NeuroImage 50 (2010) 150–161

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior

Marcel A.J. van Gerven [a,b,*], Botond Cseke [a], Floris P. de Lange [b], Tom Heskes [a,b]

[a] Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen, The Netherlands
[b] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 15 August 2009
Revised 16 November 2009
Accepted 19 November 2009
Available online 1 December 2009

Keywords:
Multivariate analysis
Bayesian inference
Expectation propagation
Logistic regression
Multivariate Laplace distribution

ABSTRACT

Bayesian logistic regression with a multivariate Laplace prior is introduced as a multivariate approach to the analysis of neuroimaging data. It is shown that, by rewriting the multivariate Laplace distribution as a scale mixture, we can incorporate spatio-temporal constraints which lead to smooth importance maps that facilitate subsequent interpretation. The posterior of interest is computed using an approximate inference method called expectation propagation and becomes feasible due to fast inversion of a sparse precision matrix. We illustrate the performance of the method on an fMRI dataset acquired while subjects were shown handwritten digits. The obtained models perform competitively in terms of predictive performance and give rise to interpretable importance maps. Estimation of the posterior of interest is shown to be feasible even for very large models with thousands of variables.

Introduction

In recent years, multivariate analysis has become a popular tool for the analysis of neuroimaging data in general (Haxby et al., 2001; Cox and Savoy, 2003; Mitchell et al., 2004; Kamitani and Tong, 2005; Norman et al., 2006; Pereira et al., 2008). The approach has the same objective as statistical parametric mapping (Friston et al., 1995) in the sense that it aims to identify those regions which show task-related activations. However, while statistical parametric mapping is a mass-univariate approach that aims to predict voxel activations from the design matrix, multivariate analysis aims to predict the structure of (part of) the design matrix from voxel activations. By using voxel activations in conjunction, experimental conditions may become easier to discriminate. We refer to Friston et al. (2008) for a lucid exposition on the differences between mass-univariate and multivariate approaches.

The goal of multivariate analysis is to learn a model that best explains the observed data, quantified in terms of model evidence (how well does the model fit the data and our prior assumptions) or predictive performance (how well does the model predict experimental condition from measured data). Once the model is learned, the obtained parameter estimates can be mapped back to native space, yielding so-called importance maps. These importance maps inform about the relative importance of features in space and/or time with respect to predicting the experimental condition in single trials.

Predictions are typically obtained using classification methods such as linear support vector machines or Gaussian naive Bayes classifiers (Norman et al., 2006). Although these methods often give high predictive performance, they are less suited for interpretation since they result in non-sparse importance maps, erroneously indicating that all voxels are of importance. It is for this reason that classification methods are typically combined with some form of feature selection (Cox and Savoy, 2003; Pereira et al., 2008), yielding sparse importance maps, that are more amenable to interpretation. However, most feature selection methods ignore the fact that importance can only be attributed to features with some degree of certainty given that inferences are based on just a small amount of data. Furthermore, one would like to be able to force the obtained models to obey anatomical constraints as identified from structural MRI and/or DTI data.

In this paper, we introduce a new Bayesian approach to multivariate analysis for the interpretation of neuroimaging data. The approach makes it possible to 1) quantify uncertainty about the relative importance of features and 2) impose constraints on the obtained models based on prior neuroscientific knowledge. Specifically, we will impose a sparsity constraint that forces parameters to have small magnitude, as well as spatio-temporal constraints that couple parameters located closely together in space and/or time. The feasibility of our approach relies on a new representation of Bayesian logistic regression with a multivariate Laplace prior, written as a scale mixture. This representation, can be used to estimate the (analytically intractable) posterior of interest using approximate Bayesian inference methods.

* Corresponding author. Institute for Computing and Information Sciences, Radboud University Nijmegen, P.O. Box 9010, 6500 GL, Nijmegen, The Netherlands. Fax: +31 24 36 52728.
E-mail address: marcelge@cs.ru.nl (M.A.J. van Gerven).

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.11.064

This paper proceeds as follows. First, Bayesian logistic regression and the scale mixture representation of the multivariate Laplace prior are introduced. Next, we show how to use this Bayesian model for the analysis of neuroimaging data. Here, we also touch upon how to estimate the posterior of interest but details of the approximate inference procedure are deferred to the Appendix A in order to improve readability of the main text. Subsequently, the fMRI dataset and experiments used to illustrate our method are introduced. The experimental results show the strengths of our method and we end the paper with a discussion of our approach to multivariate analysis.

Materials and methods

Bayesian logistic regression

Consider a dataset $\mathbf{D}=\{(y_n, \mathbf{x}_n)\}_{n=1}^N$ where the response variable y (e.g., experimental condition) depends on a small subset of the K covariates $\mathbf{x}$ (e.g., voxel activations). Logistic regression assumes that the data are Bernoulli distributed

$$y_n \sim \mathcal{B}\left(l^{-1}\left(\mathbf{x}_n^T\boldsymbol{\beta}\right)\right)$$

with regression coefficients $\boldsymbol{\beta}$ and logit link function $l(p)=\log(p/(1-p))$. Under the assumption that the data is independent and identically distributed,[1] the *likelihood* of the parameters given the data decomposes as

$$p(\mathbf{D}|\boldsymbol{\beta}) = \prod_n \mathcal{B}\left(y_n; l^{-1}\left(\mathbf{x}_n^T\boldsymbol{\beta}\right)\right).$$

Bayesian logistic regression assumes a *prior* $p(\boldsymbol{\beta}|\boldsymbol{\theta})$ over regression coefficients given fixed hyper-parameters $\boldsymbol{\theta}$. This prior distribution is used to express our a priori beliefs about the values assumed by the regression coefficients. Applying Bayes' rule, the posterior over the regression coefficients becomes

$$p(\boldsymbol{\beta}|\mathbf{D},\boldsymbol{\theta}) = \frac{p(\mathbf{D}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\theta})}{p(\mathbf{D}|\boldsymbol{\theta})} \tag{1}$$

where the model evidence

$$p(\mathbf{D}|\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{\beta}\, p(\mathbf{D}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\theta}) \tag{2}$$

captures how well our model, as parameterised by hyper-parameters $\boldsymbol{\theta}$, is in accordance with the data and our prior beliefs. Bayesian methods that make use of Eq. (1) to compute posteriors of interest are widely used in the neuroimaging community. See, for example, the work reported in Friston et al. (2006). Some of these approaches also make use of spatio-temporal priors to model dependencies between regression coefficients in mass-univariate analysis (Gössl et al., 2001; Woolrich et al., 2004; Penny et al., 2005; Brezger et al., 2007). Here, we introduce an alternative framework that is suitable for multivariate analysis and relies on a representation of the multivariate Laplace prior as a scale mixture.

Multivariate Laplace distribution

As mentioned, the prior $p(\boldsymbol{\beta}|\boldsymbol{\theta})$ allows one to incorporate a priori beliefs about model parameters. In this paper, we are interested in a prior that promotes sparsity and allows the coupling of regression coefficients in space and/or time. This is motivated by our focus on importance maps, which ideally are smooth and show localised activation in a small number of regions. As we will show, sparsity favours localised activation whereas coupling favours smooth importance maps. Sparsity is often enforced by placing a univariate Laplace prior

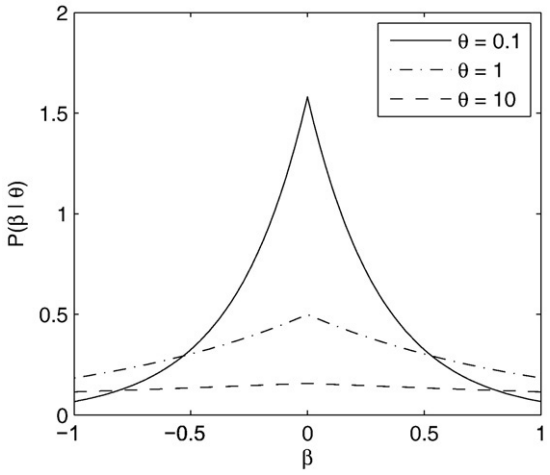

Fig. 1. The univariate Laplace prior for different values of the scale parameter θ.

$$p(\beta|\theta) \equiv \frac{1}{2\sqrt{\theta}}\exp\left(-\frac{|\beta|}{\sqrt{\theta}}\right) \tag{3}$$

with scale parameter θ on individual parameters (Williams, 1995), such that

$$p(\boldsymbol{\beta}|\mathbf{D},\theta) \propto p(\mathbf{D}|\boldsymbol{\beta}) \prod_k p(\beta_k|\theta).$$

This prior, shown in Fig. 1, has been used in the neuroscience community to obtain sparse models (Carroll et al., 2009; van Gerven et al., 2009). It is, however, important to realize that sparse solutions, with many parameters exactly equal to zero, are obtained only when one uses the maximum a posteriori (MAP) estimate for the regression coefficients:

$$\boldsymbol{\beta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\beta}} \{p(\boldsymbol{\beta}|\mathbf{D},\theta)\},$$

which is equivalent to ℓ_1 regularisation (Tibshirani, 1996). In contrast, in a Bayesian setting we are interested in the full posterior $p(\boldsymbol{\beta}|\mathbf{D}, \theta)$ which only reduces to a sparse solution in the limit of infinite data. In other words, the Bayesian approach leaves room for uncertainty in the parameter estimates, even though the prior favours sparse solutions.

A second observation is that the univariate Laplace prior implies that there is no prior coupling between parameters, which may not always be desirable. Particularly when analysing neuroimaging data, we might want to incorporate the notion that a large parameter value for some voxel will tend to be associated with large parameter values for its neighbouring voxels in space and/or time. This immediately leads to the idea of assuming a multivariate Laplace distribution, modelling dependence between regression coefficients.

Scale mixture representation

Consider again the univariate Laplace distribution (Eq. (3)). This distribution can be written as a scale mixture (Andrews and Mallows, 1974):

$$p(\beta|\theta) = \int_0^\infty \mathrm{d}w\, \mathcal{N}(\beta; 0, w)\mathcal{E}(w; 2\theta) \tag{4}$$

$$= \int_0^\infty \mathrm{d}w\, \mathcal{N}(\beta; 0, w\theta)\mathcal{E}(w; 2) \tag{5}$$

[1] Although often used for convenience, the i.i.d. assumption is debatable forfMRI time series since they are contaminated by physiological artefacts which render the time series temporally correlated [Woolrich et al., 2001].

where we made use of the Gaussian distribution $\mathcal{N}(x;\mu,\sigma^2) \equiv (2\pi\sigma^2)^{-1/2} \exp(-(x-\mu)^2/2\sigma^2)$ and exponential distribution $\mathcal{E}(x;\ \lambda) \equiv (1/\lambda)\exp(-x/\lambda)$ That is, the Laplace distribution can be written as an infinite mixture of Gaussians with variance w distributed according to an exponential distribution.

