Bayesian fMRI Data Analysis with Sparse Spatial Basis Function Priors

Guillaume Flandin [a,*,1] William D. Penny [a]

[a] *Wellcome Department of Imaging Neuroscience, UCL, London, UK*

Abstract

In previous work we have described a spatially regularised General Linear Model (GLM) for the analysis of brain functional Magnetic Resonance Imaging (fMRI) data where Posterior Probability Maps (PPMs) are used to characterise regionally specific effects. The spatial regularisation is defined over regression coefficients via a Laplacian kernel matrix and embodies prior knowledge that evoked responses are spatially contiguous and locally homogeneous. In this paper we propose to finesse this Bayesian framework by specifying spatial priors using Sparse Spatial Basis Functions (SSBFs). These are defined via a hierarchical probabilistic model which, when inverted, automatically selects an appropriate subset of basis functions. The method includes nonlinear wavelet shrinkage as a special case. As compared to Laplacian spatial priors, SSBFs allow for spatial variations in signal smoothness, are more computationally efficient and are robust to heteroscedastic noise. Results are shown on synthetic data and on data from an event-related fMRI experiment.

Key words: Variational Bayes, fMRI, Sparse spatial prior, Wavelet denoising, General linear model, Hierarchical model.

1 Introduction

Functional Magnetic Resonance Imaging (fMRI) is an established technique for making inferences about regionally specific activations in the human brain

* Corresponding author. Wellcome Department of Imaging Neuroscience, 12 Queen Square, London WC1N 3BG, UK. Fax: +44 20 7833 7478.

Email addresses: gflandin@fil.ion.ucl.ac.uk (Guillaume Flandin), wpenny@fil.ion.ucl.ac.uk (William D. Penny).

[1] G.F. is now with Service Hospitalier Frédéric Joliot, CEA/DSV/DRM, 4 Place du Général Leclerc, 91401 Orsay, France. Fax: +33 1 69 86 77 86.

(Frackowiak et al., 2003). Blood Oxygen Level-Dependent (BOLD) effects are modelled in a statistical framework such as the General Linear Model (GLM) to obtain probability maps of the underlying neuronal activations for a particular task (Friston et al., 1995b).

In the GLM framework, fMRI times series are modelled at each and every voxel by a linear combination of several regressors, defined as explanatory variables corresponding to some experimental effects. These regressors are built using a convolution model: putative neuronal signals are convolved with a set of hemodynamic basis functions such as the canonical Hemodynamic Response Function (HRF) and its latency and dispersion derivatives (Friston et al., 1998). This accounts for variability in the shape of the response from one brain region to another. The inversion of this mass univariate model yields voxel-wise estimates of the regression coefficients as well as their variance. Classical statistical inference can be performed to provide activation maps linked to a particular contrast. Random Field Theory (RFT) then provides a correction to the obtained p-values that accounts for spatial correlation in the data.

The spatial aspect of the hemodynamic response is usually taken into account indirectly, i.e. not modelled explicitly, by spatially smoothing the data with a fixed Gaussian kernel, as a preprocessing step. This corresponds to averaging – or blurring – the measured signal over a neighbourhood, which will increase the spatial correlation. The size of this neighbourhood is defined by the Full Width at Half Maximum (FWHM) of the Gaussian kernel, often chosen to be around three times the voxel width size (between 6 and 12 millimeters). The rationale for this spatial filtering is threefold. First, it helps improve the signal to noise ratio (SNR). This is because the signal of interest usually extends over several voxels. This is due both to the possibly distributed nature of neuronal sources and the spatially extended nature of the hemodynamic response. The matched-filter theorem (Worsley et al., 1996) states that one improves the SNR if one smoothes the data with a filter whose kernel equals the spatial point response function (PRF) of the process that generated them. Second, Random Field Theory has been elaborated for spatially continuous fields and appropriate smoothing ensures that discretely sampled imaging data is a good 'lattice approximation' to a continuous field. Third, activation location is known to vary across subjects and smoothing accommodates for these between-subject differences in functional anatomy.

Smoothing the data with a nonadpative fixed Gaussian kernel can however suffer from drawbacks. Indeed, the spatial hemodynamic point response function is not known and has to be guessed: over- or under specification of the FWHM will lead to a sub-optimal increase in SNR. Too much smoothing will blur activations, leading to a biased estimate of both height and location of activation peaks, while too little will leave unnecessary noise in the data. Fur-

thermore, if the PRF is nonstationary, then smoothing with a nonadaptive fixed size Gaussian kernel is clearly sub-optimal.

The other arguments advocating for smoothing the data are here irrelevant because in this paper we will work within a Bayesian inference framework (Friston and Penny, 2003), so we have no need to appeal to RFT. Furthermore, this paper describes a model for single subject analysis and there is thus no need to take into account intersubject variability. If, however, one were interested in multi-subject analysis an alternative to spatial smoothing is provided by parcellation (Thirion et al., 2006).

In the recent literature, several approaches have been proposed to replace Gaussian smoothing of the data by more elaborate denoising techniques from the image processing and computer vision research fields: anisotropic filtering (Solé et al., 2001; Kim et al., 2005), adaptive spatial filters (Friman et al., 2003), scale space analyses (Poline and Mazoyer, 1994), Markov random fields (Descombes et al., 1998), surface based analyses (Kiebel et al., 2000; Andrade et al., 2001), mixture models (Everitt and Bullmore, 1999; Hartvig and Jensen, 2000; Penny and Friston, 2003; Woolrich et al., 2005) or wavelet shrinkage (Wink and Roerdink, 2004). Many of these techniques still consider spatial modelling of the data as a preprocessing step that is applied before statistical analysis. We contend that a better approach is to have spatial features of the data as part of a probabilistic model, removing the need for preprocessing with arbitrary parameters. This is in contradistinction with preprocessing which do not allow spatial filtering strength to be automatically adapted to the data. This kind of approaches has already been introduced in some of the cited references and it motivates the work in this paper.

Spatial characteristics of fMRI can be naturally described in a Bayesian framework. Several approaches have been proposed in the recent literature to model spatial dependencies in this context (Gössl et al., 2001; Woolrich et al., 2004; Penny et al., 2005b). In particular, Penny et al. (2005b) have proposed a fully Bayesian model with spatial priors defined over the regression coefficients of a General Linear Model, using Laplacian operators or Gaussian Markov Random Fields (GMRF). Spatial regularisation is then part of the estimation procedure and smoothing the data with an arbitrary Gaussian kernel is no longer required. Results show an improvement in sensitivity compared to other spatially non-informed approaches (Penny et al., 2003). However, these kinds of priors does not handle spatial variations in smoothness arising, e.g. from regional differences in vasculature or functional anatomy.

In this work, we propose to finesse this previous approach by replacing the GMRF prior with a Sparse Spatial Basis Function (SSBF) prior in which irrelevant bases are automatically switched off using a mixture model. One of the key features of decomposing data with an appropriate basis is that deter-

ministic signal will be explained by a few large coefficients while background noise will be modelled by many very small coefficients. Setting these coefficients to zero, or at least reducing their value, performs an intrinsic shrinkage or denoising. This can be implemented automatically using a sparse prior on the spatial basis set coefficients.

Among all spatial basis sets, wavelet bases (Mallat, 1989, 1999) are of primary interest because they lead to a multiresolution decomposition that shows a natural adaptivity to nonstationary features, as well as providing decorrelation and compaction properties. The use of wavelets for fMRI has already been proposed in the literature (Ruttimann et al., 1998; Van De Ville et al., 2004; Aston et al., 2005, 2006), see (Bullmore et al., 2004) for an overview. They have been used for denoising, multiresolution hypothesis testing, linear model estimation in the wavelet domain and "wavestrapping" (data resampling in the wavelet domain). In this paper, we are primarily interested in data denoising that can be obtained via wavelet shrinkage, also referred to as nonparametric regression (Donoho and Johnstone, 1994, 1995; Antoniadis et al., 2001). The basic concept is very simple and involves three steps: first, noisy data are projected into wavelet space, then the coefficients are thresholded or shrunk, and finally data are projected back into their original space. This yields adaptively regularized nonparametric estimates of the signal underlying the data.

As compared to monoresolution Gaussian smoothing, spatial wavelet-based denoising techniques for fMRI have been shown to better preserve image sharpness and retain the original shapes of active regions (Wink and Roerdink, 2004). The relationship between Gaussian smoothing and wavelet shrinkage has also been explored in (Van De Ville et al., 2003; Fadili and Bullmore, 2004). Another alternative to monoresolution Gaussian analysis is to perform a scale space analysis using multiple Gaussian kernels of different widths (Poline and Mazoyer, 1994; Godtliebsen et al., 2004), but this has the drawback that different levels in scale space are highly correlated. This is in contradistinction to wavelet bases that have a whitening property and allow for a parsimonious representation.

The SSBF approach that we propose in this paper results in a very general Bayesian inference framework for imaging data. It contains both nonlinear wavelet shrinkage analysis and Ordinary Least Square estimation as special cases. More generally, the proposed model is a fully non-separable spatio-temporal model in which the GLM is used for a temporal decomposition with parameters that are spatially constrained by a SSBF prior.

Overview

The rest of the paper is organized as follows. In the "Theory" section we describe our probabilistic generative model of fMRI time series with a particular emphasis on the SSBF prior. We then show how a Variational Bayes approach is used to define approximate posteriors and how it provides a set of updates for the sufficient statistics of these distributions. Then, after providing implementation details, we present results obtained on synthetic data and an event-related fMRI dataset. In the "Discussion" section we outline the main qualities of our model compared to other spatio-temporal models already published in the literature and suggest starting points for further work. In Appendices A and B, we give definitions of the probabilistic density functions we use and an overview of the Variational Bayes framework. Practical details of the derivation of the approximate posterior distributions for the SSBF model are available as an online supplementary material [2] .

2 Theory

2.1 Notation

We denote a matrix in upper case, while a vector is lower case. Subscripts are used to select a particular row/column of a matrix, e.g., if X is a $M \times N$ matrix then x_n is the nth column of X while x_m^T is the mth row. Unless stated otherwise, subscripts k, l, m and n are respectively indexes over regressors, wavelet levels, mixture components and voxels. Following a `Matlab`-like notation, we define the `diag` operator which transforms a vector in a diagonal matrix and the `blkdiag` operator which concatenates several matrices to create a block diagonal one. `tr` denotes the trace of a square matrix. See Appendix A for a definition of probability distributions and standard results used throughout this article.

2.2 Hierarchical Bayesian fMRI model

Our model of fMRI time series can be described as several levels embedded in a hierarchy, where each level acts as a prior on the level underneath. Temporal modelling of the data is implemented using the General Linear Model (GLM). This level is then constrained via the wavelet transform, which implements spatial modelling of the data, in association with the sparse prior on the

[2] `http://www.fil.ion.ucl.ac.uk/spm/doc/papers/gf_sparse_vb_supp.pdf`

coefficients of that transformation. The overall probabilistic generative model is shown in Fig. B.1. In the next sections we chose to describe the formulation of the model starting from the data up to the highest priors, which means reading the graphical representation of Fig. B.1 in a bottom-up manner.

[Fig. 1 about here.]

2.2.1 Problem formulation

The standard mass-univariate method to analyse fMRI data relies on the GLM (Friston et al., 1995b). Data Y comprising N voxels with time courses of length T (stored as a $T \times N$ matrix) are explained in terms of a $T \times K$ design matrix X containing K regressors at each of N voxels

$$Y = XW + E^{(1)} \tag{1}$$

i.e. for each time course

$$y_n = Xw_n + e_n^{(1)} \tag{2}$$

where W is a $K \times N$ matrix of regression coefficients and $E^{(1)}$ is a $T \times N$ error matrix. We assume that the noise follows an independent identically distributed (i.i.d.) Gaussian distribution

$$p(e_n^{(1)}) = N\left(e_n^{(1)}; 0, \lambda_n^{-1} I_T\right) \tag{3}$$

where λ_n denotes the noise precision for voxel n. This assumption is of course approximate because of the presence of temporal autocorrelation in the data, but in this paper we focus on the signal model. We could, however, easily update the present model to deal with serial correlation using autoregressive (AR) processes as described in (Penny et al., 2003). This is referred to in Fig. B.1 which augments the probabilistic model accordingly.

2.2.2 Likelihood

Assuming conditional independence, we get the following factorisation over voxels

$$p(Y|W, \lambda) = \prod_{n=1}^{N} p(y_n|w_n, \lambda_n) \tag{4}$$

with

$$p(y_n|w_n, \lambda_n) = N(y_n; Xw_n, \lambda_n^{-1} I_T) \tag{5}$$

This linear model is the same as in classical maximum likelihood analysis but here Bayesian analysis relies upon the specification of prior expectations about the parameters of the model $\{W, \lambda\}$. One can then compute the probability of the activation given the data, i.e. the posterior density (Friston and Penny,

2003), via Bayes rule. This is precluded in classicial inference, which simply reports the probability of observing a statistic derived from the data (or more extremal value) assuming no activation.

2.3 Priors

In this section we define priors over the parameters of the GLM. The next subsection describes the spatial decomposition of the regression coefficients and following subsections describe the sparse prior defined on the basis set coefficients.

2.3.1 Regression coefficients

Each regression coefficient image w_k^T (a row of matrix W) is decomposed using a spatial basis set. This spatial decomposition is at the heart of our model and should be chosen to represent data with parsimony. In other words, the transformation of the regression coefficients images should yield a very sparse representation with many coefficients near to zero. In data compression, many basis sets have been proposed e.g. wavelets (Discrete Wavelet Transform – DWT), Fourier (Discrete Fourier Transform – DFT), cosine (Discrete Cosine Transform – DCT), Karhunen-Loève (Principal Component Analysis – PCA), Independent Component Analysis (ICA). Projecting data onto these bases is a linear operation that can be inverted without losing information whereas lossy representations can be formed by removing components. The DCT, for example, has an energy compaction property such that most of the information in natural images tends to be concentrated in a few low-frequency components. These features allow for image compression.

In the following, we will consider wavelets as the spatial basis set of choice because of their specific features that we describe below, but the same framework can be applied to any other transform. Wavelet bases have been described at length in (Mallat, 1989, 1999). They consist of a multiresolution hierarchy in which an image is represented at a number of spatial resolutions. These are known as the "coarse" levels where lower levels correspond to successively lower frequency aspects of the original image. The difference between successive coarse level images are the "detail" images. These correspond to high frequency components [3]. Overall, a wavelet decomposition transforms a d-dimensional image into a d-dimensional image of wavelet coefficients. These coefficients constitute the coarse and detail levels making up the multiresolution hierarchy as shown in Fig. B.2.

[3] In this paper only the lowest frequency coarse level is referred to as the 'coarse level'

[Fig. 2 about here.]

Importantly, the discrete wavelet transform (DWT) is orthogonal and can be implemented efficiently through quadrature mirror filterbanks (QMF). This uses an algorithm whose computational complexity is $O(N)$ where N is the number of input samples. An image can then be exactly reconstructed using a fast inverse discrete wavelet transform (IDWT). Furthermore, the decomposition easily extends to the multidimensional case by using tensor-product basis functions. In the 2D case, detail coefficients can be split into diagonal, horizontal and vertical subbands.

