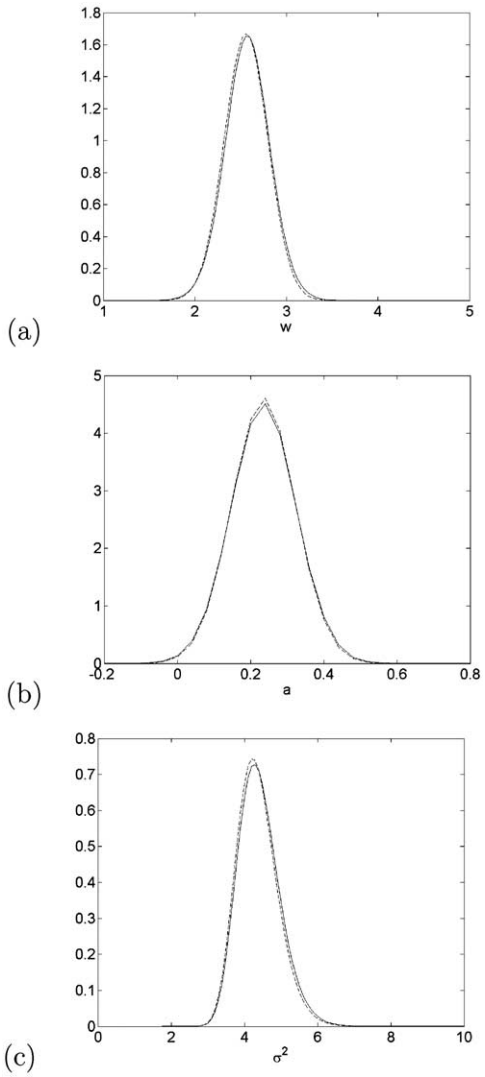

Fig. 4. The figures compare the exact (solid lines) and approximate (dashed lines) marginal posteriors (a) $p(w|Y)$ and $q(w|Y)$, (b) $p(a|Y)$ and $q(a|Y)$, (c) $p(\sigma^2|Y)$ and $q(\sigma^2|Y)$ (where $\sigma^2 = 1/\lambda$).

parameters $a_1 = 0.8$, $a_2 = -0.6$, and $a_3 = 0.4$. The noise precision was set to $1/\lambda = \mathrm{Var}(z) = \sigma^2 = 1$ and we initially generated $N = 400$ samples. An example time series produced by this process is shown in Fig. 5a.

We then generated 10 such time series and fitted GLM-AR(p) models to each using the VB algorithm. In each case the putative model order was varied between $p = 0$ and $p = 5$. Fig. 5b shows a plot of the average value of the negative free energy, *F(p)* as a function of p, indicating that the maximum occurs at the true model order. We note that the criterion *F(p)* is dependent on *KL(a)* and therefore on the chosen value of the prior precision β. We have found, however, that this dependence is very weak in that values in the range 10^{-1} to 10^6 did not change the optimal value of p.

We also generated a number of data sets containing either $N = 40$, $N = 160$, or $N = 400$ scans. At each data set size we applied the VB algorithm to a number of data sets and compared Gibbs and VB posteriors for each of the regression coefficients. For the purpose of these comparisons the model order was kept fixed at $p = 3$ for the generating models and the models inferred by Gibbs and VB. Fig. 6 shows representative results indicating a better agreement with increasing number of scans. We also note that the VB algorithm requires more iterations for fewer scans (typically 4 iterations for $N = 400$, 5 iterations for $N = 160$, and 7 iterations for $N = 40$). This is because the algorithm is initialised with the OLS solution which is closer to the VB estimate if there are a large number of scans.

Finally, we generated a number of data sets of various sizes to compare VB and OLS estimates of activation size with the true value of $w_1 = 2$. This comparison was made using a matched-pairs t test on the absolute estimation error. For $N > 100$ the VB estimation error was significantly smaller for VB than for OLS ($p < 0.05$). For $N = 160$, for example, the VB estimation error was 15% smaller than the OLS error ($p < 0.02$).

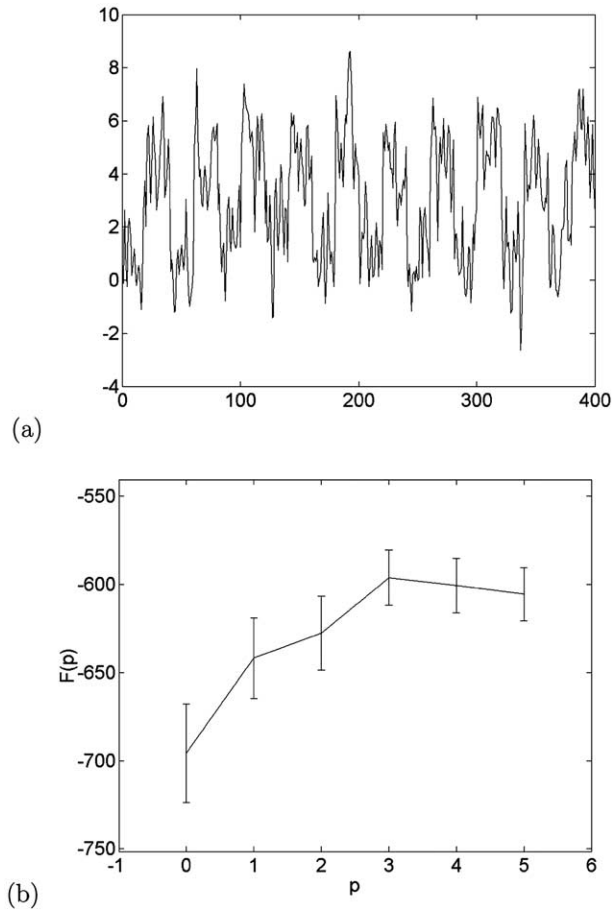

Fig. 5. The figures show (a) an example time series from a GLM-AR model with AR model order of $p = 3$ and (b) a plot of the negative free energy $F(p)$ versus p. This shows that *F(p)* picks out the correct model order.

