ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 24 (2005) 350–362

Bayesian fMRI time series analysis with spatial priors

William D. Penny,[a,*] Nelson J. Trujillo-Barreto,[b] and Karl J. Friston[a]

[a]*Wellcome Department of Imaging Neuroscience, UCL, London, UK*
[b]*Cuban Neuroscience Center, Havana, Cuba*

Received 12 March 2004; revised 5 July 2004; accepted 25 August 2004

We describe a Bayesian estimation and inference procedure for fMRI time series based on the use of General Linear Models (GLMs). Importantly, we use a spatial prior on regression coefficients which embodies our prior knowledge that evoked responses are spatially contiguous and locally homogeneous. Further, using a computationally efficient Variational Bayes framework, we are able to let the data determine the optimal amount of smoothing. We assume an arbitrary order Auto-Regressive (AR) model for the errors. Our model generalizes earlier work on voxel-wise estimation of GLM-AR models and inference in GLMs using Posterior Probability Maps (PPMs). Results are shown on simulated data and on data from an event-related fMRI experiment.

Keywords: Variational Bayes; fMRI; Spatial priors; Effect-size; General linear model; Autoregressive model; Laplacian; Smoothing

Introduction

Functional magnetic resonance imaging (fMRI) using blood oxygen level-dependent (BOLD) contrast is an established method for making inferences about regionally specific activations in the human brain (Frackowiak et al., 2004). From measurements of changes in blood oxygenation, one uses various statistical models, such as the general linear model (GLM) (Friston et al., 1995), to make inferences about task-specific changes in underlying neuronal activity.

Given an impulse of neuronal activity, the BOLD signal we measure is dispersed both in space and time according to a hemodynamic response function (HRF). The temporal characteristics of this dispersion are determined by various time and elasticity constants and hemodynamic processes related to the underlying vasculature and can be described using the Balloon model and variants thereof (Buxton et al., 1998). They can also be modeled in the GLM framework by convolving putative neuronal signals with a set of hemodynamic basis functions, for example, the so-called canonical HRF and its derivatives (Friston et al., 1998). Essentially, BOLD activity peaks 4 to 6 s after neuronal activity and experiences a marked undershoot after 10 to 12 s, returning to baseline after 20 to 30 s.

The spatial characteristics of the hemodynamic response relate to the geometry of the underlying vasculature. BOLD contrast arises mainly from oxygenation changes in small venules lying relatively close to the site of neuronal activity (Turner, 2002). It is also possible that signal can appear in larger pial veins draining activated areas. It is also of note that BOLD contrast only arises from activity in spatially extended neuronal ensembles. Additionally, there are a number of signal processing contributions to the spatial nature of the fMRI signals. These arise, for example, from realignment and spatial normalization operations that involve the use of spatial basis functions and spatial interpolation. Overall, the spatial extent of the resulting BOLD signal is of the order of several millimeters.

In the GLM framework, the spatial nature of the BOLD response is accounted for indirectly by smoothing the data in a preprocessing step (Frackowiak et al., 2004). This is implemented by smoothing using fixed-width Gaussian kernels having a Full Width at Half Maximum (FWHM) of typically 4–8 mm for single-subject analysis. The rationale for this approach comes from the matched filter theorem (Rosenfeld and Kak, 1982) which states that by changing the frequency structure of the data to that of the signal of interest one increases the Signal to Noise Ratio (SNR). For a discussion of this theorem in the context of neuroimaging, see Section 2.2 of Worsley et al. (1996).

The well-known temporal characteristics of the hemodynamic response vary across the brain and across subjects, and are accounted for in the GLM framework by using multiple basis functions per experimental manipulation. The regression coefficients corresponding to these bases are voxel- and subject-dependent allowing the peak onset times and widths to vary. The spatial characteristics of the response, however, which may also vary across the brain and across subjects are not explicitly

* Corresponding author. Wellcome Department of Imaging Neuroscience, 12 Queen Square, London WC1N 3BG, UK. Fax: +44 20 7833 7478.
E-mail address: wpenny@fil.ion.ucl.ac.uk (W.D. Penny).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2004.08.034

addressed in the standard GLM framework although a number of studies have proposed methods for dealing with this.

In the context of PET, Worsley et al. (1996) have proposed a scale-space procedure for assessing significance of activations over a range of proposed smoothings. For fMRI, Penny and Friston (2003) have proposed using a Mixture of General Linear Models (MGLMs). These procedures are motivated by the fact that smoothing images with a fixed width kernel is a sub-optimal method for increasing SNR.

In this paper, we characterize the spatial characteristics of the HRF using Bayesian inference and spatial priors over the regression coefficients. The precision with which regression coefficients, and therefore regionally specific effects, are estimated then comprises two contributions (i) the data at a given voxel and (ii) the regression coefficients at neighboring voxels. If data precision is low (e.g., due to high noise variance at that voxel), then neighboring voxels will contribute more to the estimate of the effect. This spatial regularization falls naturally out of the Bayesian framework. Moreover, we are able to use spatial regularization coefficients that can be estimated from the data. The spatial characteristics of the hemodynamic response are therefore handled in a natural and automatic way.

A further important issue in the analysis of fMRI data is the concern that successive samples are correlated in time. 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. Not all of this correlation can be removed by time-domain filtering as this would also remove much of the BOLD signal. Cardiac and respiratory components can, however, be monitored and then removed from the data (Glover et al., 2000; Hu et al., 1995). But correlations due, for example, to unmodeled neuronal activity will remain. In the GLM framework, temporal autocorrelation can be taken into account by modeling the errors as an Autoregressive (AR) process, as shown, for example, in our previous work (Penny et al., 2003). This is the approach taken in this paper.

In the recent literature, there have been a number of Baysian approaches to modeling spatial dependencies in the signal and in the noise. For example, Gossl et al. (2001) have proposed a separable spatio-temporal model where these spatial dependencies were characterized using Conditional Autoregressive (CAR) or Markov Random Field (MRF) priors. They used Bayesian inference and Markov Chain Monte Carlo (MCMC) to draw samples from the relevant posterior distributions. More recently, Woolrich et al. (2004b) described a Bayesian model of fMRI in which the noise process was characterized by separable or non-separable spatio-temporal models. Again, MCMC was used to perform posterior inference. While these approaches have clearly broken new ground in the analysis of fMRI data, their main drawback is the large amount of computer time required (several hours for a single slice (Woolrich et al., 2004b)). This motivated the more recent work in which Woolrich et al. (2004a) specified a CAR/MRF model to regularize estimation of AR coefficients using the Variational Bayes (VB) framework. This resulted in an algorithm that could process whole volumes of fMRI data in the order of minutes.

This paper also makes use of the VB approach and may be regarded as an extension of Penny et al. (2003) to include spatial priors for the regression coefficients. A key technical contribution of this paper is that we use a prior that captures dependencies across voxels but a (approximate) posterior that factorizes over voxels. This means that we can avoid the inversion of very large covariance matrices. This is made possible using the VB framework and results in an algorithm that both captures spatial dependencies and can be efficiently implemented.

In Theory, we review the GLM-AR model defined in Penny et al. (2003). We also describe the priors and show how Variational Bayes is used to define approximate posteriors and how it provides a set of update equations for the sufficient statistics of these distributions. Results present synthetic data and an event-related fMRI data set.

Notation

A multivariate normal density over $\mathbf{x}$ is written as $N(\mathbf{x}; \mathbf{m}, \boldsymbol{\Sigma})$, where $\mathbf{m}$ denotes the mean and $\boldsymbol{\Sigma}$ the covariance. A Gamma distribution over $\mathbf{x}$ is written where a and b define the density as shown in the appendix of Penny et al. (2003). We will denote matrices and vectors with bold upper case and bold lower case letters, respectively. All vectors are column vectors. Subscripts are used to name different vectors; thus $\boldsymbol{x}_n$ and $\boldsymbol{x}_k$ refer to different vectors. The operator *diag* $(\mathbf{x})$ turns a vector into a diagonal matrix, $\mathbf{x}(k)$ denotes the kth entry in a vector, $\boldsymbol{X}(j,k)$ the scalar entry in the jth row and kth column, $\otimes$ is the Kronecker product and $\bar{x}$ is used to denote the mean of x.

Theory

We write an fMRI data set consisting of T time points at N voxels as the $T \times N$ matrix $\mathbf{Y}$. In mass-univariate models, these data are explained in terms of a $T \times K$ design matrix $\mathbf{X}$, containing the values of K regressors at T time points, and a $K \times N$ matrix of regression coefficients $\mathbf{W}$, containing K regression coefficients at each of N voxels. The model is written

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{E} \tag{1}$$

where $\mathbf{E}$ is a $T \times N$ error matrix. The vector $\boldsymbol{w}_{\mathrm{n}}$, the nth column of $\mathbf{W}$, therefore contains the K regression coefficients at the nth voxel and the vector $\boldsymbol{w}_k^T$, the kth row of $\mathbf{W}$, contains an image (after appropriate reshaping) of the kth regression coefficients. We also make use of the $KN \times 1$ vector $\boldsymbol{w}_{\mathrm{v}}$ which contains all the elements of $\mathbf{W}$ ordered by voxel. Similarly we define the $KN \times 1$ vector $\boldsymbol{w}_{\mathrm{r}}$ which also contains all elements of $\mathbf{W}$ but ordered by regressor. These can both be defined using the *vec* operator which stacks columns of a matrix into one long vector

$$\boldsymbol{w}_v = vec(\mathbf{W})$$

$$\boldsymbol{w}_r = vec\left(\mathbf{W}^T\right)$$

$$\boldsymbol{w}_v = \mathbf{H}\boldsymbol{w}_r \tag{2}$$

where $\mathbf{H}$ is a $KN \times KN$ permutation matrix. It is useful to define these high-dimensional vectors as the model can then be instantiated using sparse matrix operations. This notation is used in the derivation of the VB algorithm in the appendix.