In this paper we represent the spatial wavelet basis set by a $N \times N$ matrix V. The decomposition for each regression coefficient image w_k^T is

$$w_k^T = V z_k^T + e_k^{(2)} \tag{6}$$

where z_k^T is the corresponding wavelet coefficient image. If some "basis switches" are turned off (see below), w_k^T cannot be reconstructed exactly from z_k^T. This inexactness is accounted for by the error term $e_k^{(2)}$. Equation 6 can be rephrased to deal with all the regression coefficients W at the same time

$$W = Z V^T + E^{(2)} \tag{7}$$

where Z is a $K \times N$ matrix containing the coefficients of the wavelet transform of the regression coefficients. $E^{(2)}$ denotes the residuals of this decomposition, considered as following an i.i.d. Gaussian distribution. This is true for an orthogonal basis set, which transforms i.i.d. Gaussian noise into i.i.d. Gaussian noise. The orthonormality property of the wavelet transform can be written as $V^T V = I_N$. Non orthogonal wavelet bases exist and could be used as well in our framework. We will focus here on orthogonal ones for the sake of simplicity, although a description of the model with a generic basis, orthogonal or not, is available in the supplementary material [2]. It is important to point out that orthogonal wavelet transforms suffer from missing shift invariance (translation and rotation) and exhibit Gibbs-like artefacts, that might affect their performance. A more accurate description of these problems and possible solutions to handle them will be addressed in the "Discussion" section of this article.

The prior over regression coefficients is then given by

$$p(W|Z, \alpha) = \prod_{k=1}^{K} p(w_k^T | z_k^T, \alpha_k) \tag{8}$$

with

$$p(w_k^T | z_k^T, \alpha_k) = N(w_k^T; V z_k^T, \alpha_k^{-1} I_N) \tag{9}$$

where α_k is the precision of the wavelet residuals for regressor k.

2.3.2 Wavelet coefficients

The prior defined over wavelet coefficients relies on two assumptions regarding the wavelet transform that are observed on a very broad variety of images:

- Wavelet coefficients are independent even if the original image contains spatial dependencies. This is the decorrelation feature of the wavelet transform.
- Most wavelet coefficients are very small and model noise. A few large coefficients suffice to model signal. This is the compaction feature of the wavelet transform.

The independence of the wavelet coefficients yields a factorisation of the prior over voxels. The wavelet transform gives a multiresolution hierarchical description of the data: at each level (or scale) l, an image is decomposed in two components: detail coefficients and coarse (or 'approximation' or 'smooth') coefficients. See Fig. B.2 for a 2-dimensional example. Here we propose to leave the coarse level unchanged by not specifying a prior over its coefficients. But a prior will be defined for each detail level and each subband (horizontal, vertical, diagonal) in a level. We denote L as the overall number of groups of coefficients (three times the number of "wavelet detail levels") and will loosely call it the number of levels. For each level l, we denote N_l as the number of detail coefficients, while N_c is the number of coarse coefficients. We have $\sum_{l=1}^{L} N_l = N_d$ and $N_d + N_c = N$. The splitting of the wavelet coefficients between coarse and detail levels then gives for each regressor k

$$z_k = [\underbrace{z_{k11}, \cdots, z_{k1N_1}}_{\text{detail level } l=1} | \ \cdots \ | \underbrace{z_{kL1}, \cdots, z_{kLN_L}}_{\text{detail level } l=L} | \underbrace{z^c_{k1}, \cdots, z^c_{kN_c}}_{\text{coarse level}}] = [\underbrace{z^d_k}_{\text{details}} | \underbrace{z^c_k}_{\text{coarse}}]. \tag{10}$$

The wavelet basis set matrix V is also divided into V_d and V_c such that

$$w_k^T = [V_d \ V_c] \begin{bmatrix} z_k^{dT} \\ z_k^{cT} \end{bmatrix} \tag{11}$$

The compaction feature of the wavelet transform will be used to specify the prior for each level of the spatial hierarchy. This feature can be embedded into the model through a sparse prior so that small wavelet coefficients will be explained as noise rather than signal. Probabilistic inversion of the model will then achieve signal estimation and denoising simultaneously. The sparsity property of the wavelet transform has been extensively applied for denoising with wavelet shrinkage (Donoho and Johnstone, 1994, 1995; Clyde et al., 1998; Antoniadis et al., 2001). In the Bayesian approach, following (Chipman et al., 1997), we propose to use a mixture model with M zero-mean Gaussian components ($M = 2$ here): the first Gaussian with a large variance (small precision)

describes "signal" while the second Gaussian with a variance close to zero (high precision) describes "noise". An example of such a prior is displayed in Fig. B.3.a.

[Fig. 3 about here.]

The Gaussian mixture model prior is then defined on the wavelet coefficients separately for each level

$$p(Z|S,\pi) = \prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{n=1}^{N_l} p(z_{kln}|s_{kl},\pi_{kl}) \tag{12}$$

with

$$\begin{aligned} p(z_{kln}|s_{kl},\pi_{kl}) &= \sum_{m=1}^{M} \pi_{klm}\ p(z_{kln}|s_{klm},\pi_{klm}) \\ &= \sum_{m=1}^{M} \pi_{klm}\ N(z_{kln};0,s_{klm}^{-1}) \end{aligned} \tag{13}$$

where π are the mixing proportions and S are the wavelet coefficient precisions. For each regression coefficient k, and detail level l, we have a different mixture model with different mixing proportions π_{klm} and precisions s_{klm} for $m = 1, 2$.

We introduce the latent binary variable D, indicating which component of the mixture produced each sample. We will refer to D as the wavelet switches (see below). If $d_{klnm} = 1$ and $d_{klnp} = 0$ for $p \neq m$, then sample z_{kln} was produced by the mth component. The conditional distribution of Z given D is then

$$p(Z|D,S) = \prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{n=1}^{N_l}\prod_{m=1}^{M} N(z_{kln};0,s_{klm}^{-1})^{d_{klnm}} \tag{14}$$

and the joint probability of Z and D is

$$p(Z,D|S,\pi) = p(D|\pi)p(z|D,S) = \prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{n=1}^{N_l}\prod_{m=1}^{M} \left(\pi_{klm}\ N(z_{kln};0,s_{klm}^{-1})\right)^{d_{klnm}} \tag{15}$$

This formulation as a joint distribution is useful for deriving the approximate posteriors in the Variational Bayes framework.

2.3.3 Wavelet switches

The hidden random variable D, introduced in the previous section, can be seen as a switch whereby each basis component is "switched" on or off according

to the binary value of the corresponding coefficient in D. The prior on the wavelet switches is given by

$$p(D|\pi) = \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{n=1}^{N_l} p(d_{kln}|\pi_{kl}) \tag{16}$$

with

$$p(d_{kln}|\pi_{kl}) = Mult(d_{kln}; \pi_{kl}) = \prod_{m=1}^{M} \pi_{klm}^{d_{klnm}} \tag{17}$$

where $Mult$ is a Multinomial distribution (see Appendix A).

2.3.4 Mixing proportions

The mixing proportions of the Gaussian mixture models are defined as

$$p(\pi) = \prod_{k=1}^{K} \prod_{l=1}^{L} p(\pi_{kl}) \tag{18}$$

with

$$p(\pi_{kl}) = Dir(\pi_{kl}; f_0) = \frac{1}{c(f_0)} \prod_{m=1}^{M} \pi_{klm}^{f_{0m}-1} \tag{19}$$

where Dir is a symmetric Dirichlet distribution, the conjugate prior of a multinomial distribution, with parameters f_0 as defined in Appendix A. It would also be possible to use a non-symmetric Dirichlet, which could embody prior knowledge that e.g. wavelet bases are more likely to be required at lower rather than higher spatial frequencies.

2.3.5 Precisions

Finally, we use Gamma priors on the noise precisions λ, wavelet residuals α and wavelet coefficients S, which are the standard conjugate priors for inverse variances.

$$\begin{aligned}
p(\lambda) &= \prod_{n=1}^{N} p(\lambda_n) = \prod_{n=1}^{N} G_a(\lambda_n; b_{\lambda_0}, c_{\lambda_0}), \\
p(\alpha) &= \prod_{k=1}^{K} p(\alpha_k) = \prod_{k=1}^{K} G_a(\alpha_k; b_{\alpha_0}, c_{\alpha_0}), \\
p(S) &= \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{m=1}^{M} p(s_{klm}) = \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{m=1}^{M} G_a(s_{klm}; b_{s_0}, c_{s_0}).
\end{aligned} \tag{20}$$

The quantities b_{λ_0}, c_{λ_0}, b_{α_0}, c_{α_0}, b_{s_0} and c_{s_0} are referred to as hyperparameters. These are set so as to specify vague priors (see below, in the implementation details section).

2.3.6 The generative model

The probabilistic dependencies underlying the generative model are displayed in Fig. B.1. From this graph, the joint probability can be written as

$$\begin{aligned} p(Y, \underbrace{W, \lambda, Z, \alpha, D, S, \pi}_{\Theta}) =& p(Y|W,\lambda)p(\lambda)p(W|Z,\alpha)p(\alpha)p(Z|D,S)p(S)p(D|\pi)p(\pi) \\ =& \left(\prod_{n=1}^{N} p(y_n|w_n,\lambda_n)p(\lambda_n)\right)\left(\prod_{k=1}^{K} p(w_k^T|z_k^T,\alpha_k)p(\alpha_k)\right)\times \\ & \left(\prod_{k=1}^{K}\prod_{l=1}^{L} p(\pi_{kl})\right)\left(\prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{m=1}^{M} p(s_{klm})\right)\times \\ & \left(\prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{n=1}^{N_l} p(z_{kln}|d_{kln},s_{kl})p(d_{kln}|\pi_{kl})\right) \end{aligned} \tag{21}$$

where the first term is the likelihood and the other terms are priors.

2.4 Posteriors

Given a data set Y, our aim is to estimate the distribution of the unknown parameters $\Theta = \{W, \lambda, Z, \alpha, D, S, \pi\}$ using Bayesian estimation theory. This posterior distribution is related to the likelihood of the observations and the parameter priors via Bayes rule $p(\Theta|Y) \propto p(Y|\Theta)p(\Theta)$.

However, computing the exact posterior $p(\Theta|Y)$ is an extremely difficult task because, as it can be seen in equation 21, the true posterior over regression coefficients has dependencies between voxels and regressors, resulting in very high dimensions for its definition. This would require a prohibitive amount of computer time using standard Markov Chain Monte Carlo (MCMC) methods (Woolrich et al., 2004).

A recent efficient method to bypass this problem is the Variational Bayes (VB) framework (see appendix B) (Lappalainen and Miskin, 2000; Beal, 2003), already used in previous publications (Penny et al., 2003, 2005b), which provides an approximate factorised posterior distribution $q(\Theta|Y)$, with minimal Kullback-Leibler (KL) divergence from the true posterior (see below).

In the following section, we give a short overview of Variational Bayes, a more

detailed version being available in Appendix B. Then in subsequent sections we summarise, for each parameter, the results obtained when applying VB to the SSBF model. Details of the derivations can be found in the online supplementary material [2]. The equations that appear in this paper are appropriate for an orthonormal spatial basis, whereas the more general case is presented in the supplementary document.

2.4.1 Approximate posteriors

In the VB framework, the posterior distribution over parameters is assumed to factorise

$$q(\Theta|Y) = \prod_i q(\theta_i|Y) = q(\theta_i|Y)q(\theta_{|i}|Y) \tag{22}$$

where $\theta_{|i}$ denotes those parameters not in the ith group. In practice, VB will find the factorised posterior distribution $q(\Theta|Y)$ that best matches the true posterior $p(\Theta|Y)$, in the sense of the Kullback-Leibler divergence between these two distributions. As described in Appendix B, update rules for components of the approximate posterior can be found using the formulae

$$q(\theta_i) = \frac{\exp\left[I(\theta_i)\right]}{\int \exp\left[I(\theta_i)\right] d\theta_i} \quad \text{with} \quad I(\theta_i) = \int q(\theta_{|i}) \log\left[p(Y,\theta)\right] d\theta_{|i} \tag{23}$$

Application of VB leads to a set of approximate posteriors and equations for updating their sufficient statistics. This formulation of VB is also referred to as 'free-form' VB, as the optimal form of each component (e.g. Gamma, Gaussian, Dirichlet) is also derived. This is to be contrasted with 'fixed-form' VB in which posteriors are assumed to be e.g. Gaussian. This distinction is described further in (Friston et al., 2006).

In our case, we consider the following factorisation for the approximate posterior distribution (we omit the 'conditional on Y' notation for ease of reading)

$$\begin{aligned} q(\Theta|Y) = &\left(\prod_{k=1}^{K} q(z_k^T)q(\alpha_k)\right)\left(\prod_{n=1}^{N} q(w_n)q(\lambda_n)\right)\left(\prod_{k=1}^{K}\prod_{l=1}^{L} q(\pi_{kl})\right) \times \\ &\left(\prod_{k}^{K}\prod_{l=1}^{L}\prod_{m}^{M} q(s_{klm}))\right)\left(\prod_{k}^{K}\prod_{l=1}^{L}\prod_{n=1}^{N_l} q(d_{kln})\right) \end{aligned} \tag{24}$$

Importantly, we enforce a factorisation of regression coefficients in the posterior distribution $q(W|Y)$ that will dramatically lower the dimensionality of

the problem (see next section). Because of the spatial priors, the true posterior will be correlated. But as we will see later on, update equations for the approximate factorised densities show that the effect of spatial correlation is to bias estimation of the posterior means through a top-down prediction from the wavelet stage.

2.4.2 Regression coefficients

As previously mentioned, the approximate posterior over regression coefficients is assumed to factorise over voxels

$$q(W|Y) = \prod_{n=1}^{N} q(w_n|Y) \tag{25}$$

It can then be shown that the posterior over regression coefficients at voxel n follows a Gaussian distribution whose parameters are given by

$$q(w_n|Y) = N(w_n; \bar{w}_n, \Sigma_{w_n}) \tag{26}$$

where

$$\begin{cases} \bar{w}_n &= \Sigma_{w_n} \left(\bar{\lambda}_n X^T y_n + \bar{r}_n^T \right) \\ \Sigma_{w_n} &= \left(\bar{\lambda}_n X^T X + \texttt{diag}(\bar{\alpha}) \right)^{-1} \end{cases} \tag{27}$$

and $\bar{r}_n^T$ is nth row of the $N \times K$ matrix $\bar{R}$ whose columns contain the top-down predictions from the spatial prior

$$\bar{R} = \begin{bmatrix} \vdots & \vdots & \vdots \\ \bar{\alpha}_1 V \bar{z}_1^T & \cdots & \bar{\alpha}_K V \bar{z}_K^T \\ \vdots & \vdots & \vdots \end{bmatrix} = \begin{bmatrix} \cdots & \bar{r}_1 & \cdots \\ & \vdots & \\ \cdots & \bar{r}_N & \cdots \end{bmatrix} \tag{28}$$

These equations show that the updated mean of the regression coefficients at each voxel comes from a weighted average of the data itself $X^T y_n$ and a term from the wavelet prior $\bar{r_n}^T$, the weights given by their relative precisions. This is a standard feature of Bayesian estimation where the posterior distribution is formed thanks to the likelihood on one hand and the prior on the other.