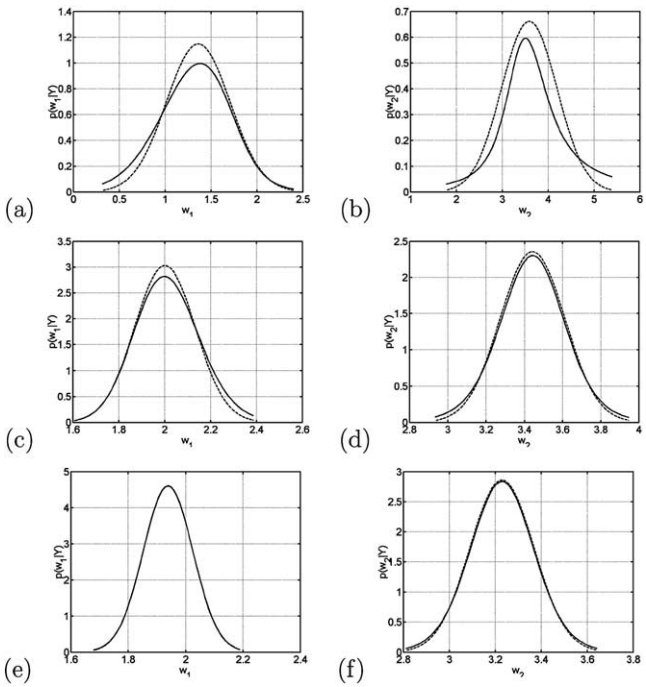

Fig. 6. The figures show the posterior distributions from Gibbs sampling (solid lines) and Variational Bayes (dashed lines) for data sets containing 40 scans (top row), 160 scans (middle row), and 400 scans (bottom row). The distributions in the left column are for the first regression coefficient (size of activation) and in the right column for the second regression coefficient (offset). The fidelity of the VB approximation increases with number of scans.

5.3. Face-repetition data

This data set[1] was recorded during an experiment concerned with the processing of images of faces (Henson et al., 2002). This was an event-related study in which gray-scale images of faces were presented for 500 ms, replacing a baseline of an oval chequerboard which was present throughout the interstimulus interval. Images were acquired from a 2T VISION system (Siemens, Erlangen, Germany) which produced T2*-weighted transverse echo-planar images (EPIs) with blood oxygen level-dependent (BOLD) contrast. Whole brain EPIs consisting of 24 transverse slices were acquired every 2-s resulting in a total of 351 scans. In this paper we restrict our analysis to a single slice at $z = -24$ mm [Talairach coordinates (Talairach and Tournoux, 1988)].

All functional images were realigned to the first functional image using a six-parameter rigid-body transformation (Friston et al., 1995a). To correct for the fact that different slices were acquired at different times, time series were interpolated to the acquisition time of the reference slice (Henson et al., 2002). Images were then spatially normalised to a standard EPI template using a nonlinear warping method (Ashburner and Friston, 1999). We then computed the global mean value, g, over all time series, excluding non-brain voxels, and scaled each time series by the factor $100/g$. After scaling by the peak magnitude of the hemodynamic response function (HRF) (see below) this makes the units of the regression coefficient values ″percentage of global mean signal.” Each time series was then high-pass-filtered using a set of discrete cosine basis functions with a filter cutoff of 120 s.

The data set was analysed using a GLM with a design matrix as shown in Fig. 7. This consists of 19 regressors. The 1st, 3rd, 5th, and 7th are indicator variables, indicating the presentation of a face image, which have been convolved with a “canonical” HRF (Friston et al., 1998). The 2nd, 4th, 6th, and 8th regressors are the corresponding HRF derivatives. Modeling the HRF in this way allows one to capture onset variability across voxels. Regressors 9 to 12 relate to performace errors and 13 to 18 to subject movement and the last regressor is an offset.

The data were then analysed using conventional least-squares SPM and the GLM-AR approach. For the SPM analysis, the images were smoothed using a Gaussian kernel of width 8 mm. For the GLM-AR analysis the images were not smoothed. The results of a standard SPM analysis showing the effect of presenting face images (using a contrast that averages the contributions from the 1st, 3rd, 5th, and 7th regressors) is shown in Fig. 8c. The corresponding structural image is shown in Fig. 8a. The SPM shows bilateral activation of fusiform cortex and earlier visual areas. We also note that many within-brain voxels did not show any BOLD effect due to T2*-signal dropout. The rest of our analysis is restricted to the non-dropout voxels.

We then applied GLM-AR(p) models to each voxel with p varying from 0 to 5. In Fig. 4b we plot a map of the optimal AR model order as computed by the VB approach.

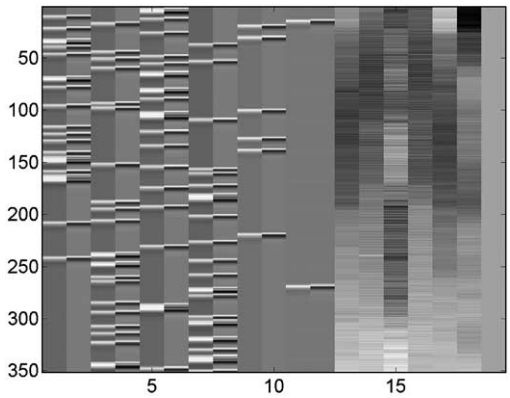

Fig. 7. Design matrix for face-repetition fMRI analysis. There are 19 regressors, 8 relating to the presentation of face images, 4 relating to performance errors, 6 relating to subject movement, and 1 being an offset. The first 12 regressors consist of indicator variables indicating the occurrence of events, such as the presentation of face images to a subject, that have been convolved with a canonical hemodynamic response function or its derivative (Friston et al., 1998).