In this paper, the errors are modeled as an autoregressive process. The overall GLM-AR model can be written

$$\boldsymbol{y}_n = \mathbf{X}\mathbf{w}_n + \boldsymbol{e}_n$$

$$\boldsymbol{e}_n = \tilde{\mathbf{E}}_n \boldsymbol{a}_n + \boldsymbol{z}_n \tag{3}$$

where, at the nth voxel, $\boldsymbol{a}_n$ is a $P \times 1$ vector of regression coefficients, $\boldsymbol{z}_n$ is a vector of zero mean Gaussian random variables each having precision λ_n and $\tilde{\mathbf{E}}_n$ is a $T \times P$ matrix of lagged prediction errors for the nth voxel as defined in Section 2 of Penny et al. (2003). Eqs. (1) and (3) define the likelihood of the data given the parameters of our model. In the following section, we describe the prior distributions over these parameters. Together, the likelihood and prior terms define our probabilistic model, which is portrayed graphically in Fig. 1.

Priors

Regression coefficients

The prior over regression coefficients is given by

$$p(\mathbf{W}) = \prod_{k=1}^{K} p\left(\boldsymbol{w}_k^T\right)$$
$$p\left(\boldsymbol{w}_k^T\right) = N\left(\boldsymbol{w}_k^T; \mathbf{0}, \alpha_k^{-1}\left(\mathbf{S}^T\mathbf{S}\right)^{-1}\right) \tag{4}$$

where we refer to $\mathbf{S}$ as an $N \times N$ spatial kernel matrix (to be defined later) and α_k is a spatial precision variable for the kth regressor. This equation shows that the prior factorizes over regressors. This means that different regression coefficients can have different smoothnesses. For the case of $\mathbf{S}$ being the Laplacian operator (see below), a sample from the prior is shown in Fig. 2. The value of α_k determines the amount of smoothness and in this paper α_k is estimated from the data. Spatial regularization is therefore fully automatic.

Spatial precisions

The precision variables α_k are collected together in the $K \times 1$ vector $\boldsymbol{\alpha}$. Although, in this paper, each component of $\boldsymbol{\alpha}$ is to be estimated from the data, in future we envisage constraining these estimates using prior information. To this end, we define a prior over $\boldsymbol{\alpha}$

$$p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} p(\alpha_k)$$
$$p(\alpha_k) = Ga(\alpha_k; q_1, q_2) \tag{5}$$

The parameters are set to $q_1 = 10$ and $q_2 = 1$ which corresponds to a Gamma density with a mean of 1 and a variance of 100. It is therefore a relatively uninformative prior reflecting our lack of knowledge about α_k. As we gather more experience applying such priors to fMRI data, we envisage, in the future, choosing q_1 and q_2 to reflect this knowledge.

Noise precisions

The observation noise precisions are defined as

$$p(\boldsymbol{\lambda}) = \prod_{n=1}^{N} p(\lambda_n)$$
$$p(\lambda_n) = Ga(\lambda_n; u_1, u_2) \tag{6}$$

The values u_1 and u_2 are set so as to make $p(\lambda_n)$ an uninformative prior as described in the previous section. The factorization in the above equation assumes that there is no correlation between the variances of neighboring voxels. While this is clearly untrue, this is nevertheless the implicit assumption underlying most GLM analyses (Friston et al., 1995), as only data at voxel n is used to estimate λ_n.

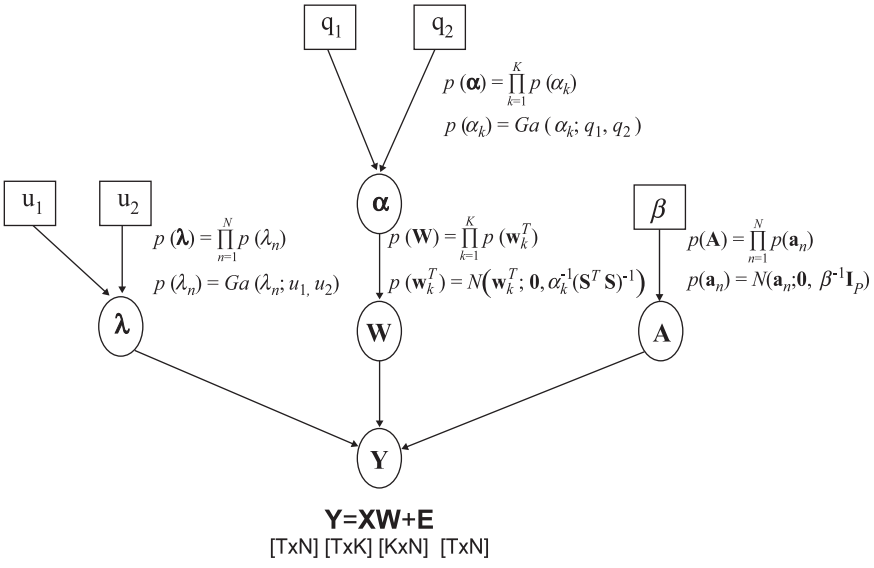

Fig. 1. The figure shows the probabilistic dependencies underlying our generative model for fMRI data. The quantities in square brackets are constants and those in circles are random variables. The spatial regularization coefficients $\boldsymbol{\alpha}$ constrain the regression coefficients $\mathbf{W}$. The parameters $\boldsymbol{\lambda}$ and $\mathbf{A}$ define the autoregressive error processes which contribute to the measurements. The graph shows that the joint probability of parameters and data can be written $p(\mathbf{Y}, \mathbf{W}, \mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = p(\mathbf{Y}|\mathbf{W}, \mathbf{A}, \boldsymbol{\lambda})\, p(\mathbf{W}|\boldsymbol{\alpha})\, p(\mathbf{A}|\beta)\, p(\boldsymbol{\lambda}|u_1, u_2)\, p(\boldsymbol{\alpha}|q_1, q_2)$, where the first term is the likelihood and the other terms are priors. The equations describe the prior distributions over model parameters.

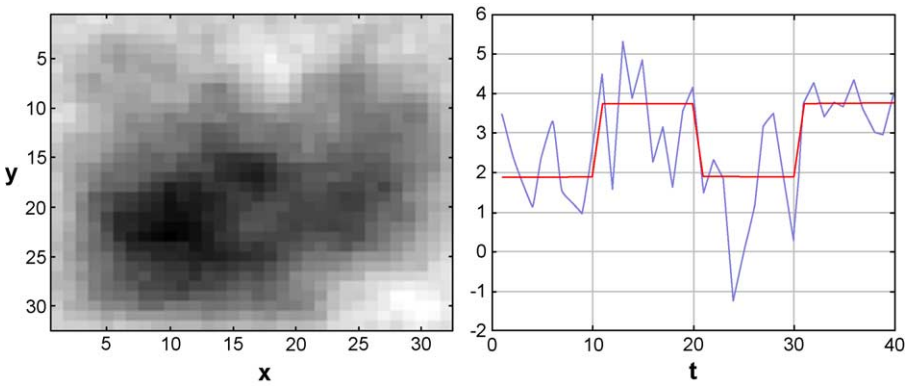

Fig. 2. The left panel shows an image of a regression coefficient generated from a Laplacian prior with $\alpha = 1$ as described in the text. These regression coefficients were then used, along with a boxar design matrix to generate time series at each voxel. The right panel shows a time series corresponding to the voxel $x = 29$, $y = 28$. The darker line indicates the model ($\mathbf{X}\mathbf{w}_n$) and the lighter line indicates the final simulated data (the model with additive Gaussian noise).

AR coefficients

The priors on the autoregressive parameters are given by

$$p(\mathbf{A}) = \prod_{n=1}^{N} p(\boldsymbol{a}_n)$$
$$p(\boldsymbol{a}_n) = N\left(\boldsymbol{a}_n; \mathbf{0}, \beta^{-1}\mathbf{I}_P\right) \quad (7)$$

where $\mathbf{A}$ is a $P \times N$ matrix of AR coefficients and $\boldsymbol{a}_n$ are the AR coefficients at the nth voxel ($\boldsymbol{a}_n$ is the nth row of $\mathbf{A}$). Again, β is chosen as described in Penny et al. (2003) to make this prior uninformative. This prior is unlikely to be optimal for fMRI due to the spatial dependence in AR values, an issue that has been addressed in a recent paper by Woolrich et al. (2004a) who show that modeling this dependence results in greater sensitivity. While this issue is clearly important, we have chosen an uninformative prior here as the focus of this paper is on modeling the signal.

Generative model

The likelihood term implicit in Eq. (3) and the priors defined in this section together define our probabilistic generative model for fMRI. This is shown graphically in Fig. 1. One benefit of defining a generative model is that by fixing certain variables and sampling from others, we can generate data from the model, as shown in Samples from the prior. These data can then be used to check the steps in the estimation algorithm that are defined in Approximate posteriors.

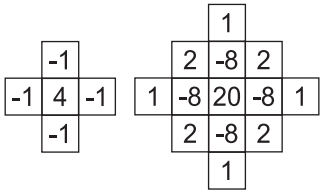

Fig. 3. The left panel shows the elements of the Laplacian operator $\mathbf{L}$. This is incorporated in the spatial prior which penalizes differences between cardinal neighbors. The right panel shows elements of the Laplacian product $\mathbf{L}^T\mathbf{L}$. The posterior mean for the kth regression coefficient at voxel n regresses toward the weighted mean of the kth regression coefficient in neighboring voxels where the weights are shown in the panel. The set of first-nearest cardinal neighbors, R_1, have value -8, the first-nearest diagonal neighbors, R_2, have value 2 and the second-nearest cardinal neighbors R_3 have value 1.

