ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 21 (2004) 1732–1747

Multilevel linear modelling for FMRI group analysis using Bayesian inference

Mark W. Woolrich,[a,b,*,1] Timothy E.J. Behrens,[a,b,1] Christian F. Beckmann,[a,b] Mark Jenkinson,[a] and Stephen M. Smith[a]

[a] *Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK*
[b] *Department of Engineering Science, University of Oxford, Oxford, UK*

Received 4 September 2003; revised 6 December 2003; accepted 9 December 2003

Functional magnetic resonance imaging studies often involve the acquisition of data from multiple sessions and/or multiple subjects. A hierarchical approach can be taken to modelling such data with a general linear model (GLM) at each level of the hierarchy introducing different random effects variance components. Inferring on these models is nontrivial with frequentist solutions being unavailable. A solution is to use a Bayesian framework. One important ingredient in this is the choice of prior on the variance components and top-level regression parameters. Due to the typically small numbers of sessions or subjects in neuroimaging, the choice of prior is critical. To alleviate this problem, we introduce to neuroimage modelling the approach of reference priors, which drives the choice of prior such that it is noninformative in an information-theoretic sense. We propose two inference techniques at the top level for multilevel hierarchies (a fast approach and a slower more accurate approach). We also demonstrate that we can infer on the top level of multilevel hierarchies by inferring on the levels of the hierarchy separately and passing summary statistics of a noncentral multivariate *t* distribution between them.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Mixed effects; Random effects; FMRI; Bayes; Reference priors; GLM

Introduction

Functional magnetic resonance imaging studies are typically used to address questions about activation effects in populations of subjects. This generally involves a multisubject and/or multisession approach where data are analysed in such a way as to allow for hypothesis tests at the group level (Holmes and Friston, 1998; Worsley et al., 2002), for example, to assess whether the observed effects are common and stable across or between groups of interest.

Calculating the level and probability of brain activation for a single subject is typically achieved using a linear model of the signal together with a Gaussian noise model for the residuals. This model is commonly called the general linear model (GLM), and much attention to date has been focussed on ways of modelling and fitting the (time series) signal and residual noise at the individual single-session level (Bullmore et al., 1996; Woolrich et al., 2001; Worsley and Friston, 1995).

To be able to generate results that extend to the population, we also need to account for the fact that the individual subjects themselves are sampled from the population and thus are random quantities with associated variances. It is exactly this step that marks the transition from a simple fixed-effects model to a mixed-effects model,[2] and it is imperative to formulate a model at the group level that allows for the explicit modelling of these additional variance terms (Holmes and Friston, 1998; Frison and Pocock, 1992).

We can formulate the problem of group statistics in neuroimaging as being hierarchical (Beckmann et al., 2003; Friston et al., 2002). For example, the different levels of the hierarchy could be separate GLMs for a session level, subject level and group level. In this paper, we attempt to deal with inference on these multilevel GLM hierarchies by utilising a fully Bayesian framework. Typically, the most important inference is at the top level of the hierarchy, for example, we may be looking for significance of a group mean. Whether we are looking to infer at the top level with the within-session FMRI time series data (Friston et al., 2002) or with summary statistic results from the level below (Holmes and Friston, 1998; Worsley et al., 2002), a fully Bayesian approach provides us with the means to assess the full uncertainty in the parameter of interest (contrasts of regression parameters) at the top level; taking into account all of the unknown variance components (fixed and random) in the model.

Bayesian statistics provides the only generic tool for inferring model parameter probability distribution functions from data. It provides strict rules for the rational and consistent adjustment of belief (in the form of probability density functions) in the presence

* Corresponding author. Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK. Fax: +44-1865-222717.
E-mail address: woolrich@fmrib.ox.ac.uk (M.W. Woolrich).
[1] These authors contributed equally to this work.
Available online on ScienceDirect (www.sciencedirect.com.)

[2] Note that in the FMRI literature, this has often been called a random-effects model. Within this paper, however, the separate fixed-effects and random-effects contributions to the mixed-effects variance are considered, thus making a clear distinction between "random" and "mixed effects" important.

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2003.12.023

of new information (Cox, 1946), which are not available in the frequentist literature. The major consequences of this are twofold. First, we may make inference about the absolute value of the parameters of interest, that is, we may ask questions of our parameters such as, "What is the probability that θ lies in the interval $[\theta_0, \theta_1]$?", a question unavailable to any frequentist technique. Frequentist statistics is typically limited to posing questions of the data under the "null hypothesis" that the parameter value is zero. Inference in a frequentist framework is then limited to the simple acceptance or rejection of this null hypothesis without being able make any statement about the parameter values. Second, Bayesian statistics gives us a tool for inferring on any model we choose and guarantees that uncertainty will be handled correctly. Only in certain special cases (not including the model presented here) is it possible to derive analytical forms for the null distributions required by frequentist statistics. In their absence, frequentist solutions rely on null distributions derived from the data (e.g., permutation tests), losing the statistical power gained from educated assumptions about, for example, the distribution of the noise.

These features of Bayesian analysis mean that we may make inference on physiological parameters of the haemodynamic response in the complex nonlinear balloon model (Friston, 2002), on spatial noise relationships in multivariate spatial autoregressive models of FMRI data (Woolrich et al., in press) or, in this paper, on higher level statistics in the presence of multiple variance components.

One important ingredient in a Bayesian approach is the choice of prior on the variance components and top-level regression parameters. Due to the typically small numbers of observations in neuroimaging above the first level (e.g., small numbers of subjects), this choice of prior is critical. To solve this problem, we introduce to neuroimage modelling the approach of reference priors, which drives the choice of prior such that it is non-informative in an information-theoretic sense. For GLMs where a frequentist solution is available, reference analysis gives the same inference as a frequentist approach. Importantly, reference analysis allows us to perform inference when frequentist solutions are unavailable.