Starting from Eq. (5), Eltoft et al. (2006) proposed the following multivariate Laplace distribution:

$$p(\boldsymbol{\beta}|\boldsymbol{\Theta}) = \int_0^\infty \mathrm{d}w \mathcal{N}(\boldsymbol{\beta};\mathbf{0},w\boldsymbol{\Theta})\mathcal{E}(w;2). \tag{6}$$

Note that this multivariate Laplace distribution couples the regression parameters themselves instead of their magnitudes. Furthermore, with diagonal $\boldsymbol{\Theta}$ it does not factorise into a product of the component probability density functions, i.e., in that sense it cannot be considered a generalisation of a product of Laplace distributions on the individual regression parameters.

In this paper, we propose an alternative definition which is similar in spirit to the scale mixture representation employed by Lyu and Simoncelli (2007). Starting from the observation that an exponential distribution can be written as a χ^2 distribution with two degrees of freedom, we can write Eq. (4) as

$$p(\beta|\theta) = \int_{-\infty}^{\infty} \mathrm{d}u\,\mathrm{d}v \mathcal{N}\left(\beta;0,u^2+v^2\right)\mathcal{N}(u;0,\theta)\mathcal{N}(v;0,\theta).$$

A product of Laplace distributions on the individual regression parameters can then be written as

$$p(\boldsymbol{\beta}|\theta) = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{v}\left(\prod_k \mathcal{N}\left(\beta_k;0,u_k^2+v_k^2\right)\right) \times \mathcal{N}(\mathbf{u};\mathbf{0},\theta\mathbf{I})\mathcal{N}(\mathbf{v};\mathbf{0},\theta\mathbf{I}),$$

with $\mathbf{I}$ the identity matrix. This representation suggests the generalisation

$$p(\boldsymbol{\beta}|\boldsymbol{\Theta}) = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{v}\left(\prod_k \mathcal{N}\left(\beta_k;0,u_k^2+v_k^2\right)\right) \times \mathcal{N}(\mathbf{u};\mathbf{0},\boldsymbol{\Theta})\mathcal{N}(\mathbf{v};\mathbf{0},\boldsymbol{\Theta}), \tag{7}$$

with a (possibly non-diagonal) covariance matrix $\boldsymbol{\Theta}$ that induces couplings between the scales. Essentially, we replace a product of univariate exponential distributions on the scales by a multivariate exponential distribution, which we defined as a generalised χ^2 distribution (Longford, 1990).

In our Eq. (7), the variances of the regression coefficients are coupled, but the regression coefficients themselves are still marginally uncorrelated, i.e., $E[\beta_i\beta_j] = E[\beta_i]E[\beta_j]$. In the alternative definition of the multivariate Laplace distribution (Eq. (6)), the regression coefficients are no longer uncorrelated; one could say that, in that case, $\boldsymbol{\Theta}$ not only introduces a dependency between the magnitudes, but also between their signs.

The scale mixture representation allows us to write the posterior over the latent variables ($\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{v}$) as

$$p(\boldsymbol{\beta},\mathbf{u},\mathbf{v}|\boldsymbol{\Theta},\mathbf{D}) \propto \mathcal{N}(\mathbf{u};\mathbf{0},\boldsymbol{\Theta})\mathcal{N}(\mathbf{v};\mathbf{0},\boldsymbol{\Theta}) \times \prod_k \mathcal{N}\left(\beta_k;0,u_k^2+v_k^2\right) \times \prod_n \mathcal{B}\left(y_n;l^{-1}\left(\mathbf{x}_n^T\boldsymbol{\beta}\right)\right). \tag{8}$$

Hence, we have defined Bayesian logistic regression with a multivariate Laplace prior by representing the prior in terms of a scale mixture using auxiliary variables ($\mathbf{u}$, $\mathbf{v}$). As will become apparent, it is the structure of the precision matrix $\boldsymbol{\Theta}^{-1}$ which determines the interactions between the auxiliary variables and, ultimately, the interactions between the regression coefficients.

We proceed by showing how to use the Bayesian model for the multivariate analysis of neuroimaging data where the goal is to infer experimental conditions from measured voxel activations instead of voxel activations from the design matrix, as is customary in mass-univariate approaches (Fig. 2). The first thing we need to consider is how to use $\boldsymbol{\Theta}$ in order to specify our prior beliefs about the regression coefficients $\boldsymbol{\beta}$. Second, while the scale mixture representation allows us to represent the posterior as in Eq. (8), we have not yet shown how to draw inferences. In this paper, this is realized by approximating the posterior with a Gaussian using expectation propagation. Finally, once we have computed the (approximate) posterior, we can use it to estimate the model evidence, to make predictions for unseen data, and to create importance maps.

Specifying the prior

In order to model the interactions between the latent variables, we need to specify the prior $p(\boldsymbol{\beta}|\boldsymbol{\Theta})$. In principle, we are allowed to use any covariance matrix $\boldsymbol{\Theta}$, but, as we will see later, approximate inference becomes doable even for a huge number of regression parameters only when the precision matrix $\boldsymbol{\Theta}^{-1}$ is sparse. The basic idea is to couple voxels only if they are close-by in space or time, i.e., we will have $(\boldsymbol{\Theta}^{-1})_{ij} \neq 0$ only if the voxels represented by i and j are neighbours (or $i=j$). For simplicity we assume here the same coupling strength between any pair of neighbours such that the prior is fully specified by hyper-parameters $\boldsymbol{\theta} = (\theta,\ s)$ with scale parameter θ and coupling strength s (our procedure easily generalises to different coupling strengths for different types of couplings, e.g., one strength for time and one for space).

Given the coupling strength we build the structure matrix $\mathbf{R}$ with elements

$$r_{i,j} = \begin{cases} -s & \text{if } i \neq j \text{ and } i \sim j \\ 1 + \sum_{k\sim i} s & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Setting $\boldsymbol{\Theta}^{-1} = \mathbf{R}$ we would get

$$p(\mathbf{u}|s) \propto \exp\left(-\frac{1}{2}\sum_i u_i^2 - s\sum_{j\sim i}\left(u_j - u_i\right)^2\right),$$

where $j \sim i$ denotes that j is a neighbour of i. That is, the probability density of the auxiliary variables is a Gaussian which prefers the u_is to be the same and small, with s regulating the relative strength of these tendencies. The scale parameter θ is now supposed to control the (absolute) amount of regularisation towards zero. We incorporate it by constructing the precision matrix from

$$\boldsymbol{\Theta}^{-1} = \frac{1}{\theta}\mathbf{VRV}$$

where $\mathbf{V}$ is a diagonal matrix with $\sqrt{\mathrm{diag}(\mathbf{R}^{-1})}$ on the diagonal. The scaling by $\mathbf{V}$ ensures that the prior variance of the auxiliary variables is independent of the strength s, which makes it easier to study the effect of introducing couplings. An alternative approach would be to use circulant structure matrices (Rue and Held, 2005), although this forces variance at the boundaries to become correlated, which is less suitable for our purposes.

Approximating the posterior

Exact inference for the posterior (Eq. (8)) is intractable. Consequently, we need to resort to approximate inference methods. Various approximate inference methods could be applied such as the Laplace

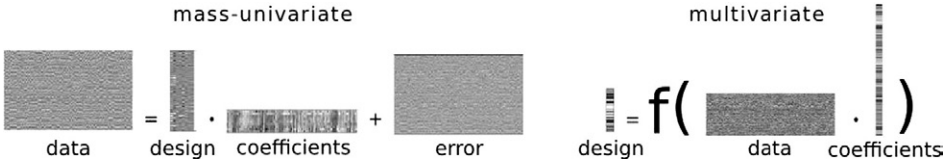

Fig. 2. Mass-univariate approaches explain K voxel time-courses, as captured in a $N \times K$ data matrix as a product of the $N \times M$ design matrix and the $M \times K$ matrix of regression coefficients plus a matrix containing residual errors. Multivariate methods, in contrast, explain conditions in the design matrix as some function f of the product of voxel activations and regression coefficients. Note also that a mass-univariate approach can be framed as a set of regression problems whereas the multivariate approach is typically understood as a classification problem since each trial induces a probability distribution over experimental conditions.

approximation (MacKay, 2004), sampling methods (Rue and Martino, 2007) or variational methods (Bishop, 2006). In this paper, we use a deterministic approximate inference method called *expectation propagation* (EP) (Minka, 2001a) which often outperforms the aforementioned methods (Minka, 2001b; Kuss and Rasmussen, 2005).

In the following, we present a general description of expectation propagation and highlight its computational benefits. EP approximates a density $p(\mathbf{z})$ with a density from the family of exponential distributions $q(\mathbf{z}) = \exp(\boldsymbol{\kappa}^T \mathbf{f}(\mathbf{z})\text{-log } Z(\boldsymbol{\kappa}))$ where $Z(\cdot)$ denotes the partition function. In our application, we are seeking a canonical form Gaussian approximation

$$q(\mathbf{z}) \propto \exp\left(\mathbf{h}^T\mathbf{z} - \frac{1}{2}\mathbf{z}^T\mathbf{K}\mathbf{z}\right)$$

of the posterior $p(\mathbf{z}|\Theta, \mathbf{D})$ with $\mathbf{z} = (\boldsymbol{\beta}, \mathbf{u}, \mathbf{v})$ (we omit the dependence on Θ and $\mathbf{D}$ in the following). The most straightforward way to achieve the approximation is to use the Kullback–Leibler divergence

$$\mathrm{D}[p \parallel q] = \int \mathrm{d}\mathbf{z}\, p(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})}$$

as an error function and minimize with regard to $\boldsymbol{\kappa}$ (which acts as a proxy for $\mathbf{h}$ and $\mathbf{K}$ in the Gaussian case), accordingly. This boils down to matching the first two moments $E[\mathbf{z}]$ and $E[\mathbf{z}\mathbf{z}^T]$ of the distributions p and q. In most cases, this problem is intractable since one has to resort to numerical integration methods in high dimensions.