Spatial kernels

The results in this paper were obtained using a spatial matrix $\mathbf{S}$ equal to the Laplacian operator, $\mathbf{L}$, as defined in Pascual-Marqui et al. (1994). This enforces smoothness by penalizing differences between neighboring voxels and is a prior that is commonly used in the analysis of EEG. In this paper, we define the Laplacian using cardinal neighbors only, as shown in Fig. 3. Extension to more general neighborhood definitions is, however, straightforward and is within the scope of the estimation algorithm defined in the following section.

If $\boldsymbol{v}_k = \mathbf{L}\mathbf{w}_k$ then $\boldsymbol{v}_k(n)$ is equal to the sum of the differences between $\boldsymbol{w}_k(n)$ and its neighbors. Each element $\boldsymbol{v}_k(n)$ is distributed as a zero-mean Gaussian with precision $\boldsymbol{\alpha}_k$. Here, $\mathbf{L}$ acts as a difference matrix and $\mathbf{L}^{-1}$ as a smoothing matrix. Data can be generated from a Laplacian prior by generating independent Gaussian variables $\boldsymbol{v}_k$ and then applying the mapping $\boldsymbol{w}_k = \mathbf{L}^{-1}\boldsymbol{v}_k$.

The Laplacian prior as defined in Pascual-Marqui et al. (1994) uses a non-singular matrix $\mathbf{L}$. It is non-singular primarily because of the boundary conditions; for 2-D data diagonal elements in $\mathbf{L}$ are fixed to 4 (see left panel of Fig. 3), even if the voxel is at an edge and therefore has fewer than 4 neighbors.

This non-singularity is important as it is a necessary condition for evaluating the lower bound on the model evidence, F (see later), which is important for comparing models. An unfortunate consequence of these boundary conditions is that they result in parameter estimates at image edges that are biased towards zero. This results in conservative estimates of effect sizes. This prior is, however, preferred to the Laplacian with unbiased boundary conditions (where diagonal entries are always equal to the number of neighbors) as such a matrix is singular.

There are a number of compelling alternative priors in the literature, though, and these could be used in the future. One example are spatial kernels based on thin-plate splines with reflective boundary conditions (Buckley, 1994). It is also possible to specify a spatial precision matrix directly as $\mathbf{D} = \mathbf{S}^T\mathbf{S}$. This is the approach taken with 'Conditional Autoregressive (CAR)' or 'Gaussian Markov Random Field' priors (Cressie, 1993). These priors have been used in functional imaging by Gossl et al. (2001) to spatially regularize regression coefficients and by Woolrich et al. (2004a) to spatially regularize AR coefficients. In Eq. (7) of Woolrich et al. (2004a), for example, diagonal elements $\mathbf{D}_{ii}$ are set to unity and, if voxel j is a cardinal neighbor of voxel i, $\mathbf{D}_{ij}$ is set to the inverse of the geometric mean of the number of neighbors for

voxels i and j. This specification also results in biased boundary conditions but is necessary to make $\mathbf{D}$ non-singular (see above). We also note that the prior defined in (Woolrich et al., 2004a), for 2-D data, specifies a $\mathbf{D}$ matrix with four neighbors per voxel, whereas the prior used in this paper (indirectly) specifies a $\mathbf{D}$ matrix with 12 neighbors (see right panel of Fig. 3).

In principle, one can use the lower bound on the model evidence (see later) to choose the appropriate prior. In this paper, we have only used the evidence to select between global (i.e., non-spatial) priors and Laplacian priors (see simulation in Gaussian blobs) as it was our aim to establish the utility of spatial priors per se. It is perfectly possible though to compare different spatial priors and this will be the topic of a future paper.

Approximate posteriors

Because the prior distribution over regression coefficients allows for dependencies between voxels, it is clear that the true posterior will also have these dependencies. Also, because design matrices are not usually orthogonal, the true posterior over regression coefficients will also have dependencies between regressors at each voxel. Assuming that these dependencies are of a Gaussian nature, then they could be described by a posterior covariance matrix. The trouble with this, however, is that the matrix would be of dimension $KN \times KN$ which, even for modern computers is prohibitively large.

To get around this problem, we apply the Variational Bayes (VB) framework. This allows one to define 'approximate posteriors' which are matched to the true posteriors by minimizing the Kullback–Liebler (KL) divergence (Cover and Thomas, 1991). In particular, we propose an approximate posterior for the regression coefficients which factorizes over voxels (but not over regressors). Application of VB will find the approximate posterior distribution, out of all possible distributions that factorize over voxels, that best matches the true posterior (in the sense of KL divergence). In practice, this leads to us only needing to estimate a $K \times K$ covariance matrix at each voxel (see Eq. (10)).

Application of the VB framework leads to a set of approximate posteriors and equations for updating their sufficient statistics. These equations are summarized in Fig. 4 and are elaborated upon below. To improve readability we drop the 'conditional on $\mathbf{Y}$' notation, e.g., $q(\mathbf{W}) \equiv q(\mathbf{W}|\mathbf{Y})$.

Regression coefficients

As described above, the posterior over regression coefficients is assumed to factorize over voxels. That is

$$q(\mathbf{W}) = \prod_{n=1}^{N} q(\boldsymbol{w}_n) \tag{8}$$

This is the key assumption of this paper and is central to the derivation of update rules for the posteriors. As shown in Appendix, this results in a posterior over regression coefficients at voxel n given by

$$q(\mathbf{W}_n) = N\left(\boldsymbol{w}_n; \hat{\boldsymbol{w}}_n, \hat{\boldsymbol{\Sigma}}_n\right) \tag{9}$$

where

$$\begin{aligned} \hat{\boldsymbol{w}}_n &= \hat{\boldsymbol{\Sigma}}_n\left(\bar{\lambda}_n \mathbf{b}_n^T + \mathbf{r}_n\right) \\ \hat{\boldsymbol{\Sigma}}_n &= \left(\bar{\lambda}_n \tilde{\mathbf{A}}_n + \mathbf{B}_{nn}\right)^{-1} \end{aligned} \tag{10}$$

$$q(\mathbf{W}, \mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\alpha} \mid \mathbf{Y}) = \left(\prod_n q(\mathbf{w}_n \mid \mathbf{Y}) q(\mathbf{a}_n \mid \mathbf{Y}) q(\lambda_n \mid \mathbf{Y})\right) q(\boldsymbol{\alpha} \mid \mathbf{Y})$$

Regression coefficients

$$\begin{aligned} &q(\mathbf{w}_n) = N\left(\mathbf{w}_n; \hat{\mathbf{w}}_n, \hat{\boldsymbol{\Sigma}}_n\right) \\ &\hat{\mathbf{w}}_n = \hat{\boldsymbol{\Sigma}}_n\left(\bar{\lambda}_n \tilde{\mathbf{b}}_n{}^T + \mathbf{r}_n\right) \\ &\hat{\boldsymbol{\Sigma}}_n = \left(\bar{\lambda}_n \tilde{\mathbf{A}}_n + \mathbf{B}_{nn}\right)^{-1} \\ &\mathbf{B} = \mathbf{H}\left(diag(\bar{\boldsymbol{\alpha}}) \otimes \mathbf{S}^T\mathbf{S}\right)\mathbf{H}^T \\ &\mathbf{r}_n = -\sum_{i=1, i\neq n}^{N} \mathbf{B}_{ni}\hat{\mathbf{w}}_i \end{aligned}$$

AR coefficients

$$\begin{aligned} &q(\mathbf{a}_n) = N(\mathbf{a}_n; \mathbf{m}_n, \mathbf{V}_n) \\ &\mathbf{V}_n = \left(\bar{\lambda}_n \tilde{\mathbf{C}}_n + \beta \mathbf{I}_p\right)^{-1} \\ &\mathbf{m}_n = \bar{\lambda}_n \tilde{\mathbf{D}}_n \mathbf{V}_n \end{aligned}$$

Spatial precisions

$$\begin{aligned} &q(\boldsymbol{\alpha}) = \prod_{k=1}^{K} q(\alpha_k) \\ &q(\alpha_k) = Ga(\alpha_k; g_k, h_k) \\ &\frac{1}{g_k} = \frac{1}{2}\left[Tr\left(\hat{\boldsymbol{\Sigma}}_k \mathbf{S}^T\mathbf{S}\right) + \hat{\mathbf{w}}_k{}^T \mathbf{S}^T\mathbf{S}\hat{\mathbf{w}}_k\right] + \frac{1}{q_1} \\ &h_k = \frac{N}{2} + q_2 \\ &\bar{\alpha}_k = g_k h_k \end{aligned}$$

Observation noise

$$\begin{aligned} &q(\lambda_n) = Ga(\lambda_n; b_n, c_n) \\ &\frac{1}{b_n} = \frac{\tilde{G}_n}{2} + \frac{1}{u_1} \\ &c_n = \frac{T}{2} + u_2 \end{aligned}$$

Fig. 4. The figure shows the approximate posteriors and update equations for their sufficient statistics. The top equation describes the full approximate posterior with each component in a box below.

The estimated regression coefficients at the nth voxel are described by the $K \times 1$ vector $\hat{\mathbf{w}}_n$. The approximate covariance at the nth voxel is described by the $K \times K$ matrix, $\hat{\boldsymbol{\Sigma}}_n$. We emphasize that, because the approximate posterior factorizes over voxels, this matrix is of dimension $K \times K$, rather than $KN \times KN$. Matrix operation and storage is therefore straightforward.

