Bayesian Analysis of fMRI data with Spatial Priors

Will Penny and Guillaume Flandin
Wellcome Department of Imaging Neuroscience,
University College, London WC1N 3BG.

KEY WORDS: Bayesian, fMRI, spatial prior, GLM, variational

1. 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 [7]. From measurements of changes in blood oxygenation one can use various statistical models, such as the General Linear Model (GLM) [8], to make inferences about task-specific changes in underlying neuronal activity.

In previous work [21, 23, 22] we have developed a spatially regularised General Linear Model (GLM) for the analysis of fMRI data which allows for the characterisation of regionally specific effects using Posterior Probability Maps (PPMs). This spatial regularisation has been shown [23] to increase the sensitivity of inferences one can make.

This paper reviews our body of work on spatially regularised GLMs and describes two new developments. These are (i) an approach for assessing multivariate contrasts and (ii) a method for choosing the thresholds that generate PPMs. The paper is structured as follows. Section 2 reviews the theoretical development of the algorithm. This includes a description of a Variational Bayesian algorithm in which inference is based on an approximation to the posterior distribution that has minimal KL-divergence from the true posterior. Sections 3 and 4 describe the new approaches for assessing multivariate contrasts and PPM thresholding. In section 5 we present results on null fMRI data, synthetic data and fMRI from functional activation studies of auditory and face processing. The paper finishes with a discussion in section 6.

2. Theory

We write an fMRI data set consisting of T time points at N voxels as the $T \times N$ matrix Y. In mass-univariate models [8], these data are explained in terms of a $T \times K$ design matrix X, containing the values of K regressors at T time points, and a $K \times N$ matrix of regression coefficients W, containing K regression coefficients at each of the N voxels. The model is written

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

where E is a $T \times N$ error matrix.

It is well known that fMRI data is contaminated with artifacts. These stem primarily from low-frequency drifts due to hardware instabilities, aliased cardiac pulsation and respiratory sources, unmodelled neuronal activity and residual motion artifacts not accounted for by rigid body registration methods [25]. This results in the residuals of an fMRI analysis being temporally autocorrelated.

In previous work we have shown that, after removal of low-frequency drifts using Discrete Cosine Transform (DCT) basis sets, low-order voxel-wise autoregressive (AR) models are sufficient for modelling this autocorrelation [21]. It is important to model these noise processes as parameter estimation becomes less biased [11] and more accurate [21]. Together, DCT and AR modelling can account for long-memory noise processes. Alternative procedures for removing low-frequency drifts include the use of running-line smoothers or polynomial expansions [17].

2.1 Model likelihood

We now describe the approach taken in our previous work. For a Pth-order AR model, the likelihood of the data is given by

$$p(Y|W,A,\lambda) = \prod_{t=P+1}^{T}\prod_{n=1}^{N} \mathsf{N}(y_{tn} - x_t w_n;\ (d_{tn} - X_t w_n)^T a_n, \lambda_n^{-1}) \tag{2}$$

where $\mathsf{N}(x;m,C)$ is a uni/multivariate Normal density with mean m and variance/covariance C, n indexes the nth voxel, a_n is a $P \times 1$ vector of autoregressive coefficients, w_n is a $K \times 1$ vector of regression coefficients and λ_n is the observation noise precision. The vector x_t is the tth row of the design matrix and X_t is a $P \times K$ matrix containing the previous P rows of X prior to time point t. The scalar y_{tn} is the fMRI scan at the tth time point and nth

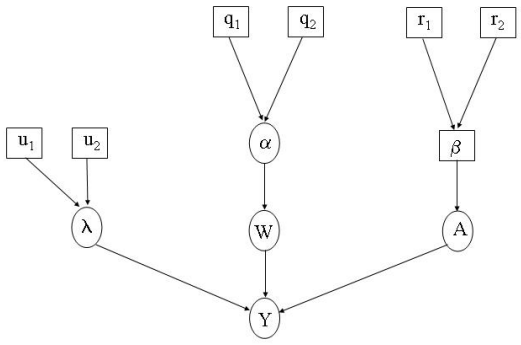

Figure 1: *The figure shows the probabilistic dependencies underlying our generative model for fMRI data. The quantities in square brackets are constants and those in circles are random variables. The spatial regularisation coefficients α constrain the regression coefficients W. The spatial regularisation coefficients β constrain the AR coefficients A. The parameters λ and A define the autoregressive error processes which contribute to the measurements.*

voxel and $d_{tn} = [y_{t-1,n}, y_{t-2,n}, ..., y_{t-P,n}]^T$. Because d_{tn} depends on data P time steps before, the likelihood is evaluated starting at time point $P+1$, thus ignoring the GLM fit at the first P time points.

Equation 2 shows that higher model likelihoods are obtained when the prediction error $y_{tn} - x_t w_n$ is closer to what is expected from the AR estimate of prediction error.

The voxel wise parameters w_n and a_n are contained in the matrices W and A and the voxel-wise precisions λ_n are contained in λ. The next section describes the prior distributions over these parameters. Together, the likelihood and priors define the probabilistic generative model, which is portrayed graphically in Figure 1.

2.2 Priors

The graph in Figure 1 shows that the joint probability of parameters and data can be written

$$\begin{aligned} p(Y, W, A, \lambda, \alpha, \beta) &= p(Y|W, A, \lambda) p(W|\alpha) \quad (3) \\ & \quad p(A|\beta) p(\lambda|u_1, u_2) \\ & \quad p(\alpha|q_1, q_2) p(\beta|r_1, r_2) \end{aligned}$$

where the first term is the likelihood and the other terms are the priors. The likelihood is given in equation 2 and the priors are described below.