Using fully Bayesian reference analysis, we propose two approaches to inferring at the top level; these are a fast approximation to the marginal posterior and a slower approach utilising Markov Chain Monte Carlo (MCMC) followed by a multivariate noncentral t distribution fit to the MCMC chains.

In Friston et al. (2002), the hierarchical model is solved "all in one" using the within-session FMRI time series data as input. However, in neuroimaging, where the human and computational costs involved in data analysis are relatively high, it is desirable to be able to make top-level inferences using the results of separate lower level analyses without the need to reanalyse any of the lower level data; an approach commonly called the summary statistics approach to FMRI analysis (Holmes and Friston, 1998). Within such a summary statistic split-level approach, group parameters of interest can easily be refined as more data become available.

In Holmes and Friston (1998), when inferring at the top level, this summary statistic split-level approach is shown to be equivalent to inferring all in one under certain conditions (e.g., the approach in Holmes and Friston, 1998, requires balanced designs). Beckmann et al. (2003) show that top-level inference using the split-level summary statistics approach can be made equivalent to the all-in-one approach with no restrictions, if we pass up the correct summary statistics (in particular, the covariances from previous levels). Furthermore, Beckmann et al. (2003) demonstrate that by taking into account lower level covariance heterogeneity, a substantial increase in higher level z statistic is possible. However, Beckmann et al. (2003) only show that this is the case when all variance components are known. Independently, in this paper, using the fully Bayesian approach, we show this equivalence for when the variance components (excluding autocorrelation) are unknown. The equivalence relies on the assumption that the summary statistics, which correspond to the marginal distributions of the GLM regressions parameters, can be represented as a multivariate noncentral t distributions. Between the first level (within session) and the second level, this can be shown analytically. For summary statistics at higher levels, this is an assumption which we test empirically using artificial data.

In summary, there are three main contributions presented in this paper. Firstly, we introduce reference analysis to neuroimaging. Secondly, we propose two inference techniques at the top level for multilevel hierarchies (a fast approach and a slower more accurate approach). Thirdly, we demonstrate that we can infer on the top level of multilevel hierarchies by inferring on the split levels separately and passing summary statistics between them.

Paper overview

We start in the Model section by considering the traditional two-level model. In the Inference section, using the reference analysis fully Bayesian framework, we show how inference on the two-level model can be split into separate inference on the two levels with the summary statistics of a multivariate noncentral t distribution being passed between the two levels of inference. We then propose two approaches to inferring at the top level. In Higher level models, we discuss how we can extend the split model inference approach to higher level models than the two-level model. In Multiple group variances, we also discuss how we can deal with multiple group variances under certain conditions. In the Artificial data section, we validate the crucial assumption of the marginal distribution of the GLM regressions parameters being a multivariate noncentral t distribution at levels higher than the first using artificial data. Finally, in the FMRI data section, we go on to show results on FMRI data.

Model

To begin with, we consider the familiar two-level univariate GLM for FMRI. For example, the model that in the first level deals with individual sessions for individual subjects, relating time series to activation, and in the second level deals with a group of subjects or sessions (or both), relating the combined individual activation estimates to some group parameter, such as mean activation level. Note that all models and inference in this paper are mass univariate, that is, each voxel is modelled and processed independently of the others in the data.

Two-level GLM

Consider an experiment where there are N_K first-level sessions and that for each first-level session, k, the preprocessed FMRI data are a $T \times 1$ vector $\boldsymbol{Y}_k$, the $T \times P_K$ design matrix is $\mathbf{X}_k$ and $\boldsymbol{\beta}_k$ is a $P_K \times 1$ vector of parameter estimates ($k = 1,\ldots, N_K$). The

preprocessed FMRI data, $\boldsymbol{Y}_k$, is assumed to have been prewhitened (Bullmore et al., 1996; Woolrich et al., 2001). An individual GLM relates first-level parameters to the N_k individual data sets:

$$\boldsymbol{Y}_k = \mathbf{X}_k\boldsymbol{\beta}_k + e_k, \tag{1}$$

where $e_k \sim N(0,\sigma_k^2 I)$. In this paper, we consider the variance components as unknown with the exception of the first-level FMRI time series autocorrelation. The residuals e_k are assumed to be prewhitened data and as a result are uncorrelated. This inherently means that we assume that the autocorrelation is known with no uncertainty, an assumption that is commonly made in FMRI time series analysis (Bullmore et al., 1996; Friston et al., 2000; Woolrich et al., 2001). Note that the first-level design matrices, $\mathbf{X}_k$, do not need to be the same for all k.

Using the block diagonal forms, that is, with

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_{N_K} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2 & & 0 \\ \vdots & & O & \vdots \\ 0 & \cdots & 0 & \mathbf{X}_{N_K} \end{bmatrix},$$

$$\boldsymbol{\beta}_K = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_{N_K} \end{bmatrix} \text{ and } e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{N_K} \end{bmatrix}$$

the two-level model is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}_K + e_K \tag{2}$$

$$\boldsymbol{\beta}_K = \mathbf{X}_g\boldsymbol{\beta}_g + e_g \tag{3}$$

where $\mathbf{X}_g$ is the $N_K \times P_G$ second-level design matrix (e.g., separating controls from normals or modelling different sessions for subjects), $\boldsymbol{\beta}_g$ is the $P_G \times 1$ vector of second-level parameters, and $e_g \sim N(0, \sigma_g^2 I)$ and where $e_K \sim N(0, V_K)$ with V denoting the diagonal form of first-level covariance matrices $\sigma_k^2 I$. We call σ_g^2 the random effects variance.