The quantity $\bar{\lambda}_n$ in Eq. (10) is the estimated noise precision at the nth voxel defined in Eq. (18). The matrix $\tilde{\mathbf{A}}_n$ is related to the data precision at the nth voxel and the vector $\tilde{\mathbf{b}}_n$ is related to the data at the nth voxel projected onto the design matrix. These last two quantities are identical to those defined in Eqs. (63) and (64) in Penny et al. (2003). The matrix $\mathbf{B}$ is the $KN \times KN$ spatial precision matrix (with entries ordered by voxel—hence the permutation matrix $\mathbf{H}$ in the following equation) and is given by

$$\mathbf{B} = \mathbf{H}\left(diag(\bar{\alpha}) \otimes \mathbf{S}^T\mathbf{S}\right)\mathbf{H}^T \tag{11}$$

The quantity $\boldsymbol{r}_n$ is given by

$$\mathbf{r}_n = -\sum_{i=1, i\neq n}^{N} \mathbf{B}_{ni}\hat{\boldsymbol{w}}_i \tag{12}$$

The above equation is implemented using 'old' $\hat{\mathbf{w}}_i$ values. This results in 'new' $\mathbf{r}_n$ values and consequently 'new' $\hat{\mathbf{w}}_n$ values in Eq. (10). The subscripts in $\mathbf{B}_{ni}$ denote those entries in $\mathbf{B}$ pertaining to voxels n and i. The matrix $\mathbf{B}_{nn}$, for example, is of dimension $K \times K$ and contains entries in the spatial precision matrix relevant to voxel n. An intuitive description of Eq. (10) is given in Special cases.

Spatial precisions

The posteriors over the spatial precision variables can be shown to be (the derivation is similar to that for the regression coefficients and the derivations in the Appendix of Penny et al. (2003)

$$q(\boldsymbol{\alpha}) = \prod_{k=1}^{K} q(\alpha_k)$$

$$q(\alpha_k) = Ga(\alpha_k; g_k, h_k)$$

$$\frac{1}{g_k} = \frac{1}{2}\left[Tr\left(\hat{\boldsymbol{\Sigma}}_k \mathbf{S}^T\mathbf{S}\right) + \hat{\boldsymbol{w}}_k^T \mathbf{S}^T\mathbf{S}\hat{\boldsymbol{w}}_k\right] + \frac{1}{q_1} \tag{13}$$

$$h_k = \frac{N}{2} + q_2$$

$$\overline{\alpha}_k = g_k h_k$$

where $\boldsymbol{\Sigma}_k$ is a $N \times N$ diagonal matrix with nth entry $\boldsymbol{\Sigma}_n(k,k)$. The posteriors over the noise precision and autoregressive coefficients are identical to those defined in previous work (Penny et al., 2003). For completeness, they are given in the Appendices A–C.

Model evidence

The lower bound on the model evidence, F, can be computed as shown in the Appendices A–C. This bound can be used for model selection as shown in Penny et al. (2003), where it was used to find the optimal AR model order. In principle, it can be used to tune whatever aspect of the model is of special interest, e.g., the design matrix, spatial kernel. In this paper, it is used to compare spatial versus non-spatial priors (see Gaussian blobs). The quantity F also doubles as the objective function of the VB algorithm.

Special cases

Gaussian errors

The equations for the approximate posteriors are perhaps best understood by looking at special cases. For the special case of Gaussian errors rather than AR(P) errors, the posterior distribution over regression coefficients is given by

$$\begin{aligned}\hat{\boldsymbol{w}}_n &= \hat{\boldsymbol{\Sigma}}_n\left(\overline{\lambda}_n \mathbf{X}^T\mathbf{y}_n + r_n\right)\\ \hat{\boldsymbol{\Sigma}}_n &= \left(\overline{\lambda}_n \mathbf{X}^T\mathbf{X} + \mathbf{B}_{nn}\right)^{-1}\end{aligned} \tag{14}$$

No spatial prior

In the absence of any spatial prior, the above estimate reduces to the least squares solution

$$\hat{\boldsymbol{w}}_n = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}_n \tag{15}$$

Zero data precision

When there is a spatial prior, however, the regression coefficients regress towards a linear combination of the values of regression coefficients in neighboring voxels. This can be seen most clearly in the limit of zero data precision, $\bar{\lambda}_n \to 0$. For our Laplacian spatial kernel, the kth regression coefficient at the nth voxel then becomes

$$\hat{\boldsymbol{w}}_n(k) = \frac{1}{20}\left(8\sum_{i\in R_1}\hat{\boldsymbol{w}}_i(k) - 2\sum_{i\in R_2}\hat{\boldsymbol{w}}_i(k) - \sum_{i\in R_3}\hat{\boldsymbol{w}}_i(k)\right) \tag{16}$$

where R_1 denotes the set of first-nearest cardinal neighbors of n, R_2 denotes the first-nearest diagonal neighbors, and R_3 denotes the second-nearest cardinal neighbors. As $\bar{\lambda}_n \to 0$, the regression coefficients become equal to a weighted mean of neighboring coefficients. This is shown graphically in the right panel of Fig. 3.

Global shrinkage prior

If the spatial kernel $\mathbf{S}$ is set to the identity matrix, then our prior reduces to the 'global-shrinkage' prior proposed in Friston and Penny (2003). The parameter α_k then corresponds to the (inverse) prior variance of the kth regression coefficient. Note α_k in this paper is then equivalent to λ_m in Friston and Penny (2003). Our algorithm therefore generalizes the approach in that work. In following sections, we will compare global-shrinkage priors ('G-priors') to Laplacian priors ('L-priors') using the model evidence.

Uninformative prior

We will also compare the L-prior to the use of uninformative priors ('U'-priors)—see, e.g., (Penny et al., 2003)—applied to smoothed data. This will allow us to compare our results against the more common method of taking into account the spatial characteristics of the signal—namely, smoothing. We use an uninformative prior in conjunction with smoothing as it allows us to look at the effects of smoothing in the absence of other effects—such as shrinkage. We instantiate U-priors by setting $\mathbf{S} = \mathbf{I}$ and $\alpha_k = 1e^{-6}$.

Coefficient RESELS

For the general case of the algorithm described in Approximate posteriors, the effect of the prior is determined by the ratio of the data precision to the prior precision. As these quantities are both estimated from the data, spatial regularization is fully automatic.

The ratio of the data precision to the posterior precision, which for the kth coefficient at the nth voxel we will label γ_{nk}, plays an interesting role in our model. Firstly, we note that because the posterior precision is equal to the data precision plus the prior precision γ_{nk} can be evaluated as 1 minus the prior precision divided by the posterior precision

$$\gamma_{nk} = 1 - \hat{\boldsymbol{\Sigma}}_n(k,k)\mathbf{B}_{\mathrm{nn}}(k,k) \tag{17}$$

where $\hat{\boldsymbol{\Sigma}}_n$ is the approximate posterior covariance matrix for the nth voxel (defined in Eq. (10)) and $\mathbf{B}_{\mathrm{nn}}$ contains those elements in the prior spatial precision matrix pertinent to voxel n. If we then sum this over voxels, $\gamma_k = \sum_n \gamma_{nk}$, then γ_k represents the number of voxels for which the kth regression coefficient has been determined by the data (rather than the prior). For the definition of such a quantity in the context of Bayesian regression, see Mackay (1992). This quantity plays a role that is somewhat analogous to the concept of Resolution Elements (RESELs) in Random Field Theory (Worsley et al., 2004). A key difference is that γ_k refers to particular regression coefficients that describe regionally specific effects, whereas RESELs correspond to spatial degrees of freedom in images of residuals. Nonetheless, the concept is similar and it may be useful to think of γ_k as a 'Coefficient RESEL'. Typically, as we shall see, this is much less than N and gives a measure of how regionally specific an effect is.

Practicalities

The estimation algorithm is implemented by updating Eqs. (10), (13), (18), and (19) until convergence which is defined as less than

a 1% increase in F, the objective function, which is defined in Appendices A–C. In practice, convergence is very often seen to occur within four iterations (see, for example, Fig. 6). Because it is expensive to compute F on the larger models (e.g., when whole volumes are analyzed) an alternative is to simply run the algorithm for four iterations.

The algorithm is initialized using Ordinary Least Squares (OLS) estimates for regression and autoregressive parameters as described in Penny et al. (2003). Although, in this paper, the order of the autoregressive model is chosen apriori it can be estimated using F as described in Penny et al. (2003). In this paper, we present results where slices are analyzed separately and a 2-D Laplacian prior was used. The extension to 3-D priors is straightforward.

Results

Samples from the prior

Our estimation algorithm and the spatial prior underlying it are perhaps best understood with an example in which we compare VB with an L-prior to OLS. To this end, we generated data from our model as follows. We used a design matrix comprising two regressors, the first being a boxcar with a period of 20 scans and the second a constant. The design matrix, **X**, is therefore of dimension T × 2 and we chose $T = 40$ scans. We used spatial precision parameters $\alpha_1 = \alpha_2 = 1$ and generated two $N = 1024$ dimensional regression coefficient vectors. These were then reshaped into 32 × 32 images for display purposes. An image of the first regressor is shown in Fig. 2a. We then used the regression coefficients, the design matrix and additive Gaussian noise having precision $\lambda_n = \lambda = 0.5$ to generate the $T \times N$ data matrix **Y**. The time series at one of the voxels is shown in Fig. 2b. We did not add temporally autocorrelated errors as the focus of this simulation is the spatial domain.

We then fitted these data firstly using voxel-wise GLMs with parameters estimated via Ordinary Least Squares (OLS) and secondly using the VB algorithm with the L-prior described in this paper (the model from which the data were generated).

OLS and VB estimates of the first regression coefficient are shown in Fig. 5. Not surprisingly, VB model fits were superior. This can be quantified by computing the squared error between the estimated parameters and the true parameters. VB fits had 71% less error than OLS fits. Fig. 6 plots the log-evidence as a function of iteration number indicating that the algorithm has converged after five iterations of the update equations. The spatial precision for the first regression coefficient was estimated to be $\alpha = 0.51$. The number of RESELS for the first regression coefficient was found to be 366.