2.2.1 Regression coefficients

For the regressions coefficients we have

$$\begin{aligned} p(W|\alpha) &= \prod_{k=1}^{K} p(w_k^T|\alpha_k) \quad (4) \\ p(w_k^T|\alpha_k) &= N(w_k^T; 0, \alpha_k^{-1} D_w^{-1}) \end{aligned}$$

where D_w is a spatial precision matrix. This can be set to correspond to eg. a Low Resolution Tomography (LORETA) prior, a Gaussian Markov Random Field (GMRF) prior or a Minimum Norm (MN) prior ($D_w = I$) [9] as described in earlier work [23]. These priors are specified separately for each slice of data. Specification of 3-dimensional spatial priors (ie. over multiple slices) is desirable from a modelling perspective but is computationally too demanding for current computer technology.

We can also write $w_v = vec(W)$, $w_r = \mathsf{vec}(W^T)$, $w_v = H_w w_r$ where H_w is a permutation matrix. This leads to

$$\begin{aligned} p(W|\alpha) &= p(w_v|\alpha) \quad (5) \\ &= N(w_v; 0, B^{-1}) \end{aligned}$$

where B is an augmented spatial precision matrix given by

$$B = H_w(\mathsf{diag}(\alpha) \otimes D_w) H_w^T \quad (6)$$

This form of the prior is useful as our specification of approximate posteriors is based on similar quantities.

The above Gaussian priors underly GMRFs and LORETA and have been used previously in fMRI [26] and EEG [18]. They are by no means, however, the optimal choice for imaging data. In EEG, for example, much interest has focussed on the use of L^p-norm priors [3] instead of the L^2-norm implicit in the Gaussian assumption. Additionally, we are currently investigating the use of wavelet priors. This is an active area of research and will be the topic of future publications.

2.2.2 AR coefficients

We also define a spatial prior for the AR coefficients so that they too can be spatially regularised. We have

$$\begin{aligned} p(A|\beta) &= \prod_{p=1}^{P} p(a_p|\beta_p) \quad (7) \\ p(a_p|\beta_p) &= N(a_p; 0, \beta_p^{-1} D_a^{-1}) \end{aligned}$$

Again, D_a is a user-defined spatial precision matrix, $a_v = \mathsf{vec}(A)$, $a_r = \mathsf{vec}(A^T)$ and $a_v = H_a a_r$ where H_a is a permutation matrix. We can write

$$\begin{aligned} p(A|\beta) &= p(a_v|\beta) \\ &= N(a_v; 0, J^{-1}) \end{aligned} \tag{8}$$

where J is an augmented spatial precision matrix

$$J = H_a \left(\mathsf{diag}(\beta) \otimes D_a\right) H_a^T \tag{9}$$

This form of the prior is useful as our specification of approximate posteriors is based on similar quantities.

We have also investigated 'Tissue-type' priors which constrain AR estimates to be similar for voxels in the same tissue-type eg. gray matter, white matter or cerebro-spinal fluid. Bayesian model selection [22], however, favours the smoothly varying priors defined in equation 7.

2.2.3 Precisions

We use Gamma priors on the precisions α, β and λ

$$\begin{aligned} p(\lambda|u_1, u_2) &= \prod_{n=1}^{N} \mathsf{Ga}(\lambda_n; u_1, u_2) \\ p(\alpha|q_1, q_2) &= \prod_{k=1}^{K} \mathsf{Ga}(\alpha_k; q_1, q_2) \\ p(\beta|r_1, r_2) &= \prod_{p=1}^{P} \mathsf{Ga}(\beta_p; r_1, r_2) \end{aligned} \tag{10}$$

where the Gamma density is defined as

$$Ga(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right) \tag{11}$$

Gamma priors were chosen as they are the conjugate priors for Gaussian error models. The parameters are set to $q_1 = r_1 = u_1 = 10$ and $q_2 = r_2 = u_2 = 0.1$. These parameters produce Gamma densities with a mean of 1 and a variance of 10. The robustness of, for example, model selection to the choice of these parameters is discussed in [22].

2.3 Variational Bayes

The central quantity of interest in Bayesian learning is the posterior distribution $p(\theta|Y)$. This implies estimation both of the parameters θ and the uncertainties associated with their estimation. This can be achieved using standard Markov Chain Monte Carlo (MCMC) [12] procedures to produce samples from the posterior. Brain imaging data sets are, however, prohibitively large (typically $N = 50,000$, $T = 200$) making MCMC impractical for routine use. We have therefore adopted an approximate inference procedure which is computationally efficient. This allows inferences to be made within minutes rather than hours. The approach is called Variational Bayes (VB), a full tutorial on which is given in [16]. 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

$$\begin{aligned} \log p(Y) &= \int q(\theta|Y) \log p(Y) d\theta \\ &= \int q(\theta|Y) \log \frac{p(Y,\theta)}{p(\theta|Y)} d\theta \\ &= \int q(\theta|Y) \log \left[\frac{q(\theta|Y) p(Y,\theta)}{p(\theta|Y) q(\theta|Y)}\right] d\theta \\ &= F + KL. \end{aligned} \tag{12}$$

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

$$F = \int q(\theta|Y) \log \frac{p(Y,\theta)}{q(\theta|Y)} d\theta, \tag{13}$$

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

$$KL = \int q(\theta|Y) \log \frac{q(\theta|Y)}{p(\theta|Y)} d\theta \tag{14}$$

is the KL-divergence [6] between the density $q(\theta|Y)$ and the true posterior $p(\theta|Y)$.