Inference

There are no solutions in the frequentist literature to this model when the variance components are unknown. Furthermore, inference is highly sensitive to any assumptions made due to the low number of observations typically available at the subject level in FMRI.

Friston et al. (2002) have proposed an approximate Bayesian solution for the model all in one by assuming that the posterior over the regression parameters is multivariate normal. However, this does not fully incorporate the full uncertainty of the variance components into the parameters of interest (the regression parameters) at the top level. Indeed, the marginal posterior over the regression parameters turns out to be multivariate t distributed.

In this section, we start by introducing the Bayesian inference framework. However, when using a Bayesian framework, we also need to choose priors for the parameters in our model. In particular, we need to choose priors on the top-level regression and variance parameters. Hence, in the next part of this section, we describe how we can use reference priors as noninformative priors.

We could proceed to infer on the full model all in one. Instead, by using the fully Bayesian approach with reference priors, we go on to show how we can use summary statistics (from inferring on the first-level model in isolation) as the input into a second level. We show that this gives the same inference as we would obtain from using the full model all in one.

Bayesian inference

The two rules at the heart of Bayesian learning techniques are conceptually very simple. The first tells us how (for a model $\mathcal{M}$), we should use the data $\mathbf{Y}$ to update our prior belief in the values of the parameters Θ, $p(\Theta|\mathcal{M})$ to a posterior distribution of the parameter values $p(\Theta|\mathbf{Y}, \mathcal{M})$. This is known as Bayes' rule:

$$p(\Theta \mid \mathbf{Y}, \mathcal{M}) = \frac{p(\mathbf{Y} \mid \Theta, \mathcal{M})p(\Theta \mid \mathcal{M})}{p(\mathbf{Y} \mid \mathcal{M})} \tag{4}$$

Unfortunately, calculating this posterior pdf is seldom straightforward. The denominator in Eq. (4) is:

$$p(\mathbf{Y} \mid \mathcal{M}) = \int_{\Theta} p(\mathbf{Y} \mid \Theta, \mathcal{M})p(\Theta \mid \mathcal{M})d\Theta \tag{5}$$

an integral that is often not tractable analytically. To make matters worse, this joint posterior pdf on all parameters is often not the distribution, which we are most interested in. We are often interested in the posterior pdf on a single parameter or an interesting subset of parameters. Obtaining these marginal distributions again involves performing large integrals,

$$p(\Theta_I \mid \mathbf{Y}, \mathcal{M}) = \int_{\Theta_{-I}} p(\Theta \mid \mathbf{Y}, \mathcal{M})d\Theta_{-I} \tag{6}$$

where Θ_I are the parameters of interest and Θ_{-I} are all other parameters. Again, these integrals are seldom tractable analytically.

One solution is to use approximations to the marginal distributions. This is the approach we take in the Fast posterior approximation section. Another solution is to draw samples in parameter space from the joint posterior distribution, implicitly performing the integrals numerically. For example, we may repetitively choose random sets of parameter values and choose to accept or reject these samples according to a criterion based on the value of the numerator in Eq. (4). It can be shown (e.g., Gilks et al., 1996) that a correct choice of this criterion will result in the accepted samples being distributed according to the joint posterior pdf (Eq. (4)). Schemes such as this are rejection sampling and importance sampling, which generate independent samples from the posterior. Any marginal distributions may then be generated by examining the samples from only the parameters of interest. However, these kinds of sampling schemes tend to be very slow, particularly in high dimensional parameter spaces, as samples are proposed at random, and thus each has a very small chance of being accepted.

Markov Chain Monte Carlo (MCMC) (for texts on MCMC, see Gamerman, 1997; Gilks et al., 1996) is a sampling technique

that addresses this problem by proposing samples preferentially in areas of high probability. Samples drawn from the posterior are no longer independent of one another, but the high probability of accepting samples allows for many samples to be drawn and, in many cases, for the posterior pdf to be built in a relatively short period of time. This is the approach we take in the MCMC section.

Priors and reference analysis

In the fully Bayesian framework, the choice of prior is critical to the inference we perform. In group statistics for FMRI, the number of observations we have is typically so small as to make the influence of the priors significant. As we have no prior information, we want the priors we use to be in some sense "noninformative", that is, we want to "let the data speak for itself". Reference priors are priors that attempt to reflect such prior ignorance. For an overview, see Bernardo and Smith (2000) and Kass and Wasserman (1996).

An intuitive approach would be to choose the prior of θ to be $\pi(\theta) = 1$. However, the resulting posteriors can change significantly depending on the parameterisation used. This is because a constant prior for one parameter will not typically transform into a constant prior for another. To overcome this reparameterisation problem, the Jeffreys prior was introduced for one-dimensional problems (Kass and Wasserman, 1996):

$$\pi(\theta) \propto \det(H(\theta))^{1/2} \tag{7}$$

where $H(\theta)$ is the Fisher information. However, this has difficulties dealing with multidimensional problems, that is, $\vec{\Theta} = (\theta_1 \ldots \theta_m)$. The Berger–Bernardo method (Bernardo and Smith, 2000) of reference analysis overcomes this by determining reference priors using information-theoretical ideas that maximise the amount of expected "information" from the data. See Appendix E for the derivation of the reference priors used in this paper.