If the wavelet residual precision α is set to zero, then $\bar{R}$ is the null matrix and the posterior mean becomes $\bar{w}_n = (X^T X)^{-1} X^T y_n$, which is the Ordinary Least Square (OLS) estimate for a GLM. Indeed, setting α to zero means that there is no error in the wavelet stage of the hierarchical model, which means that no shrinkage is performed on the wavelet coefficients.

2.4.3 Wavelet coefficients

The approximate posterior over wavelet coefficients Z is assumed to factorise over regressors. With the supplementary assumption that we use an orthonormal basis, we have $V^T V = I_N$ and this leads to a further factorisation of the posterior over wavelet levels and wavelet coefficients in each level

$$q(Z) = \prod_{k=1}^{K} q(z_k^{dT}) = \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{n=1}^{N_l} q(z_{kln}) \tag{29}$$

$$q(z_{kln}) = N(z_{kln}; \bar{z}_{kln}, \sigma^2_{z_{kln}}) \tag{30}$$

where

$$\begin{cases} \sigma^2_{z_{kln}} &= \left(\bar{\alpha}_k + \sum_{m=1}^{M} \bar{s}_{klm}\gamma_{klnm}\right)^{-1} \\ \bar{z}_{kln} &= \frac{\bar{\alpha}_k V_{ln}^T \bar{w}_k^T}{\bar{\alpha}_k + \sum_{m=1}^{M} \bar{s}_{klm}\gamma_{klnm}} \end{cases} \tag{31}$$

and V_{ln} is the wavelet basis for the nth element of the lth detail level.

Equation 31 plays a central role as it embodies the wavelet shrinkage procedure through the sparse prior. Indeed, we can see that the estimate of the posterior mean of a particular wavelet coefficient $\bar{z}_{kln}$ is proportional to the corresponding bottom-up estimate $V_{ln}^T \bar{w}_k^T$ (wavelet transform of the regression coefficient $\bar{w}_k^T$). The multiplicative term, $\left(1 + \frac{1}{\bar{\alpha}_k} \sum_{m=1}^{M} \bar{s}_{klm}\gamma_{klnm}\right)^{-1}$, determines the amount of shrinkage. If the corresponding wavelet coefficient $\bar{z}_{kln}$ is large, it probably belongs to the Gaussian component m_1 modelling signal so that $\gamma_{klnm_1} \simeq 1$. Because this component has low precision ($\bar{s}_{klm_1}$ is small), the multiplicative term is close to 1 and the wavelet coefficient is preserved. In the alternative case, $\gamma_{klnm_2} \simeq 1$ which means that the wavelet coefficient is modelling noise ($\bar{s}_{klm_2}$ is very large) and the multiplicative term is thus very small, shrinking the estimate of the posterior mean towards zero. Fig. B.3.b shows the profile of the posterior mean of the wavelet coefficients as a function of their bottom-up estimates from the regression coefficients (i.e. without mixture prior). This highlights the nonlinear shrinkage produced by the sparse prior.

2.4.4 Wavelet switches

The approximate posterior over wavelet switches D factorises over regressors, wavelet levels and voxels

$$q(D) = \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{n=1}^{N_l} q(d_{kln}) \tag{32}$$

The VB framework gives the following updates

$$q(d_{kln}) = Mult(d_{kln}; \gamma_{kln}) \tag{33}$$

where

$$\gamma_{klnm} = \frac{\tilde{\gamma}_{klnm}}{\sum_{m'} \tilde{\gamma}_{klnm'}} \quad \text{and} \quad \tilde{\gamma}_{klnm} = \tilde{\pi}_{klm} \tilde{s}_{klm}^{1/2} \exp\left(-\frac{\bar{s}_{klm}}{2}(\bar{z}_{kln}^2 + \sigma_{z_{kln}}^2)\right) \tag{34}$$

with

$$\log \tilde{\pi}_{klm} = \int q(\pi_{klm}) \log \pi_{klm} d\pi_{klm} \quad \text{and} \quad \log \tilde{s}_{klm} = \int q(s_{klm}) \log s_{klm} ds_{klm} \tag{35}$$

Closed form equations for computing $\log \tilde{\pi}_{klm}$ and $\log \tilde{s}_{klm}$ are given in the following sections. The term γ_{klnm} is the posterior probability that component m is responsible for data point z_{kln}. The equations we obtain here are similar to those presented in (Attias, 2000; Penny, 2001) for a Gaussian mixture model with zero mean components.

2.4.5 Mixing proportions

The approximate posterior over mixing proportions is given by

$$q(\pi) = \prod_{k=1}^{K} \prod_{l=1}^{L} q(\pi_{kl}) \tag{36}$$

$$q(\pi_{kl}) = Dir(\pi_{kl}; f_{kl}) \tag{37}$$

where

$$f_{klm} = \bar{N}_{klm} + f_{0m} \quad \text{with 'data counts'} \quad \bar{N}_{klm} = \sum_{n=1}^{N_l} \gamma_{klnm} \tag{38}$$

We also use the following result

$$\log \tilde{\pi}_{klm} = \int q(\pi_{klm}) \log \pi_{klm} d\pi_{klm} = \Psi(f_{klm}) - \Psi\left(\sum_{m'=1}^{M} f_{klm'}\right) \tag{39}$$

2.4.6 Noise precisions

The approximate posterior over noise precisions is given by

$$q(\lambda) = \prod_{n=1}^{N} q(\lambda_n) \tag{40}$$

$$q(\lambda_n) = G_a(\lambda_n; b_{\lambda_n}, c_{\lambda_n}) \tag{41}$$

where

$$\begin{cases} \frac{1}{b_{\lambda_n}} = \frac{1}{2}\left[(y_n - X\bar{w}_n)^T(y_n - X\bar{w}_n) + \mathtt{tr}(\Sigma_{w_n} X^T X)\right] + \frac{1}{b_{\lambda_0}} \\ c_{\lambda_n} = \frac{T}{2} + c_{\lambda_0} \end{cases} \tag{42}$$

The expectation of λ_n is then given by $\bar{\lambda}_n = b_{\lambda_n} c_{\lambda_n}$.

2.4.7 Wavelet residual precisions

The approximate posterior over wavelet residual precisions is given by

$$q(\alpha) = \prod_{k=1}^{K} q(\alpha_k) \tag{43}$$

$$q(\alpha_k) = G_a(\alpha_k; b_{\alpha_k}, c_{\alpha_k}) \tag{44}$$

where

$$\begin{cases} \frac{1}{b_{\alpha_k}} = \frac{1}{2}\left[\mathtt{tr}(\Sigma_{w_k}) + \sum_{l=1}^{L}\sum_{n=1}^{N_l} \sigma_{z^2_{kln}} + (\bar{w}_k^T - V\bar{z}_k^T)^T(\bar{w}_k^T - V\bar{z}_k^T)\right] + \frac{1}{b_{\alpha_0}} \\ c_{\alpha_k} = \frac{N}{2} + c_{\alpha_0} \end{cases} \tag{45}$$

The term $\mathtt{tr}(\Sigma_{w_k})$ can be obtained from $\{\Sigma_{w_n}\}_{n=1}^{N}$ using

$$\mathtt{tr}(\Sigma_{w_k}) = \sum_{n=1}^{N} \Sigma_{w_n}[k,k] \tag{46}$$

where $\Sigma_{w_n}[k,k]$ is the kth diagonal term of the covariance matrix Σ_{w_n}. The expectation of α_k is given by $\bar{\alpha}_k = b_{\alpha_k} c_{\alpha_k}$.

2.4.8 Wavelet coefficient precisions

The approximate posterior over wavelet coefficient precisions is given by

$$q(S) = \prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{m=1}^{M} q(s_{klm}) \tag{47}$$

$$q(s_{klm}) = G_a(s_{klm}; b_{s_{klm}}, c_{s_{klm}}) \tag{48}$$

where

$$\begin{cases} \frac{1}{b_{s_{klm}}} = \frac{1}{2}\left[\sum_{n=1}^{N_l} \gamma_{klnm}(\sigma^2_{z_{kln}} + \bar{z}^2_{kln})\right] + \frac{1}{b_{s_0}} \\ c_{s_{klm}} = \frac{\bar{N}_{klm}}{2} + c_{s_0} \text{ with } \bar{N}_{klm} = \sum_{n=1}^{N_l} \gamma_{klnm} \end{cases} \tag{49}$$

The expectation of s_{klm} is given by $\bar{s}_{klm} = b_{s_{klm}} c_{s_{klm}}$. We also use the following result

$$\log \tilde{s}_{klm} = \int q(s_{klm}) \log s_{klm} ds_{klm} = \Psi(c_{s_{klm}}) + \log b_{s_{klm}} \tag{50}$$

2.4.9 Implementation details

This section describes the choice of hyper-parameters, algorithm implementation and initialisation.

Hyper-parameters. We set vague priors ($b = 1000$ and $c = 0.1$) for precision distributions λ and α: this corresponds to Gamma densities with mean 100 and variance 100,000. For the wavelet coefficient precisions S, we set the same vague prior for the 'signal' component but a much peakier one for the 'noise' component ($b = 10$ and $c = 0.1$). These settings implement the sparsity constraints. For all the Dirichlet distributions, we use a scale factor $f_0 = 1$.

The number of wavelet levels to be governed by the sparse prior is set using the following formula

$$L = \log_2 \left(\log\left(\sqrt{N}\right)\right) + 1 \tag{51}$$

which has been prescribed asymptotically for wavelet shrinkage in (Antoniadis et al., 2001). This choice was found to provide a good trade-off between specificity and sensitivity (Fadili and Bullmore, 2004). The number of detail levels can, in principle, also be tested via Bayesian model comparison.

Initialisation. The algorithm is initialised using Ordinary Least Square (OLS) estimates for regression parameters, on a voxel by voxel basis. The posterior means of the mixing proportions corresponding to the signal component are set to $\exp(l)/\sum \exp(l)$ for each wavelet level l, to model the fact that we expect to remove most of the coefficients for the most detailed levels and less for coarser ones. Wavelet switches are initialised such that all detail coefficients are switched off to start with.

[Fig. 4 about here.]

VB algorithm. The Variational Bayes algorithm is then an iterative procedure, which updates the summary statistics of each posterior distribution, using equations 26, 30, 33, 36, 41, 44 and 47. Fig. B.4 provides an overview of all the update equations for all the parameters of the model. In the iterative scheme, we first update Z, then W, π, α, λ, D and finally S. One important remark about the implementation is that we never have to explicitly construct the $N \times N$ matrix V containing the wavelet basis set. Indeed, when using an orthonormal basis set, matrix V only appears when applying the wavelet transform or its inverse to regression coefficients or wavelet coefficients: $V\bar{z}_k^T$ for $q(W)$ and $q(\alpha)$, $V_d^T\bar{w}_k^T$ for $q(Z)$. It is then possible to apply the very efficient Discrete Wavelet Transform (DWT), or its inverse IDWT, on the corresponding images whenever such computations are required. This yields a relatively fast algorithm because operations related to spatial priors that are traditionally very time consuming are here replaced by a dedicated algorithm with a complexity in $O(N)$ (Mallat, 1999).

In our `Matlab` (The MathWorks, Inc.) implementation, we used the free software wavelet toolbox `WaveLab` (WaveLab, 1999). More precisely, we used the Battle-Lemarié wavelet family, which is a symmetric orthogonal spline basis set that displays good smoothness properties (Ruttimann et al., 1998). Indeed, Battle-Lemarié wavelets are cubic spline wavelets that correspond to a good trade-off between number of zero moments and compact support. These wavelets are symmetric so that they do not introduce phase differences between decomposition levels, orthogonal so that they transform white noise into white noise and are well localised in both the spatial and frequency domains.

All computations are performed slice by slice, using two dimensional transforms, to reduce the amount of memory required. However, there is no theoretical limitation that prevents the model being estimated using data from all slices at the same time, in a 3D fashion, as the DWT can be extended to multiple dimensions using tensor product basis functions. It should be pointed out, though, that by using a separable wavelet transform, one makes the assumption that fMRI data have an isotropic spatial resolution. Raw data have often non cubic voxel sizes, the slice thickness being bigger than the in-plane one. But, after preprocessing – normalisation in particular – fMRI data are traditionally reinterpolated to an isotropic resolution so this is not really a problem. Wavelet bases well suited to non-isotropic sampling are described in (Van De Ville et al., 2003). Another issue comes from the way volumes are acquired: in a standard Echo Planar Imaging (EPI) sequence, each 3D volume is acquired slice by slice yielding time delays between slices. A slice timing correction can be used to reinterpolate data so as they look as if they were acquired at the same time for all the slices of each volume. Applying such a correction is essential if a 3D prior is applied to the data using a same design matrix and its robustness would need to be evaluated but this is out of the topic of this paper.

Convergence. The overall criterion of the Variational Bayes algorithm is the log evidence of the data, for which a lower bound is given by the so called negative free energy (see Appendix B). Model fitting is then terminated when the relative change in free energy (equations not provided) drops below 0.01%. However, the computation of the free energy at each iteration of the algorithm is time consuming and we explicitly set up a fixed number of iterations, 8, as we noticed that only a few iterations were required to attain convergence.

3 Results

In this section, we present results obtained from synthetic data and an event-related fMRI dataset.

3.1 Synthetic data

In a first experiment, we illustrate our method on a synthetic Gaussian blobs dataset where we generated noisy data from a known GLM. The design matrix describing the temporal part of the simulation is shown in Fig. B.5 on the left. It comprises two regressors, the first being a boxcar with a period of 20 scans and the second a constant. The length of the time series was chosen to be T = 40. Two identical 32×32 images of regression coefficients (N = 1024) were formed by placing Dirac 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), see Fig. B.5 on the right. White Gaussian noise of precision $\lambda = 10$ was added to generate the $T \times N$ data matrix, using the previously defined design matrix and regression coefficient images.

[Fig. 5 about here.]

Note that by creating blobs with different Gaussian kernel widths, we put ourselves in the presence of spatial variations in smoothness of the data. Thus, there is no optimal smoothing filter, according to the matched-filter theorem (Worsley et al., 1996), that will improve the signal to noise ratio for all three blobs. If the data were to be smoothed, a Gaussian kernel of FWHM = 3 pixels would provide optimal signal recovery for the third blob, but the estimates of the regression coefficients for the first two would be underestimated. A wavelet shrinkage procedure, however, embedded in a spatial prior allows us to parsimoniously model this non uniformity in the spatial correlation.

We have fitted the SSBF model to this simulated dataset and the results are presented in Fig. B.6. If we compare the Ordinary Least Square estimate of the

first regressor and the one obtained via our model, we can observe the intrinsic spatial smoothing performed by the shrinkage procedure. Importantly, signal intensities are largely preserved. The posterior estimate of the mean of the wavelet coefficients after convergence (not shown), shows that most of the coefficients are indeed zero and only a very small subset of them has been kept nearly unchanged (see also next example).