Equation 12 is the fundamental equation of the VB-framework. Importantly, because the KL-divergence is always positive [6], 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. 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:

$$\begin{aligned} q(\theta|Y) &= \prod_i q(\theta_i|Y) \\ &= q(\theta_i) q(\theta^{\setminus i}) \end{aligned} \tag{15}$$

where θ_i is the ith group of parameters and $\theta^{\setminus i}$ denotes all parameters *not* in the ith group. The distributions which maximise F can then, via the calculus of variations, be shown to be [16]

$$q(\theta_i|Y) = \frac{\exp[I(\theta_i)]}{\int \exp[I(\theta_i)]d\theta_i} \quad (16)$$

where

$$I(\theta_i) = \int q(\theta^{\setminus i}|Y) \log p(Y, \theta) d\theta^{\setminus i} \quad (17)$$

Note that, importantly, this means we are able to determine the optimal analytic *form* of the component posteriors (using equation 16). This is to be contrasted with Laplace approximations where we fix the form of the component posteriors to be Gaussian about the maximum posterior solution [20].

The above principles lead to a set of coupled update rules for the *parameters* of the component posteriors, iterated application of which leads to the desired maximisation. Further, by computing F for different models, we can perform model selection. This provides a mechanism for fine-tuning models. For example, the choice of hemodynamic basis set [22] or the order of the autoregressive models [21].

It is also possible to approximate the model evidence using sampling methods [12, 4]. In the very general context of probabilistic graphical models, Beal and Ghahramani [4] have shown that the VB approximation of model evidence is considerably more accurate than the Bayesian Information Criterion (BIC) whilst incurring little extra computational cost. Moreover, model selection using VB is of comparable accuracy to a much more computationally demanding method based on Annealed Importance Sampling (AIS) [4].

2.4 Approximate Posteriors

This paper uses the Variational Bayes framework [19] for estimation and inference. We describe the algorithm developed in previous work [23] in which we assumed that the approximate posterior factorises over voxels and subsets of parameters.

Because of the spatial priors, the regression coefficients in the true posterior $p(W|Y)$ will clearly be correlated. Our perspective, however, is that this is too computationally burdensome for current personal computers to take account of. Moreover, as we shall see in section 2.4.1, updates for our approximate factorised densities $q(w_n)$ do encourage the approximate posterior means to be similar at nearby voxels, thereby achieving the desired effect of the prior.

Our approximate posterior is given by

$$q(W, A, \lambda, \alpha, \beta) = \prod_n q(w_n)q(a_n)q(\lambda_n) \prod_k q(\alpha_k) \prod_p q(\beta_p) \quad (18)$$

and each component of the approximate posterior is described below. These update equations are self-contained except for a number of quantities that are marked out using the 'tilde' notation. These are $\tilde{A}_n, \tilde{b}_n, \tilde{C}_n, \tilde{d}_n$ and $\tilde{G}_n$ which are all defined in Appendix B of [21].

2.4.1 Regression coefficients

We have

$$\begin{aligned} q(w_n) &= \mathsf{N}(w_n; \hat{w}_n, \hat{\Sigma}_n) \quad (19) \\ \hat{\Sigma}_n &= \left(\bar{\lambda}_n \tilde{A}_n + \bar{B}_{nn}\right)^{-1} \\ \hat{w}_n &= \hat{\Sigma}_n \left(\bar{\lambda}_n \tilde{b}_n^T + r_n\right) \\ r_n &= -\sum_{i=1, i\neq n}^{N} \bar{B}_{ni} \hat{w}_i \end{aligned}$$

where $\bar{B}$ is defined as in equation 6 but uses $\bar{\alpha}$ instead of α. The quantities $\tilde{A}_n$ and $\tilde{b}_n$ are expectations related to autoregressive processes and are defined in Appendix B of [21]. In the absence of temporal autocorrelation we have $\tilde{A}_n = X^T X$ and $\tilde{b}_n^T = X^T y_n$.

2.4.2 AR coefficients

We have

$$q(a_n) = \mathsf{N}(a_n; m_n, V_n)$$

where

$$\begin{aligned} V_n &= \left(\bar{\lambda}_n \tilde{C}_n + \bar{J}_{nn}\right)^{-1} \quad (20) \\ m_n &= V_n(\bar{\lambda}_n \tilde{d}_n + j_n) \\ j_n &= -\sum_{i=1, i\neq n}^{N} \bar{J}_{ni} m_i \end{aligned}$$

and $\bar{J}$ is defined as in equation 9 but $\bar{\beta}$ is used instead of β. The subscripts in $\bar{J}_{ni}$ denote that part of $\bar{J}$ relevant to the nth and ith voxels. The quantities $\tilde{C}_n$ and $\tilde{d}_n$ are expectations that are defined in equation 50 of [21].

2.4.3 Precisions

The approximate posteriors over the precision variables are Gamma densities. For the precisions on the observation noise we have

$$\begin{aligned} q(\lambda_n) &= \mathsf{Ga}(\lambda_n; b_n, c_n) \\ \frac{1}{b_n} &= \frac{\tilde{G}_n}{2} + \frac{1}{u_1} \\ c_n &= \frac{T}{2} + u_2 \\ \bar{\lambda}_n &= b_n c_n \end{aligned} \tag{21}$$

where $\tilde{G}_n$ is the expected prediction error defined in Appendix B of [21]. In the abscence of temporal autocorrelation we have

$$\tilde{G}_n = (y_n - X\hat{w}_n)^T(y_n - X\hat{w}_n) + \mathsf{Tr}(\hat{\Sigma}X^TX) \tag{22}$$

For the precisions of the regression coefficients we have

$$\begin{aligned} q(\alpha_k) &= \mathsf{Ga}(\alpha_k; g_k, h_k) \\ \frac{1}{g_k} &= \frac{1}{2}\left(\mathsf{Tr}(\hat{\Sigma}_k D_w) + \hat{w}_k^T D_w \hat{w}_k\right) + \frac{1}{q_1} \\ h_k &= \frac{N}{2} + q_2 \\ \bar{\alpha}_k &= g_k h_k \end{aligned} \tag{23}$$

For the precisions of the AR coefficients we have

$$\begin{aligned} q(\beta_p) &= \mathsf{Ga}(\beta_p; r_{1p}, r_{2p}) \\ \frac{1}{r_{1p}} &= \frac{1}{2}\left(\mathsf{Tr}(V_p D_a) + m_p^T D_a m_p\right) + \frac{1}{r_1} \\ r_{2p} &= \frac{N}{2} + r_2 \\ \bar{\beta}_p &= r_{1p} r_{2p} \end{aligned} \tag{24}$$

2.5 Practicalities

For the empirical work in this paper we set up the spatial precision matrices D_a and D_w, defined in section 2.2, to produce GMRF priors. We also used AR models of order $P = 3$.