The use of reference priors can be justified by consideration of the information theory that underpins them (Bernardo and Smith, 2000). However, whilst in this paper the null hypothesis frequentist inference is generally unknown, it is interesting to note that the Berger–Bernardo reference priors give the same inference as frequentist null hypothesis testing for cases of GLM inference for which the frequentist null hypothesis test is known. For example, in frequentist inference on the GLM, we typically examine the probability of attaining statistics for linear combinations (contrasts) of regression parameters under the null hypothesis. In cases where the null distribution on the GLM is known analytically, the Berger–Bernardo reference priors give the same probabilities when we test the probability that a contrast is greater than zero.

First level

Here we consider the first level in isolation and derive the marginal posterior distribution for $\boldsymbol{\beta}_k$, the vector of GLM height parameters for the first-level model fit. Eq. (1) gives us the likelihood for a first-level model in isolation, $p(\boldsymbol{Y}_k|\boldsymbol{\beta}_k, \sigma_k^2)$. The joint posterior on all parameters in this model is then:

$$p(\boldsymbol{\beta}_k, \sigma_k^2 \mid \boldsymbol{Y}_k) \propto p(\boldsymbol{Y}_k \mid \boldsymbol{\beta}_k, \sigma_k^2) p(\boldsymbol{\beta}_k, \sigma_k^2) \tag{8}$$

where $p(\boldsymbol{\beta}_k,\sigma_k^2)$ is the prior distribution on the regression and variance parameters. We use the Berger–Bernardo reference prior (see Priors and reference analysis), which is:

$$p(\boldsymbol{\beta}_k, \sigma_k^2) = 1/\sigma_k^2. \tag{9}$$

However, Eq. (8) does not give the distribution of interest for inference. We would like to infer on the posterior distribution on the activation height parameters $\boldsymbol{\beta}_k$ when the effect of estimating σ_k^2 is accounted for, that is, we would like to infer on $p(\boldsymbol{\beta}_k|\boldsymbol{Y}_k)$. To get this distribution, we must marginalise the joint posterior (Eq. (8)) over the parameter of no interest σ_k^2. This integral gives a multivariate noncentral t distribution for the posterior distribution on $\boldsymbol{\beta}_k$ (Lee, 1997):

$$p(\boldsymbol{\beta}_k \mid \boldsymbol{Y}_k) \propto \int p(\boldsymbol{Y}_k \mid \boldsymbol{\beta}_k, \sigma_k^2)/\sigma_k^2 d\sigma_k^2 = T(\boldsymbol{\beta}_k; \mu_{\beta_k}, \sigma_{\beta_k}^2 \Sigma_{\beta_k}, \nu_{\beta_k}), \tag{10}$$

where

$$\mu_{\beta_k} = (\mathbf{X}_k^T \mathbf{X}_k)^{-1} \mathbf{X}_k^T \boldsymbol{Y}_k$$

$$\sigma_{\beta_k}^2 = (\boldsymbol{Y}_k - \mathbf{X}_k \mu_{\beta_k})^T (\boldsymbol{Y}_k - \mathbf{X}_k \mu_{\beta_k})/(T - P_K)$$

$$\Sigma_{\beta_k} = (\mathbf{X}_k^T \mathbf{X}_k)^{-1}$$

$$\nu_{\beta_k} = T - P_K. \tag{11}$$

Note that if inference is performed in the frequentist framework, the null distribution on $\boldsymbol{\beta}_k$ is the multivariate central t distribution with the exact same covariance structure, $\sigma_{\beta_k}^2 \Sigma_{\beta_k}$, and degrees of freedom, ν_{β_k}, and the maximum likelihood estimate for $\boldsymbol{\beta}_k$ is exactly μ_{β_k}, the mean of the posterior distribution in the Bayesian framework.

Two-level

Here we consider the full two-level model laid out in Eqs. (2) and (3), applying the same ideas as in the previous section to infer on the second-level GLM height parameters $\boldsymbol{\beta}_g$. We will substitute into the posterior for the full two-level model the summary result of the first-level model derived in the previous section. This will provide us with the way of inferring on the full two-level model using just the summary result of the first level, that is, without reusing the data $\boldsymbol{Y}$.

Considering Eqs. (2) and (3). The full joint posterior for the two-level model is:

$$p(\boldsymbol{\beta}_g, \sigma_g^2, \boldsymbol{\beta}_{\boldsymbol{K}}, \boldsymbol{\sigma}_{\boldsymbol{K}}^2 \mid \boldsymbol{Y}) \propto \prod_k \{p(\boldsymbol{Y}_k \mid \boldsymbol{\beta}_k, \sigma_k^2)\} p(\boldsymbol{\beta}_{\boldsymbol{K}} \mid \boldsymbol{\beta}_g, \sigma_g^2) \times p(\boldsymbol{\beta}_g, \sigma_g^2, \boldsymbol{\sigma}_{\boldsymbol{K}}^2), \tag{12}$$

where $\boldsymbol{\sigma}_{\boldsymbol{K}}^2$ is the $(K \times 1)$ vector of first level variances σ_k^2, and $\boldsymbol{\beta}_{\boldsymbol{K}}$ is the $(K \times 1)$ vector of first level regression parameters $\boldsymbol{\beta}_k$ (for $k = 1 \ldots K$). We set the prior to be the Berger–Bernardo reference prior for this full two-level model (see Priors and reference analysis):

$$p(\boldsymbol{\beta}_g, \sigma_g^2, \boldsymbol{\sigma}_{\boldsymbol{K}}^2) = \frac{1}{\sigma_g^2} \prod_k \frac{1}{\sigma_k^2}. \tag{13}$$