[Fig. 6 about here.]

We also used synthetic data to compute Receiver Operating Characteristic (ROC) curves. These are plots of sensitivity versus one minus specificity under the variation of a parameter. This curve allows us to see if a more sensitive method conserves specificity: the more on the top left of the figure, the better. To do so, we generated another dataset with the same known GLM as before (same dimensions, same design matrix) but with different images of regression coefficients, see Fig. B.7.a. These are obviously non Gaussian activation shapes. Estimated effects corresponding to the first regressor are displayed in Figs. B.7.b, B.7.c and B.7.d with, respectively, Ordinary Least Squares (OLS), Gaussian Markov Random Field (GMRF) prior and SSBF prior. The sum of square errors (difference between estimated and true effect) are respectively 101.1 (OLS), 30.08 (GMRF) and 26.65 (SSBF). We can see that GMRF and SSBF priors remove most of the background noise observed with OLS. Furthermore, SSBF yields better estimates as signal intensities are closer to their true simulated values. Fig. B.7.e shows the estimated wavelet coefficients z_1 corresponding to the first regressor, after convergence of the Variational Bayes algorithm. We can see that only a few coefficients are non zero (about 8% here), thanks to the sparse prior. The inverse wavelet transform of Fig. B.7.e is displayed in Fig. B.7.f: this is the top-down estimate of the effect. Equation 27 shows that the estimated regression coefficient (Fig. B.7.d) is indeed obtained from a bottom-up estimate from the data (Fig. B.7.b) and a top-down estimate from the SSBF prior (Fig. B.7.f). We display in Fig. B.8 the ROC curve obtained when declaring a voxel to be active if the effect size was larger than some arbitrary threshold. We varied this threshold between 0.1 and 0.9 to produce each point in the curve. Here, we compared estimates from (i) SSBF prior, (ii) GMRF prior and (iii) OLS. We observe that OLS is the least sensitive method, while GMRF and SSBF give good results, with an advantage for SSBF, due to the highly non-Gaussian pattern of activations. We can conclude that SSBF is the superior method.

[Fig. 7 about here.]

[Fig. 8 about here.]

We performed another simulation to study the performance of the model in the presence of heteroscedastic noise. Indeed, in real fMRI data, one often

observe regions of high noise variance which can produce false positives in these regions. To illustrate this phenomenon, Fig. B.9 displays a typical slice of variance of the residuals estimated from an fMRI dataset using a GLM. This points out that noise is highly non uniform, with important fluctuations between brain tissues and from region to region. To quantify the robustness of our model towards spatially varying noise, i.e. heteroscedasticity, we created a synthetic dataset containing only noise time series, but where its standard deviation was 1 in the background and 10 in a square in the middle of the image (32×32) (see Figs. B.10.c and B.10.f). We fitted our SSBF prior model and the GMRF prior model described in (Penny et al., 2005b) on these data using a random design matrix: it consisted of two regressors, the first one being an event-related one where events where chosen randomly across time and the second a constant. Such a GLM should reveal no activation, and we are thus testing for false positive rates. Fig. B.10 displays estimates of the parameters of the two models. Figs. B.10.a and B.10.d show the effect size estimates and B.10.c and B.10.f the noise standard deviation estimates. We can see that while the estimates of the standard deviation of the noise are similar and precise (we find back what has been simulated), we observe some important difference in the estimates of the effect sizes. If we compute a Posterior Probability Map (PPM) using the thresholds as detailed in (Penny and Flandin, 2005), i.e. $q(w_n > 0) > 1 - 1/N$, we obtain the PPMs presented in Figs. B.10.b and B.10.e. All surviving voxels are false positives and we can see that there are clearly more false positive with the GMRF prior than with the SSBF prior described in this paper. We replicated this experiment many times and the same conclusion was met each time: SSBF yields zero or very few false positives while the GMRF prior leads to a significantly non-zero false positive rate in regions of high noise variance. This is displayed in Fig. B.11 where we generated 1000 simulated datasets and computed the number of false positives with GMRF and SSBF. The histogram in Fig. B.11.b confirms the overall better performance, in terms of false positive rate, of SSBF. We conclude that a nonstationary spatial prior is useful for dealing with noise heteroscedasticity.

[Fig. 9 about here.]

[Fig. 10 about here.]

[Fig. 11 about here.]

Finally, we compared the computer time required to estimate parameters of the model using two different priors: SSBF or GMRF. To do so, we fixed the number of iterations to 4 in both cases and simulated data with different image sizes. Parameters were T = 40, K = 2, and N = [16^2 32^2 64^2 128^2]. Results are presented in Fig. B.12 on a logarithmic scale. We can see that the GMRF prior leads to a dramatic increase of computer time with an increase of the

dimension of the data. On the contrary, the time increase for the SSBF prior is nearly linear with data size. Speed is thus a great benefit of the method presented in this paper.

[Fig. 12 about here.]

3.2 Event-related fMRI data

This dataset [4] corresponds to functional and anatomical images from a subject that participated in an experiment studying a repetition priming effect for famous and non famous faces (Henson et al., 2002). This was a 2 x 2 factorial event-related fMRI experiment where famous and non famous grayscale photographs, randomly interleaved, were presented twice (first and second presentation) against a baseline of an oval checkerboard which was present throughout the interstimulus interval. Images were acquired using a continuous Echo-Planar Imaging (EPI) sequence on a 2T VISION system (Siemens, Erlangen, Germany) with TE = 40ms and TR = 2s, producing 351 T_2^*-weighted full-brain covered scans. Each volume is composed of 24 transverse slices of dimension 64×64 with a voxel size of $3\times3\times3\text{mm}^3$ and a 1.5mm gap between slices. During the same experiment, a T_1-weighted volume was also acquired to get anatomical information from the subject.

We performed the standard preprocessing steps to analyse this dataset using SPM5 (SPM, 2005), following the tutorial attached to the data. First, all functional images were spatially realigned to the first image using a six-parameter rigid-body transformation (Friston et al., 1995a). To correct for the fact that different slices were acquired at different times, time series were interpolated to the acquisition of the reference slice (slice 12, in the middle of the volume) (Henson et al., 2002). The mean functional image and the structural were then coregistered using a mutual information criterion (Friston et al., 1995a). A non linear transformation was then estimated to register the anatomical image with a standard T1-weighted template image in MNI space. This was implemented in the unified segmentation procedure that iteratively corrects for intensity inhomogeneities in the image, segments gray matter, white matter and cerebrospinal fluid using default tissue probability maps as prior and performs a nonlinear warping method (Ashburner and Friston, 2005). The estimated deformation field was then applied to the 351 realigned, slice-timing corrected functional scans. Importantly, and contrary to the standard pipeline of preprocessing, we did not perform any spatial smoothing of the data, this step being replaced by a spatial prior.

[4] This dataset, a full description of the experiment and a tutorial to analyse it with SPM5 are available from `http://www.fil.ion.ucl.ac.uk/spm/data/face_rep_SPM5.html`

A scaling factor was then applied to each volume, computed as being 100 over the global mean value over all time series, excluding non-brain voxels. By doing so, the regression coefficient values become meaningful and correspond to "percentage of global mean signal". Also, each time series was high-pass filtered using a set of discrete cosine basis functions with a cutoff of 120 seconds, to remove slow drifts.

The paradigm was then modelled via a GLM with the design matrix shown in Fig. B.13. There are five regressors relating to the 4 types of event – first or second presentation of images of famous and non-famous faces, the last regressor being an offset. Each regressor has been built from the corresponding onsets and convolved with a canonical hemodynamic response function.

[Fig. 13 about here.]

Figs. B.14.a and B.14.b depict the histograms of the first regression coefficients (estimated by OLS) and of its wavelet transform. This highlights the compaction feature of the wavelet transform. Indeed, histogram B.14.b is much narrower towards zero than B.14.a, and its distribution can be efficiently modelled by a zero mean Gaussian mixture model, as displayed in Fig. B.14.b.

[Fig. 14 about here.]

We then fitted the SSBF model and the following results correspond to a single slice at $z = -18$ mm in MNI space coordinates. We used the orthogonal cubic Battle-Lemarié wavelet transform with 3 detail levels (L=3). Fig. B.15 shows the contrast image for the main effect of faces, obtained by applying the contrast weight vector $c^T = \frac{1}{4}\,[11110]$, using different estimators: (i) with OLS on raw data, (ii) with SSBF on raw data and (iii) with OLS on data smoothed by a Gaussian kernel of FWHM=8mm. This corresponds to the standard intra-subject amount of smoothing that is applied as a preprocessing step. The same contrast images have been thresholded at 5% of the global mean and displayed in Fig. B.16.

[Fig. 15 about here.]

[Fig. 16 about here.]

If we compare the raw-OLS and SSBF estimates, we can see that we obtain much smoother estimates with SSBF, while maintaining a distinction between activation and noise. The OLS estimate on smoothed data incorporates a uniform and isotropic smoothing which introduces blurring of the activations between gray and white matter and an under-estimation of their amplitude. On the contrary, SSBF allows, via the sparse wavelet shrinkage prior on the regression coefficients, a non-uniform and anisotropic smoothing of the data so that signal intensities can be preserved. Also, the computational

burden is maintained to a manageable amount thanks to the use of the discrete wavelet transform. For example, the 8 iterations used for fitting the SSBF model (T=351, K=5, L=3, N=128^2) on this dataset took about 100 seconds on a standard desktop computer. This is an important improvement compared to the Laplacian/GMRF prior implemented in (Penny et al., 2005b), where the update step for lower resolution images (N=64^2) and fewer iterations (4 instead of 8), took 132 seconds. For this data, SSBF is approximately 8 times faster.

4 Discussion

In this paper, we presented a statistical framework where spatial correlation in the data is captured as part of the model and is not left to a non probabilistic pre- or postprocessing step. The main idea is to decompose the data on an appropriate spatial basis set where signal and noise can be easily separated. Wavelet bases are well suited for that purpose as they allow modelling of transient, non-stationary or spatial varying phenomenon. The sparsity of the transformed data is a key feature of the model. We used a sparse prior defined by a mixture of two Gaussian components. This sparse spatial basis set prior is embedded in a spatio-temporal model where temporal decomposition of the data is performed through the usual General Linear Model. The model is inverted via a variational Bayes scheme that yields an efficient algorithm for computing the posterior distributions.

The benefits of our approach are that (i) we can capture non-uniform spatial regularities in the effects of interest, thus providing Posterior Probabilities Maps with an increased sensitivity, (ii) fast algorithms are available for the spatial transform involved in the VB framework (e.g. Discrete Wavelet Transform), which allows for a very computationally efficient alternative to Laplacian or GMRF priors and (iii) we are robust to heteroscedastic noise. Furthermore, as only a few wavelet coefficients are required, it becomes possible to reduce the size of the data that have to be dealt with through the iterations of the VB algorithm.

The VB framework not only allows one to build an efficient algorithm to estimate the posterior densities of the model, but also allows for Bayesian model comparison. Indeed, the objective function is the negative free energy $F(M)$, which is a lower bound on the log evidence $\log p(Y|M)$, for model M. Comparing negative free energies obtained when inferring different models on the same dataset can then be used for model selection. This has been done in (Penny et al., 2005a) to compare noise models, hemodynamic basis sets and to perform Bayesian ANOVAs. Here, it would be very interesting to compare the relative performance of different basis sets, wavelets versus DCT

for example, but also to compare different wavelet bases, different number of wavelet levels in the decomposition, the separability or not of the wavelet detail coefficients in horizontal, vertical, diagonal subbands, etc. The computation of the negative free energy and its application to model selection will be presented in a subsequent paper.

The ability of the wavelet transform to concentrate signal information in a few large coefficients is behind the success of wavelet based denoising algorithms. However, wavelet denoising with an orthogonal transform suffers two main drawbacks: (i) it is not shift invariant and (ii) it exhibits Gibbs phenomenon around discontinuities. These artefactual oscillations around edges are due to the shift-variance of the orthogonal wavelet transform and to the fact that wavelet coefficients are thresholded independently, so the dependency between wavelet coefficients is not exploited.

Shift variance of the orthogonal DWT means that wavelet coefficients of a shifted or translated signal are not the same as the shifted coefficients of the original signal. An example of the dramatic effect of that can be found in (Van De Ville et al., 2006). Also, multi-dimensional wavelet transforms implemented using a separable orthogonal basis are also non optimal as they have some directional selectivity: for instance, 2D transforms have a directionality to horizontal, diagonal and vertical features. Singularities, such as edges, might not be modelled efficiently using an orthogonal transform. A solution is to use redundant wavelet transforms (also called undecimated spatial wavelet transform, shift invariant wavelet transform, complex wavelet transform, etc). This requires a modification of the DWT algorithm to remove the subsampling step and keep all the wavelet coefficients at each level of the transformation. This yields a shift-invariant transform at the expense of an increase in the number of coefficients (and loss of the orthogonality property).

Our framework is generic enough to allow such overcomplete spatial transforms, even if, as detailed in the supplementary material [2], the VB update equations are more computationally expensive. This will be investigated in future work. Van De Ville et al. (2006) compared redundant DWT and multiple non-redundant DWTs where the same model is applied to data shifted by a small amount and conclude that the combination of multiple non-redundant DWTs allows shift invariant analysis while keeping the demand for storage and computation to a manageable level.

Also, even if the DWT tends to decorrelate the data, wavelet coefficients are not independent and one observes some correlation between coefficients of different levels. Indeed large "detail" parent coefficients tend to have large "children" in the same location at coarser levels. This is not taken into account in our hierarchical model, where wavelet coefficients are assumed to be independent and factorise over voxels. A solution would be to model the

statistical dependencies between wavelet coefficients in the definition of the prior. This has been proposed in (Sendur et al., 2005) for fMRI data by using a Bayesian bivariate shrinkage operator, that introduces parent-child relationships between wavelet coefficients to model inter-scale dependencies. A similar empirical Bayes approach is taken in (Turkheimer et al., 2006) in the context of PET data. Modifying our proposed model to handle these correlations is an interesting area of future research. Importantly, we will be able to assess their usefulness for imaging data using Bayesian model comparison (Penny et al., 2005a).

Acknowledgements

G. Flandin was funded by an INRIA grant and W.D. Penny is supported by the Wellcome Trust. The authors would like to thank Rik Henson for making his face processing dataset available to the community.

Appendix

A Densities, Divergences and Expectations

We include here definitions of the densities used throughout this article. We also include standard formulae used in the derivation of the approximate posteriors.

A.1 Multivariate Normal Density

The multivariate Normal density for d-dimensional variable x with mean m and variance Σ is given by

$$N(x; m, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right) \tag{A.1}$$

A quadratic expectation of a Normal random variable $x \sim N(m, \Sigma)$ is

$$E[x^T A x] = \mathtt{tr}(A\Sigma) + m^T A m \tag{A.2}$$

A.2 Gamma Density

The Gamma density for variable x with parameters b and c is defined by

$$G_a(x; b, c) = \frac{1}{\Gamma(c)}\frac{x^{c-1}}{b^c}\exp\left(\frac{-x}{b}\right) \quad \text{for} \quad x \geq 0 \quad \text{and} \quad 0 \quad \text{otherwise} \tag{A.3}$$

where $\Gamma(x)$ is the Gamma function. Mean and variance are respectively given by bc and b^2c.