[1] This data set and a full description of the experiments and data preprocessing are available from http://www.fil.ion.ucl.ac.uk/spm/data.

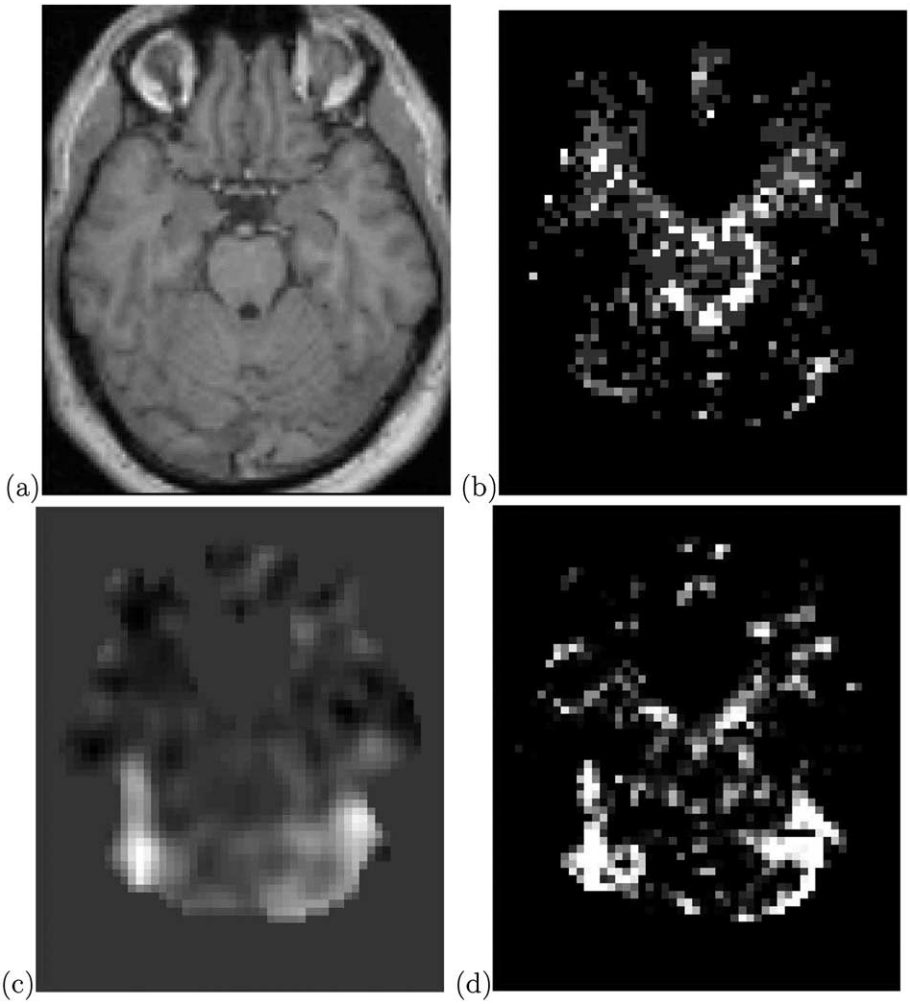

Fig. 8. The figures show (a) a structural MRI image, (b) a map of the optimal AR model order with black being 0 and white being 3, (c) a statistical parametric map of the *t* statistic from an SPM analysis, the background gray shade indicating non-brain voxels and areas of fMRI signal dropout, and (d) a posterior probability map showing the effect of presenting face images. The map shows the probability that the peak effect is greater than 0.5% of the global mean value.

Matching this figure with the structural image in Fig. 4a we see that cerebro-spinal fluid (CSF) voxels typically have a higher model order than gray or white matter voxels. To investigate this further we segmented the structural image into gray, white, and CSF voxels using the algorithm described in (Ashburner and Friston, 2000) and computed histograms of optimal AR model order. These are shown in Fig. 9. We note that, overall, a model order of $p = 3$ is sufficient for all voxels.

In a similar vein, Fig. 10 shows a map of the AR(1) coefficient from the GLM-AR(1) models. This shows a similar pattern to that of the optimal model order map. Tissue-specific boxplots of the AR(1) coefficients in Fig. 11 confirm that, as was also observed in (Bullmore et al., 1996), temporal correlation is stronger in CSF than in gray or white matter.

We then compared VB posteriors with posteriors derived from Gibbs sampling. For this comparison only the first 8 columns of the design matrix (plus an offset) were used in order to reduce the computation time required by the Gibbs sampler. Fig. 12 shows the posteriors for an activated voxel in right fusiform cortex for the four coefficients relating to the presentation of face images. The model order was set to the maximum of $F(p)$ for that voxel ($p = 1$) for both the Gibbs sampler and the VB algorithm. These results are typical of the data set as a whole indicating a very close agreement between Gibbs and VB.

We then took the fitted GLM-AR(1) model for that voxel and generated 100 different data sets from it using different realisations of the noise process. A comparison of VB versus OLS parameter estimates showed that, on average, all 8 of the regression coefficients were estimated more accurately using VB and 5 of them significantly so ($p < 0.05$). This was repeated for a number of voxels with similar results, the improvement being commensurate with strength of correlation.