Expectation propagation is a heuristic method that tries to circumvent this problem by iteratively minimising the Kullback–Leibler divergence between appropriately chosen distributions. It does not compute the moment matching distribution, but the approximation it provides is often close to that. The starting assumption of EP is that the distribution which is to be approximated can be written in a factorised form

$$p(\mathbf{z}) \propto \prod_j t_j(\mathbf{z}).$$

EP approximates p by a distribution q which has a similar form as p:

$$q(\mathbf{z}) \propto \prod_j \tilde{t}_j(\mathbf{z})$$

where the term approximations $\tilde{t}_j$ are defined as Gaussian functions $\tilde{t}_j(\mathbf{z}) = \exp(\mathbf{h}_j^T\mathbf{z} - \mathbf{z}^T\mathbf{K}_j\mathbf{z}/2)$. That is, they have the same exponential form as q and they are not constrained to be normalisable. In our application, the terms t_j either depend on a subset of variables (i.e., the terms which couple variables (β_k, u_k, v_k) or on a linear transformation of the variables $\mathbf{U}_j\mathbf{z}$ (i.e., the likelihood terms for the logistic regression which depend only on $\mathbf{x}_n^T\boldsymbol{\beta}$). As we will see later on, this leads to significant simplifications in the representation of $\tilde{t}_j$.

In order to approximate p, EP proceeds as follows:

1. Remove a term approximation $\tilde{t}_i$ from q and form the function

$$q^{\setminus i}(\mathbf{z}) \propto \prod_{\setminus i} \tilde{t}_j(\mathbf{z}).$$

2. Append t_i to $q^{\setminus i}$ and find $\tilde{t}_i^*$ that minimises

$$\mathrm{D}\left[\frac{1}{Z_i} t_i(\mathbf{z}) q^{\setminus i}(\mathbf{z}) \parallel \frac{1}{\tilde{Z}_i^*} \tilde{t}_i^*(\mathbf{z}) q^{\setminus i}(\mathbf{z})\right]. \tag{9}$$

(this step boils down to choosing $\mathbf{h}_i^*$ and $\mathbf{K}_i^*$ such that the first two moments of both distributions are equal).

3. Repeat the first two steps by cycling through all term approximations $\tilde{t}_i$ until convergence.

Apparently, EP provides no computational advantage over the approximation $\mathrm{D}[p||q]$ since the moment computation in Eq. (9) still requires high dimensional integrals. However, when the terms t_j depend only on a small subset or linear transformation of the variables, the Gaussian approximation to the posterior simplifies substantially since the moment matching in Eq. (9) boils down to computing low dimensional integrals.

In order to approximate our posterior (Eq. (8)) using EP, we need to be able to compute updates for the data terms $\mathcal{B}(y_n; l^{-1}(\mathbf{x}_n^T\boldsymbol{\beta}))$ and auxiliary variable terms $\mathcal{N}(\beta_k; 0, u_k^2 + v_k^2)$ in an efficient manner (the remaining terms are already in Gaussian form). Details concerning the Gaussian approximation using EP and how to apply EP to Bayesian logistic regression with a multivariate Laplace prior are given in Appendix A.

Briefly, we adopt the following strategy: we use a Gaussian approximation $q(\mathbf{z})$ in canonical form and update the terms in parallel. That is, we simultaneously update the parameters of all terms t_j, add them to form the new canonical parameters $\mathbf{h}^* = \Sigma_j \mathbf{h}_j^*$ and $\mathbf{K}^* = \Sigma_j \mathbf{K}_j^*$, and finally compute the moment parameters which are needed for the moment matching in Eq. (9). Canonical form parameters $(\mathbf{h}, \mathbf{K})$ are related to the moment form parameters $(\mathbf{m}, \mathbf{C})$ by $\mathbf{m} = \mathbf{K}^{-1}\mathbf{h}$ and $\mathbf{C} = \mathbf{K}^{-1}$. Although the change in representation seems expensive, it can be very efficient in practice due to reasons described in Appendix A. For stability, we will perform power EP (Minka, 2004) which slightly changes the term updates as we remove and update $\tilde{t}_i^{\alpha}$ with $\alpha < 1$. We chose $\alpha = 0.9$ since this gives more stable behaviour as compared with standard EP where $\alpha = 1$ (Seeger, 2008).

Determining model quality

Next to an approximation of the posterior over latent variables, we would like to have an objective measure of model quality. In this paper, we use the *model evidence* as well as the *predictive performance* as such measures. The model evidence, already encountered in Eq. (2), reads

$$\begin{aligned} p(\mathbf{D}|\Theta) = \int \mathrm{d}\mathbf{z}\, &\mathcal{N}(\mathbf{u}; \mathbf{0}, \Theta)\mathcal{N}(\mathbf{v}; \mathbf{0}, \Theta) \\ &\times \prod_k \mathcal{N}\left(\beta_k; 0, u_k^2 + v_k^2\right) \\ &\times \prod_n \mathcal{B}\left(y_n; l^{-1}\left(\mathbf{x}_n^T\boldsymbol{\beta}\right)\right) \end{aligned}$$

and captures how well our model is in accordance with the data and our prior beliefs. An approximation to the log model evidence within

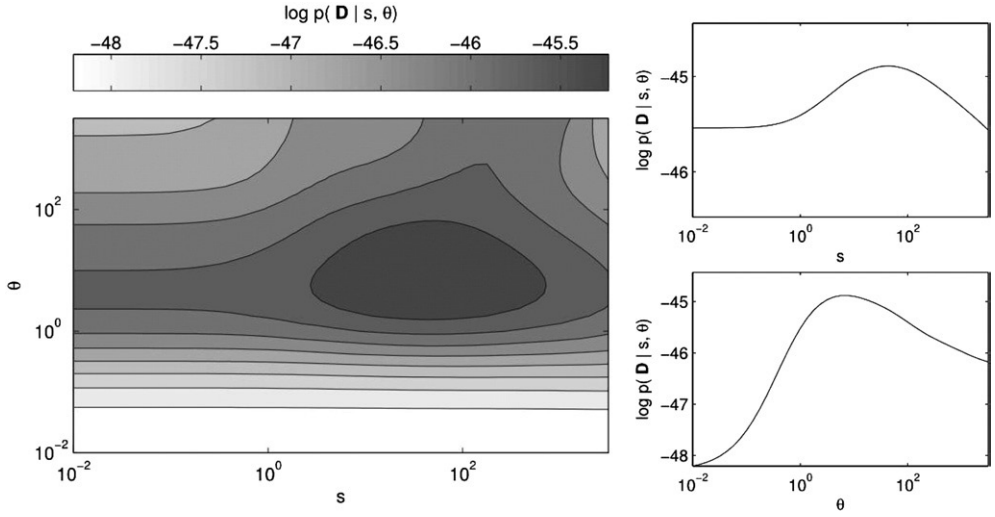

Fig. 3. Plot of the approximate model evidence. A set of 64 samples was generated using a design matrix **X** formed by normalised random vectors as columns. Observations **y** were generated by thresholding $\sigma(\mathbf{X}^T\beta)$ at 0.5 using the sigmoid function σ and a vector of regression coefficients β consisting of a concatenation of 4 ones, 64 zeros, 4 minus ones, 64 zeros, and 4 ones. Neighbouring elements were coupled using a coupling strength s.

the EP framework is given in Appendix A. Given the approximate log model evidences L_1 and L_2 for two competing models M_1 and M_2, we favour M_1 over M_2 whenever $L_1 > L_2$. We can also adjust our hyper-parameters (the coupling strength s and scale parameter θ) in order to maximise the model evidence. This is known in the literature as *empirical Bayes* (Bernardo and Smith, 1994) or *type II maximum likelihood* (Berger, 1985). Fig. 3 provides an example.

Another approach to determining model quality is to determine how well the model predicts the response variable y from measured covariates via the predictive density:

$$p(y_n \mid \mathbf{x}_n, \Theta, \mathbf{D}) = \int \mathrm{d}\mathbf{z}\, p(y_n \mid \mathbf{x}_n, \boldsymbol{\beta}) p(\mathbf{z} \mid \Theta, \mathbf{D}) .$$

This allows us to determine how well the model generalises to previously unseen data. Furthermore, it enables the prediction of experimental condition for single trials, which has applications in real-time analysis of functional imaging data (DeCharms, 2008). Note that the predictive density is impossible to compute exactly since the posterior (Eq. (8)) is intractable. We therefore use the Gaussian approximation to the posterior which reads $q(\mathbf{z}) = \mathcal{N}(\mathbf{z};\ \mathbf{m},\ \mathbf{C})$ in moment form. This leads to

$$p(y_n \mid \mathbf{x}_n) \approx \int d\boldsymbol{\beta}\, \mathcal{B}\left(y_n; l^{-1}\left(\mathbf{x}_n^T \boldsymbol{\beta}\right)\right) \mathcal{N}\left(\boldsymbol{\beta}; \mathbf{m}_\beta, \mathbf{C}_\beta\right)$$

where we marginalised out the auxiliary variables **u** and **v**, and omit Θ and **D** from the notation. In order to compute this integral, we need to make use of some tricks. Introducing notation $\mathbf{a}_n \equiv (2y_n - 1)\mathbf{x}_n$, we can write g_n in the alternative form $\sigma(z_n) = 1/(1 + \exp(-z_n))$ which shows that g_n only depends on $z_n \equiv \mathbf{a}_n^T\boldsymbol{\beta}$. In that case, we may write (Bishop, 2006):

$$p(y_n \mid \mathbf{x}_n) \approx \int dz_n\, \sigma(z_n) \mathcal{N}\left(z_n; \mathbf{a}_n^T \mathbf{m}_\beta, \mathbf{a}_n^T \mathbf{C}_\beta \mathbf{a}_n\right) . \quad (10)$$

The integral over z_n cannot be evaluated analytically but we can obtain a good approximation by making use of the close similarity between the logistic function σ and the probit function, as described in (Bishop, 2006). Here, instead, we choose to numerically approximate Eq. (10) using Gaussian quadrature (Press et al., 2007) in order to be able to control the quality of the approximation; these ideas are reused to increase the efficiency of expectation propagation in the Appendix A. In this way, we can use Eq. (10) to obtain an estimate of the predictive performance, typically expressed in terms of accuracy (proportion of correctly classified trials).