An expectation formula used in the Variational Bayes framework is

$$E[\log x] = \Psi(c) + \log b \tag{A.4}$$

where Ψ is the digamma function (logarithmic derivative of the Gamma function Γ).

A.3 Multinomial Density

The density of a Multinomial discrete distribution for variable $x = \{x_1, \ldots, x_N\}$ with parameters $\pi = \{\pi_1, \ldots, \pi_N\}$ is defined by

$$Mult(x; \pi) = \frac{M!}{\prod_{i=1}^{N} x_i!} \prod_{i=1}^{N} \pi_i^{x_i} \tag{A.5}$$

$$\text{where } \begin{cases} x_i \geq 0 \\ \sum_{i=1}^{N} x_i = M \end{cases} \text{ and } \begin{cases} \pi_i > 0 \\ \sum_{i=1}^{N} \pi_i = 1 \end{cases} \tag{A.6}$$

A.4 Dirichlet Density

The probabilistic density function of the N-state Dirichlet distribution for variable $\pi = \{\pi_1, \ldots, \pi_N\}$ with parameters $f = \{f_1, \ldots, f_N\}$ is defined by

$$Dir(\pi; f) = \frac{\Gamma(\sum_{i=1}^{N} f_i)}{\prod_{i=1}^{N} \Gamma(f_i)} \prod_{i=1}^{N} \pi_i^{f_i - 1} \tag{A.7}$$

where $\Gamma(x)$ is, as before, the Gamma function. Restrictions on variable π and parameter f are the following

$$\pi_i \geq 0 \ \ \forall i, \ \ \sum_{i=1}^{N} \pi_i = 1 \ \ \text{and} \ \ \ f_i > 0 \ \ \forall i$$

Parameters f_i are *prior observation counts* for events governed by π_i. The Dirichlet distribution is the conjugate prior of the parameters of a multinomial distribution. One special case is the symmetric Dirichlet distribution where $f_i = f_0 \ \forall i$. In this case, the density is given by

$$Dir(\pi; f_0) = \frac{1}{c(f_0)} \prod_{i=1}^{N} \pi_i^{f_0 - 1} \tag{A.8}$$

where $c(f_0)$ is a normalizing factor depending only on N and f_0.

We also use the following expectation

$$E[\log \pi_i] = \Psi(f_i) - \Psi\left(\sum_{i'=1}^{N} f_{i'}\right) \tag{A.9}$$

B Variational Bayes Learning

Parameter estimation within the Bayesian framework reduces to an inference problem, that of evaluating the posterior probability $p(\Theta|Y)$ over the parameters Θ given the observed data Y.

In many cases, this results in a high-dimensional problem involving computationally intensive multidimensional integrals over a large number of random variables. Thus, the exact computation of posterior probabilities in such models is prohibitive. Approximate but computationally efficient learning methods are therefore of special interest and Variational Bayes is one of those.

The main idea is to find an analytical and simple distribution $q(\Theta|Y)$ to approximate the complicated posterior probability $p(\Theta|Y)$, such that the Kullback-Leibler (KL) divergence of these two distributions is minimized. Indeed, the log likelihood of the data can be written as

$$\begin{aligned}
\log p(Y) &= \int q(\Theta|Y) \log p(Y) d\Theta \\
&= \int q(\Theta|Y) \log \frac{p(Y,\Theta)}{p(\Theta|Y)} d\Theta \\
&= \int q(\Theta|Y) \log \frac{p(Y,\Theta) q(\Theta|Y)}{q(\Theta|Y) p(\Theta|Y)} d\Theta \\
&= \underbrace{\int q(\Theta|Y) \log \frac{p(Y,\Theta)}{q(\Theta|Y)} d\Theta}_{F(q(\Theta|Y))} + \underbrace{\int q(\Theta|Y) \log \frac{q(\Theta|Y)}{p(\Theta|Y)} d\Theta}_{KL(q(\Theta|Y)||p(\Theta|Y)) \geq 0}
\end{aligned} \tag{B.1}$$

The first term $F(q)$ is also known as the negative variational free energy, while the second term is the Kullback-Leibler divergence between the approximate density $q(\Theta|Y)$ and the true posterior $p(\Theta|Y)$. Furthermore, it is easy to see that the log evidence is lower bounded by $F(q)$ since the KL divergence is always nonnegative. Then, by maximizing the lower bound $F(q)$ with regard to q, an optimal approximation of $p(\Theta|Y)$ can be obtained with $q^*(\Theta)$, and a closest value of the log evidence $\log p(Y)$ by $F(q^*)$. The first output will be used for Bayesian inference while the second one, F, can be used for Bayesian Model Comparison (BMC) (Penny et al., 2005a).

The main idea of the VB framework is to find the best approximation $q(\Theta)$ of $p(\Theta|Y)$ within a family of densities which will yield good properties, such as analytical approximations. Factorized forms prove useful and these correspond to a *mean field* approximation. Thus, parameters are split into several groups $\Theta = \{\theta_i\}$ and it is assumed that the approximate posterior density factorises

over these groups of parameters

$$q(\Theta|Y) = \prod_i q(\theta_i|Y) \tag{B.2}$$
$$= q(\theta_i|Y)q(\theta_{|i}|Y)$$

where $\theta_{|i}$ indicates components of Θ other than θ_i.

The maximisation of the negative free energy F, lower bound of the log likelihood, over the posterior density q using Lagrange multipliers yields

$$q(\theta_i|Y) = \frac{\exp\left(\int q(\theta_{|i}|Y)\log p(Y,\Theta)d\theta_{|i}\right)}{\int \exp\left(\int q(\theta_{|i}|Y)\log p(Y,\Theta)d\theta_{|i}\right)d\theta_i} = \frac{\exp\left(I(\theta_i)\right)}{\int \exp\left(I(\theta_i)\right)d\theta_i} \tag{B.3}$$

where

$$I(\theta_i) = \int q(\theta_{|i}|Y)\log p(Y,\Theta)d\theta_{|i} = \langle \log p(Y,\Theta)\rangle_{q(\theta_{|i})} \tag{B.4}$$

Note that the latter integral need only to contain terms dependent on θ_i: this is the feature which lowers the dimensionality of the Bayesian inference. Furthermore, this provides a way to analytically obtain approximate equations for the posterior distributions

$$q(\theta_i|Y) \propto \exp\left(I(\theta_i)\right) \tag{B.5}$$

Using well behaved priors, such as conjugate priors, the approximate posterior distributions will have a known form, giving closed form equations for the updates of their sufficient statistics. The VB algorithm is then simply the iterative computation of these update equations, that will converge so as to maximize the objective function, the negative free energy, by finding the approximate posterior distribution, in the factorized family, that will best match the true posterior in the sense of KL divergence.

References

Andrade, A., Kherif, F., Mangin, J.-F., Worsley, K., Paradis, A.-L., Simon, O., Dehaene, S., Poline, J.-B., 2001. Detection of fMRI activation using cortical surface mapping. Human Brain Mapping 12, 79–93.

Antoniadis, A., Bigot, J., Sapatinas, T., 2001. Wavelet estimators in nonparametric regression: A comparative simulation study. Journal of Statistical Software 6 (6), 1–83.

Ashburner, J., Friston, K., 2005. Unified segmentation. NeuroImage 26, 839–851.

Aston, J., Gunn, R., Hinz, R., Turkheimer, F., 2005. Wavelet variance components in image space for spatiotemporal neuroimaging data. NeuroImage 25, 159–168.

Aston, J., Turkheimer, F., Brett, M., 2006. HBM functional imaging analysis contest data analysis in wavelet space. Human Brain Mapping 27, 372–379.

Attias, H., 2000. A variational Bayesian framework for graphical models. In: Advances in Neural Information Processing Systems 12. MIT Press, pp. 209–215.

Beal, M., 2003. Variational algorithms for approximate Bayesian inference. Ph.D. thesis, University College London.

Bullmore, E., Fadili, M., Maxim, V., Sendur, L., Whitcher, B., Suckling, J., Brammer, M., Breakspear, M., 2004. Wavelets and functional magnetic resonance imaging of the human brain. NeuroImage 23, 234–249.

Chipman, H., McCulloch, R., Kolaczyk, E., 1997. Adaptive Bayesian wavelet shrinkage. Journal of the American Statistical Association 92, 1413–1421.

Clyde, M., Parmigiani, G., Vidakovic, B., 1998. Multiple shrinkage and subset selection in Wavelets. Biometrika 85 (2), 391–401.

Descombes, X., Kruggel, F., von Cramon, D. Y., 1998. fMRI signal restoration using a spatio-temporal Markov random field preserving transitions. NeuroImage 8, 340–349.

Donoho, D., Johnstone, I., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81 (3), 425–455.

Donoho, D., Johnstone, I., 1995. Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association 90 (432), 1200–1224.

Everitt, B., Bullmore, E., 1999. Mixture model mapping of brain activation in functional magnetic resonance images. Human Brain Mapping 8, 340–349.

Fadili, M., Bullmore, E., 2004. A comparative evaluation of wavelet-based methods for hypothesis testing of brain activation maps. NeuroImage 23 (3), 1112–1128.

Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., Penny, W., 2003. Human Brain Function, 2nd Edition. Academic Press.

Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2003. Adaptive analysis of fMRI data. NeuroImage 19 (3), 837–845.

Friston, K., Ashburner, J., Frith, C., Poline, J., Heather, J. D., Frackowiak,

R., 1995a. Spatial registration and normalization of images. Human Brain Mapping 2, 165–189.

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

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., Frackowiak, R., 1995b. Statistical parametric maps in functional imaging: A general linear approach. Human Brain Mapping 2, 189–210.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2006. Variational free energy and the Laplace approximation. NeuroImage Submitted.

Friston, K., Penny, W., 2003. Posterior probability maps and SPMs. NeuroImage 19 (3), 1240–1249.

Godtliebsen, F., Marron, J., Chaudhuri, P., 2004. Statistical significance of features in digital images. Image and Vision Computing 22 (13), 1093–1104.

Gössl, C., Auer, D., Fahrmeir, L., 2001. Bayesian spatio-temporal inference in functional magnetic resonance imaging. Biometrics , 554–562.

Hartvig, N., Jensen, J., 2000. Spatial mixture modelling of fMRI data. Human Brain Mapping 11, 233–248.

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

Kiebel, S., Goebel, R., Friston, K., 2000. Anatomically informed basis functions. NeuroImage 11 (6), 656–667.

Kim, H., Giacomantone, J., Cho, Z., 2005. Robust anisotropic diffusion to produce enhanced statistical parametric map from noisy fMRI. Computer Vision and Image Understanding 99 (3), 435–452.

Lappalainen, H., Miskin, J., 2000. Ensemble learning. In: Advances in Independent Component Analysis. Springer Verlag, pp. 75–92.

Mallat, S., 1989. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7), 674–693.

Mallat, S., 1999. A wavelet tour of signal processing. Academic Press.

Penny, W., 2001. Variational Bayes for d-dimensional Gaussian mixture models. Tech. rep., University College London.

Penny, W., Flandin, G., 2005. Bayesian analysis of fMRI data with spatial priors. In: Proceedings of the Joint Statistical Meeting (JSM), Section on Statistical Graphics. Alexandria, VA: American Statistical Association.

Penny, W., Flandin, G., Trujillo-Barreto, N., 2005a. Bayesian comparison of spatially regularised general linear models. Human Brain Mapping Accepted for publication.

Penny, W., Friston, K., 2003. Mixtures of general linear models for functional neuroimaging. IEEE Transactions on Medical Imaging 22 (4), 504–514.

Penny, W., Kiebel, S., Friston, K., 2003. Variational Bayesian inference for fMRI time series. NeuroImage 19 (3), 727–741.

Penny, W., Trujillo-Bareto, N., Friston, K., 2005b. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24 (2), 350–362.

Poline, J.-B., Mazoyer, B., 1994. Analysis of individual brain activation maps using hierarchical description and multiscale detection. IEEE Transactions on Medical Imaging 13, 702–710.

Ruttimann, U., Unser, M., Rawlings, R., Rio, D., Ramsey, N., Mattay, V., Hommer, D., Frank, J., Weinberger, D., 1998. Searching scale space for activation in PET images. IEEE Transactions on Medical Imaging 17, 142–154.

Sendur, L., Maxim, V., Whitcher, B., Bullmore, E., 2005. Multiple hypothesis mapping of functional MRI data in orthogonal and complex wavelet domains. IEEE Transactions in Signal Processing 53 (9), 3413–3426.

Solé, A., Ngan, S., Sapiro, G., Hu, X., López, A., 2001. Anisotropic 2D and 3D averaging of fMRI signals. IEEE Transactions on Medical Imaging 20 (2), 86–93.

SPM, 2005. Wellcome Department of Imaging Neuroscience. Available from `http://www.fil.ion.ucl.ac.uk/spm/`.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.-B., 2006. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. Human Brain Mapping 27 (8), 678–693.

Turkheimer, F., Aston, J., Asselin, M., Hinz, R., 2006. Multi-resolution Bayesian regression in PET dynamic studies using wavelets. NeuroImage 32 (1), 111–121.

Van De Ville, D., Blu, T., Unser, M., 2003. Wavelets versus resels in the context of fMRI: establishing the link with SPM. In: Wavelets X: part of SPIE's Symposium on Optical Science and Technology. San Diego, USA, pp. 417–428.

Van De Ville, D., Blu, T., Unser, M., 2004. Integrated wavelet processing and spatial spatistical testing of fMRI data. NeuroImage 23, 1472–1485.

Van De Ville, D., Blu, T., Unser, M., 2006. WSPM or how to obtain statistical parametric maps using shift-invariant wavelet processing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. V. Toulouse, France, pp. 1101–1104.

WaveLab, 1999. D. Donoho, M. Duncan, X. Huo, and O. Levi. Available from `http://www-stat.stanford.edu/~wavelab/`.

Wink, A., Roerdink, J., 2004. Denoising functional MR images: a comparison of wavelet denoising and Gaussian smoothing. IEEE Transactions on Medical Imaging 23, 375–387.

Woolrich, M., Behrens, T., Beckmann, C., Smith, S., 2005. Mixture models with adaptive sptial regularization for segmentation with an application to fMRI data. IEEE Transactions on Medical Imaging 24 (1), 1–11.

Woolrich, M., Jenkinson, M., Brady, J., Smith, S., 2004. Fully Bayesian spatio-temporal modeling of fMRI data. IEEE Transactions on Medical Imaging 23, 213–231.

Worsley, K., Marret, S., Neelin, P., Evans, A., 1996. Searching scale space for

activation in PET images. Human Brain Mapping 4, 74–90.