Finally, in Fig. 8d we plot a posterior probability map (see Friston et al., 2002b) of the effect of presenting images

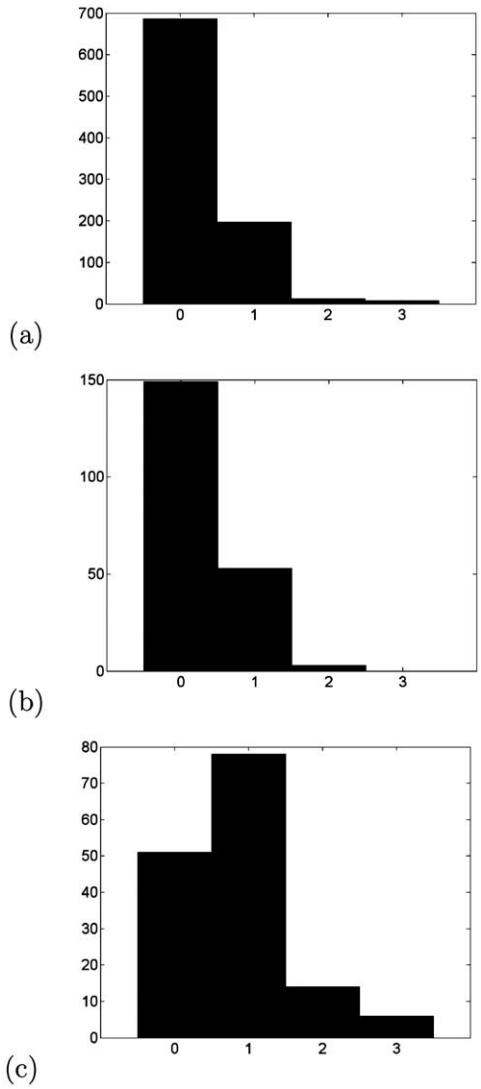

Fig. 9. The figures show histograms of optimal AR model order for (a) gray matter, (b) white matter, and (c) CSF.

of faces. The map shows the probability that the effect is greater than 0.5% of the global mean value. This is similar to the SPM in Fig. 8c in having highly activated voxels in bilateral fusiform cortex. For a discussion of the relation between PPMs and SPMs see Friston et al. (2002b).

6. Discussion

We have described a Bayesian estimation and inference procedure for General Linear Models with Autoregressive error processes of arbitrary order. The algorithm makes use of the VB framework which approximates the true posterior density with a factorised density. The fidelity of this approximation was verified via Gibbs sampling. With low numbers of scans and a high degree of serial correlation the posterior density over regression components is highly non-

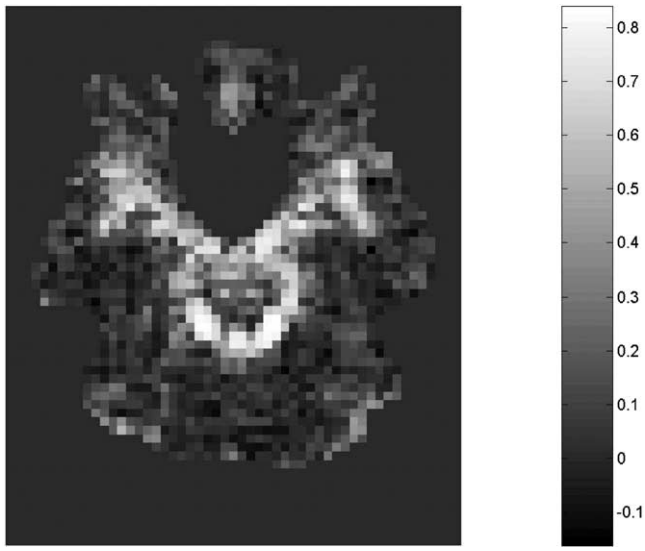

Fig. 10. The figures show a map of the AR(1) coefficient in GLM-AR(1) models of the face data set.

Gaussian showing dependence between autoregressive coefficients and regression coefficients. The corresponding VB posterior is the best matching multivariate Gaussian without such dependence. With the numbers of scans used in current fMRI studies (typically > 100) the true posteriors are well approximated by the VB posteriors. This good agreement has been found on both synthetic and real data.

Although the VB posterior over regression coefficients is Gaussian it is not the same Gaussian as would correspond to the OLS solution. First, the centre of the Gaussian is reestimated to take the autocorrelation into account. This results in consistently better estimates of the true regression coefficients. Secondly, the width of the OLS-Gaussian is a

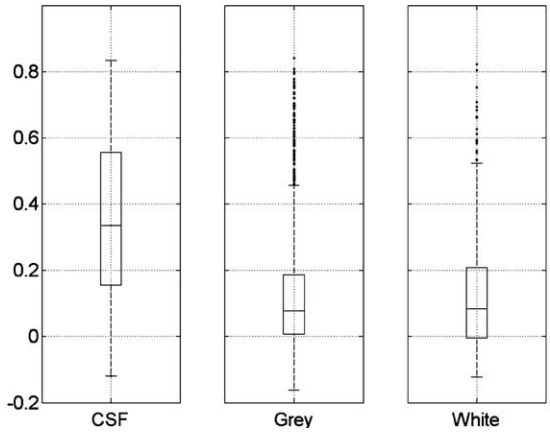

Fig. 11. Box and whisker plots of the autoregressive coefficient from GLM-AR(1) models applied to the face data set for CSF and gray and white matter. The boxes have lines at the lower, median, and upper quartile values. The whiskers extend out to the most extreme value within a distance of one and a half times the interquartile range from the box. Data points outside of the whiskers are drawn as dots.