The VB algorithm is initialised using Ordinary Least Square (OLS) estimates for regression and autoregressive parameters as described in [21]. Quantities are then updated using equations 19,20,21,23,24. These equations can then be iterated until convergence, which is defined as less than a 1% increase in F, the objective function. For the empirical work in this paper, however, we fixed the number of iterations to 4.

Expressions for computing F are given in [22]. This is an important quantity as it can also be used for model comparison. This is described at length in [22].

The algorithm we have described is implemented in SPM version 5 and can be downloaded from [1]. Computation of a number of quantites (eg. $\tilde{C}_n$, $\tilde{d}_n$ and $\tilde{G}_n$) is now much more efficient than in previous versions [23]. These improvements are described in a separate document [27]. To analyse a single session of data (eg. the face fMRI data) takes about 30 minutes on a typical modern PC.

2.6 Spatio-temporal deconvolution

The central quantity of interest in fMRI analysis is our estimate of effect sizes, embodied in contrasts of regression coefficients. A key update equation in our VB scheme is, therefore, the approximate posterior for the regression coefficients. This is given by equation 19. For the special case of temporally uncorrelated data we have

$$\begin{aligned} \hat{\Sigma}_n &= \left(\bar{\lambda}_n X^T X + \bar{B}_{nn}\right)^{-1} \\ \hat{w}_n &= \hat{\Sigma}_n\left(\bar{\lambda}_n X^T y_n + r_n\right) \end{aligned} \tag{25}$$

where $\bar{B}$ is a spatial precision matrix and r_n is the weighted sum of neighboring regression coefficient estimates.

This update indicates that the regression coefficient estimate at a given voxel regresses towards those at nearby voxels. This is the desired effect of the spatial prior and it is preserved despite the factoristion over voxels in the approximate posterior (see equation 18). Equation 25 can be thought of as the combination of a temporal prediction $X^T y_n$ and a spatial prediction from r_n. Each prediction is weighted by its relative precision to produce the optimal estimate $\hat{w}_n$. In this sense, the VB update rules provide a spatio-temporal deconvolution of fMRI data. Moreover, the parameters controlling the relative precisions, $\bar{\lambda}_n$ and $\bar{\alpha}$ are estimated from the data. This means that our effect size estimates derive from an automatically regularised spatio-temporal deconvolution.

2.7 Global scaling

Before statistical analysis, brain imaging data are usually scaled in some way. In the SPM software package [1], for example, imaging data are, by default, scaled according to the following procedure. Firstly, the global mean value is computed

$$g = \frac{1}{TN}\sum_{t=1}^{T}\sum_{n=1}^{N} u_{nt} \tag{26}$$

where u_{nt} are fMRI values at voxel n and time point t. Scaled images are then computed as

$$y_{nt} = \frac{100}{g} u_{nt} \tag{27}$$

If we express an activated voxel as

$$u_a = u_b + \Delta u \tag{28}$$

where u_b is the baseline value and Δu is the absolute amount of activation then the difference in scaled image data is then given by

$$\begin{aligned} y_a - y_b &= \frac{100(u_b + \Delta u)}{g} - \frac{100 u_b}{g} \\ &= \frac{100 \Delta u}{g} \end{aligned} \tag{29}$$

So, differences in scaled image values correspond to changes in the original data that are expressed as percentages of the global mean. Because estimated regression coefficients are just estimates of differences in scaled image values, eg. the difference between conditions, then they too have this interpretation. The regression coefficients in GLMs (and contrasts of them, see next section) reflect effect sizes as a percentage of g. This is important as these effect sizes can be plotted in Posterior Probability Maps (see section 5).

3. Contrasts

After having estimated a model, we will be interested in characterising a particular effect, c, which can usually be expressed as a linear function or 'contrast' of parameters, w. That is,

$$c = C^T w \tag{30}$$

where C is a contrast vector or matrix. For example, the contrast vector $c^T = [1 - 1]$ will allow us to look at the difference between two experimental conditions.

Our statistical inferences will be based on the approximate distribution. Because this factorises over voxels we can write

$$q(c) = \prod_{n=1}^{N} q(c_n) \tag{31}$$

where c_n is the effect size at voxel n. Given a contrast matrix C we have

$$q(c_n) = \mathsf{N}(c_n; m_n, V_n) \tag{32}$$

with mean and covariance

$$\begin{aligned} m_n &= C \hat{w}_n \\ V_n &= C^T \hat{\Sigma}_n C \end{aligned} \tag{33}$$

Bayesian inference based on this posterior can then take place using confidence intervals [5]. For univariate contrasts we have suggested the use of Posterior Probability Maps (PPMs). Before discussing this at length in section 4, we describe a new approach that allows us to make inferences about multivariate contrasts. That is, where c_n is a vector.