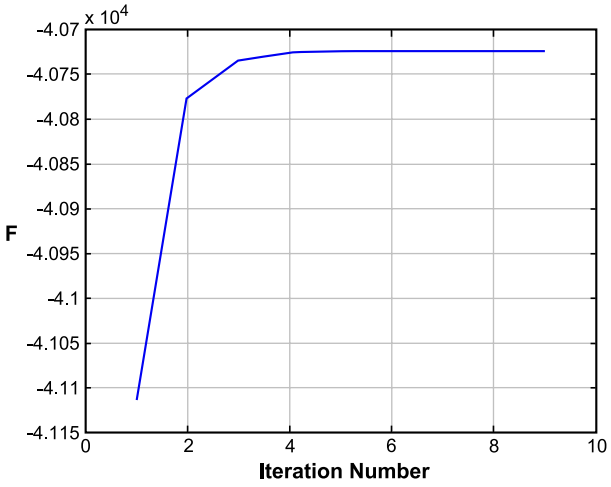

Fig. 6. The plot shows the lower bound on the log model evidence, F, as a function of iteration number. This indicates that the algorithm has converged after five VB iterations.

The above example shows the main features of the algorithm for a particular setting of model parameters (α, T, λ). Generally, as compared to OLS, the algorithm is more beneficial for smoother data (larger α) and lower data precision (smaller T and λ).

Gaussian blobs

In this second set of simulations, we compare the use of L-priors with G-priors and with the standard approach of smoothing the data. Again, the AR aspect of the model was neglected as we wished to focus on the spatial domain. We used the same design matrix as in the previous section. We then generated two 32 × 32 images of regression coefficients each containing Gaussian blobs. The image of the first regression coefficient is shown in Fig. 7a (the second coefficient is identical). This was formed by placing delta functions at three locations and then smoothing with Gaussian kernels having FWHMs of 2, 3, and 4 pixels (going clockwise from the top-left blob). The images were rescaled to make the peaks unity. We then used the regression coefficients, the design matrix and additive Gaussian noise of precision $\lambda = 10$ to

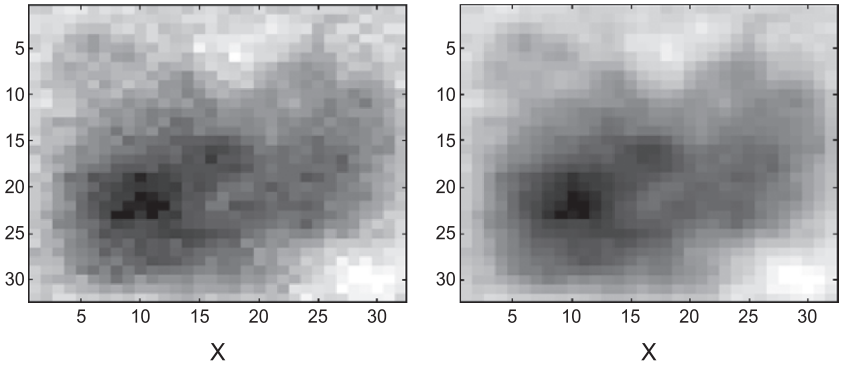

Fig. 5. The images show estimated regression coefficients using OLS (left panel) and VB with an L-prior (right panel). The VB images are clearly closer to the true regression coefficients (Fig. 2, left panel).

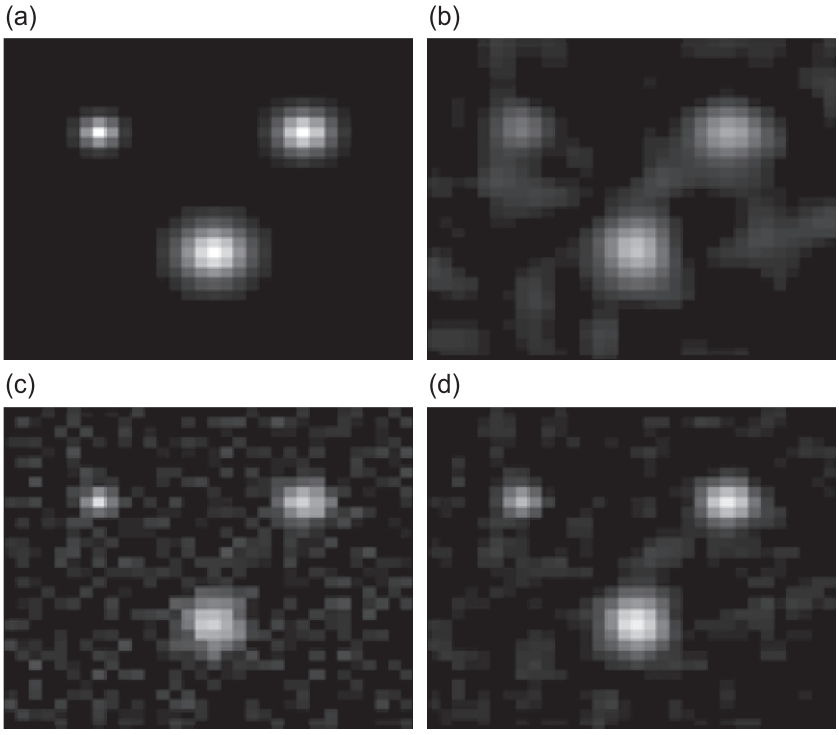

Fig. 7. The images are (a) true regression coefficients and regression coefficients estimated using (b) smoothing and an uninformative prior, (c) a global-shrinkage prior and (d) a Laplacian prior. Black corresponds to zero and white to one.

generate the $T \times N$ data matrix $\mathbf{Y}$ where $T = 40$ and $N = 1024$. We also created another data set, $\mathbf{Y}_s$, in which data images at each time point were smoothed using a Gaussian kernel having FWHM = 3 pixels. The smoothing was implemented so as to preserve the power (variance) in the signal (see later).

We then fitted (i) smoothed data with a model having an uninformative prior, unsmoothed data with (ii) a global-shrinkage prior and (iii) a Laplacian prior. Images of the estimated first regression coefficient are shown in Fig. 7. The quality of the estimates was then quantified by computing the squared error between the estimated parameters and the true parameters. Use of the L-prior resulted in parameter estimates with 66% less error than the U-prior on unsmoothed data and 64% less error than the G-prior.

Compared to the L-prior, the other schemes tend to under-estimate effect sizes in activated regions (the blobs are much grayer in Fig. 7b and c than in 7d). At the center of the upper right-hand blob, for example, the true effect size is 1 and the estimated effect sizes were 0.92 using the L-prior, 0.61 using the U-prior on smoothed data and 0.79 using the G-prior.

We also tried smoothing as it is implemented in SPM (Frackowiak et al., 2004), where the coefficients of the Gaussian kernels sum to unity. This does not preserve the variance of the signal—in fact, it reduces it. This led to estimated effect sizes (images not shown) that were even smaller. For example, in the center of the upper right blob, the estimated effect size was 0.51. The overall quality of estimates, however, was better than with variance preservation because truly non-activated areas were modeled better. But the L-prior still had 47% less estimation error.

That the L-prior produces better parameter estimates than the G-prior is also reflected in the model evidence which was many orders of magnitude higher (the difference in log-evidence was 857). The smoothness of the parameter estimates is reflected in the number of coefficient RESELS, which for the first regression coefficient was estimated to be 109 using the L-prior and 714 using the G-prior. Use of the G-prior leads to a mistaken inference that the effect is less regionally specific than it actually is. Note also the patches of false weak activations in the smoothed data in Fig. 7b. This also implies a lack of regional specificity.

Another perspective on these simulation results is given by the Receiver Operating Characteristic (ROC) curve which is a plot of the sensitivity versus 1 minus the specificity. This curve, shown in Fig. 8, was generated by declaring a voxel to be active if the effect size was larger than some arbitrary threshold. This threshold was

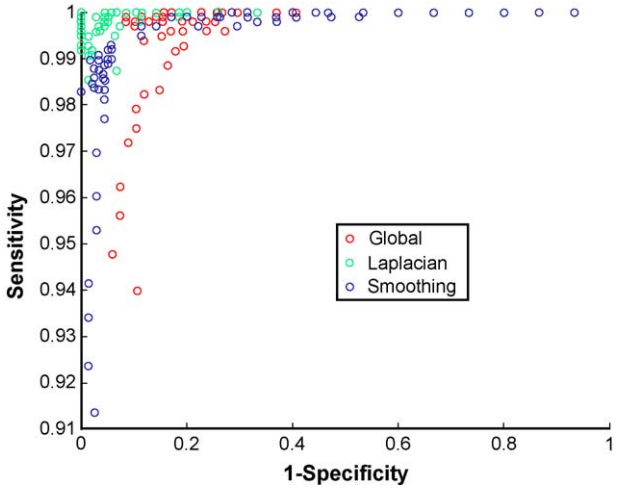

Fig. 8. ROC curve for Gaussian blob data. The preponderance of green points in the upper left corner, the region of high specificity and high sensitivity, mark out the Laplacian prior as the superior method.

then varied over the range 0.1 to 0.7 to produce the points in each curve. The preponderance of green points in the upper left corner, the region of high specificity and high sensitivity, mark out the L-prior as the superior method. This is important as it indicates that increased sensitivity can be achieved while maintaining high specificity.

Face-repetition data

This data set and a full description of the experiments and data pre-processing are available from http://www.fil.ion.ucl.ac.uk/spm/data. The data were acquired during an experiment concerned with the processing of images of faces (Henson et al., 2002). This was an event-related study in which greyscale 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 BOLD contrast. Whole brain EPIs consisting of 24 transverse slices were acquired every 2 s resulting in a total of $T = 351$ scans.

All functional images were realigned to the first functional image using a six-parameter rigid-body transformation. 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. Images were then spatially normalized to a standard EPI template using a nonlinear warping method. This set of images constitutes an 'unsmoothed' data set. We also created a 'smoothed' data set by convolving the images with an isotropic Gaussian kernel having FWHM = 8 mm as was implemented in the original paper (Henson et al., 2002). This was implemented in SPM (note that this implies the variance of the images was not preserved—see previous section).