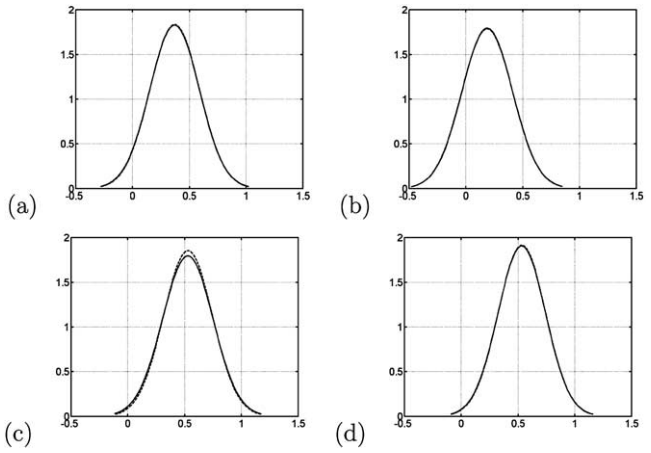

Fig. 12. The figures compare the exact posterior, $p(w_i|Y)$ (solid lines), as computed from Gibbs' sampling with the approximate posterior, $q(w_i|Y)$ (dashed lines), as used in VB for regression coefficients (a) w_1, (b) w_3, (c) w_5, and (d) w_7 from a GLM belonging to a single voxel in the right fusiform activation area of the face-repetition data set.

consistent underestimate of the true width of the posterior, whereas this is not the case with VB (see, e.g., Fig. 6b).

Experiments comparing the accuracy with which VB and OLS estimate activation effects showed VB to be significantly more accurate in data sets with at least 100–200 scans. This shows that it is worthwhile modeling the error autocorrelation and correcting the estimated regression coefficients accordingly. It also shows that a certain minimum amount of data is required in order for the AR coefficients to be estimated well enough for this correction to be beneficial. This improvement over OLS will also be shared by other iterative algorithms such as the Expectation-Maximisation (EM) algorithms described in Friston et al. (2002b) and Worsley et al. (2002).

The VB approach provides a natural extension to these algorithms, however, in that the variability of hyperparameter estimation is also taken into account. This is achieved with little additional computational effort. Specifically, the objective function which is maximised during model fitting contains a penalty term consisting of the KL divergence between the prior over hyperparameters and the approximate posterior. In this way, model overfitting is prevented. This constitutes the VB solution to the overconfidence problem (Friston et al., 2002a). Further, VB allows for automatic selection of AR order.

In an exploratory analysis of event-related fMRI data the optimal AR order was seen to be higher in CSF than in gray or white matter. Overall, an AR(3) model was seen to be sufficient for all voxels. Also, the magnitude of the first AR coefficient was seen to be higher in CSF. This confirms earlier observations (Bullmore et al., 1996; Worsley et al., 2002) that the values of AR coefficients have spatial structure. On other data sets Woolrich et al. (2001) observed stronger correlation in gray matter than in white matter or CSF. These observations confirm that there is a physiological component to the autocorrelation, whereas earlier investigations using phantoms suggested that this correlation might be purely due to the physics of the measurement process (Zarahn et al., 1997).

In this paper we have used the order criterion furnished by VB for "model selection." Using it we have established that GLM models with low-order AR error processes are suitable for fMRI data analysis. We also note that the order criterion could be used for "model averaging" (Gelman et al., 1995) in which, rather than selecting the "best" model order, we average over model orders using the criterion as a weighting factor. This approach is, for example, used routinely in Bayesian wavelet analysis (Clyde et al., 1998).

Currently the VB algorithm is implemented in MATLAB (Mathworks, Inc.) and requires 30 min on a high-end computer to analyse a single slice of data. This is an order of magnitude faster than Gibbs sampling. With an optimised compiled software implementation this could be reduced further.

This paper has focused on voxel-wise Bayesian modeling of fMRI time series in which priors over the regression and autoregressive coefficients were set to be vague. The next step is to tie together the voxel-wise models using informative priors where voxel-wise parameter estimates will be informed by data from other voxels and other subjects (see, e.g., Friston et al., 2002a; Worsley et al., 2002). In this way, quantities such as the prior precisions on regression and autoregressive parameters (α and β) can be estimated rather than set to arbitrary values. This will ultimately lead to a multiple-subject random effects model with Bayesian regularisation.

Acknowledgments

All authors are supported by the Wellcome Trust. We also thank Rik Henson for providing data and advice.

Appendix A: Gamma density

We define the Gamma density

$$p(x) = \mathsf{Ga}(b, c) \tag{36}$$

as

$$\mathsf{Ga}(b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right). \tag{37}$$

In the derivations that follow in the next section we will refer to the log of a gamma density

$$\log p(x) = -\log \Gamma(c) - c \log b + (c-1)\log x - \frac{x}{b}. \tag{38}$$

Note that the mean and variance of a Gamma variate are bc and b^2c.

Appendix B: Derivation of VB algorithm

B.1. Autoregressive coefficients

We first note that the log-likelihood in Eq. (10) can be expressed as a quadratic function of a and dropping all terms not dependent on a we get

$$\log p(Y|w, a, \lambda) = -\frac{\lambda}{2}(aC(w)a^T - 2D(w)a^T), \tag{39}$$

where

$$C(w) = \sum_t (d_t - \tilde{X}_t w)\,(d_t - \tilde{X}_t w)^T$$
$$D(w) = \sum_t (y_t - x_t w)\,(d_t - \tilde{X}_t w)^T. \tag{40}$$

From Eq. (20) we see that

$$q(a|Y) \propto \exp[I(a)], \tag{41}$$

where

$$I(a) = \int\int q(\theta|Y)\,(\log p(Y|\theta) + \log p(\theta))d\theta. \tag{42}$$