Creating importance maps

We propose the following way of creating importance maps, inspired by our representation of the Laplace prior as a scale mixture distribution. In this representation, the auxiliary variables u_k and v_k, and more specifically their squared sum $u_k^2 + v_k^2$, set the width of the Gaussian prior on the regression parameter β_k. The larger $u_k^2 + v_k^2$, the wider this Gaussian prior, the weaker the regularisation, and the more important the regression parameter. The other way around, a very small scale $u_k^2 + v_k^2$ amounts to a narrow Gaussian prior (close to a Dirac delta), strong regularisation towards zero, and tends to make the regression parameter irrelevant. Applying the same reasoning to the posterior distribution given all observations, we then propose to use the average $E[u_k^2 + v_k^2 \mid \mathbf{D}]$ as a measure of importance of variable x_k. More specifically, we normalise this measure with respect to the importance induced by the prior and simplify to (since $E[u_k^2 \mid \mathbf{D}] = E[v_k^2 \mid \mathbf{D}]$ and $E[u_k \mid \mathbf{D}] = E[v_k \mid \mathbf{D}] = 0$):

$$\begin{aligned} \textbf{importance}(x_k) &= \mathrm{Var}[u_k \mid \mathrm{D}] - \mathrm{Var}[u_k] \\ &= \left(\Theta^{-1} + \mathbf{K}_u^f\right)_{kk}^{-1} - \Theta_{kk} \end{aligned}$$

with $\mathbf{K}_u^f$ being the contribution of the data to the posterior variance as defined in Appendix A.

Stimuli

In order to illustrate our method, we use a dataset which has been collected in the context of another ongoing project. The goal is to examine whether the class to which individual handwritten characters belong can be predicted from measured BOLD response. The stimuli for our experiment consisted of fifty handwritten sixes and fifty handwritten nines, selected at random from the MNIST database of handwritten digits (http://yann.lecun.com/exdb/mnist). The 28×28 pixel grey-scale digits were interpolated to 256×256 pixel images, and subtended a visual angle of 7.8°. Stimuli were presented using Psychtoolbox 3 (Brainard, 1997). Note the large differences between individual instances in our stimulus set which makes the classification problem non-trivial (Fig. 4).

Experimental design

Data was collected for one subject. In each trial, a handwritten six or nine was presented which remained visible for 12500 ms. Stimuli

Fig. 4. Example of the variations within the set of handwritten sixes and nines.

flickered at a rate of 6 Hz on a black background. The experimental task was to maintain fixation to a fixation dot and the detect a brief (30 ms) change in colour from red to green and back occurring once and randomly within a trial. Detection was indicated by pressing a button with the right hand thumb as fast as possible. One-hundred trials were collected which were separated by a 12500 ms inter-trial interval. The total duration of the experiment was 42 min.

Data collection

Functional images were acquired using a Siemens 3 T MRI system using a 32 channel coil for signal reception. Blood oxygenation level dependent (BOLD) sensitive functional images were acquired using a single shot gradient EPI sequence, with a repetition time (TR) of 2500 ms, echo time (TE) of 30 ms, isotropic voxel size of $2\times2\times2$ mm, acquired in 42 axial slices in ascending order. A high-resolution anatomical image was acquired using an MP-RAGE sequence (TE/TR = 3.39/2250 ms; 176 sagittal slices, with isotropic voxel size of $1\times1\times1$ mm).

Data analysis

Functional data were preprocessed within the framework of SPM5 (Statistical Parametric Mapping, www.fil.ion.ucl.ac.uk/spm). Functional brain volumes were realigned to the mean image in order to correct for motion and the anatomical image was coregistered with the mean of the functional images. In order to restrict the number of considered voxels, a grey-matter mask was applied (threshold $p>0.5$, voxel size $4\times4\times4$ mm). Functional data was high-pass filtered and detrended. The volumes acquired up to 15 s after trial onset were collected in order to obtain an estimate of the response in individual voxels. The trials were normalised such that responses for each voxel had mean zero and a standard deviation of one. Trials were subsequently used as input to the multivariate analysis. Predictive performance and model evidence was computed using a ten-fold cross validation scheme while the scale parameter was varied between 10^{-6} and 10^{4}. A binomial test was used to determine whether predictive performance was significantly different from chance level performance (Salzberg, 1997). The cross validation scheme ensured that training and test data remained isolated. Importance maps were obtained by learning a model based on all data while the scale parameter was chosen such that predictive performance was maximal according to the cross validation results.

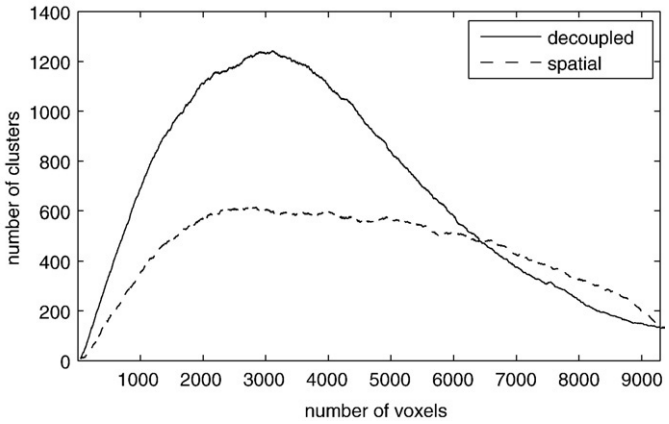

Fig. 6. Number of clusters obtained when varying the number of included voxels sorted according to importance for the decoupled model and the spatial model.

We tested both a spatial model where the prior couples neighbouring voxels in the x, y and z directions with a coupling strength of $s=10$ and a temporal model where the prior couples the same voxel over multiple volumes with a coupling strength of $s=10$. For the spatial model, we used the average (steady-state) response 10 to 15 s after trial onset in order to remove the temporal dimension. For the temporal model, we used all volumes acquired up to 15 s after trial onset. For each model, predictive performance and model evidence was compared with the performance obtained by a model that employed a decoupled prior.

Results

Spatial model

Fig. 5 shows that the predictive performance of the spatial model is marginally better than that of the decoupled model although the difference is not significant. Furthermore, the results clearly show that optimal performance is reached when $\theta\approx0.01$, indicating the improvement over the unregularized model that is obtained in the limit when θ goes to infinity. Likewise, too much regularisation is also detrimental to predictive performance. Predictive performance was

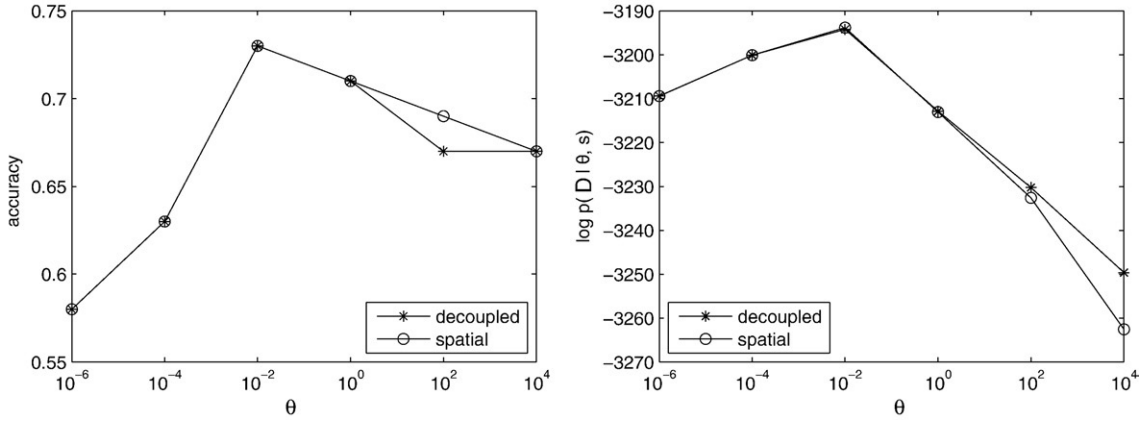

Fig. 5. Proportion of correctly classified trials and approximate log model evidence for the decoupled model and the spatial model.

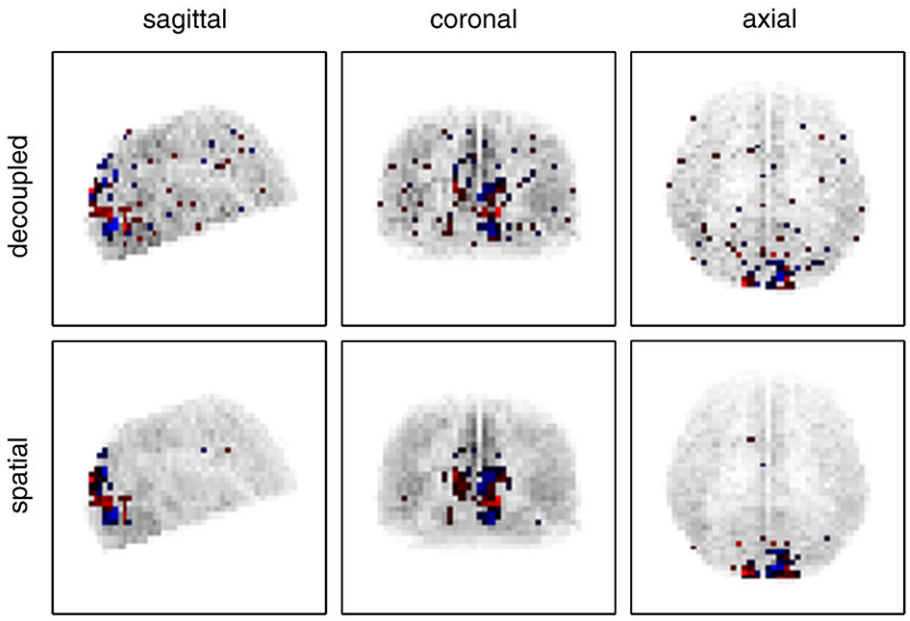

Fig. 7. Glass-brain view of importance values for the decoupled versus the spatial prior. Means of the regression coefficients of the 100 most important voxels are colour-coded with red standing for positive and blue for negative values.

significant ($p<0.05$) for all models with $\theta=0.01$ and $\theta=1$. Log model evidence was slightly larger for the spatial model (log $p(\mathbf{D}|\theta=0.01, s=10)\approx-3913.84$) as compared with the decoupled model (log $p(\mathbf{D}|\theta=0.01, s=0)\approx-3194.21$).

Throughout this paper, we emphasised the improved interpretability that is obtained when using informed priors. Fig. 6 establishes this claim by showing the number of included voxels sorted according to importance versus the number of clusters obtained, where a cluster is defined as a connected component in the measured brain volume. The spatial prior leads to a much lower number of clusters compared to the decoupled prior. The absolute number of clusters remains relatively large due to the grey-matter mask, which selects a non-contiguous subset of voxels from the measured volume.