Note that this model specification gives the posterior distribution, not only on the second-level parameters ($\boldsymbol{\beta}_g$,σ_g^2) but also on the parameters from all of the first-level models ($\boldsymbol{\beta}_{\boldsymbol{K}}$,$\boldsymbol{\sigma}_{\boldsymbol{K}}^2$). However, if we are only interested in the top or second-level parameters, we may substitute the summary result from the first level into this two-level model and marginalise over $\boldsymbol{\beta}_{\boldsymbol{K}}$ and $\boldsymbol{\sigma}_{\boldsymbol{K}}^2$ (see Appendix F), showing that the marginal distribution on $\boldsymbol{\beta}_g$ and σ_g^2 does not depend on the original data, but on the summary parameters from the first level, that is, μ_{β_k} and $\sigma_{\beta_k}^2\Sigma_{\beta_k}$:

$$p(\boldsymbol{\beta}_g, \sigma_g^2, \boldsymbol{\tau}_{\boldsymbol{K}} \mid \boldsymbol{Y}) \propto \prod_k \{\mathcal{N}(\mu_{\beta_k}; \boldsymbol{X}_{gk}\boldsymbol{\beta}_g, (\sigma_{\beta_k}^2 \Sigma_{\beta_k}/\tau_k) + \sigma_g^2 I) \times \Gamma(\tau_k; \nu_{\beta_k}/2, \nu_{\beta_k}/2)\} 1/\sigma_g^2 \quad (14)$$

where $\boldsymbol{X}_{gk}$ is the kth row vector of the second-level design matrix $\boldsymbol{X}_g$, and $\boldsymbol{\tau}_{\boldsymbol{K}}$ is a ($K \times 1$) vector of latent variables τ_k for $k = 1 \ldots K$ introduced for mathematical convenience (see Appendix F).

A special case of Eq. (14) is when the variances, $\sigma_{\beta_k}^2\Sigma_{\beta_k}$, on the first-level GLM parameters are known with very high degrees of freedom ($\nu_k \rightarrow \infty$). This is equivalent to $p(\boldsymbol{\beta}_k|\boldsymbol{Y}_k)$ in Eq. (10) being a normal distribution instead of a t distribution. In this case, the prior distribution on $\boldsymbol{\tau}_{\boldsymbol{K}}$ reduces to a delta function centered on $\boldsymbol{\tau}_{\boldsymbol{K}} = 1$ and the joint posterior distribution on the second-level parameters reduces to:

$$p(\boldsymbol{\beta}_g, \sigma_g^2 \mid \boldsymbol{Y}) \propto \prod_k \{\mathcal{N}(\mu_{\beta_k}; \boldsymbol{X}_{gk}\boldsymbol{\beta}_g, (\sigma_{\beta_k}^2 \Sigma_{\beta_k}) + \sigma_g^2 I)\} 1/\sigma_g^2. \quad (15)$$

Eq. (14) (or, in the special case, Eq. (15)) gives us the joint posterior distributions of $\boldsymbol{\beta}_g$, σ_g^2 and $\boldsymbol{\tau}_{\boldsymbol{K}}$. However, as in the first-level model, we are actually interested in inferring upon the marginal distribution over the GLM height parameters, $\boldsymbol{\beta}_g$. This marginal posterior $p(\boldsymbol{\beta}_g|\boldsymbol{Y})$ cannot be obtained analytically. Therefore, we consider two approaches, a fast posterior approximation and a slower but more accurate approach using Markov Chain Monte Carlo (MCMC) sampling. Crucially, in both approaches, we are going to assume that $p(\boldsymbol{\beta}_g|\boldsymbol{Y})$ is a multivariate noncentral t distribution:

$$p(\boldsymbol{\beta}_g \mid \boldsymbol{Y}) \propto \int p(\boldsymbol{\beta}_g, \sigma_g^2, \boldsymbol{\tau}_{\boldsymbol{K}} \mid \boldsymbol{Y}) d\sigma_g^2 d\boldsymbol{\tau}_{\boldsymbol{K}} \quad (16)$$

$$\approx T(\boldsymbol{\beta}_g; \mu_{\beta_g}, \sigma_{\beta_g}^2 \Sigma_{\beta_g}, \nu_{\beta_g}) \quad (17)$$

This assumption is crucial to the idea of being able to split hierarchies into inference on different levels for higher order models as we shall see in Higher level models. We shall test the validity of this assumption later. The fast posterior approximation or MCMC approaches are the means by which we get the distribution parameters μ_{β_g}, $\sigma_{\beta g}^2\Sigma_{\beta_g}$ and ν_{β_g}.

Fast posterior approximation

Here we propose a fast but approximate approach for estimating the distribution parameters, μ_{β_g},$\sigma_{\beta_g}^2\Sigma_{\beta_g}$ and ν_{β_g}, in Eq. (16). First, we assume high degrees of freedom at the first level, that is, $\tau_k = 1$ for all k. We then obtain a point estimate of σ_g^2 and use this point estimate to compute a point estimate of $\boldsymbol{\beta}_g$. For the details of how we obtain these point estimates $\widehat{\sigma_g^2}$ and $\mu_{\beta g}$, see Appendix G.

We then make the assumption that the effect of uncertainty in σ_g^2 is the same as the effect of uncertainty in σ_k^2 in a first-level model. This means that $p(\boldsymbol{\beta}_g|\boldsymbol{Y})$ is a multivariate noncentral t distribution:

$$\mathcal{N}(\boldsymbol{\beta}_g; \hat{\boldsymbol{\beta}}_g, (\mathbf{X}_G^T U^{-1} \mathbf{X}_G)^{-1}, \nu), \quad (18)$$

where U is a diagonal matrix with the kth diagonal element given by $S_k = (\sigma_{\beta_k}^2 \Sigma_{\beta_k}/\tau_k) + \sigma_g^2 I$.