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. This makes the units of the regression coefficients

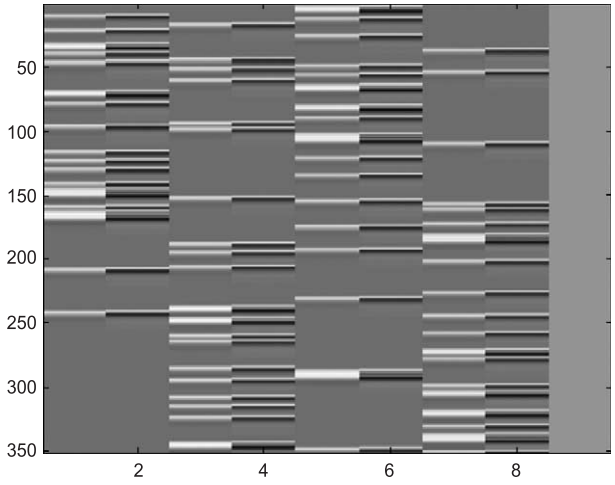

Fig. 9. Design matrix for face-repetition fMRI analysis. There are 9 regressors relating to 4 types of event 'N1', 'N2', 'F1' and 'F2' which are the first or second (1/2) presentations of images of famous 'F' or non-famous 'N' faces. These comprise the 1st and 2nd, 3rd and 4th, 5th and 6th, and 7th and 8th columns, respectively. Regressors 1, 3, 5 and 7 have been convolved with a canonical hemodynamic response function and regressors 2, 4, 6, and 8 with its temporal derivative. The last regressor is a constant term.

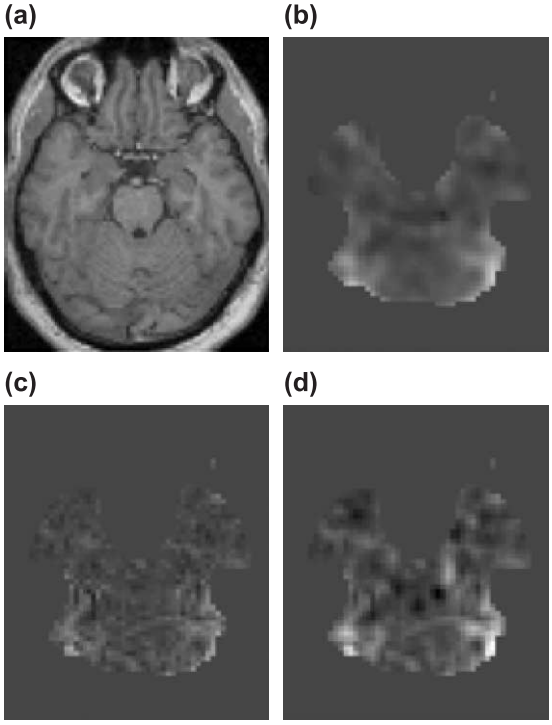

Fig. 10. Contrasts for the main effect of faces. Slice $z = -24$ mm (a) Normalized structural image and images of estimated contrasts for (b) smoothed data and a U-prior, (c) unsmoothed data and a G-prior and (d) unsmoothed data and an L-prior. For plots (b), (c) and (d) black denotes an effect size of −1 and white denotes 2.7 (in units of percentage global mean value).

'percentage of global mean value'. Each time series was then high-pass filtered using a set of discrete cosine basis functions with a filter cut-off of 128 s.

The data set was analyzed using a GLM with a design matrix as shown in Fig. 9. There are nine regressors relating to 4 types of event 'N1', 'N2', 'F1' and 'F2' which are the first or second (1/2) presentations of images of famous 'F' or non-famous 'N' faces. These comprise the 1st and 2nd, 3rd and 4th, 5th and 6th, and 7th and 8th columns, respectively. Regressors 1, 3, 5, and 7 have been convolved with a canonical hemodynamic response function and regressors 2, 4, 6, and 8 with its temporal derivative. Modeling the HRF in this way allows one to capture onset variability across voxels. The last regressor is a constant term.

We then fitted GLM-AR models to the smoothed and unsmoothed data sets. We chose an AR model order of $P = 3$ as this was shown to be sufficient in a previous analysis (Penny et al., 2003). If we were to regress out cardiac and respiratory activity as in Glover et al. (2000) and Hu et al. (1995), the optimal AR model order is likely to be smaller. These were, however, not measured for this data set.

For the smoothed data, we used a U-prior on the regression coefficients. This is because the spatial characteristics of the data have already been accounted for so it would not make sense to use an L-prior. Also, this allows us to assess the effect of smoothing on effect size independently of the effect of informative priors (such as G or L). For the unsmoothed data, we applied both a G-prior and an L-prior.

These three different approaches are compared, firstly, by looking at the main effect of faces. This is assessed by applying the

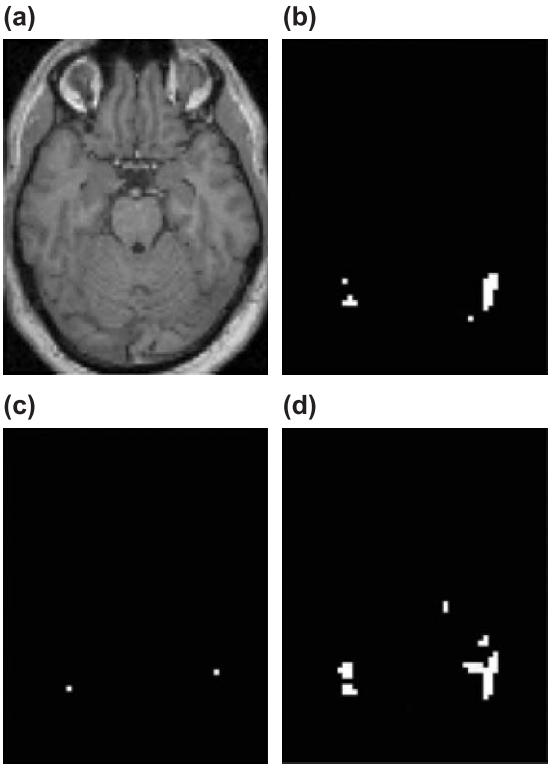

Fig. 11. PPMs for the main effect of faces. Slice $z = -24$ mm (a) Normalized structural image and Posterior Probability Maps of the effect size being greater than 1% of the global mean value for (b) smoothed data and a U-prior, (c) unsmoothed data and a G-prior and (d) unsmoothed data and an L-prior. For plots (b), (c) and (d) black denotes a probability of 0 and white denotes 1. Only voxels with probabilities greater than 0.9 are shown.

contrast weight vector $\boldsymbol{c}^T = 0.25k[1,0,1,0,1,0,1,0,0]$ to the estimated regression coefficients where k is the peak hemodynamic response to a single event (this depends on the TR and for our data $k = 2.1$). This contrast shows the average peak response to the presentation of a face image. Maps of the estimated contrast are shown for a single slice at $z = -24$ mm (MNI coordinates) in Fig. 10. These maps show large responses in bilateral fusiform and occipital cortex as previously reported (Henson et al., 2002). However, use of the Laplacian prior results in much higher estimates of effect size in these regions. A typical voxel ($x = 45$, $y = -66$, $z = -24$ mm) in the activated region in the right hemisphere, for example, has effect sizes of 1.9% for the U-prior on smoothed data, 1.2% for the G-prior and 2.6% for the L-prior on unsmoothed data.

These differences in effect size are also reflected in the Posterior Probability Maps (PPMs) (Friston and Penny, 2003) in Fig. 11. Here, we plot the probability that the contrast is greater than 1%. Bilateral activations of this size or greater have highest probability using the L-prior on unsmoothed data. Use of the L-prior also shows up an additional area, the parahippocampal gyrus ($x = 18$, $y = -27$, $z = -24$ mm).

We also looked at the main effect of fame-judgement. This is assessed by applying the contrast weight vector $\boldsymbol{c}^T = 0.25k[-1,0,-1,0,1,0,1,0,0]$ to the estimated regression coefficients. This contrast shows the average difference in peak response to famous versus non-famous faces. Maps of the contrast are shown for a single slice at $z = -18$ mm in Fig. 12. These maps show little variation in response over the slice except for the contrast map obtained with the L-prior which has a cluster of large values in left anterior temporal cortex. Damage to this part of the brain is often associated with loss of personal knowledge, e.g., an inability to name famous faces (Damasio et al., 1996). The voxel in the center of this activated region ($x = -42$, $y = -3$, $z = -18$ mm) has an estimated effect size of 1.9%. The corresponding value for the G-prior is 0.6% and for the U-prior on smoothed data it is 0.8%. These differences in effect size are reflected in the PPMs in Fig. 13. Here, we plot the probability that the contrast is greater than 1%. Only the L-prior allows us to infer, with high probability, that there are two focal activations in this slice, the second being at ($x = 0$, $y = -6$, $z = -18$ mm).

Discussion

We have proposed a Bayesian estimation and inference procedure for fMRI time series based on the use of GLM-AR(P) models. The novel contribution of this paper has been the incorporation of a spatial prior over regression coefficients which embodies our prior knowledge that evoked responses are spatially contiguous and locally homogeneous. Further, we have been able to let the data determine the spatial regularization coefficients. Our model generalizes earlier work on voxel-wise estimation of GLM-AR models that used uninformative priors (Penny et al., 2003) and inference in GLMs using Posterior Probability Maps (PPMs) based on global-shrinkage priors (Friston and Penny, 2003).