Fig. 7 provides a visualisation of the resulting models. The spatial prior leads to the selection of clusters of important voxels. Spatial regularisation can also be interpreted as a form of noise reduction since spatially segregated voxels are less likely to have large importance values. The mean regression coefficients have a large magnitude for the most important voxels. Note the strong agreement between the uncoupled and the spatial model regarding these voxels. It can also be seen from Fig 7 that the most important voxels are to some extent scattered throughout the brain in the unconstrained model, while they are almost exclusively observed in the occipital lobe in the spatially constrained model, encompassing Brodmann Areas 17 and 18. This pattern of results appears neurobiologically plausible, in view of the only difference between conditions being visual in nature.

Temporal model

Fig. 8 shows that the predictive performance of the temporal model is again somewhat better than that of the decoupled model although the difference is not significant. Optimal performance is reached when $\theta=1$. Predictive performance was significant ($p<0.05$) only for this setting of the scale parameter. Log model evidence was slightly larger for the decoupled model (log $p(\mathbf{D}|\theta=10^{-4}, s=0)\approx-13629.79$) as compared with the temporal model (log $p(\mathbf{D}|\theta=10^{-4}, s=10)\approx-13629.93$). Note the disagreement between the optimum according to predictive performance and according to model evidence; it is well-known that model evidence optimisation does not always lead to the best predicting models. Note further that predictive performance of the temporal model was significantly lower than that of the spatial model due to the inclusion of volumes for which the task-related response is likely to be negligible.

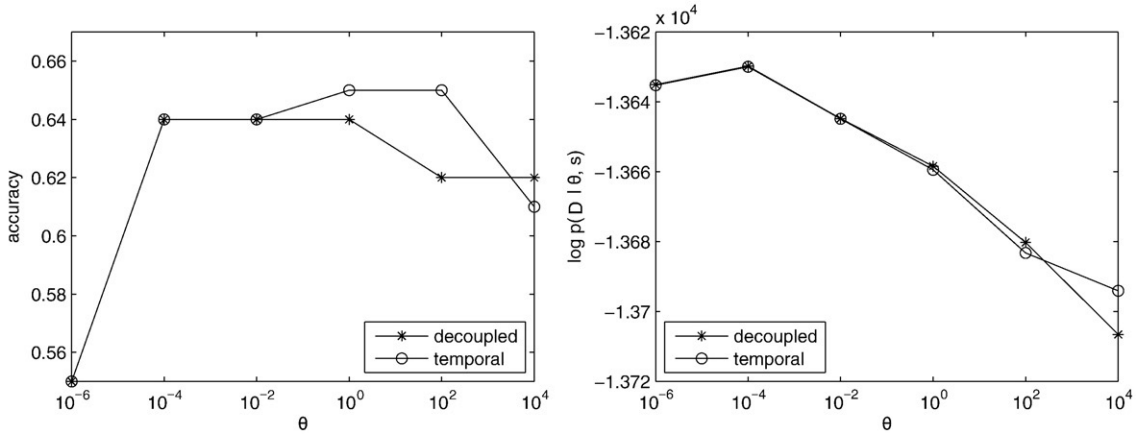

Fig. 8. Proportion of correctly classified trials and approximate log model evidence for the decoupled model and the temporal model.

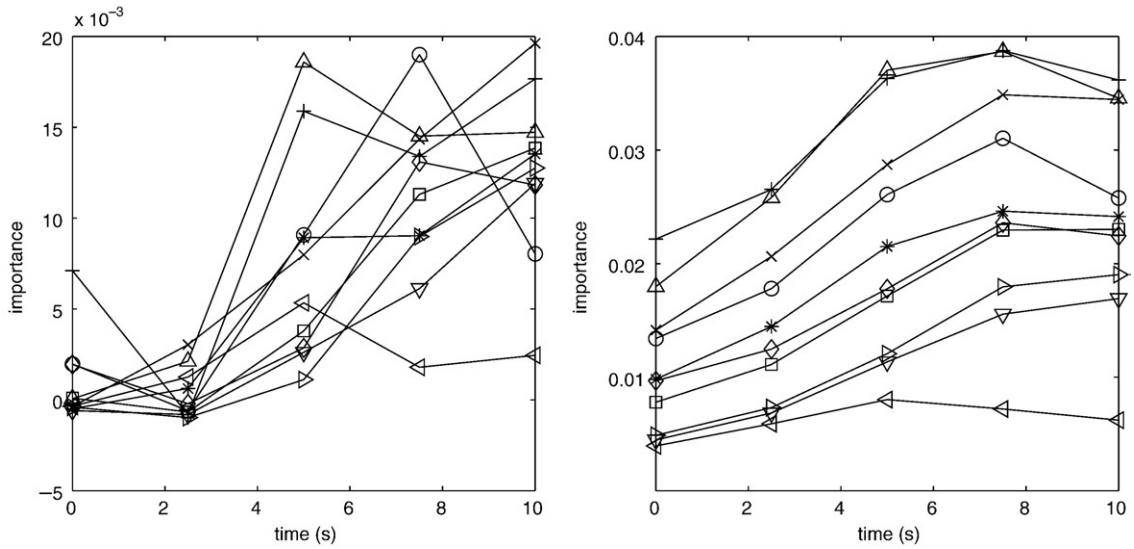

Fig. 9. Importance values for ten voxels whose BOLD response was acquired over five consecutive volumes using the decoupled model (left) and the temporal model (right).

Fig. 9 depicts the importance values for five consecutive volumes for ten voxels that were considered to be most important by the spatial model. The temporal model leads to temporal smoothing of the importance values. As a result, it becomes clear that the last few volumes carry more task-related information, which is in agreement with the (lagged) BOLD response.

Complexity analysis

The previous analysis has been performed for a fixed number of voxels. Here, we examine how the approximate inference algorithm scales with an increasing number of K voxels. With a small number of trials and a large number of features, the bottleneck of our algorithm is the inversion of the $K \times K$ precision matrix $\mathbf{K}$ when converting from canonical to moment form before each parallel EP update. As explained in the Appendix A, this inversion is obtained through the Takahashi equation (Takahashi et al., 1973; Erisman and Tinney, 1975), which operates on the lower-triangular matrix $\mathbf{L}$ resulting from a Cholesky decomposition of $\mathbf{K}$.

Although application of the Takahashi equation has worst-case complexity K^3, in practice, the computational complexity as well as the required storage depends on the resulting number of non-zero elements in $\mathbf{L}$. Typically $\mathbf{L}$ contains more non-zero elements than the sparse $\mathbf{K}$, because the computation of $\mathbf{L}$ introduces fill-in zeros. However, by reordering the rows and columns of a matrix, it is often possible to reduce the amount of fill-in created by factorisation, thereby reducing computation time and storage cost. The approximate minimum degree ordering (AMD) algorithm (Amestoy et al., 2004) seems to be a good general purpose method. Fig. 10 shows the precision matrix for a $4 \times 4 \times 4 \times 4$ volume with different priors and the corresponding matrix $\mathbf{L}$ after applying a Cholesky decomposition and a reordering based on the AMD algorithm. Note that the relative number of non-zero elements increases as more complex priors are used.

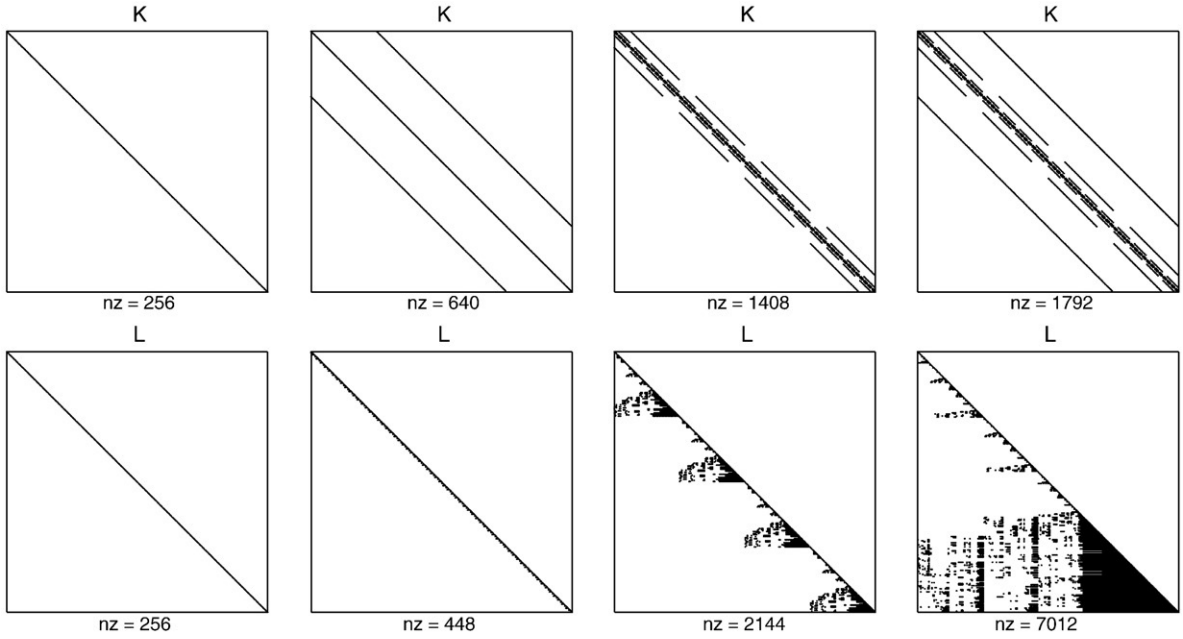

Fig. 10. Sparsity pattern of precision matrices (top row) and lower-triangular matrices $\mathbf{L}$ (bottom row) after a Cholesky decomposition and reordering using the AMD algorithm for a $4 \times 4 \times 4 \times 4$ volume. From left to right, a sparseprior, temporal prior, spatial prior, and spatio-temporal prior were used. The number of non-zero elements is shown below each matrix.