This gives

$$I(a) = \int q(w|Y)q(\lambda|Y)\log p(Y|w, a, \lambda)dwd\lambda + \log p(a|\beta) \tag{43}$$
$$= -\frac{\bar{\lambda}}{2}(a\tilde{C}a^T - 2\tilde{D}a^T) + \beta aa^T + \cdots \tag{44}$$

where

$$\tilde{C} = \int q(w|Y)C(w)dw$$
$$\tilde{D} = \int q(w|Y)\,D(w)dw. \tag{45}$$

Given that $I(a)$ can be expressed as the quadratic

$$I(a) = -\frac{1}{2}[a(\bar{\lambda}\tilde{C} + \beta I)a^T - 2\bar{\lambda}\tilde{D}a^T] + \cdots \tag{46}$$

and that the log of a Gaussian density $p(a)$ with mean m (a row vector) and covariance V, including only a-dependent terms, is

$$\log p(a) = -\frac{1}{2}[aV^{-1}a^T - 2mV^{-1}a^T], \tag{47}$$

we see that

$$q(a|Y) = \mathsf{N}(m, V), \tag{48}$$

where

$$V = (\bar{\lambda}\tilde{C} + \beta I)^{-1}$$
$$m = \bar{\lambda}\tilde{D}V. \tag{49}$$

It now remains to compute the integrals $\tilde{C}$ and $\tilde{D}$ which are given as follows

$$\tilde{C} = \sum_t d_t d_t^T + \tilde{X}_t(\hat{w}\hat{w}^T + \hat{\sum})\tilde{X}_t^T - d_t\hat{w}^T\tilde{X}_t^T - \tilde{X}_t\hat{w}d_t^T$$
$$\tilde{D} = \sum_t y_t d_t^T - x_t\hat{w}d_t^T - y_t\hat{w}^T\tilde{X}^T + x_t(\hat{w}\hat{w}^T + \hat{\Sigma})\tilde{X}_t^T. \tag{50}$$

If instead of integrating out the dependence on $q(w|Y)$ we simply use the point estimate $\hat{w}$, then $\tilde{C} = \tilde{E}\tilde{E}^T$ and $\tilde{D} = \tilde{E}E$ where the elements of $\tilde{E}$ are now given by $\tilde{E}_t = d_t - \tilde{x}_t\hat{w}$. If we also have no prior on the AR coefficients, i.e., $\beta = 0$, we then recover the ML update (see Eq. (26)]

$$m = (\tilde{E}\tilde{E}^T)^{-1}\tilde{E}E. \tag{51}$$

B.2. Regression coefficients

The regression coefficients are derived in much the same way. We first note that the log-likelihood in Eq. (9) can be expressed as a quadratic function of w and dropping all terms not dependent on w we get

$$\log p(Y|w, a, \lambda) = -\frac{\lambda}{2}(w^T A(a)w - 2B(a)w), \tag{52}$$

where

$$A(a) = \sum_t (x_t - a\tilde{X}_t)(x_t - a\tilde{X}_t)^T$$
$$B(a) = \sum_t (y_t - ad_t)(x_t - a\tilde{X}_t)^T. \tag{53}$$

From Eq. (20) we see that

$$q(w|Y) \propto \exp[I(w)], \tag{54}$$

where

$$I(w) = \int q(\theta|Y)\,(\log p(Y|\theta) + \log p(\theta))\,d\theta. \tag{55}$$

This gives

$$I(w) = \int\int q(a|Y)q(\lambda|Y)\,\log p(Y|w,\,a,\,\lambda)dad\lambda + \log p(w|\alpha) \quad (56)$$

$$= -\frac{\bar{\lambda}}{2}(w^T\tilde{A}w - 2\tilde{B}w) + \alpha w^T w + \cdots, \quad (57)$$

where

$$\tilde{A} = \int q(a|Y)A(a)da$$

$$\tilde{B} = \int q(a|Y)B(a)da. \quad (58)$$

Given that $I(w)$ can be expressed as the quadratic

$$I(w) = -\frac{1}{2}[w^T(\bar{\lambda}\tilde{A} + \alpha I)w - 2\bar{\lambda}\tilde{B}w] + \cdots \quad (59)$$

and that the log of a Gaussian density $p(w)$ with mean $\hat{w}$ (a column vector) and covariance $\hat{\Sigma}$, including only w-dependent terms, is

$$\log p(w) = -\frac{1}{2}[w^T\hat{\Sigma}^{-1}w - 2\hat{w}^T\hat{\Sigma}^{-1}w], \quad (60)$$

we see that

$$q(w|Y) = \mathsf{N}(\hat{w},\,\hat{\Sigma}), \quad (61)$$

where

$$\hat{\Sigma} = (\bar{\lambda}\tilde{A} + \alpha I)^{-1}$$

$$\hat{w} = \hat{\Sigma}\bar{\lambda}\tilde{B}^T. \quad (62)$$

It now remains to compute the integrals $\tilde{A}$ and $\tilde{B}$ which are given by

$$\tilde{A} = \sum_t x_t^T x_t + \tilde{X}_t^T(m^T m + V)\tilde{X}_t - x_t^T m\tilde{X}_t - \tilde{X}_t^T m^T x_t \quad (63)$$

$$\tilde{B} = \sum_t y_t x_t - m d_t x_t - y_t m\tilde{X}_t + d_t^T(m^T m + V)\tilde{X}_t. \quad (64)$$