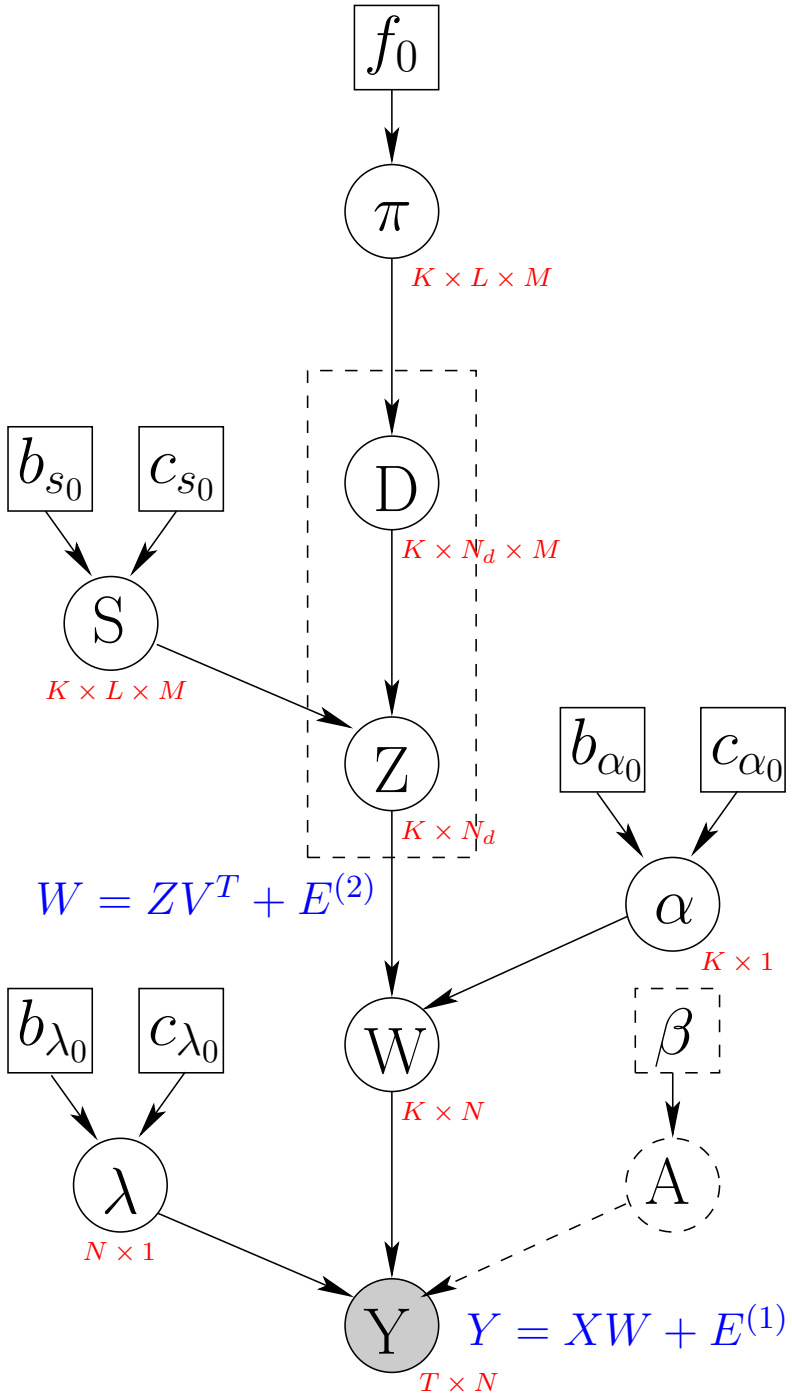


Fig. B.1. Graphical representation of the probabilistic generative model for fMRI with SSBF priors. Each parameter of the model is a node in the graph, whose links correspond to directed probabilistic dependencies. Circles are used to represent unknown quantities to be inferred, and squares for fixed values. Dashed variables A and β are parameters of an autoregressive model, not presented in this article, that could be incorporated here to deal with short term temporal correlation of the noise, as described in (Penny et al., 2003).

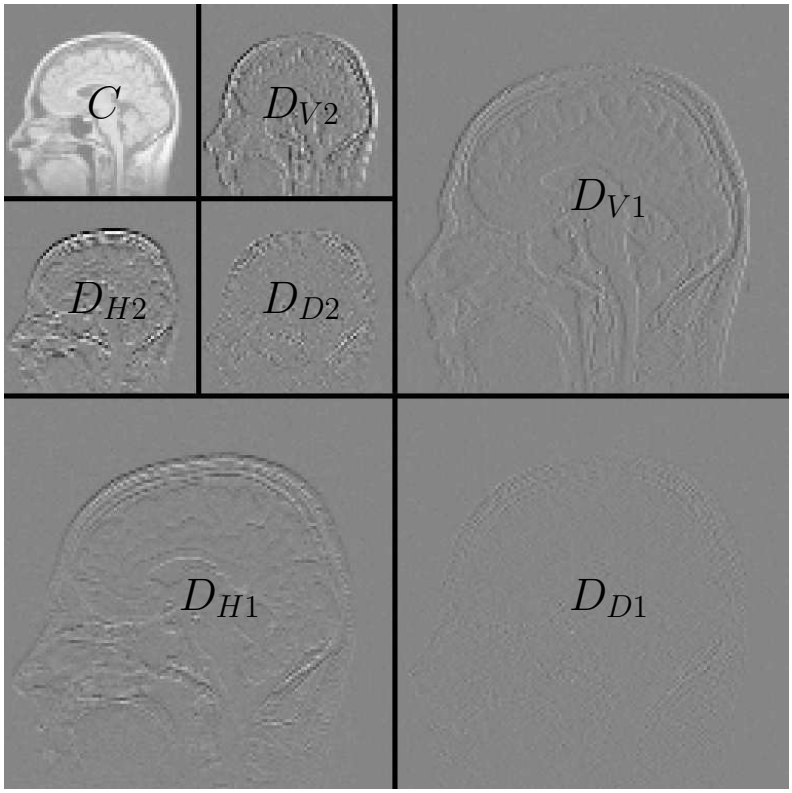
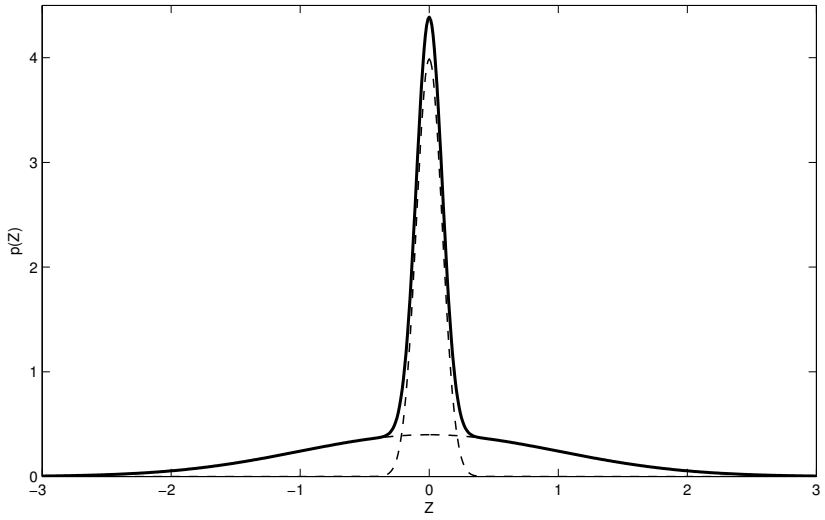
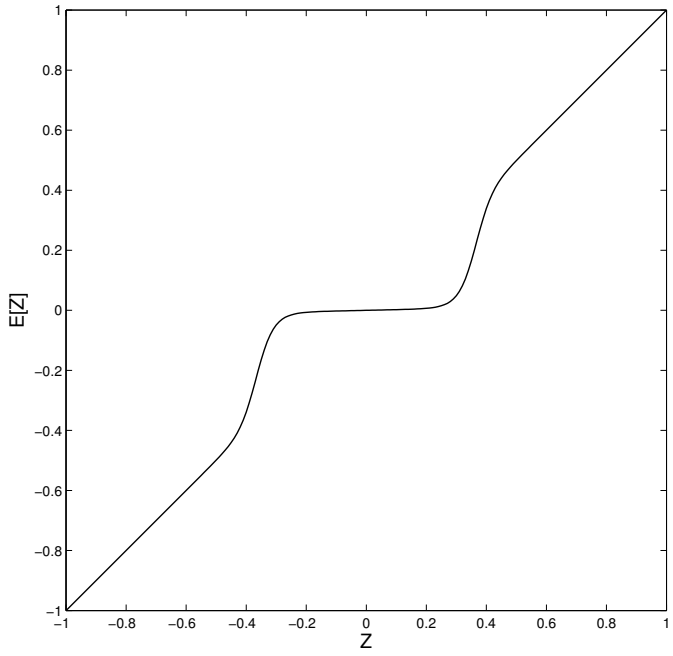


Fig. B.2. Wavelet representation using three resolution levels of a sagittal (2D) slice of a structural MRI (following (Mallat, 1989)). Each detail level ($l = \{1; 2\}$ here) contains diagonal D_{Dl}, horizontal D_{Hl} and vertical D_{Vl} orientation coefficients. The top left box C contains the coarse level.



(a)



(b)

Fig. B.3. (a) Sparse prior: A two component Gaussian mixture enforces sparsity over wavelet coefficients z_k. Each Gaussian is zero mean and the precisions are such that one component has low precision, modelling signal, and the other high precision, modelling noise. (b) Posterior mean of wavelet coefficients, as a function of the bottom-up estimate (i.e. without mixture prior), highlighting the shrinkage effect.

$$q(\Theta|Y) = \left(\prod_{k=1}^{K} q(z_k^T)q(\alpha_k)\right)\left(\prod_{n=1}^{N} q(w_n)q(\lambda_n)\right)\left(\prod_{k,l=1}^{K,L} q(\pi_{kl})\right)\left(\prod_{k,l,m=1}^{K,L,M} q(s_{klm})\right)\left(\prod_{k,l,n=1}^{K,L,N_l} q(d_{kln})\right)$$

Regression coefficients

$$q(W) = \prod_{n=1}^{N} q(w_n)$$
$$q(w_n) = N(w_n; \bar{w}_n, \Sigma_{w_n})$$
$$\bar{w}_n = \Sigma_{w_n}\left(\bar{\lambda}_n X^T y_n + \bar{r}_n^T\right)$$
$$\Sigma_{w_n} = \left(\bar{\lambda}_n X^T X + \mathtt{diag}(\bar{\alpha})\right)^{-1}$$
$$\text{with } \bar{R} = \begin{bmatrix} \vdots & & \vdots \\ \bar{\alpha}_1 V \bar{z}_1^T & \cdots & \bar{\alpha}_K V \bar{z}_K^T \\ \vdots & & \vdots \end{bmatrix}$$

Observation noise precisions

$$q(\lambda) = \prod_{n=1}^{N} q(\lambda_n)$$
$$q(\lambda_n) = G_a(\lambda_n; b_{\lambda_n}, c_{\lambda_n})$$
$$\frac{1}{b_{\lambda_n}} = \frac{1}{2}\left[(y_n - X\bar{w}_n)^T(y_n - X\bar{w}_n) + \mathtt{tr}(\Sigma_{w_n} X^T X)\right] + \frac{1}{b_{\lambda_0}}$$
$$c_{\lambda_n} = \frac{T}{2} + c_{\lambda_0}$$

Wavelet coefficient precisions

$$q(S) = \prod_{k,l,m=1}^{K,L,M} s_{klm}$$
$$q(s_{klm}) = G_a(s_{klm}; b_{s_{klm}}, c_{s_{klm}})$$
$$\frac{1}{b_{s_{klm}}} = \frac{1}{2}\left[\mathtt{tr}(\bar{\Gamma}_{klm}\Sigma_{z_{kl}}) + \bar{z}_{kl}\bar{\Gamma}_{klm}\bar{z}_{kl}^T\right] + \frac{1}{b_{s_0}}$$
$$c_{s_{klm}} = \frac{\bar{N}_{klm}}{2} + c_{s_0} \text{ with } \bar{N}_{klm} = \sum_{n=1}^{N_l} \gamma_{klnm}$$
$$\text{and } \bar{\Gamma}_{klm} = \mathtt{diag}(\gamma_{kl1m}, \cdots, \gamma_{klN_lm})$$

Wavelet coefficients

$$q(Z) = \prod_{k=1}^{K} q(z_k^{dT})$$
$$q(z_k^{dT}) = N(z_k^{dT}; \bar{z}_k^{dT}, \Sigma_{z_k^d})$$
$$\bar{z}_{kn}^{dT} = \Sigma_{z_k^d}\bar{\alpha}_k V_d^T \bar{w}_k^T$$
$$\Sigma_{z_k^d} = \left(\bar{\alpha}_k V_d^T V_d + \sum_{m=1}^{M} \bar{\Lambda}_{km}\right)^{-1}$$
$$\bar{\Lambda}_{km} = \mathtt{blkdiag}(\bar{s}_{klm}\bar{\Gamma}_{klm})_{l=1}^{L}$$

Wavelet residual precisions

$$q(\alpha) = \prod_{k=1}^{K} q(\alpha_k)$$
$$q(\alpha_k) = G_a(\alpha_k; b_{\alpha_k}, c_{\alpha_k})$$
$$\frac{1}{b_{\alpha_k}} = \frac{1}{2}\left[\mathtt{tr}(\Sigma_{w_k}) + \mathtt{tr}(V_d^T V_d \Sigma_{z_k^d}) + (\bar{w}_k^T - V\bar{z}_k^T)^T(\bar{w}_k^T - V\bar{z}_k^T)\right] + \frac{1}{b_{\alpha_0}}$$
$$c_{\alpha_k} = \frac{N}{2} + c_{\alpha_0}$$

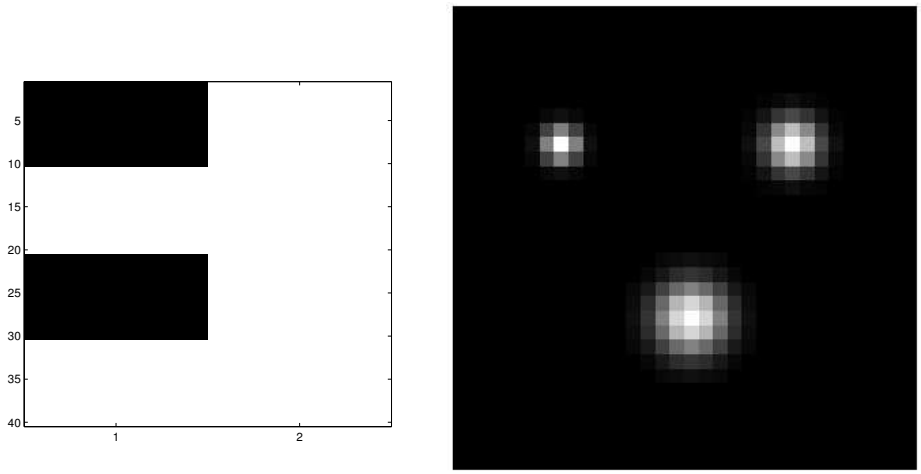
Wavelet switches

$$q(D) = \prod_{k,l,n=1}^{K,L,N_l} q(d_{kln})$$
$$q(d_{kln}) = Mult(d_{kln}; \gamma_{kln})$$
$$\gamma_{klnm} = \frac{\tilde{\gamma}_{klnm}}{\sum_{m'} \tilde{\gamma}_{klnm'}}$$
$$\tilde{\gamma}_{klnm} = \tilde{\pi}_{klm}\tilde{s}_{klm}^{1/2}\exp\left(-\frac{\bar{s}_{klm}}{2}(\tilde{z}_{kln}^2 + \sigma_{z_{kln}}^2)\right)$$