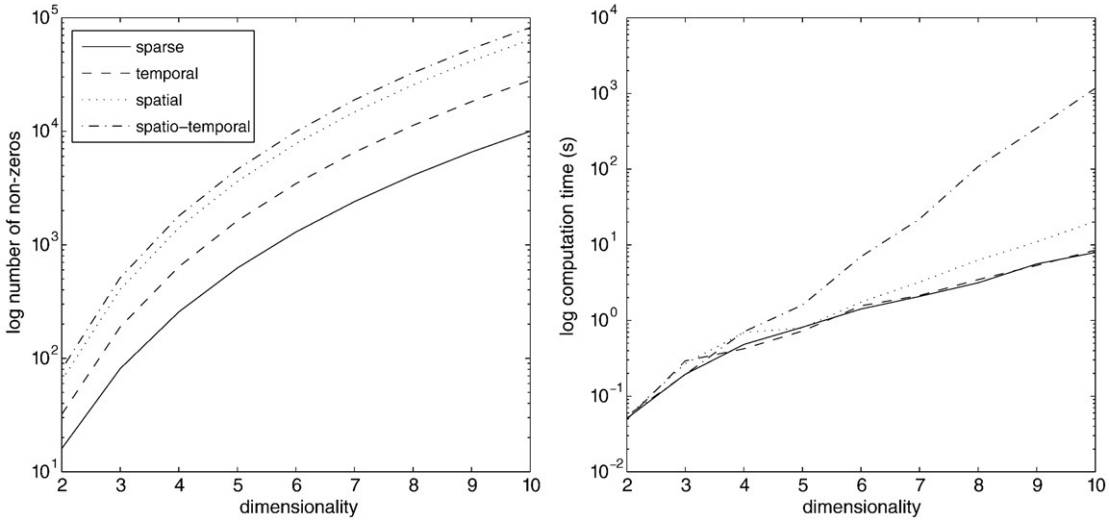

Fig. 11. Number of non-zero elements in **K** and computation time of the EP algorithm as a function of the dimensionality M of the volume when using a spatial prior (64 bit Intel Xeon CPU, 2.83 GHz, 16 GB internal memory).

In order to get an empirical estimate of computation time, we ran the EP algorithm on $M\times M\times M\times M$ volumes with M ranging from one to ten. We used forty trials per condition and filled the volumes with random voxels from the original fMRI dataset. Fig. 11 shows the number of non-zeros and computation time for the different priors as we vary volume dimensions. Even though we have an exponential increase in the number of non-zero elements and computation time with dimensionality, we are still able to handle very large models with the less complex priors. For the spatio-temporal prior, estimating a $10\times 10\times 10\times 10$ model took 20 min. In contrast, for the spatial prior, estimating the same model took just 20 s. Note that it is not the number of non-zero elements in **K** per se but rather the reduced sparsity structure of **L**, due to the disproportionate increase in the number of non-zero elements in **L**, which increases computation time for the more complex priors.

Discussion

In this paper, we introduced a Bayesian approach to multivariate fMRI analysis. The novelty of the approach lies in our formulation of the multivariate Laplace distribution in terms of a scale mixture. This representation can be used to specify spatio-temporal constraints on the magnitudes of the regression coefficients, which leads to smoothed importance maps that are therefore easier to interpret. Furthermore, we have demonstrated that good predictive performance is obtained with our method. Predictive performance was good for all regularised models, independent of the spatio-temporal coupling chosen. Fig. 3 already indicated that for strong regularisation, the model becomes relatively insensitive to the coupling strength. Still, from the point of view of interpretability, particular spatio-temporal priors will be preferred.

Our use of the multivariate Laplace distribution was motivated by the common use of the univariate Laplace distribution to obtain sparse solutions (Carroll et al., 2009; van Gerven et al., 2009). As mentioned, in a Bayesian setting, we lose these sparsifying properties since the posterior marginals will always have non-zero variance. In our opinion, this is more fair since the Laplace distribution can never warrant the redundancy of particular regressors given a finite amount of data. Still, in order to facilitate interpretability it could be of use to focus on sparse solutions where most regression coefficients are exactly equal to zero. We remark that such sparse behaviour can be approximated in our framework by replacing the posterior with $p(\beta|\mathbf{D},\theta)^{1/T}$. In the limit, as $T\rightarrow 0$, we recover the MAP estimate since the posterior becomes more strongly peaked around the mode of the posterior. An alternative to using the multivariate Laplace would be to use Bayesian models with true sparsifying properties such as the spike and slab prior (Mitchell and Beauchamp, 1988). However, as yet, it is computationally too demanding to apply these models to large-scale problems.

The multivariate Laplace distribution allows the specification of spatio-temporal interactions between variables by coupling the magnitudes of the regression coefficients through the auxiliary variables. In this paper, for the purpose of demonstration, we used a constant coupling between neighbouring elements in space or time. The prior could be more informed by taking into account the shape of the haemodynamic response function as well as anatomical constraints that can be derived from structural MRI. With minor modifications, our approach will also be suitable for multivariate analysis on the group level, where a coupling can be induced between voxels in different subjects. We leave these extensions as a topic for further research.

Our representation of the multivariate Laplace distribution can be contrasted with that of (Eltoft et al., 2006). As mentioned, in their representation, the regression coefficients are coupled not only in terms of the magnitude, but also in terms of the signs. This representation could be used to embody the prior knowledge that responses are spatially contiguous and locally homogeneous, as in (Penny et al., 2005). However, in our experience, regression coefficients that are close-by in space and/or time are typically correlated in terms of their magnitude but not necessarily in terms of their signs (cf. Fig. 7). A possible explanation, put forward in (van Gerven et al., 2009), is that these differences in signs could act as local filters, making the classifier more sensitive to relative instead of absolute changes in activity. Such behaviour could occur when multiple voxels are affected in the same way by some noise component (e.g., movement artifacts).

An important advantage of multivariate methods over mass-univariate methods in general is that weighted combinations of voxel activations are used to predict experimental condition. This increases sensitivity and allows the detection of patterns that are undetectable by mass-univariate methods (Norman et al., 2006). Furthermore, multivariate methods can be used to instantaneously predict experimental condition from voxel activations, which has applications in real-time analysis, as required for instance in brain–computer interfacing (Wolpaw et al., 2002). The specific merits of our approach are that we have derived an efficient and fully Bayesian approach to multivariate fMRI analysis where the employed prior has sparsifying properties and allows for the specification of spatio-temporal interactions that are known to occur throughout the brain proper.

Acknowledgments

The authors gratefully acknowledge the support of the Dutch technology foundation STW (project number 07050), the Netherlands Organization for Scientific Research NWO (Vici grant 639.023.604) and the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science.

Appendix A

Expectation propagation for Gaussian approximations

In this section we present the details of the expectation propagation method for Gaussian approximations. We show that when the terms t_i depend on a linear transformation of the variables, then both the representation of the term approximations $\tilde{t}_i$ as well as their updating simplifies significantly. We keep the presentation fairly general and specialise it to our application in the following section. We assume that the distribution p has the form

$$p(\mathbf{z}) = p_0(\mathbf{z}) \prod_j t_j\left(\mathbf{U}_j \mathbf{z}\right)$$

where $\mathbf{z} = (\boldsymbol{\beta}^T, \mathbf{u}^T, \mathbf{v}^T)^T$, $\mathbf{U}_j$ is a linear transformation, p_0 stands for the Gaussian priors on $\mathbf{u}$ and $\mathbf{v}$, which do not have to be approximated, and t_j terms denote the non-Gaussian factors. This form includes both the representation when $\mathbf{U}_i\mathbf{z} = \mathbf{x}_i^T\boldsymbol{\beta}$ for the data terms, as well as the representation when t_i depends only on a subset of parameters, that is, $\mathbf{U}_i\mathbf{z} = (\beta_i, u_i, v_i)^T$ for the auxiliary variable terms.

Term updates

The updates $\tilde{t}_i^*$ of Eq. (9) can be computed by $\tilde{t}_i^* \propto q^*/q^{\backslash i}$, where q^* is the Gaussian distribution that has the same first two moments as $t_i q^{\backslash i}$. Let us use the notations $q^*(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{m}^*, \mathbf{C}^*)$, $q^{\backslash i}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{m}^{\backslash i}, \mathbf{C}^{\backslash i})$, $\hat{\mathbf{m}}_i \equiv \mathbf{U}_i\mathbf{m}^{\backslash i}$ and $\hat{\mathbf{C}}_i \equiv \mathbf{U}_i\mathbf{C}^{\backslash i}\mathbf{U}_i^T$. After some calculus one finds that q^* is defined by

$$\begin{aligned} \mathbf{m}^* &= \mathbf{m}^{\backslash i} + \mathbf{C}^{\backslash i}\mathbf{U}_i^T \hat{\mathbf{C}}^{-1}(\mathrm{E}[\mathbf{s}_i] - \hat{\mathbf{m}}_i) \\ \mathbf{C}^* &= \mathbf{C}^{\backslash i} + \mathbf{C}^{\backslash i}\mathbf{U}_i^T \hat{\mathbf{C}}_i^{-1}\left(\mathrm{Var}[\mathbf{s}_i] - \hat{\mathbf{C}}_i\right)\hat{\mathbf{C}}_i^{-1}\mathbf{U}_i\mathbf{C}^{\backslash i} \end{aligned}$$

where $\mathbf{s}_i \sim t_i^{\alpha}(\mathbf{s}_i)\mathcal{N}(\mathbf{s}_i|\hat{\mathbf{m}}_i, \hat{\mathbf{C}}_i)$. Dividing $\mathcal{N}(\mathbf{z}|\mathbf{m}^*, \mathbf{C}^*)$ by $\mathcal{N}(\mathbf{z}|\mathbf{m}^{\backslash i}, \mathbf{C}^{\backslash i})$ we get

$$\begin{aligned} \left(\mathbf{C}^*\right)^{-1} - \left(\mathbf{C}^{\backslash i}\right)^{-1} &= \mathbf{U}_i^T\left(\mathrm{Var}[\mathbf{s}_i]^{-1} - \hat{\mathbf{C}}_i^{-1}\right)\mathbf{U}_i \\ \left(\mathbf{C}^*\right)^{-1}\mathbf{m}^* - \left(\mathbf{C}^{\backslash i}\right)^{-1}\mathbf{m}^{\backslash i} &= \mathbf{U}_i^T\left(\mathrm{Var}[\mathbf{s}_i]^{-1}\mathrm{E}[\mathbf{s}_i] - \hat{\mathbf{C}}_i\hat{\mathbf{m}}_i\right) \end{aligned} \tag{11}$$

leading to a low rank representation of $\tilde{t}_i$ with the form

$$\tilde{t}_i(\mathbf{z}) = \exp\left((\mathbf{U}_i\mathbf{z})^T\mathbf{h}_i - \frac{1}{2}(\mathbf{U}_i\mathbf{z})^T\mathbf{K}_i(\mathbf{U}_i\mathbf{z})\right)$$

where $\mathbf{h}_i$ and $\mathbf{K}_i$ are given by the quantities in Eq. (11) (divided by α when using power EP). Based on this representation, we can define the approximating distribution q as a Gaussian with canonical parameters

$$\begin{aligned} \mathbf{K} &= \mathbf{K}_0 + \textstyle\sum_i \mathbf{U}_i^T\mathbf{K}_i\mathbf{U}_i \\ \mathbf{h} &= \mathbf{h}_0 + \textstyle\sum_i \mathbf{U}_i^T\mathbf{h}_i, \end{aligned}$$

that is, the sum of the parameters of $\tilde{t}_i$ added to the parameters $\mathbf{h}_0$ and $\mathbf{K}_0$ of p_0. Now that we know the form of $\tilde{t}_i$, we can proceed to find a way to compute the quantities $\hat{\mathbf{C}}_i$ and $\hat{\mathbf{m}}_i$ needed for the updates.