3.1 Multivariate contrasts

The probability α that the zero vector lies on the $1 - \alpha$ confidence region of the posterior distribution at each voxel is computed as follows. We first note that this probability is the same as the probability that the vector m_n lies on the edge of the $1 - \alpha$ confidence region for the distribution $\mathsf{N}(m_n; 0, V_n)$. This latter probability can be computed by forming the test statistic

$$d_n = m_n^T V_n^{-1} m_n \tag{34}$$

which will be the sum of $v_n = rank(V_n)$ independent, squared Gaussian variables. As such it has a χ^2 distribution

$$p(d_n) = \chi^2(v_n) \tag{35}$$

This procedure is identical to that used for making inferences in Bayesian Multivariate Autoregressive Models [13]. We can also use this procedure to test for two-sided effects, that is, activations or deactivations. Although these contrasts are univariate we will use the term 'multivariate contrasts' to also include the assessment of these two-sided effects.

4. Thresholding

In previous work [9] we have suggested deriving Posterior Probability Maps (PPMs) by applying two thresholds to the posterior distributions (i) an effect size threshold, γ, and (ii) a probability threshold p_T. Voxel n is then included in the PPM if $q(c_n > \gamma) > p_T$.

If voxel n is to be included then the posterior exceedance probability $q(c_n > \gamma)$ is plotted. In this paper, we instead propose plotting the effect size itself, c_n.

We also propose the following procedure for exploring the posterior distribution of effect sizes. Firstly, plot a map of effect sizes using the thresholds

$\gamma = 0$ and $p_T = 1 - 1/N$ where N is the number of voxels. We refer to these values as the default thresholds. Then, after visual inspection of the resulting map use a non-zero γ, the value of which reflects effect sizes in areas of interest. It will then be possible to reduce p_T to a value such as 0.95.

Of course, if previous imaging analyses have indicated what effect sizes are physiologically relevant then this exploratory procedure is unnecessary. Alternatively, one could stick with the default thresholds.

4.1 False positive rates

If we partition effect-size values into two hypothesis spaces $H0 : c \leq \gamma$ and $H1 : c > \gamma$ then we can characterise the sensitivity and specificity of our algorithm. This is different to classical inference which uses $H0 : c = 0$. A False Positive (FP) occurs if we accept $H1$ when $H0$ is true.

If we use the default threshold and the approximate posterior were exact then the distribution of FPs is binomial with rate $p = 1/N$. The mean and variance of the binomial distribution are Np and $Np(1-p)$. The expected number of false positives in each PPM is therefore $N \times 1/N = 1$. The variance is $N \times 1/N \times (1 - 1/N)$ which is approximately 1. We would therefore expect 0, 1 or 2 false positives per PPM. We suggest that this is sufficiently high specificity for brain imaging analyses.

Of course, the above result only holds if the approximate posterior is equal to the true posterior. But given that all of our computational effort (see section 2.4) is aimed at this goal it would not be too surprising if the above analysis were indicative of actual FP rates. This issue will be investigated using Null fMRI data in the next section.

5. Results

5.1 Null data

This section describes the analysis of a Null data set to find out how many false positives are obtained using PPMs with default thresholds.

Images were acquired from a 1.5T Sonata(Siemens, Erlangen Germany) which produced T2*-weighted transverse Echo-Planar Images (EPIs) with BOLD contrast. Whole brain EPIs consisting of 48 transverse slices were acquired every TR=4.32s resulting in a total of $T = 98$ scans. The voxel size is $3 \times 3 \times 3$mm. All images were realigned to the first image using a six-parameter rigid-body transformation to account for subject movement.

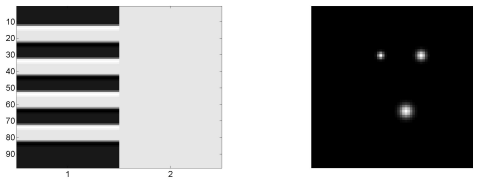

Figure 2: *Left: Design matrix for null fMRI data. The first column models a boxcar activation and the second column models the mean. There are* $n = 1..98$ *rows corresponding to the 98 scans. Right: Image of regression coefficients corresponding to synthetic activation added to null data. In this image black is 0 and white is 1.*

We then implemented a standard whole volume analysis on images comprising $N = 59,945$ voxels. We used the design matrix shown in the left panel of Figure 2. This design is used in the following section. Use of the default thresholds resulted in no spurious activations in the PPM. We then repeated the above analysis but with a number of different design matrices.

These were based on the epoch design in Figure 2 but each epoch onset was jittered by a number between plus or minus 9 scans, sampled from a uniform distribution, and the epoch durations were drawn from a uniform distribution between 4 and 10 scans. Ten such designs were created and VB models fitted to the null data. For designs 1 to 10, the numbers of false positives were 0,0,1,2,4,0,0,1,0 and 0. This is close to what we'd expect from the binomial analysis described in section 4.1.

5.2 Synthetic data

We then added three synthetic activations to a slice of null data ($z = -13$mm). These were created using the design matrix and regression coefficient image shown in Figure 2 (the two regression coefficient images, ie. for the activation and the mean, were identical). These images were 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). Images were then rescaled to make the peaks unity.

In principle, smoothing with a Gaussian kernel renders the true effect size greater than zero everywhere because a Gaussian has infinite spatial support. In practice, however, when implemented on a digital computer with finite numerical precision most voxels will be numerically zero. Indeed, our simulated data contained 299 'activated' voxels ie.

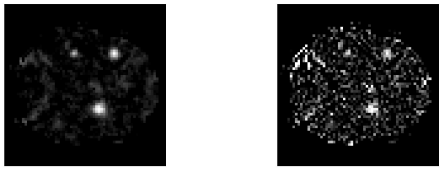

Figure 3: *Left: Effect as estimated using VB Right: Effect as estimated using OLS. The true effect is shown in the right plot in Figure 2. In these images, black denotes 0 and white 1.*

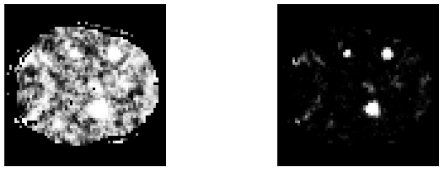

Figure 4: *Plots of exceedance probabilities $p(c_n > \gamma)$ for two thresholds. Left: $\gamma = 0$. Right: $\gamma = 0.3$. In these images, black denotes 0 and white 1.*

voxels with effect sizes numerically greater than zero.