However, we do not know the degrees of freedom (DOF), ν. We might expect the DOF to be within the range, $N_K - P_G \leq \nu \leq \infty$. In the validation section, we will look at using $\nu = N_K - P_G$ (lower estimate) and $\nu = \infty$ (upper estimate). The accuracy of these assumptions is examined with simulations in the Artificial data section.

Markov Chain Monte Carlo

Here, we use Markov Chain Monte Carlo (MCMC) to sample from the full joint posterior distribution given in Eq. (14). This also automatically provides us with samples from the marginal posterior distribution, $p(\boldsymbol{\beta}_g|\boldsymbol{Y})$.

We use single-component Metropolis–Hastings jumps (i.e., we propose separate jumps for each of the parameters in turn) for all parameters. We use separate normal proposal distributions for each parameter, with the mean fixed on the current value and with a scale parameter σ_p for the pth parameter that is updated every 30 jumps. At the jth update, σ_p is updated according to:

$$\sigma_p^{j+1} = \sigma_p^j \tilde{R} \frac{(1 + A + R)}{(1 + R)} \quad (19)$$

where A and R are the number of accepted and rejected jumps since the last σ_p update, respectively, $\tilde{R}$ is the desired rejection rate, which we fix at 0.5.

We require a good initialisation of the parameters in the model purely to reduce the required burn-in of the MCMC chains (the burn-in is the part of the MCMC chain, which is used to ensure that the chain has converged to be sampling from the true distribution). To initialise, we use the fast approximation approach described in Fast posterior approximation.

BIDET

MCMC can be used to directly obtain samples from $p(\boldsymbol{\beta}_g|\boldsymbol{Y})$. However, we would need to get lots of samples well into the tail of the distribution, and MCMC sampling is computationally intensive. Hence, we avoid the need for many samples by assuming that $p(\boldsymbol{\beta}_g|\boldsymbol{Y})$ is a multivariate noncentral t distribution. Recall that assuming a multivariate noncentral t distribution is also important to the idea of being able to split hierarchies into inference on different levels. Therefore, we clean up the samples of the posterior using Bayesian inference with distribution estimation using a T fit (BIDET).

BIDET fits a multivariate noncentral t distribution to the MCMC samples of $p(\boldsymbol{\beta}_g|\boldsymbol{Y})$ as described in Appendix D. Fig. 1 shows the result of using the multivariate noncentral t distribution fit to an MCMC chain obtained (see Markov Chain Monte Carlo) on a voxel in data set 2 described in Artificial data.

Contrasts

Whether from the fast approximation approach or from MCMC plus BIDET, the output from the analysis at any level

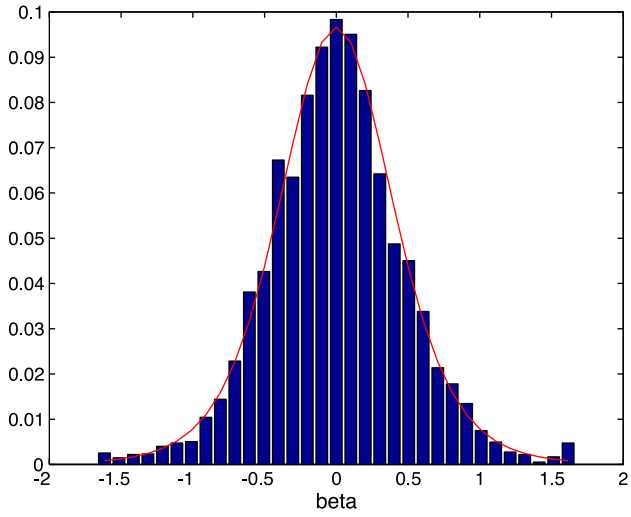

Fig. 1. The t fit (in this case one-dimensional) obtained on the MCMC samples from a voxel in data set 1.

in the hierarchy gives us a multivariate noncentral t distribution (Eq. (16)). As in the frequentist framework, we can ask questions about linear combinations (or contrasts) $c^T\boldsymbol{\beta}_g$ of the parameters in $\boldsymbol{\beta}_g$.

If c is a $P \times 1$ vector representing a t contrast, we can use Eq. (16) to give us the univariate noncentral t distribution over $c^T\boldsymbol{\beta}_g$

$$p(c^T\boldsymbol{\beta}_g \mid \boldsymbol{Y}) = T(c^T\boldsymbol{\beta}_g; c^T\mu_{\beta_g}, \sigma^2_{\beta_g} c^T\Sigma_{\beta_g}c, \nu_{\beta_g}) \tag{20}$$

We can then look at the $p(c^T\boldsymbol{\beta}_g > 0|\boldsymbol{Y})$. Note that this is equal to the probability of getting a t value greater than the t statistic:

$$t = c^T\mu_{\beta_g} \Big/ \sqrt{\sigma^2_{\beta_g} c^T\Sigma_{\beta_g}c} \tag{21}$$

under a central t distribution with degrees of freedom ν_{β_g}.

Higher level models

An increasing number of studies have three levels, in particular, a within session level, a session level and a subject level. With multiple sessions for multiple subjects, it becomes possible to model the between-session variance separately from the between-subject variance, and hence one can benefit from the improvements in sensitivity (due to heterogeneity of variance) this produces.

In the Two-level section, we showed that we could infer on the full two-level model using just the summary result of the first level without using the data $\boldsymbol{Y}$. We can use a similar argument to show that we can infer on a full three-level model using the summary result of the two-level model (given by Eq. (16)) without using the data $\boldsymbol{Y}$. The resulting distribution is similar to that in Eq. (14). Hence, we similarly assume that the marginal posterior is a multivariate noncentral t distribution equivalent to Eq. (16), and again we can use the fast posterior approximation or MCMC approaches to get the distribution parameters.