Computing $q^{\backslash i}$

One can compute the mean and covariance of $q^{\backslash i}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{m}^{\backslash i}, \mathbf{C}^{\backslash i})$ from $q^{\backslash i} = \tilde{t}_i^{(1-\alpha)}\Pi_{\backslash i}\tilde{t}_j$. After some calculus, one can show that

$$\hat{\mathbf{C}}_i = \hat{\mathbf{K}}_i\left(\mathrm{I} - \alpha\mathbf{K}_i\hat{\mathbf{K}}_i\right)^{-1} \tag{12}$$

$$\hat{\mathbf{m}}_i = \left(\mathrm{I} - \alpha\mathbf{K}_i\hat{\mathbf{K}}_i\right)^{-1}\left(\hat{\mathbf{h}}_i - \alpha\hat{\mathbf{K}}_i\mathbf{h}_i\right) \tag{13}$$

using shorthand notation $\hat{\mathbf{K}}_i \equiv (\mathbf{U}_i\mathbf{K}^{-1}\mathbf{U}_i^T)$ and $\hat{\mathbf{h}}_i \equiv \mathbf{U}_i\mathbf{K}^{-1}\mathbf{h}$. In order to compute the update of $\tilde{t}_i$ given in Eq. (11) one needs to compute $\hat{\mathbf{K}}_i$ and $\mathbf{h}_i$. The computational bottleneck of EP reduces to the computation of these quantities since they are typically more expensive than the low-dimensional (numerical) computations of $E[\mathbf{s}_i]$ and $Var[\mathbf{s}_i]$.

EP applied to Bayesian logistic regression

In this section, we will apply the results of the previous section to the Bayesian logistic regression model in Eq. (8). I.e., we specialise it to the data terms and auxiliary variable terms. For both types of terms, we will (1) identify $\mathbf{U}_i$ and $\mathbf{s}_i$ and derive the form of the term $\tilde{t}_i$, (2) compute the quantities $E[\mathbf{s}_i]$ and $Var[\mathbf{s}_i]$ from Eq. (11) and finally (3) compute the quantities in Eq. (12) and Eq. (13).

Data term approximations

The data terms $g_n(y_n, \mathbf{x}_n, \mathcal{B}) = \mathcal{B}(y_n, l^{-1}(\mathbf{x}_n^T\boldsymbol{\beta}))$ arise from the logistic regression likelihood terms in Eq. (8). This means that $\mathbf{U}_n = \mathbf{x}_n^T\mathbf{G}^T$, where we define $\mathbf{G} = [\mathbf{I}_K, \mathbf{0}_K, \mathbf{0}_K]^T$ as the $3K \times K$ matrix that selects the variables $\boldsymbol{\beta}$ in $\mathbf{z}$. Therefore, the term approximations $\tilde{g}_n$ are characterised by two scalar parameters h_β^n and K_β^n and can be written as

$$\tilde{g}_n(\mathbf{z}) = \exp\left(\mathbf{z}^T\mathbf{G}\mathbf{x}_n h_\beta^n - \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{x}_n K_\beta^n \mathbf{x}_n^T\mathbf{G}^T\mathbf{z}\right)$$

such that

$$\prod_{n=1}^{N} \tilde{g}_n(\mathbf{z}) = \exp\left(\mathbf{z}^T\mathbf{G}\mathbf{X}^T\mathbf{h}_\beta^g - \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{X}^T\mathbf{K}_\beta^g\mathbf{X}\mathbf{G}^T\mathbf{z}\right)$$

where $\mathbf{h}_\beta^g$ is a vector formed by h_β^n and $\mathbf{K}_\beta^g$ is a diagonal matrix formed by K_β^n. The product over the $\tilde{g}_n$s influences only the $\boldsymbol{\beta}$ part of $\mathbf{z}$. Moreover, it has a low rank representation characterised by the $N \times K$ design matrix $\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_N]^T$. The variables $\mathbf{s}_n$ are scalar variables and they are distributed according to the unnormalised distribution

$$s_n \sim \mathcal{B}\left(y_n; l^{-1}(s_n)\right)\mathcal{N}\left(s_n|\mathbf{x}_n^T\mathbf{G}^T\mathbf{m}^{\backslash n}, \mathbf{x}_n^T\mathbf{G}^T\mathbf{C}^{\backslash n}\mathbf{G}\mathbf{x}_n\right).$$

The quantities $E[s_n]$ and $\mathrm{Var}[s_n]$ can be computed numerically using the one-dimensional Gaussian–Hermite quadrature rule.

Auxiliary variable term approximations

The auxiliary variable terms $f_k(\beta_k, u_k, v_k) \equiv \mathcal{N}(\beta_k; 0, u_k^2 + v_k^2)$ arise from the scale mixture representation of the Laplace distribution that couples the variables β_k, u_k and v_k. In this case, we identify $\mathbf{U}_k$ with the $3 \times 3K$ projection matrix $\mathbf{F}_k$ that selects these variables from $\mathbf{z}$. As a result, the term approximations $\tilde{f}_k$ are characterised by three-dimensional vectors $\mathbf{h}_k^f$ and matrices $\mathbf{K}_k^f$, and can be written as

$$\tilde{f}_k(\mathbf{z}) = \exp\left(\mathbf{z}^T\mathbf{F}_k^T\mathbf{h}_k^f - \frac{1}{2}\mathbf{z}^T\mathbf{F}_k^T\mathbf{K}_k^F\mathbf{F}_k\mathbf{z}\right).$$

Since the only coupling between the β_k and u_k, v_k is in f_k and since u_k and v_k are a priori independent and identically distributed with zero mean, we assume that $\mathbf{F}_k\mathbf{C}^{\setminus k}\mathbf{F}_k^T$ is diagonal and the last two components of $\mathbf{F}_k\mathbf{m}^{\setminus k}$ are zero. This assumption is valid for the prior and it remains valid after each EP update. With a slight abuse in notation, we identify $\mathbf{s}_k = (\beta_k,\ u_k,\ v_k)$ with the random variable distributed as

$$\mathbf{s}_k \sim \mathcal{N}(\beta_k|0,U_k)^{\alpha}\mathcal{N}\left(\beta_k|m_k,\sigma_k^2\right)\mathcal{N}\left(u_k|0,\gamma_k^2\right)\mathcal{N}\left(v_k|0,\gamma_k^2\right) \tag{14}$$

where we use shorthand notation $U_k \equiv u_k^2 + v_k^2$. The quantities m_k, σ_k^2 and γ_k^2 denote the marginal means and variances of $q_f^{\setminus i}(\boldsymbol{\beta}, \mathbf{u}, \mathbf{v}) \propto q(\boldsymbol{\beta}, \mathbf{u}, \mathbf{v})/\tilde{f}_k(\beta_k, u_k, v_k)$. Now we can compute $\mathrm{E}[\mathbf{s}_k]$ and $Var[\mathbf{s}_k]$ and show that, due to the symmetry of $u_k^2 + v_k^2$, Eq. (14) keeps β_k, u_k and v_k uncoupled in $\tilde{f}_k$. After regrouping the terms in (14), one finds that

$$\beta_k|u_k,v_k \sim \mathcal{N}\left(\beta_k|\frac{m_k U_k}{\alpha\sigma_k^2 + U_k}, \frac{\sigma_k^2 U_k}{\alpha\sigma_k^2 + U_k}\right) \tag{15}$$

$$\begin{aligned}(u_k,v_k) \sim U_k^{(1-\alpha)/2}\mathcal{N}\left(\sqrt{\alpha}m_k|0,\alpha\sigma_k^2 + U_k\right) \\ \times \mathcal{N}\left(u_k|0,\gamma_k^2\right)\mathcal{N}\left(v_k|0,\gamma_k^2\right).\end{aligned} \tag{16}$$

This implies that the marginal distribution of $(u_k,\ v_k)^T$ given by Eq. (16) depends only on $u_k^2 + v_k^2$ and is symmetric around 0. Therefore, $E[u_k] = \mathrm{E}[v_k] = 0, E[u_k v_k] = 0$ and $\mathrm{Var}[u_k] = \mathrm{Var}[v_k]$. The mean value $E[\beta_k]$ and the variance $\mathrm{Var}[\beta_k]$ can be computed by averaging the mean $\mathrm{E}[\beta_k|u_k, v_k]$ and variance $\mathrm{Var}[\beta_k|u_k, v_k]$ parameters in Eq. (15) over the joint distribution of u_k and v_k given in Eq. (16). Using the symmetry of Eq. (16), one can show that $E[\beta_k u_k] = E[\beta_k v_k] = 0$ results in a diagonal $\mathbf{K}_k^f$ and an $\mathbf{h}_k^f$ with its last two components set to zero (the components corresponding to u_k and v_k). This is a substantial computational advantage because it keeps $\mathbf{K}$ block diagonal. Since both Eqs. (15) and (16) depend on u_k and v_k only through U_k we can once more make use of the exponential distribution and approximate the required integrals using the one-dimensional Gauss–Laguerre quadrature rule.