As compared to the standard approach based on smoothing the data, our simulations show that the use of our VB algorithm with

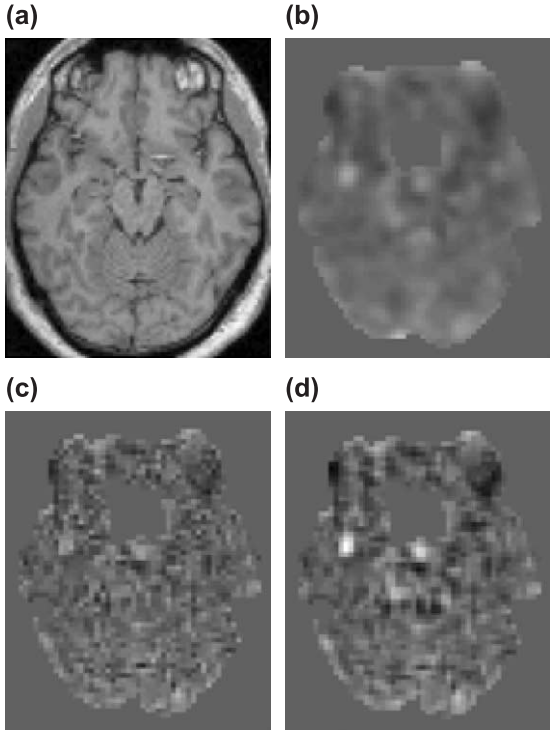

Fig. 12. Contrasts for the main effect of fame Slice $z = -18$ mm (a) Normalized structural image and images of estimated contrasts for (b) smoothed data with a U-prior, (c) unsmoothed data with a G-prior and (d) unsmoothed data with an L-prior. For plots (b), (c) and (d) black denotes an effect size of -1.2 and white denotes 1.9 (in units of percentage global mean value).

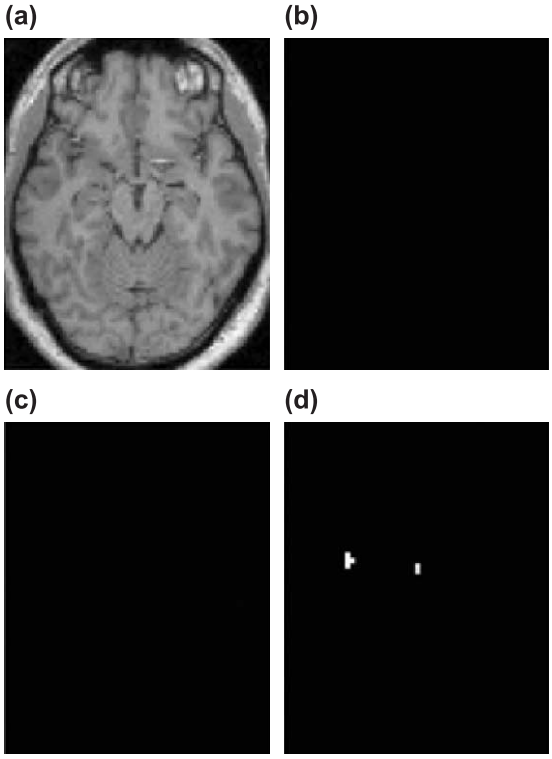

Fig. 13. PPMs for the main effect of fame. Slice $z = -18$ mm (a) Normalized structural image and PPMs of the effect size being greater than 1% of the global mean value for (b) smoothed data with a U-prior, (c) unsmoothed data with a G-prior and (d) unsmoothed data and an L-prior. For plots (b), (c) and (d) black denotes a probability of 0 and white denotes 1. Only voxels with probabilities greater than 0.9 are shown.

Laplacian priors has higher sensitivity for any given high level of specificity. Results on event-related fMRI data show that the smoothing approach underestimates the effect size of fsneurophysiologically interpretable activations. On simulated data, the Laplacian prior was shown to be superior to the use of a global shrinkage prior in terms of Bayesian evidence, estimation accuracy and area under the ROC curve. On the event-related fMRI data, the global shrinkage prior also appeared to underestimate effect size. While effect size is not of concern in classical inference, it is of central interest to Bayesian inference (Friston and Penny, 2003). From our results, we conclude that currently the best way to estimate effect size accurately is via the use of Laplacian spatial priors.

One attractive property of our spatial regularization scheme is that different regression coefficients have different regularizes. So, for example, if one models the hemodynamic response to an event with two regressors, e.g., a 'canonical' response and its temporal derivative (as in Face-repetition data), then the canonical regressor can be smoothed a different amount than the temporal derivative regressor. The optimal amount of smoothing will be determined separately for each regressor using Eq. (13). Allowing different signal components to have different smoothnesses is a further reason why this spatial regularization procedure is superior to simply smoothing the data (which smooths all components equally).

Previous use of spatial priors in fMRI has either relied on the use of hard (i.e., non-adaptive) priors and projection of the data onto a subspace spanned by these priors (see, for example, Section 3 of (Friston et al., 2002)) or used adaptive priors but the computationally expensive MCMC algorithm to sample from the relevant posterior distributions (Gossl et al., 2001). In this paper, we have shown how to use a fast analytic approximation to the posterior distribution which was derived using the VB framework. Woolrich et al. (2004a) have shown that inference using VB is orders of magnitude quicker than with similar MCMC approaches (Woolrich et al., 2004b).

In the application of VB to regularizing AR coefficients Woolrich et al. (2004a) still faced a key technical challenge which was to invert an $N \times N$ spatial precision matrix, where N is the number of voxels. The solution was found by first noticing that the inverse (say, $\mathbf{A}^{-1}$) only ever occurred when postmultiplied by a positive symmetric definite matrix (say $\mathbf{B}$, i.e., $\mathbf{x} = \mathbf{A}^{-1}\mathbf{B}$). A conjugate gradient method was then used to find the vector $\boldsymbol{x}$ that minimized $\mathbf{Ax-B}$ (see Appendix B.3 of Woolrich et al. (2004a). In this paper, we have presented a different solution. We specified a prior that captures dependencies between voxels, but an approximate posterior that factorizes over voxels. In other words, we have spatial dependencies in the prior but not in the (approximate) posterior. This factorization is essential to the development of a tractable learning algorithm and relies upon the VB framework where such factorizations play a central role. In practice this means that, in an fMRI model with K regressors and N voxels, we only have to invert $K \times K$ matrices, rather than $KN \times KN$ matrices.

While we favor the use of Laplacian priors over uninformative and non-spatial priors and note their widespread application in localizing sources underlying EEG data, they could yet be improved upon. One simple extension would allow separate spatial regularization coefficients for the x and y (phase-encode) directions. The motivation for this is that, for Echo Planar Imaging, the BOLD signal is more blurred in the phase-encode direction than in the read-out direction (O. Josephs, personal communication). As this would entail more than one spatial regularization coefficient per regressor this would require an extension, albeit a minor one, to the current framework.

One problem with Laplacian priors is that they conflate magnitude and smoothness. That is, there is a single parameter α that determines both. A promising alternative is the use of wavelet priors where coarse levels determine 'magnitude', and 'detail' levels determine smoothness. Also, Laplacian priors have no notion of location or spatial frequency. They are therefore unable to reflect spatial variations in smoothness arising from regional differences in vasculature or functional anatomy. Again, wavelets (or other basis set decompositions) provide a promising alternative. Preliminary work in this direction, where wavelets were used to smooth fMRI contrast images (Penny, 2002), indicates they may be a natural choice. This family of approaches, however, would use multiple regularization coefficients per regression coefficient and so would require an extension to the current framework.

Another solution to the problem of non-stationary smoothness estimation is variable resolution tomography (VARETA) which is used in EEG source estimation (Valdes-Sosa et al., 2000). VARETA specifies a spatial precision variable for each voxel, $\alpha(n)$, but this vector of spatial precisions is itself regularized using a Laplacian operator. Incorporation of this method into the current VB framework via 'hyperpriors' is another possible direction for future research.

The work in Penny et al. (2003) which used the VB framework and assumed uninformative priors over regression and AR

coefficients has since been extended by Woolrich et al. (2004a) to account for spatial dependence in AR coefficients. In this paper, we have extended it to account for spatial dependence in regression coefficients. A natural next step is to allow for spatial dependence in the regression coefficients, AR coefficients and possibly noise precisions.

A further extension to the model relates to the error process. In this paper and in previous work, we have characterized the error process using arbitrary order autoregressive models as our aim has been to analyze single subject fMRI data. If one wished to model multiple subject data, however, and employ the computationally efficient summary statistic approach (Holmes and Friston, 1998), whereby possibly multiple contrast images from multiple subjects form the data for a 'second-level' of analysis then a different model of the error process would be appropriate. The use of spatial priors at the second level would also need to be modified as one must take into account between-subject differences in functional anatomy.

Acknowledgments

Will Penny is supported by the Wellcome Trust. The authors would also like to thank Rik Henson for commenting on the manuscript and for his advice on analyzing the face processing data.

Appendix A. Derivation of approximate posterior for regression coefficients

If the approximate posterior factorizes as $q(\theta) = \Pi_i q(\theta_i)$, where θ are the parameters of the model, then the components of the approximate posteriors that maximise a lower bound on the model evidence (or equivalently minimize the KL-divergence between the true posterior and the approximate posterior) are given by Roberts and Penny (2002).

$$q(\boldsymbol{\theta}_i) = \frac{\exp[I(\boldsymbol{\theta}_i)]}{\int \exp[I(\boldsymbol{\theta}_i)]\mathrm{d}\boldsymbol{\theta}_i}$$

where

$$I(\boldsymbol{\theta}_i) = \int q(\boldsymbol{\theta}_{/i}) \log p(\mathbf{Y},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}_{/i}$$

and $\boldsymbol{\theta}_{/i}$ indicates components of $\boldsymbol{\theta}$ other than $\boldsymbol{\theta}_i$. This latter integral need only contain terms dependent on $\boldsymbol{\theta}_i$. For the model in this paper the parameters are $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\alpha}\}$ and the joint probability of the data and parameters, $p(\mathbf{Y}, \boldsymbol{\theta})$, is given in the caption to Fig. 1.