This slice of data was then analysed using VB. The contrast $c = [1, 0]^T$ was then used to look at the estimated activation effect. This is shown in the left panel of Figure 3. For comparison, we also show the effect as estimated using OLS. Clearly, OLS estimates are much noisier than VB estimates.

Figure 4 shows plots of the exceedance probabilities for two different effect-size thresholds, $\gamma = 0$ and $\gamma = 0.3$. Figure 5 shows thresholded versions of these images. These are PPMs. Neither of these PPMs contain any false positives. That is, the true effect size is greater than zero wherever a white voxel occurs. This shows, informally, that use of the default thresholds provides good specificity whilst retaining reasonable sensitivity. Also, a combination of non-zero effect-size thresholds and more liberal probability thresholds can do the same.

Figure 5: *PPMs for two thresholds. Left: The default thresholds ($\gamma = 0$, $p_T = 1 - 1/N$) Right: The thresholds $\gamma = 0.3$, $p_T = 0.95$. In these images, black denotes 0 and white 1.*

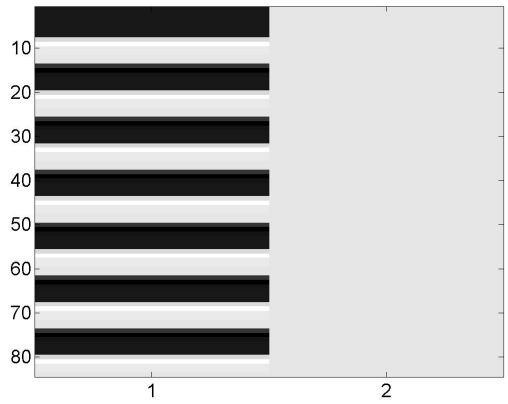

Figure 6: *Design matrix for analysis of the auditory data. The first column models epochs of auditory stimulation and the second models the mean response.*

5.3 Auditory data

This section describes the use of multivariate contrasts for an auditory fMRI data set. This data set comprises whole brain BOLD/EPI images acquired on a modified 2T Siemens Vision system. Each acquisition consisted of 64 contiguous slices (64x64x64 3mm x 3mm x 3mm voxels). A time series of 96 images was acquired with TR=7s from a single subject.

This was an epoch fMRI experiment in which the condition for successive epochs alternated between rest and auditory stimulation, starting with rest. Auditory stimulation was bi-syllabic words presented binaurally at a rate of 60 per minute.

These data were analysed using VB with the design matrix shown in Figure 6. To look for voxels that increase activity in response to auditory stimulation we used the univariate contrast $c = [1, 0]^T$. Figure 7 shows a PPM that maps effect-sizes of above threshold voxels.

To look for either increases or decreases in activity we use the multivariate contrast $c = [1, 0]^T$. This inference uses the χ^2 approach described earlier. Figures 8 shows the PPM obtained using default thresholds.

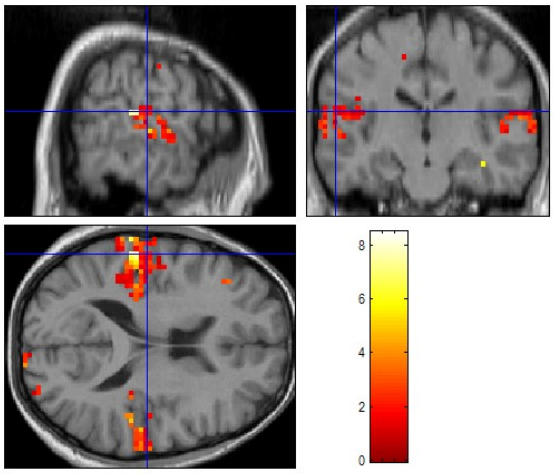

Figure 7: *PPM for positive auditory activation. Overlay of effect-size, in units of percentage of global mean, on subjects MRI for above threshold voxels. The default thresholds were used, that is, we plot c_n for voxels which satisfy $p(c_n > 0) > 1 - 1/N$.*

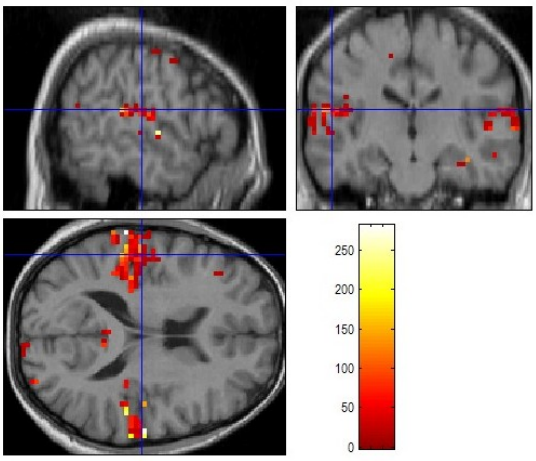

Figure 8: *PPM for positive or negative auditory activation. Overlay of χ^2 statistic on subjects MRI for above threshold voxels. The default thresholds were used, that is, we plot χ^2_n for voxels which satisfy $p(c_n > 0) > 1 - 1/N$.*

5.4 Face data

This is an event-related fMRI data set acquired by Henson et al. [15]. The data were acquired during an experiment concerned with the processing of images of faces [15]. This was an event-related study in which greyscale images of faces were presented for 500ms, replacing a baseline of an oval chequerboard which was present throughout the interstimulus interval. Some faces were of famous people and were therefore familiar to the subject and others were not. Each face in the database was presented twice. This paradigm is a two-by-two factorial design where the factors are familiarity and repetition. The four experimental conditions are 'U1', 'U2', 'F1' and 'F2' which are the first or second (1/2) presentations of images of familiar 'F' or unfamiliar 'U' faces.