Higher level models can be considered using exactly the same argument. This is because after the first level, outputs and inputs for subsequent levels can be summarised as a multivariate noncentral t distribution. Hence, the assumption that the marginal distribution in Eq. (14) is a multivariate noncentral t distribution is integral to the idea of being able to split inference on multiple-level models into inference on the different levels. We shall test the validity of this assumption later.

Multiple group variances

We can use the framework we have described to work with multiple group variances at any level after the first level. An example of when this would be useful is when we might expect different between-subject variances for a patient group and a control group. We can easily deal with such multiple group variances if we limit ourselves to design matrices, which are "separable" with respect to the variance groupings.

We define a subdesign matrix as the part of the design matrix belonging to a group of observations for which we want to have a separate variance group. A design matrix would be "separable" with respect to the variance groupings if the subdesign matrices could be inferred upon using separate GLMs to give the same result as inferring on one GLM using the full design matrix.

We define a "group regressor" as that part of an regressor that belongs to a particular group variance:

Variance group	Regressor 1	Regressor 2
1	1	0
1	1	0
1	1	0
2	0	1
2	0	1
2	0	1

The group regressor for regressor 1 and for group 1 is $[1, 1, 1]^T$. The group regressor for regressor 1, group 2 is $[0, 0, 0]^T$.

We can check if our design matrix is "separable" by checking that within each regressor, only one group regressor has nonzero values in it. An example of a design matrix that violates this is:

Variance group	Regressor 1	Regressor 2
1	1	1
1	1	1
1	1	1
2	1	−1
2	1	−1
2	1	−1

Simulations have shown that if this constraint is not met then the resulting $\boldsymbol{\beta}_g$ vector is not generally multivariate t distributed. Whilst MCMC could deal with it, this violation prohibits the use of BIDET. This would require the use of longer MCMC chains and would also prohibit carrying the output to higher levels as the output from a level with these properties could not be summarised as a multivariate t distribution. Hence, we need in practice to ensure that our designs are "separable" with respect to the variance groupings.

These "separable" multiple group variance designs can then be implemented by inferring on separate GLMs using the fast approximation or MCMC plus BIDET. The results for different variance groups are pooled into one multivariate t distribution

for β_g. We can then proceed to the contrast stage and ask questions within or across variance groupings.

Artificial data

Methods

In the Two-level section, we showed that the two-level model can be inferred upon using the summary statistics of the first-level model inference (Eq. (14)). This means that all-in-one and split-level inferences are equivalent when we infer on the top-level regression parameters. Here we use four different null artificial data sets from the two-level model for 400 voxels to validate the fast approximation and MCMC or BIDET inference we perform on Eq. (14).

Inference approaches

The different inference approaches are all different ways of obtaining a z statistic for the t contrast of interest. The different inference approaches considered are as follows:

- MCMC: We sample from $p(\boldsymbol{\beta_g}|\boldsymbol{Y})$ to get an MCMC chain of 200,000 samples (with a burn-in of 1000 samples) using the approach described in Markov Chain Monte Carlo, we directly calculate the $p(c^T\boldsymbol{\beta}_g > 0|\boldsymbol{Y})$ from the MCMC samples of the marginal posterior of $p(c^T\boldsymbol{\beta}_g|\boldsymbol{Y})$. We can then use a p to z transform to calculate a z statistic at each voxel.
- BIDET: We fit a noncentral t distribution to an MCMC chain of 200,000 samples (with a burn-in of 1000 samples) using the approach described in BIDET. We can then use a t to p to z transform to calculate a z statistic at each voxel.
- LOWER: We use the lower bound from the fast approximation approach described in Fast posterior approximation to get an approximate noncentral t distribution. The lower bound is obtained when we assume DOF, $v = N_g - P_G$. We can then use a t to p to z transform to calculate a z statistic at each voxel.
- UPPER: We use the upper bound from the fast approximation approach described in Fast posterior approximation to get an approximate noncentral t distribution. The upper bound is obtained when we assume DOF, $v = \infty$.
- OLS: This is the standard frequentist approach (described at the start of Inference section) of estimating the total mixed effects variance. This ignores $\sigma^2_{\beta_k}$. Using the total mixed effects variance estimate, frequentist theory gives that the normalised OLS estimate of $c^T\boldsymbol{\beta}_g$ is t distributed with DOF, $v = N_g - P_G$. We can then use a t to p to z transform to calculate a z statistic at each voxel.

z Statistics

We want to be able to compare the resulting inference of these different approaches. It is difficult to compare different t statistics with different DOF. Therefore, for each of the different inference approaches, we convert to the probability of the contrast being greater than zero. This provides us with a measure that we can compare directly between the different approaches. We represent this probability as a z statistic by ensuring that the area under one tail of a standardised (zero mean and standard deviation of one) normal distribution corresponds to that probability. In Relating fully Bayesian inference to frequentist inference, we will explore the possibility of using these z statistics to mimic null hypothesis frequentist inference.

Relating the MCMC approach to OLS

It is important to appreciate that there are two different ways in which the z statistic can be changed between OLS and MCMC. The first was demonstrated in Beckmann et al. (2003), in that by taking into account lower level covariances and their heterogeneity, a substantial increase in higher level z statistic is possible. This is because the heterogeneity of the lower level covariances is effectively used to weight the summary statistic data to give more efficient estimates (resulting in reduced top-level regression parameter variance). This is analogous to the way in which prewhitening is used in first-level analyses to weight the regression parameter estimation to give more efficient estimators (Woolrich et al., 2001).