Mixing proportions

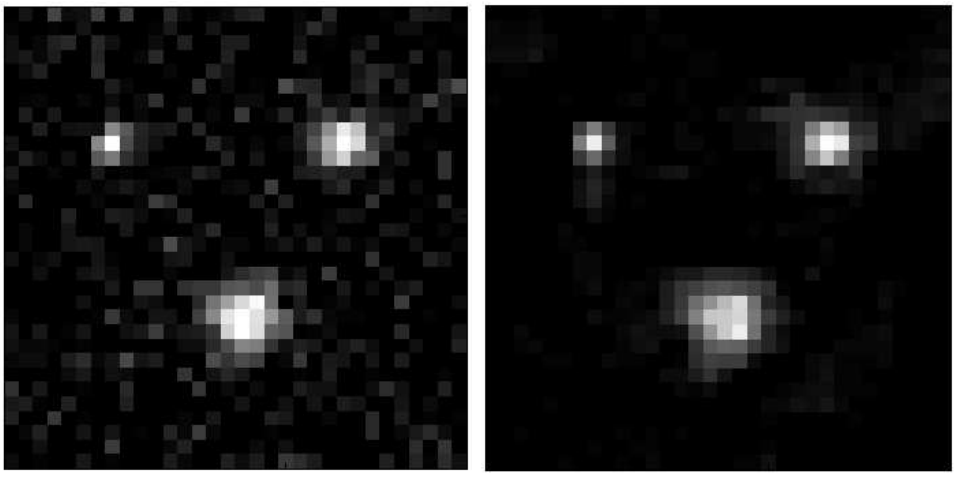
$$q(\pi) = \prod_{k,l=1}^{K,L} q(\pi_{kl})$$
$$q(\pi_{kl}) = Dir(\pi_{kl}; f_{kl})$$
$$f_{klm} = \bar{N}_{klm} + f_{0m}$$
$$\bar{N}_{klm} = \sum_{n=1}^{N_l} \gamma_{klnm}$$

Fig. B.4. Approximate posteriors and update equations for their sufficient statistics. The top equation describes the full approximate posterior with each component in a box below. These equations are valid for any spatial basis V wheras equations in the main text contain simplifications obtained for an orthonormal basis.



(a) (b)

Fig. B.5. Material for the generation of the synthetic data: (a) Design matrix X constituted of two regressors – a boxcar and a constant of dimension T = 40 and (b) 32 × 32 image of regression coefficients. This is the same for both regressors.



(a) (b)

Fig. B.6. (a) Effect as estimated using OLS. (b) Effect as estimated using SSBF. In these images, black denotes 0 and white 1.

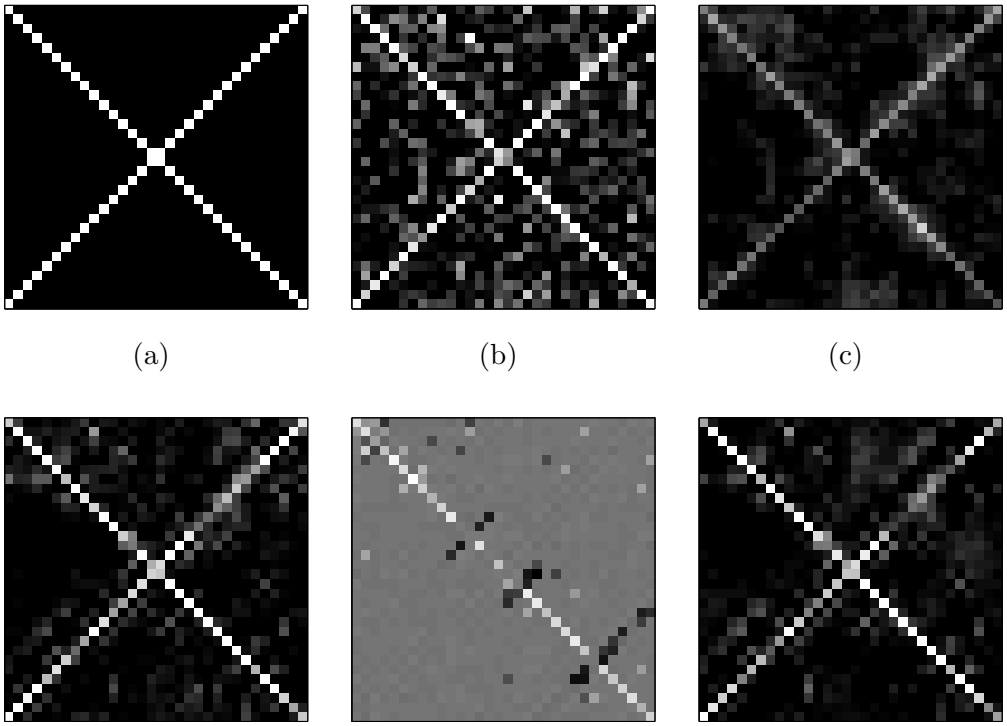


Fig. B.7. (a) 32×32 image of regression coefficients used to generate synthetic data with the design matrix displayed in figure B.5.a. Estimated effects using (b) Ordinary Least Square (OLS), (c) Gaussian Markov Random Fields (GMRFs) and (d) Sparse Spatial Basis Function (SSBF) prior. The estimated wavelet coefficients after convergence of the VB algorithm are displayed in (e). Its inverse wavelet transform is displayed in (f).

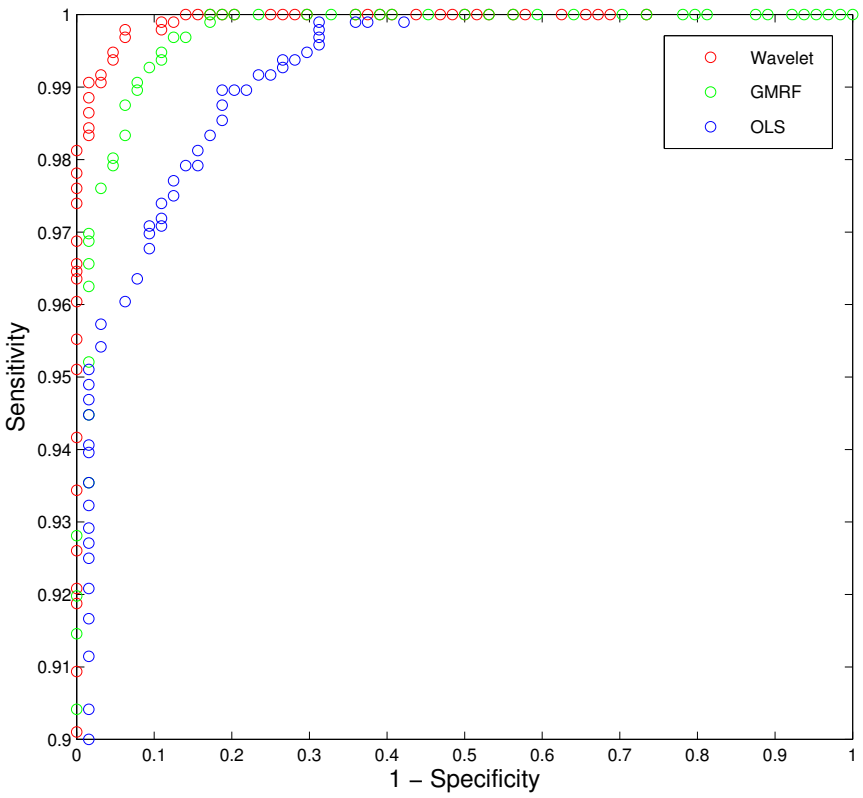


Fig. B.8. ROC curve for the synthetic data of figure B.7 with (i) the SSBF prior model, (ii) the GMRF prior model and (iii) OLS estimates.

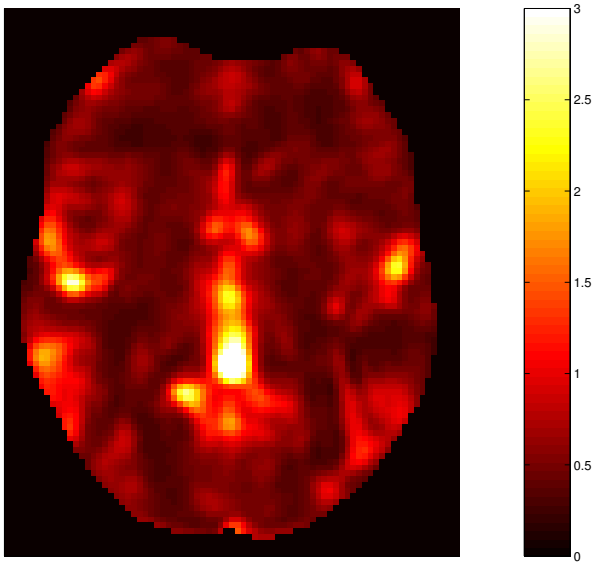


Fig. B.9. Typical axial slice image of residual variances ($\frac{1}{\lambda_n}$), after a GLM model has been estimated on fMRI data. We observe that the noise is far from being uniform across the brain with variations between gray and white matter but also varying inside gray matter.

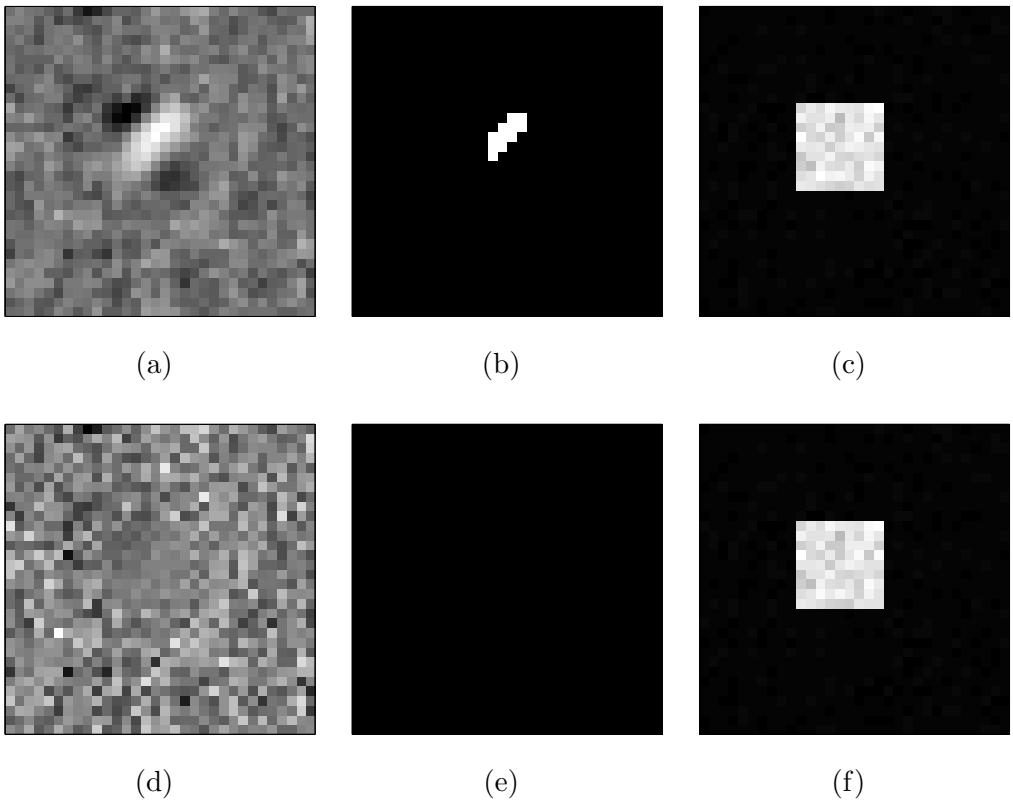
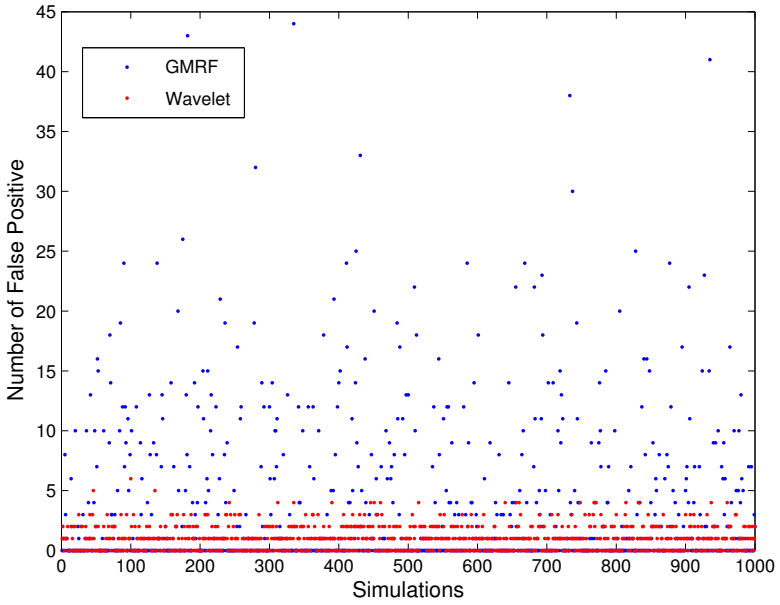
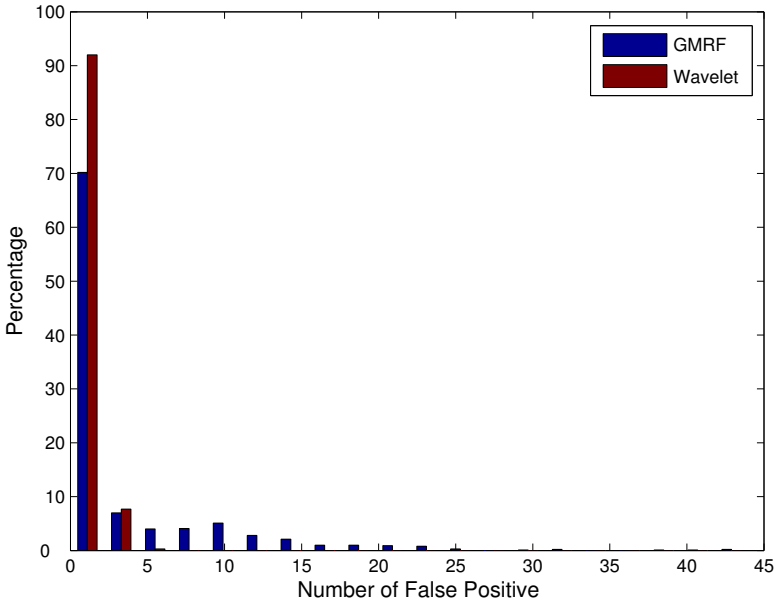


Fig. B.10. Simulated null dataset with heteroscedastic noise (see text). Standard deviation of the noise in a square in the middle of the image is 10 times higher than the one in the background. (a) and (d) show estimated effect using GMRF and SSBF priors. (b) and (e) display the corresponding PPMs and the last column shows the estimated standard deviations for the two models.