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 two seconds 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 [2]. Each time series was then high-pass filtered using a set of discrete cosine basis functions with a filter cut-off of 128 seconds.

The data were then analysed using the design matrix shown in Figure 9. The first 8 columns contain stimulus related regressors. These correspond to the four experimental conditions, where each stimulus train has been convolved with two different hemodynamic bases (i) the canonical Hemodynamic Response Function (HRF) and (ii) the time derivative of the canonical [14]. The next 6 regressors in the design matrix describe movement of the subject in the scanner and the final column models the mean response.

The model was then fitted using the VB algorithm. Figure 10 plots a map of the first autoregressive component as estimated using VB. This shows a good deal of heterogeneity and justifies our assumption that that AR coefficients are spatially varying. The estimated spatial variation is smooth, however, due to the spatial prior.

Figure 11 shows a PPM for 'Any effect of faces'. This was obtained using the multivariate contrast matrix shown in Figure 9.

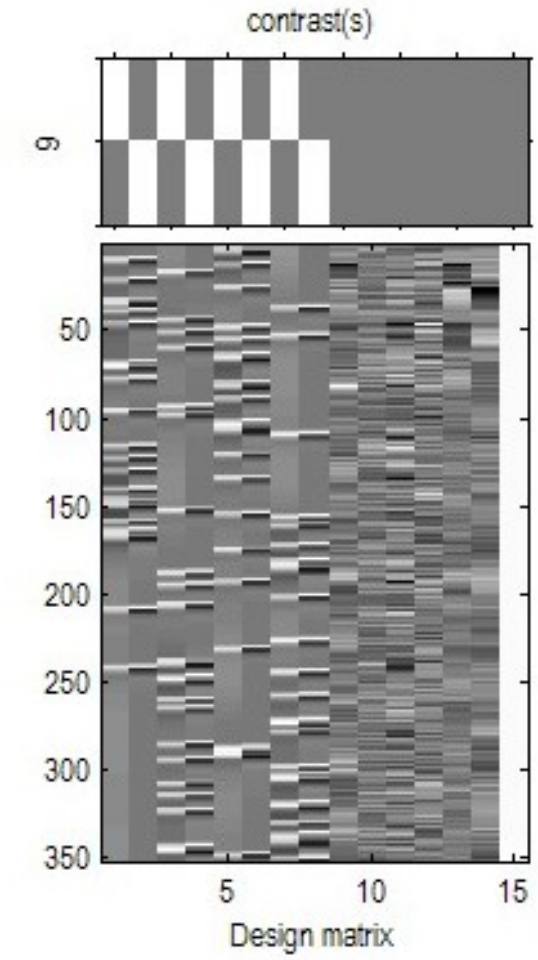

Figure 9: *Lower part: Design matrix for analysis of face data, Upper part: Multivariate contrast used to test for any effect of faces.*

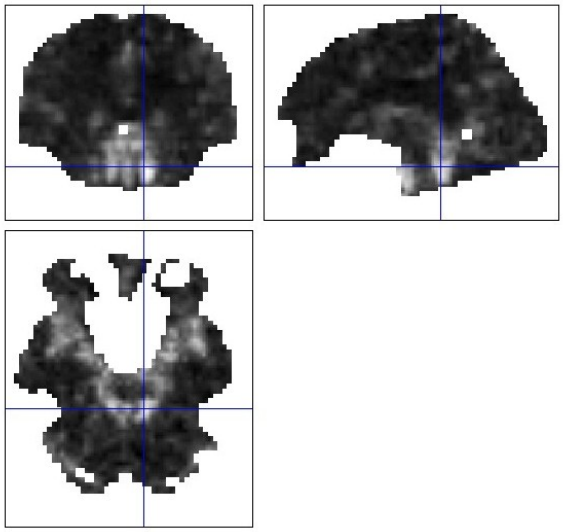

Figure 10: *Image of the first autoregressive coefficient estimated from the face fMRI data. In these images, black denotes 0 and white 1.*

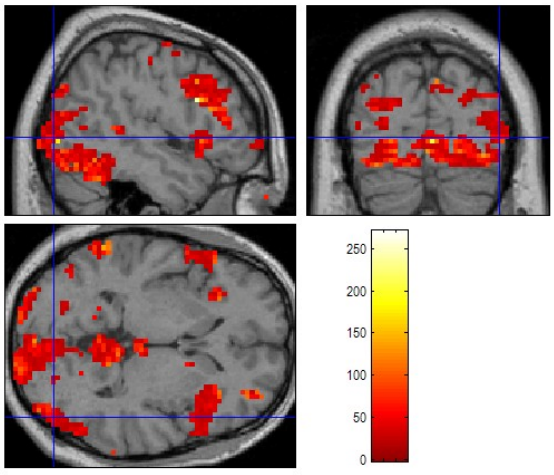

Figure 11: *PPM showing above threshold χ^2 statistics for any effect of faces*

6. Discussion

In previous work[10], we have compared the sensitivity and specificity of classical inference to Bayesian inference with Minimum Norm (MN) priors. This comparison was made possible by looking at the expected properties of the Bayesian estimators across an ensemble of data sets, and showed that the Bayesian inference was no more sensitive than the classical inference.

Bayesian inferences can be used, however, to look for effects greater than a physiologically relevant size eg. 0.5% of the global mean. These inferences can be presented visually using Posterior Probability Maps (PPMs) and have high intrinsic specificity. Also, one can use Bayesian inference to assess the absence of an experimental effect [10].

We have since developed a Bayesian inference framework based on spatial priors which has been reviewed in this paper. This prior embodies our knowledge that evoked responses are spatially homogeneous and locally contiguous. The approach may be viewed as an automatically regularized spatio-temporal deconvolution algorithm.