Note that for the special case in which the errors E are uncorrelated, i.e., $m = 0$, we have $\tilde{A} = X^T X$ and $\tilde{B} = X^T Y$. If we also have no prior on the regression coefficients, i.e., $\alpha = 0$, we then recover the OLS update [see Eq. (24)]

$$\hat{w} = (X^T X)^{-1}X^T Y. \quad (65)$$

B.3. Noise precision

We write the log-likelihood in Eq. (9), dropping all terms not dependent on λ, as

$$\log p(Y|\theta) = -\frac{\lambda}{2}G(w,\,a) + \frac{N-p}{2}\log\lambda, \quad (66)$$

where

$$G(w,\,a) = \sum_t ((y_t - ad_t) - (x_t w - a\tilde{X}_t w))^2. \quad (67)$$

From Eq. (20) we see that

$$q(\lambda|Y) \propto \exp[I(\lambda)], \quad (68)$$

where

$$I(\lambda) = \int q(\theta|Y)(\log p(Y|\theta) + \log p(\theta))d\theta. \quad (69)$$

This gives

$$I(\lambda) = -\frac{\lambda}{2}\int\int q(w|Y)q(a|Y)G(w,\,a)dwda + \frac{N-p}{2}\log\lambda + \log p(\lambda)$$

$$= -\frac{\lambda}{2}\tilde{G} + \frac{N-p}{2}\log\lambda + \log p(\lambda), \quad (70)$$

where

$$\tilde{G} = \int\int q(w|Y)q(a|Y)G(w,\,a)dwda. \quad (71)$$

Substituting for $\log p(\lambda)$ and keeping only λ-dependent terms give

$$I(\lambda) = -\frac{\lambda}{2}\tilde{G} + \frac{N-p}{2}\log\lambda + (c_0 - 1)\log\lambda - \frac{\lambda}{b_0} \quad (72)$$

$$= -\lambda\left(\frac{\tilde{G}}{2} + \frac{1}{b_0}\right) + \left(\frac{N-p}{2} + c_0 - 1\right)\log\lambda. \quad (73)$$

Comparing this with the log of a gamma density in appendix A we see that

$$q(\lambda|Y) = \mathsf{Ga}(b_\lambda,\,c_\lambda), \quad (74)$$

where

$$\frac{1}{b_\lambda} = \frac{\tilde{G}}{2} + \frac{1}{b_0}$$

$$c_\lambda = \frac{N-p}{2} + c_0. \tag{75}$$

Note that the mean of this density is

$$\bar{\lambda} = b_\lambda c_\lambda. \tag{76}$$

It now remains to compute the integral

$$\begin{aligned}\tilde{G} &= \sum_t \int\int q(w|Y)q(a|Y)((y_t - ad_t) \\ &\quad - (x_t w - a\tilde{X}_t w))^2 dwda \\ &= \tilde{G}_1 + \tilde{G}_2 + \tilde{G}_3,\end{aligned} \tag{77}$$

where

$$\tilde{G}_1 = \sum_t \int q(a|Y)(y_t - ad_t)^2 da$$

$$\tilde{G}_2 = \sum_t \int\int q(w|Y)q(a|Y)(x_t w - a\tilde{X}_t w)^2 dwda$$

$$\begin{aligned}\tilde{G}_3 = &-2\sum_t \int\int q(w|Y)q(a|Y)(y_t - ad_t) \\ &\times (x_t w - a\tilde{X}_t w) dwda.\end{aligned} \tag{78}$$

These integrals can be evaluated as

$$\tilde{G}_1 = \sum_t y_t^2 + d_t^T(m^T m + V)d_t - 2y_t d_t^T m \tag{79}$$

$$\begin{aligned}\tilde{G}_2 = &\sum_t x_t(\hat{w}\hat{w}^T + \hat{\sum})x_t^T + \mathrm{Tr}(\tilde{X}_t^T(m^T m + V)\tilde{X}_t\hat{\sum}) \\ &+ \hat{w}^T\tilde{X}_t^T(m^T m + V)\tilde{X}_t\hat{w} - 2x_t(\hat{w}\hat{w}^T + \hat{\sum})\tilde{X}_t m^T\end{aligned} \tag{80}$$

$$\begin{aligned}\tilde{G}_3 = &\sum_t -2y_t x_t\hat{w} + 2md_t x_t\hat{w} \\ &+ 2y_t m\tilde{X}_t\hat{w} - 2d_t^T(m^T m + V)\tilde{X}_t\hat{w}.\end{aligned} \tag{81}$$

B.4. Negative free energy

From Eq. (17) we have

$$\begin{aligned}F &= \int q(\theta|Y)\log\frac{p(Y,\theta)}{q(\theta|Y)}d\theta \\ &= L_{av} - KL_{prior},\end{aligned} \tag{82}$$

where

$$L_{av} = \int q(\theta|Y)\log p(Y|\theta)d\theta \tag{83}$$

$$KL_{prior} = \int q(\theta|Y)\log\frac{q(\theta|Y)}{p(\theta)}d\theta. \tag{84}$$

Now, from Eq. (15) we have

$$p(\theta) = p(w|\alpha)p(a|\beta)p(\lambda) \tag{85}$$

and from Eq. (22)

$$q(\theta|Y) = q(w|Y)q(a|Y)q(\lambda|Y). \tag{86}$$

Hence

$$KL_{prior} = KL(w) + KL(a) + KL(\lambda), \tag{87}$$

where, for a generic parameter θ_i, $KL(\theta_i)$ denotes the KL divergence between the approximate posterior $q(\theta_i|Y)$ and the prior $p(\theta_i)$. Expressions for the KL divergences for the various densities are given in (Roberts and Penny, 2002). The average log-likelihood is given by

$$\begin{aligned}L_{av} = &\int\int\int q(w|Y)q(a|Y)q(\lambda|Y) \\ &\times \log p(Y|w, a, \lambda)dwdad\lambda.\end{aligned} \tag{88}$$