Computing $q_g^{\setminus n}$ and $q_f^{\setminus k}$

Based on the results presented in the previous two subsections, the approximating distribution q has a canonical form $q(\mathbf{z}) \propto \exp(\mathbf{h}^T\mathbf{z} - \mathbf{z}^T\mathbf{K}\mathbf{z}/2)$ characterised by

$$\mathbf{h} = \begin{bmatrix} \mathbf{X}^T\mathbf{h}_\beta^g + \mathbf{h}_\beta^f \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{K} = \begin{bmatrix} \mathbf{X}^T\mathbf{K}_\beta^g\mathbf{X} + \mathbf{K}_\beta^f & 0_K & 0_K \\ 0_K & \Theta^{-1} + \mathbf{K}_u^f & 0_K \\ 0_K & 0_K & \Theta^{-1} + \mathbf{K}_v^f \end{bmatrix}$$

where $\mathbf{h}_\beta^f$ is the vector formed by the components of $\mathbf{h}_k^f$, $k = 1, \ldots, K$ corresponding to β_k, $k = 1, \ldots, K$ and $\mathbf{K}_\beta^f$, $\mathbf{K}_u^f$ and $\mathbf{K}_v^f$ are the diagonal matrices formed by the corresponding components of $\mathbf{K}_k^f$, $k = 1, \ldots K$. Since $\mathbf{K}$ is block diagonal, in order to compute the quantities in Eq. (12) and Eq. (13), one has to compute the following:

1. For the cavity distribution $q_g^{\setminus n}$ one needs to compute both $\mathbf{x}_n^T\mathbf{C}_\beta^{\setminus n}\mathbf{x}_n^T = \mathbf{x}_n^T(\mathbf{X}^T\mathbf{K}_\beta^g\mathbf{X} + \mathbf{K}_\beta^f)^{-1}\mathbf{x}_n$ and $\mathbf{x}_n^T\mathbf{m}_\beta^{\setminus n} = \mathbf{x}_n^T(\mathbf{X}^T\mathbf{K}_\beta^g\mathbf{X} + \mathbf{K}_\beta^f)^{-1}(\mathbf{X}^T\mathbf{h}_\beta^g + \mathbf{h}_\beta^f)$ for $n = 1, \ldots, N$. This can be done efficiently using the matrix inversion lemma, replacing the inversion of a $K \times K$ matrix with that of an $N \times N$ matrix. This is useful in the degenerate case when the number of features is much larger than the number of samples, as is typically the case for neuroimaging data.
2. For the cavity distribution $q_f^{\setminus k}$ one needs to find the diagonal elements of $(\mathbf{X}^T\mathbf{K}_\beta^g\mathbf{X} + \mathbf{K}_\beta^f)^{-1}$, $(\Theta^{-1} + \mathbf{K}_u^f)^{-1}$ and $(\Theta^{-1} + \mathbf{K}_v^f)^{-1}$. None of these operations require the full inversion of the corresponding matrices as they reduce to (1) computing the Cholesky factorisation of a dense $N \times N$ matrix and its multiplication with an $N \times K$ matrix (matrix inversion lemma), and (2) computing certain elements of the inverse of a sparse $K \times K$ matrix. The latter can be realized efficiently via the Takahashi equation (Takahashi et al., 1973; Erisman and Tinney, 1975).

Computing the model evidence

The model evidence

$$p(\mathbf{D}|\Theta) = \int \mathrm{d}\mathbf{z}\,\mathcal{N}(\mathbf{u};\mathbf{0},\Theta)\mathcal{N}(\mathbf{v};\mathbf{0},\Theta)\prod_k f_k(\mathbf{z}_k)\prod_n g_n(y_n,\mathbf{x}_n,\boldsymbol{\beta})$$

with $f_k(\beta_k,\ u_k,\ v_k) \equiv \mathcal{N}(\beta_k;\ 0,\ u_k^2 + v_k^2)$ and $g_n(y_n,\ \mathbf{x}_n,\ \boldsymbol{\beta}) \equiv \mathcal{B}(y_n;\ l^{-1}(\mathbf{x}_n^T\boldsymbol{\beta}))$ as before, may be approximated within the EP framework by plugging in the normalised term approximations $c_k\tilde{f}_k$ and $c'_n\tilde{g}_n$ (Seeger, 2008):

$$p(\mathbf{D}|\Theta) \approx \prod_k c_k \prod_n c'_n \int \mathrm{d}\mathbf{z}\,\mathcal{N}(\mathbf{u};\mathbf{0},\Theta)\mathcal{N}(\mathbf{v};\mathbf{0},\Theta)\prod_k \tilde{f}_k(\mathbf{z}_k)\prod_n \tilde{g}_n(y_n,\mathbf{x}_n,\boldsymbol{\beta}).$$

The quantities c_k and c'_n are given by replacing t with the corresponding terms f_k and g_n in

$$c = \left(\frac{Z}{\tilde{Z}}\right)^{\frac{1}{\alpha}} = \left(\frac{\int \mathrm{d}\mathbf{z}\, t(\mathbf{z})^\alpha q^{\setminus i}(\mathbf{z})}{\int \mathrm{d}\mathbf{z}\,\tilde{t}(\mathbf{z})^\alpha q^{\setminus i}(\mathbf{z})}\right)^{\frac{1}{\alpha}}.$$

The integrals are relatively easy to approximate since they boil down to one-dimensional numerical quadratures due to the forms of g_n and f_k. Let $\log\Phi_{\mathcal{C}}(\mathbf{h},\mathbf{K}) = -\log|2\pi\mathbf{K}|/2 + \mathbf{h}^T\mathbf{K}^{-1}\mathbf{h}/2$ denote the log partition function of a multivariate Gaussian distribution in canonical form. We then obtain the following approximation to the log model evidence:

$$\log p(\mathbf{D}|\Theta) \approx \sum_k \log c_k + \sum_n \log c'_n + \log\Phi_{\mathcal{C}}(\mathbf{h},\mathbf{K}) - 2\log\Phi_{\mathcal{C}}\left(0,\Theta^{-1}\right).$$

The log partition function $\log\Phi_{\mathcal{C}}(\mathbf{h},\mathbf{K})$ is computed in an efficient manner by making use of the matrix determinant lemma.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2009.11.064.

References

Amestoy, P.R., Davis, T.A., Duff, I.S., 2004. Algorithm 837: Amd, an approximate minimum degree ordering algorithm. ACM T. Math. Software 30 (3), 381–388.

Andrews, D.F., Mallows, C.L., 1974. Scale mixtures of normal distributions. J. Roy. Statistical Society Series B 36 (1), 99–102.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis 2nd Ed. Springer.

Bernardo, J.M., Smith, J.F.M., 1994. Bayesian Theory. Wiley.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning 1st Ed. Springer, New York, NY.

Brainard, D.H., 1997. The psychophysics toolbox. Spatial Vision 10, 433–436.

Brezger, A., Fahrmeir, L., Hennerfeind, A., 2007. Adaptive Gaussian Markov random fields with applications in human brain mapping. Appl. Statist. 56 (3), 327–345.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R., 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44 (1), 112–122.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270.

DeCharms, R.C., 2008. Applications of real-time fMRI. Nat. Rev. Neurosci. 9, 720–729.
Eltoft, T., Kim, T., Lee, T., 2006. On the multivariate Laplace distribution. IEEE Signal Proc. Let. 13 (5), 300–303.
Erisman, A.M., Tinney, W.F., 1975. On computing certain elements of the inverse of a sparse matrix. Commun. ACM 18 (3), 177–179.
Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.
Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., Penny, W.D. (Eds.), 2006. Statistical Parametric Mapping: The Analysis of Functional Brain Images, 1st Ed. Academic Press, London, UK.
Friston, K., Chu, C., Mourão Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. NeuroImage 39, 181–205.
Gössl, C., Auer, D.P., Fahrmeir, L., 2001. Bayesian spatiotemporal inference in functional magnetic resonance imaging. Biometrics 57, 554–562.
Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8 (5), 679–685.
Kuss, M., Rasmussen, C.E., 2005. Assessing approximate inference for binary Gaussian process classification. JMLR 6, 1679–1704.
Longford, N.T., 1990. Classes of multivariate exponential and multivariate geometric distributions derived from Markov processes. In: Block, H.W., Sampson, A.R., Savits, T.H. (Eds.), Topics in statistical dependence. Vol. 16 of IMS Lecture Notes Monograph Series. IMS Business Office, Hayward, CA, pp. 359–369.
Lyu, S., Simoncelli, E.P., 2007. Statistical modeling of images with fields of Gaussian scale mixtures. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), Advances in Neural Information Processing Systems, 19. MIT Press, Cambridge, MA, pp. 945–952.
MacKay, D.J.C., 2004. Information Theory, Inference and Learning Algorithms3rd Ed. Cambridge Univ. Press, Cambridge, UK.
Minka, T., 2001a. Expectation propagation for approximate Bayesian inference. In: Breese, J., Koller, D. (Eds.), Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, pp. 362–369.
Minka, T., 2001b. A family of algorithms for approximate Bayesian inference. PhD thesis, MIT.
Minka, T., 2004. Power EP. Tech. rep., Microsoft Research, Cambridge.
Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression (with discussion). J. Amer. Statist. Assoc. 83, 1023–1036.
Mitchell, T.M., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to decode cognitive states from brain images. Mach. Learn. 57 (1–2), 145–175.
Norman, K.A., Polyn, S.M., Detra, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10 (9), 424–430.
Penny, W., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24, 350–362.
Pereira, F., Mitchell, T.M., Botvinick, M., 2008. Machine learning classiffiers and fMRI: a tutorial overview. NeuroImage 45 (1), S199–S209.
Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2007. Numerical Recipes in C3rd Ed. Cambridge University Press.
Rue, H., Held, L., 2005. Gaussian Markov random fields: theory and applications, Monographs on Statistics and Applied Probability1st Ed. Chapman and Hall/CRC, Boca Raton, FL.
Rue, H., Martino, S., 2007. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. J. Stat. Plan. Inference 137 (10), 3177–3192.
Salzberg, S.L., 1997. On comparing classiffiers: pitfalls to avoid and a recommended approach. Data Min. Knowl. Disc. 1, 317–327.
Seeger, M.W., 2008. Bayesian inference and optimal design for the sparse linear model. JMLR 9, 759–813.
Takahashi, K., Fagan, J., Chen, M.S., 1973. Formation of a sparse bus-impedance matrix and its application to short circuit study. 8th IEEE PICA Conference. Minneapolis, MN, pp. 63–69.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. Roy. Statistical Society Series B 58, 267–288.
van Gerven, M.A.J., Hesse, C., Jensen, O., Heskes, T., 2009. Interpreting single trial data using groupwise regularisation. NeuroImage 46, 665–676.
Williams, P.M., 1995. Bayesian regularisation and pruning using a Laplace prior. Neural Comput. 7 (1), 117–143.
Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M., 2002. Brain–computer interfaces for communication and control. Clin. Neurophysiol. 113, 767–791.
Woolrich, M.W., Behrens, T.E., Smith, S.M., 2004. Constrained linear basis sets for HRF modelling using variational Bayes. NeuroImage 21, 1748–1761.
Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of fMRI data. NeuroImage 14 (6), 1370–1386.