As compared to standard approaches based on spatially smoothing the imaging data itself, the spatial regularisation procedure has been shown to result in inferences with higher sensitivity [23].

In this paper we have also described a new PPM procedure for making inferences about multivariate contrasts. This allows us to make inferences about (i) hemodynamic responses that are characterised by multiple basis functions (ii) main effects and interactions in factorial fMRI designs and (iii) two-sided effects. The procedure uses the same χ^2 approach that has previously been used in the context of Multivariate Autoregressive (MAR) models.

PPMs provide a visual representation of the posterior distribution of effect sizes across the brain. They are generated using an effect size threshold and a probability threshold. These two thresholds convert a Gaussian posterior distribution at each voxel, specified by two quantities - the mean and variance, into a single quantity that can be mapped eg. the exceedance probability, effect size or statistic value. One can create PPMs using various effect-size thresholds, which can be chosen on the basis of prior knowledge about what is physiologically revelant in a given experimental context. One can also vary the probability thresholds used, although 0.95 would be a typical value. The use of these different thresholds allows for a visual exploration of the posterior distribution.

However, it is also useful to categorize responses into regions which are or are not activated. Such categorization is usually unavoidable when reporting neuroimaging results because one has to choose which regions to report in tables and, indeed, discuss. In this paper we have therefore described a simple new procedure for setting the thresholds that generate PPMs. These 'default thresholds' comprise an effect size threshold of zero and a probability threshold of $1 - 1/N$ where N is the number of voxels in the volume. Use of PPMs with 'default thresholds' resulted in low false positive rates for null fMRI data, and physiologically plausible activations for auditory and face fMRI data sets.

A comprehensive Bayesian thresholding approach has been implemented by Woolrich et al. [24]. This work uses explicit models of the null and alternative hypotheses based on Gaussian and Gamma variates. This requires a further computationally expensive stage of model-fitting, based on spatially regularised discrete Markov Random Fields, but has the benefit that false-positive and true-positive rates can be controlled explicitly.

Acknowledgements

W.D. Penny is funded by the Wellcome Trust. We would also like to thank Chloe Hutton for providing the Null fMRI data and Karl Friston, Tom Nichols and John Aston for commenting on earlier drafts of this manuscript.

References

[1] SPM, Wellcome Department of Imaging Neuroscience. Available from http://www.fil.ion.ucl.ac.uk/spm/software/, 2002.

[2] J. Ashburner and K.J. Friston. Spatial normalization using basis functions. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

[3] T. Auranen, A. Nummenmaa, M. Hammalainen, I. Jaaskelainen, J. Lampinen, A. Vehtari, and M. Sams. Bayesian analysis of the neuromagnetic inverse problem with l^p norm priors. *Neuroimage*, 2005. In Press.

[4] M. Beal and Z. Ghahramani. The Variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In J.M. Bernardo, M.J. Bayarri,

J.O. Berger, A.P. Dawid, D. Heckerman, A.F. Smith, and M. West, editors, *Bayesian Statistics*, volume 7. Oxford University Press, 2003.

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

[6] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[7] R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny. *Human Brain Function*. Academic Press, 2nd edition, 2003.

[8] K.J. Friston, A.P. Holmes, K.J. Worsley, J.B. Poline, C. Frith, and R.S.J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

[9] K.J. Friston and W.D. Penny. Posterior probability maps and SPMs. *NeuroImage*, 19(3):1240–1249, 2003.

[10] K.J. Friston, W.D. Penny, C. Phillips, S.J. Kiebel, G. Hinton, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage*, 16:465–483, 2002.

[11] T. Gautama and M.M. Van Hulle. Optimal spatial regularisation of autocorrelation estimates in fMRI analysis. *NeuroImage*, (23):1203–1216, 2004.

[12] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, Boca Raton, 1995.

[13] L. Harrison, W.D. Penny, and K.J. Friston. Multivariate autoregressive modelling of fMRI time series. *NeuroImage*, 19(4):1477–1491, 2003.

[14] R.N.A. Henson. Analysis of fMRI time series. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

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

[16] H. Lappalainen and J.W. Miskin. Ensemble Learning. In M.Girolami, editor, *Advances in Independent Component Analysis*. Springer-Verlag, 2000.

[17] J. Marchini and B. Ripley. A new statistical approach to detecting significant activation in functional MRI. *Neuroimage*, 12:168–193, 2000.

[18] R. Pascual Marqui, C. Michel, and D. Lehman. Low resolution electromagnetic tomography: a new method for localizing electrical activity of the brain. *International Journal of Psychophysiology*, pages 49–65, 1994.

[19] M.Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

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

[21] W.D. Penny, S.J. Kiebel, and K.J. Friston. Variational Bayesian Inference for fMRI time series. *NeuroImage*, 19(3):727–741, 2003.

[22] W.D. Penny and N. Trujillo-Bareto. Bayesian comparison of spatially regularised general linear models. *Human Brain Mapping*, 2005. Accepted.

[23] W.D. Penny, N. Trujillo-Barreto, and K.J. Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.

[24] M. W. Woolrich, T.E.J. Behrens, C.J. Beckman, and S. M. Smith. Mixture models with adaptive spatial regularisation for segmentation with an application to fMRI data. *IEEE Transactions on Medical Imaging*, 14(6):1370–1386, December 2004.

[25] M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith. Temporal autocorrelation in univariate linear modelling of fMRI data. *NeuroImage*, 14(6):1370–1386, December 2001.

[26] M.W. Woolrich, T.E. Behrens, and S.M. Smith. Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, 21:1748–1761, 2004.

[27] W.Penny and G. Flandin. Bayesian analysis of single-subject fMRI: SPM implementation. Technical report, Wellcome Department of Imaging Neuroscience, 2005.