From Eq. (9)

$$\log p(Y|w, a, \lambda) = -\frac{\lambda}{2}G(w,a) + \frac{N-p}{2}\log\lambda, \tag{89}$$

where

$$G(w, a) = \sum_t((y_t - ad_t) - (x_t w - a\tilde{X}_t w))^2. \tag{90}$$

Hence,

$$\begin{aligned}L_{av} = &-\frac{\bar{\lambda}}{2}\tilde{G} + \frac{N-p}{2}\int q(\lambda|Y)\log\lambda d\lambda \\ &+ \frac{N-p}{2}\log 2\pi \\ = &\frac{N-p}{2}\log\bar{\lambda} - \frac{\bar{\lambda}}{2}\tilde{G} - \frac{N-p}{2}\log 2\pi,\end{aligned} \tag{91}$$

where $\tilde{G}$ is given in Eq. (77) and

$$\begin{aligned}\log\tilde{\lambda} &= \int q(\lambda|Y)\log\lambda d\lambda \\ &= \psi(c_\lambda) + \log b_\lambda\end{aligned} \tag{92}$$

and $\psi()$ is the digamma function.

References

Ashburner, J., Friston, K.J., 1999. Nonlinear spatial normalization using basis functions. Hum. Brain Mapp. 7, 254–266.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11, 805–821.

Attias, H., 2000. A variational Bayesian framework for graphical models, in: Leen, T., et al. (Eds.), NIPS 12. MIT Press, Cambridge, MA.

Barnett, V., 1999. Comparative Statistical Inference. Wiley, New York.

Bernardo, J.M., Smith, A.F.M., 2000. Bayesian Theory. Wiley, New York.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford Univ. Press, Oxford.

Blake, A., Isard, M., 1998. Active Contours. Springer-Verlag, Berlin.

Box, G.E.P., Tiao, G.C., 1992. Bayesian Inference in Statistical Analysis. Wiley, New York.

Bullmore, E., Brammer, M., Williams, S., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P., 1996. Statistical methods of estimation and inference for functional MR image analysis. Magn. Reson. Med. 35, 261–277.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall, London.

Casella, G., Berger, R., 1990. Statistical Inference. Duxbury, N. Scituate, MA.

Clyde, M., Parmigiani, G., Vidakovic, B., 1998. Multiple shrinkage and subset selection selection in wavelets. Biometrika 85, 391–402.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley, New York.

Friston, K.J., Ashburner, J., Frith, C.D., Poline, J.-B., Heather, J.D., Frackowiak, R.S.J., 1995a. Spatial registration and normalization of images. Hum. Brain Mapp. 2, 165–189.

Friston, K.J., Holmes, A.P., Poline, J.-B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S.J., Turner, R., 1995b. Analysis of fMRI time series revisited. NeuroImage 2, 45–53.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-B., Frith, C.D., Frackowiak, R.S.J., 1995c. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-related fMRI: characterizing differential responses. NeuroImage 7, 30–40.

Friston, K.J., Glaser, D., Henson, R., Kiebel, S., Phillips, C., Ashburner, J., 2002a. Classical and Bayesian inference in neuroimaging: applications. NeuroImage 16, 484–512.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002b. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16, 465–483.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall, London.

Henson, R.N.A., Shallice, T., Gorno-Tempini, M.L., Dolan, R.J., 2002. Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cereb. Cortex 12, 178–186.

Jordan, M.I., 1999. Learning in Graphical Models. MIT Press, Cambridge, MA, (Ed.).

Kiebel, S.J. Glaser, D. Friston, K.J. A heuristic for the degrees of freedom of statistics based on multiple hyperparameters. Technical report, manuscript in preparation, 2002a.

Kiebel, S.J. Penny, W.D. Friston, K.J. Application of the Gibbs sampler to fMRI data. Manuscript in preparation, 2002b.

Lappalainen, H., Miskin, J.W., 2000. Ensemble Learning, in: Girolami, M. (Ed.), Advances in Independent Component Analysis. Springer-Verlag, Berlin.

Locascio, J., Jennings, P., Moore, C., Corkin, S., 1997. Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. Hum. Brain Mapp. 5, 168–193.

Neumaier, A. Schneider, T. 2000. Estimation of parameters and eigenmodes of multivariate autoregressive models. Submitted for publication.

O'Ruaniaidh, J.J.K., Fitzgerald, W.J., 1996. Numerical Bayesian Methods Applied to Signal Processing. Springer, Berlin.

Purdon, P.L., Weisskoff, R., 1998. Effect of temporal autocorrelations due to physiological noise stimulus paradigm on voxel-level false positive rates in fMRI. Hum. Brain Mapp. 6, 239–249.

Roberts, S.J., Penny, W.D., 2002. Variational Bayes for generalised autoregressive models. IEEE Trans. Signal Process 50, 2245–2257.

Talairach, J., Tournoux, P., 1988. Coplanar Stereotaxic Atlas of the Human Brain. Thieme Medical, New York.

Woolrich, Mark W., Ripley, Brian D., Brady, Michael, Smith, Stephen M., 2001. Temporal autocorrelation in univariate linear modelling of fMRI data. NeuroImage 14, 1370–1386.

Worsley, K.J. Liao, C.H. Aston, J. Petre, V. Duncan, G.H. Morales, F. Evans, A.C. A general statistical analysis for fMRI data. NeuroImage 15.2002.

Yandell, B.S., 1997. Practical Data Analysis for Designed Experiments. Chapman & Hall, London.

Zarahn, E., Aguirre, G.K., D'Esposito, M., 1997. Empirical analysis of BOLD fMRI statistics. 1. Spatially unsmoothed data collected under null-hypothesis conditions. NeuroImage 5, 179–197.