To derive an expression for the approximate posterior for the regression coefficients at voxel n, $q(\boldsymbol{w}_n)$, the relevant integral is

$$I(\boldsymbol{w}_n) = \int q(\lambda_n) q(\boldsymbol{\alpha}) q(\boldsymbol{w}_{/\ln}) \log(p(\mathbf{y}_n|\lambda_n, \mathbf{a}_n, \boldsymbol{w}_n) p(\mathbf{w}|\boldsymbol{\alpha}))\mathrm{d}\lambda_n \mathrm{d}\boldsymbol{\alpha}\mathrm{d}\boldsymbol{w}_{/ln}$$

where

$$\log p(\mathbf{y}_n|\lambda_n, \mathbf{a}_n, \boldsymbol{w}_n) = -\frac{\lambda_n}{2}\left(\boldsymbol{w}_n^T \mathbf{A}(\mathbf{a}_n)\boldsymbol{w}_n - 2\mathbf{b}(\mathbf{a}_n)^T \boldsymbol{w}_n\right) + \ldots$$

and the quantities $\mathbf{A}(\mathbf{a}_n)$ and $\mathbf{b}(\mathbf{a}_n)$ depend on the autoregressive coefficients and have been defined in Eq. (53) of Penny et al. (2003). The second relevant log-probability is

$$\begin{aligned}\log p(\mathbf{W}|\boldsymbol{\alpha}) &= -\frac{1}{2}\sum_k \boldsymbol{\alpha}_k \boldsymbol{w}_k^T \mathbf{S}^T \mathbf{S} \boldsymbol{w}_k + \ldots \\ &= -\frac{1}{2}\boldsymbol{w}_r^T \left(diag(\boldsymbol{\alpha}) \otimes \mathbf{S}^T\mathbf{S}\right)\boldsymbol{w}_r + \ldots \\ &= -\frac{1}{2}\boldsymbol{w}_r^T \mathbf{H}\left(diag(\boldsymbol{\alpha}) \otimes \mathbf{S}^T\mathbf{S}\right)\mathbf{H}^T\boldsymbol{w}_v + \ldots\end{aligned}$$

where the second and third lines involve the $KN \times 1$ vectors $\boldsymbol{w}_r$ and $\boldsymbol{w}_v$ and the permutation matrix $\mathbf{H}$ defined in Eq. (2). Substituting these quantities into the earlier integral equation gives

$$I(\boldsymbol{w}_n) = -\frac{1}{2}\boldsymbol{w}_n^T\left(\overline{\lambda}_n \tilde{\mathbf{A}}_n + \mathbf{B}_{nn}\right)\boldsymbol{w}_n + \boldsymbol{w}_n^T\left(\overline{\lambda}_n \tilde{\mathbf{b}}_n^T + \mathbf{r}_n\right)$$

where

$$\mathbf{B} = \mathbf{H}\left(diag(\overline{\alpha}) \otimes \mathbf{S}^T\mathbf{S}\right)\mathbf{H}^T$$

$\mathbf{B}_{nn}$ contains those entries in $\mathbf{B}$ relevant to voxel n and

$$\mathbf{r}_n = -\sum_{i=1,\, i\neq n}^{N} \mathbf{B}_{ni}\hat{\mathbf{w}}_i$$

The quantities $\tilde{\mathbf{A}}_n$ and $\tilde{\mathbf{b}}_n$ in the above expression for $I(\boldsymbol{w}_n)$ are equivalent to $\mathbf{A}(\boldsymbol{a}_n)$ and $\mathbf{b}(\boldsymbol{a}_n)$ integrated over $q(\boldsymbol{a}_n)$ and are given in Eqs. (63) and (64) of Penny et al. (2003). By noting that the log of a Gaussian density is given by

$$\log N(\mathbf{x};\mathbf{m},\boldsymbol{\Sigma}) = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}$$

equating terms $I(\boldsymbol{w}_n)$ with shows that is $q(\boldsymbol{w}_n)$ Gaussian with mean and covariance given by

$$\hat{\mathbf{w}}_n = \hat{\boldsymbol{\Sigma}}_n\left(\overline{\lambda}_n \tilde{\mathbf{b}}_n^T + \mathbf{r}_n\right)$$

$$\hat{\boldsymbol{\Sigma}}_n = \left(\overline{\lambda}_n \tilde{\mathbf{A}}_n + \mathbf{B}_{nn}\right)^{-1}$$

Appendix B. Noise and AR updates

The posteriors over the noise precision and autoregressive coefficients are identical to those defined in previous work (Penny et al., 2003). For the noise precision, we have

$$\begin{aligned} q(\lambda_n) &= Ga(\lambda_n; b_n, c_n) \\ \frac{1}{b_n} &= \frac{\tilde{G}_n}{2} + \frac{1}{u_1} \\ c_n &= \frac{T}{2} + u_2 \\ \overline{\lambda}_n &= b_n c_n \end{aligned} \tag{18}$$

where $\tilde{G}_n$ is related to the GLM prediction error and is defined for a single voxel in Eq. (77) in the appendix of Penny et al. (2003). For the autoregressive coefficients, we have

$$\begin{aligned} q(\mathbf{a}_n) &= N(\mathbf{a}_n; \mathbf{m}_n, \mathbf{V}_n) \\ \mathbf{V}_n &= \left(\overline{\lambda}_n \tilde{\mathbf{C}}_n + \beta \mathbf{I}_p\right)^{-1} \\ \mathbf{m}_n &= \overline{\lambda}_n \tilde{\mathbf{D}}_n \mathbf{V}_n \end{aligned} \tag{19}$$

where $\tilde{C}_n$ and $\tilde{D}_n$ are quantities related to AR prediction error and are defined in Eq. (50) in the appendix of Penny et al. (2003).

Appendix C. Model evidence

The objective function for the algorithm is the lower bound on the logarithm of the model evidence which for our model is given by

$$F = L_{av} - (\mathrm{KL}(\mathbf{W}) + \mathrm{KL}(\mathbf{a}) + \mathrm{KL}(\boldsymbol{\alpha}) + \mathrm{KL}(\boldsymbol{\lambda})) \quad (20)$$

where L_{av} is the average log-likelihood and the KL terms are the Kullback–Liebler divergences between the priors and approximate posteriors. These are computed using standard results for KL-divergences for Gamma and Normal distributions given in Roberts and Penny (2002). The average log-likelihood is given by

$$L_{av} = \sum_n \frac{T}{2}(\varphi(c_n) + \log(b_n)) - \frac{\overline{\lambda_n}}{2}\tilde{G}_n \quad (21)$$

where $\varphi()$ is the digamma function (Press et al., 1992).

References

Buckley, M.J., 1994. Fast computation of a discretized thin-plate spline for image data. Biometrika 81, 247–258.

Buxton, R.B., Wong, E.C., Frank, L.R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magn. Reson. Med. 39, 855–864.

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

Cressie, N., 1993. Statistics for Spatial Data. Wiley, New York.

Damasio, H., Grabowski, T.J., Tranel, D., Hichwa, R.D., Damasio, A.R., 1996. A neural basis for lexical retrieval. Nature 380, 499–505.

Frackowiak, R.S., Friston, K.J., Frith, C.D., Dolan, R., Ashburner, J., Price, C., Penny, W., Zeki, S., 2004. Human Brain Function. Academic Press.

Friston, K.J., Penny, W., 2003. Posterior probability maps and SPMs. NeuroImage 19, 1240–1249.

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

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.E., Henson, R.N., Kiebel, S., Phillips, C., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: applications. NeuroImage 16, 484–512.

Glover, G.H., Li, T.Q., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med. 44, 162–167.

Gossl, C., Auer, D.P., Fahrmeir, L., 2001. Bayesian spatiotemporal inference in functional magnetic resonance imaging. Biometrics 57, 554–562.

Henson, R.N., 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.

Holmes, A.P., Friston, K.J., 1998. Generalisability, random effects and population inference. NeuroImage 7, 754.

Hu, X., Le, T.H., Parrish, T., Erhard, P., 1995. Retrospective estimation and correction of physiological fluctuation in functional MRI. Magn. Reson. Med. 34, 201–212.

Mackay, D.J.C., 1992. Bayesian interpolation. Neural Comput. 4, 415–447.

Pascual-Marqui, R.D., Michel, C.M., Lehmann, D., 1994. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. Int. J. Psychophysiol. 18, 49–65.

Penny, W., 2002. Wavelet Smoothing of fMRI Activation Images. Technical Report. Wellcome Department of Imaging Neuroscience, UCL, UK.

Penny, W., Friston, K., 2003. Mixtures of general linear models for functional neuroimaging. IEEE Trans. Med. Imaging 22, 504–514.

Penny, W., Kiebel, S., Friston, K., 2003. Variational Bayesian inference for fMRI time series. NeuroImage 19, 727–741.

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

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

Rosenfeld, A., Kak, A.C., 1982. Digital Picture Processing. Academic Press.

Turner, R., 2002. How much cortex can a vein drain? Downstream dilution of activation-related cerebral blood oxygenation changes. NeuroImage 16, 1062–1067.

Valdes-Sosa, P.A., Marti, F., Garica, F., Casanova, R., 2000. Variable resolution electric-magnetic tomography. Proceedings of the Tenth International Conference on Biomagnetism, vol. 2. pp. 373–376. Ref Type: Conference Proceeding.

Woolrich, M.W., Behrens, T.E., Smith, S.M., 2004a. Constrained linear basis sets for HRF modelling using Variational Bayes. NeuroImage 21, 1748–1761.

Woolrich, M.W., Jenkinson, M., Brady, J.M., Smith, S.M., 2004b. Fully Bayesian spatio-temporal modeling of fMRI data. IEEE Trans. Med. Imaging 23, 213–231.

Worsley, K.J., Marret, S., Neelin, P., Evans, A.C., 1995. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. 4, 58–73.

Worsley, K.J., Marrett, S., Neelin, P., Evans, A.C., 1996. Searching scale space for activation in PET images. Hum. Brain Mapp. 4, 74–90.