(a)



(b)

Fig. B.11. 1000 simulated null datasets with heteroscedastic noise (as in figure B.10) have been generated. (a) displays the number of false positives observed on each simulation with either GMRF or wavelet-based prior. (b) displays the histogram of these observations. Use of the SSBF prior leads to a much lower false positive rate.

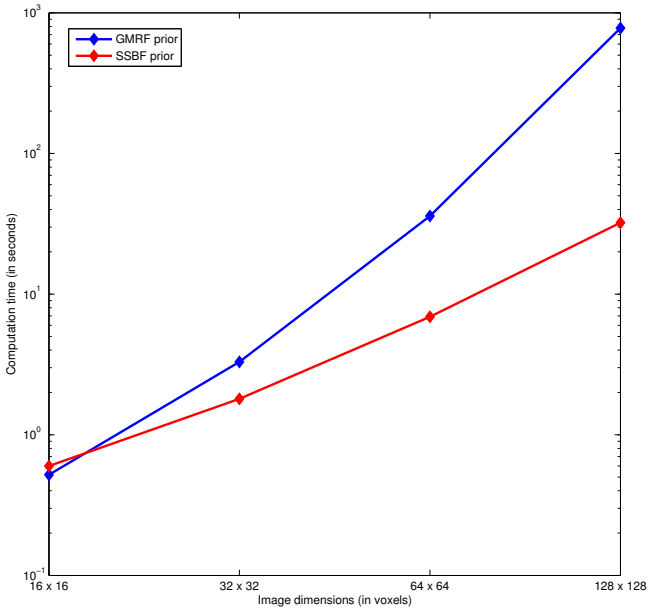


Fig. B.12. Computation time as a function of the dimension of the data. Bayesian estimation was performed with an SSBF prior and a GMRF prior on synthetic images of sizes 16×16, 32×32, 64×64 and 128×128, using the same number of iterations in both cases. The y axis uses a logarithmic scale.

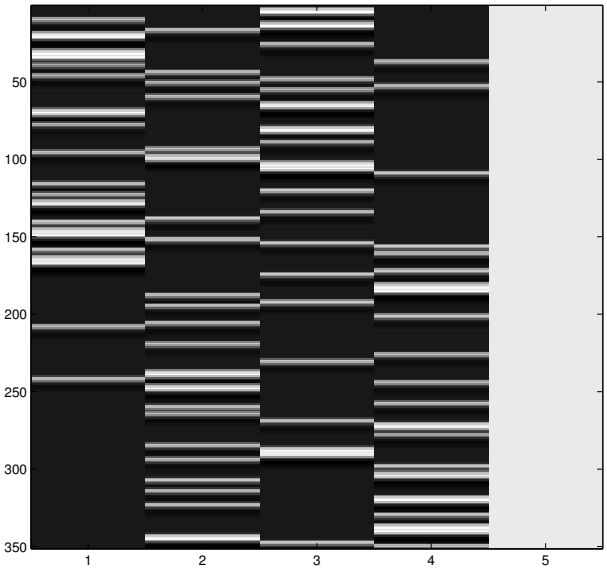
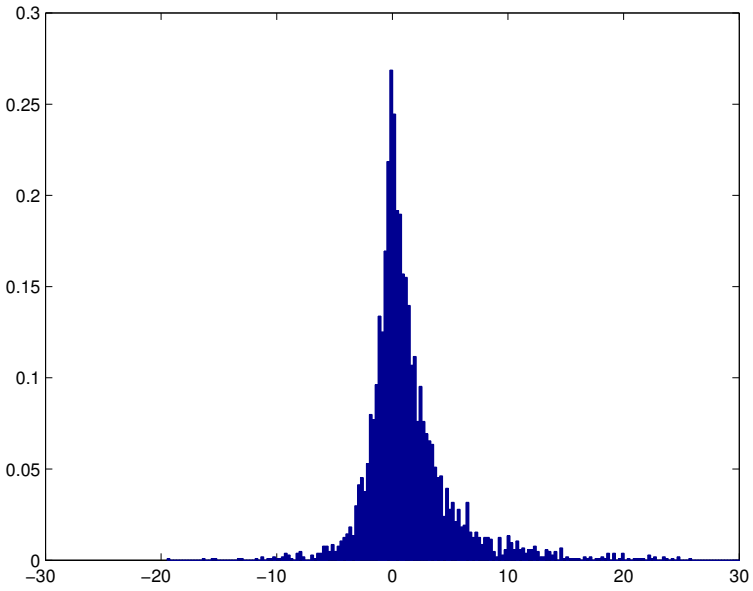
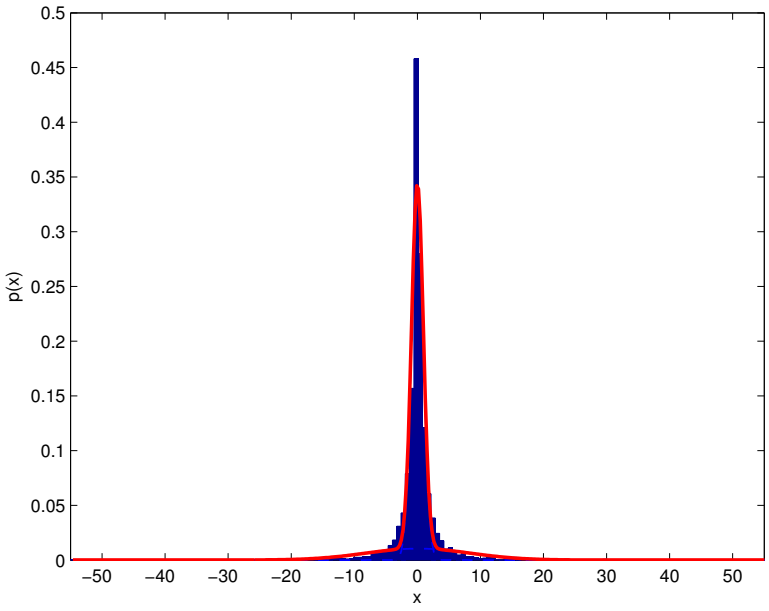


Fig. B.13. Design matrix for the analysis of the event-related fMRI dataset. The first four columns model the four conditions of the factorial design (fame vs. repetition), while the last column models the mean response.

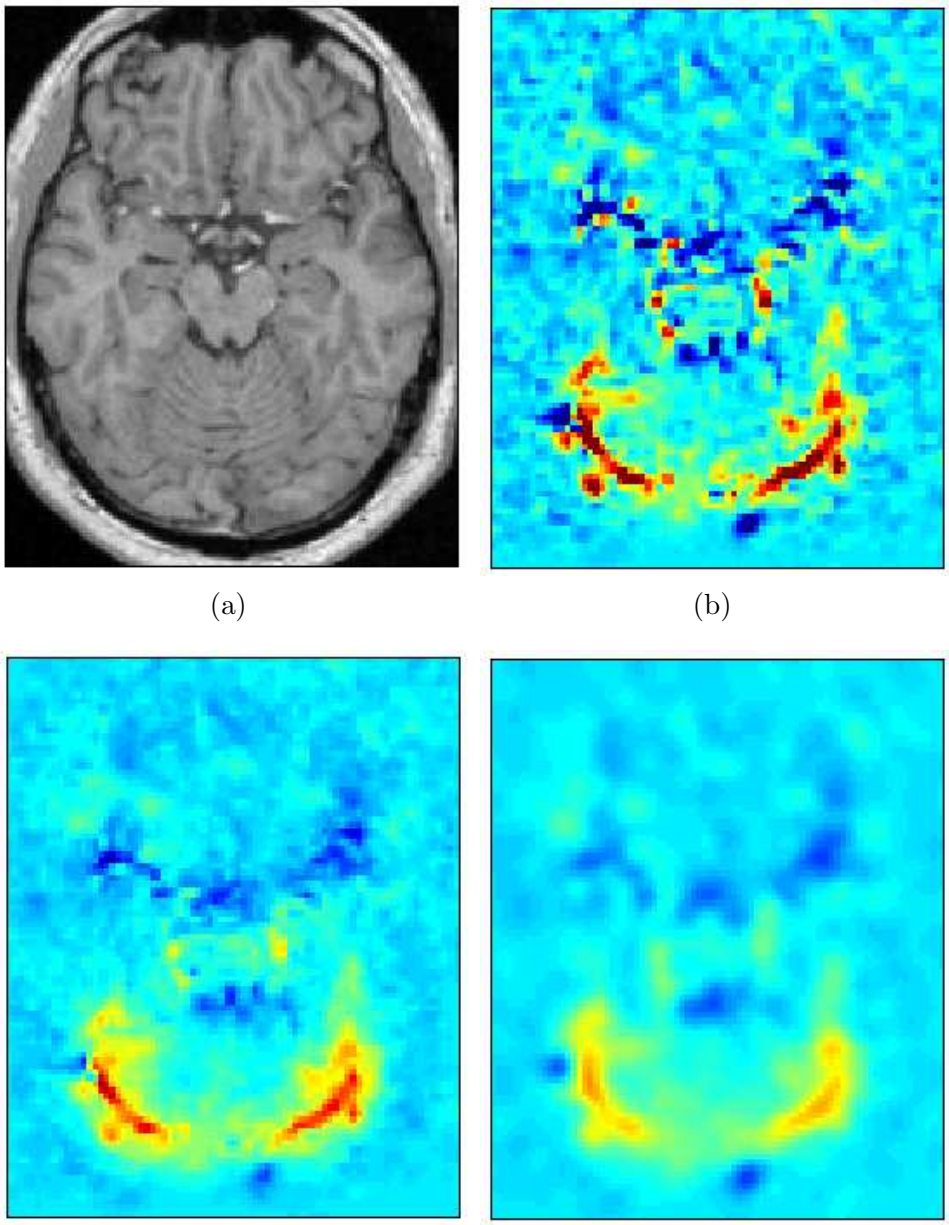


(a)



(b)

Fig. B.14. Histograms of (a) regression coefficients w_1^T of the first column of the design matrix for the event-related fMRI dataset and (b) corresponding wavelet coefficients z_1^T with a fitted zero mean Gaussian mixture model (red curve) superimposed.



(a) (b)

(c) (d)

Fig. B.15. Contrasts images for the main effect of faces, obtained by applying the contrast weight vector $c^T = \frac{1}{4}[11110]$ on an axial slice. (a) Normalized structural scan and images of estimated contrasts using (b) Ordinary Least Square (OLS), (c) a Sparse Spatial Basis Function (SSBF) prior and (d) OLS on images smoothed by a Gaussian kernel of 8mm FWHM.

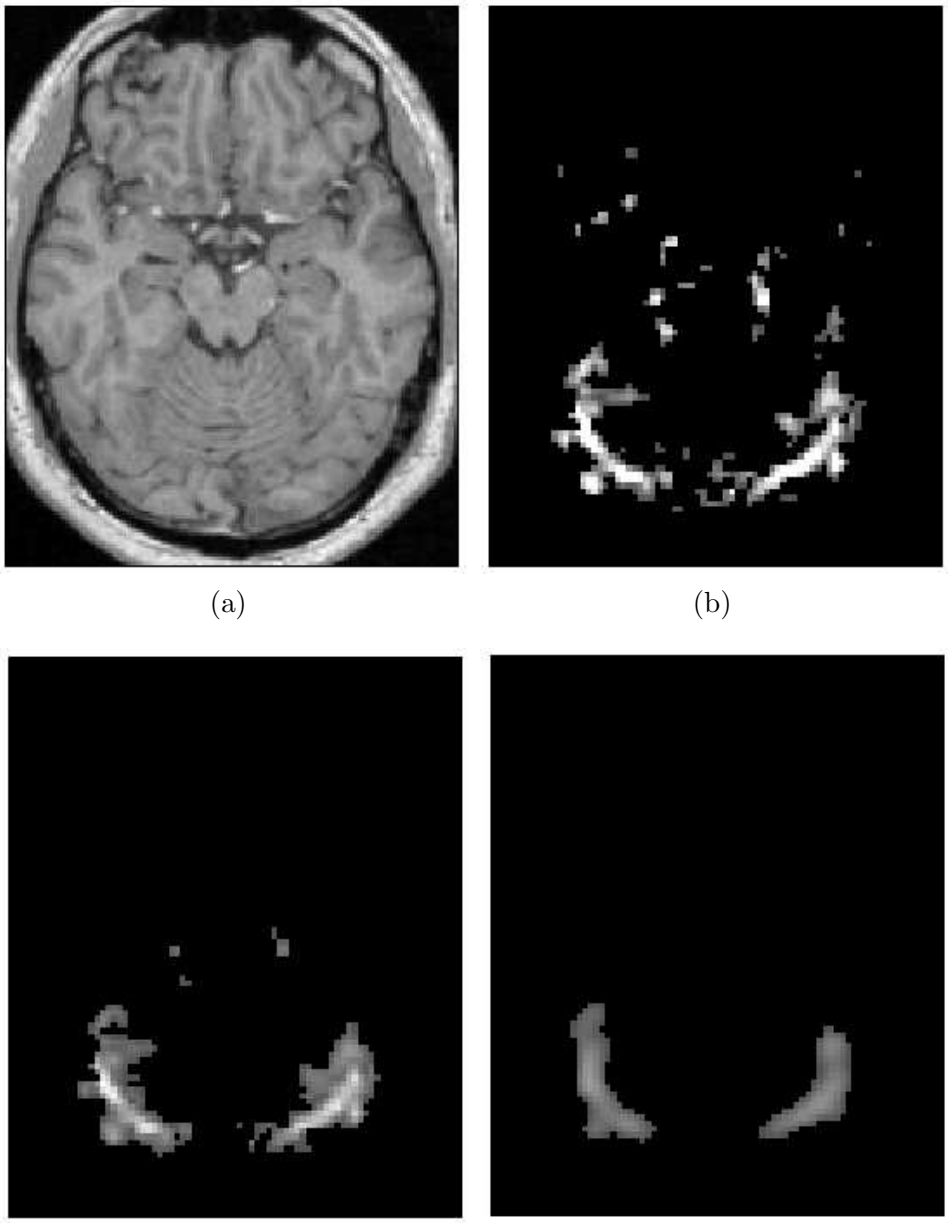


Fig. B.16. Thresholded contrasts images for the main effect of faces, obtained by applying the contrast weight vector $c^T = \frac{1}{4}[11110]$ and thresholding at 5% of the global mean value on an axial slice. (a) Normalized structural scan and thresholded images of estimated contrasts using (b) Ordinary Least Square (OLS), (c) a Sparse Spatial Basis Function (SSBF) prior and (d) OLS on images smoothed by a Gaussian kernel of 8mm FWHM.