Beckmann et al. (2003) were unable to demonstrate the second way in which the z statistic can be changed between OLS and MCMC because they assumed that variances were known. In this paper, when we estimate the higher level variances, they are constrained to be positive. This overcomes the well-known "negative variance" problem in OLS (Leibovici and Smith, 2001) by forcing the total variance to be greater than it would be in the OLS case. This increased variance translates into lower z statistics in voxels that would have suffered from this problem.

In summary, we have two ways in which z statistics can change between OLS and MCMC. Firstly, they can increase due to increased efficiency from using lower level variance heterogeneity. Secondly, they can decrease due to the higher level variance being constrained to be positive.

Data sets

To avoid unnecessary consideration of first-level design matrices and because we are only looking to validate the inference on Eq. (14), we do not generate artificial first-level data $\boldsymbol{Y}$. Instead, we directly generate second-level summary "data", μ_{β_k}, via Eq. (14). To do this, we specify that we want null data by setting $\boldsymbol{\beta}_g = 0$ and then choose values for σ_{β_k} and σ_g. As a result, μ_{β_k}, v_k and σ_{β_k} form the summary statistic data we then use in the second-level inference.

To generate artificial data, we need to decide on our values for σ_{β_k} (for $k = 1 \ldots k$) and σ_g. Our choice is governed by the variance ratios we want between the top level and the lower levels. In relating the MCMC approach to OLS, we discussed two ways in which we would expect differences between OLS and MCMC inference. However, we would expect this difference in z statistics to be less and less substantial as the top-level variance dominates over the lower level variance. Beckmann et al. (2003) demonstrated that at a 10:1 ratio of between-session or -subject variance to within-session variance, the increase in higher-level z statistic (due to taking into account variance heterogeneity) is negligible. One of our data sets (data set 4) utilises a 10:1 variance ratio to explore if the combination of the two possible effects discussed in Relating the MCMC approach to OLS shows any difference in z statistics between OLS and MCMC.

However, we also consider variance ratios of the order of 1:1. The widely reported existence of the negative variance problem in FMRI (Leibovici and Smith, 2001; Worsley et al., 2002), along with the effects seen in the real FMRI data later in this paper, demonstrates that such low group to first-level variance ratios do exist in FMRI data. We need such a ratio to reproduce data that will suffer from the well-reported "negative variance" problem when using traditional OLS estimation (Leibovici and Smith, 2001). Furthermore, we need to consider the case of three-level hierarchies, which are popular in neuroimaging studies (e.g., hierarchies

containing within session levels, session levels and subject levels). When one uses the summary statistics from the output of the second level to infer on the third level, the variance ratio we are concerned with is between session variance to between subject variance, for which a ratio of the order of 1:1 is realistic.

The four data sets are as follows:

- Data set 1: A group mean design, with $N_K = 8$ subjects and $\sigma^2_{\beta_k} \approx 0$, $\sigma^2_g = 1$. The second level design matrix is: $\mathbf{X}_g = [1, 1, 1, 1, 1, 1, 1, 1]^T$ with t contrast $c = [1]$.
- Data set 2: A group mean design, with $N_K = 8$ subjects and $\sigma^2_{\beta_k} \sim$ uniform (0.1, 1.9), $v_k = 8$ and random effects variance $\sigma^2_g = 1$. The design matrix and t contrast is the same as for data set 1.
- Data set 3: A paired t test design with five subjects under two conditions (giving $N_K = 10$) and $\sigma^2_{\beta_k} \sim$ uniform (0.1; 1.9), $v_k = 8$ and random effects variance $\sigma^2_g = 0.5$. The second level design matrix is:

$$\mathbf{X}_g = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

  with t contrast $c = [1, 0, 0, 0, 0, 0]^T$.
- Data set 4: A group mean design, with $N_K = 8$ subjects and $\sigma^2_{\beta_k} \approx\sim$ uniform (0.1, 1.9), $\sigma^2_g = 10$. The second level design matrix is: $\mathbf{X}_g = [1, 1, 1, 1, 1, 1, 1, 1]^T$ with t contrast $c = [1]$.

Results

Fig. 2 show box plots of the difference in z statistics between those obtained from a long MCMC chain of 200,000 samples and those obtained from the different inference approaches considered. The intention is to consider the inference from a very long MCMC time series as a ''gold standard''. To help validate this assumption, the first box plot (labelled MCMC) compares this ''gold standard'' inference with another equally long MCMC chain but with a different random seed. This allows us to assess the inaccuracies in the ''gold standard'' due to the finite length of the MCMC chain. In all four data sets, the difference in z statistics for this is of the order of 0.01.

The second box plot (labelled BIDET) compares our ''gold standard'' to the inference obtained when we fit the noncentral multivariate t distribution to the long MCMC chain with a different random seed. This allows us to validate one of the strongest assumptions that we make in this paper. That is that the marginal posterior in Eq. (14) is a noncentral multivariate t distribution. This is crucial to the idea of being able to split hierarchies into inference on different levels. By making this distributional assumption, it also allows us to infer on shorter MCMC chains and gives us some basis for the fast approximation approach. This assumption is well supported by these BIDET box plots with the difference in z statistics being of the order of 0.01 for all four data sets.

Fig. 2 also shows box plots for the fast approximation approaches. We show box plots for the upper bound (labelled UPPER) and lower bound (labelled LOWER). Of particular interest is how good these bounds are at actually bounding the ''gold standard'' MCMC. Hence, a third box plot (labelled BOUND) shows the how far outside the bound the ''gold standard'' is. This shows a z statistic difference of up to 0.2 for data set 2. This z statistic difference of up to 0.2 between the fast approximation bounds and the MCMC ''gold standard'' will be used later as part of the HYBRID inference approach (see Hybrid inference approach).

The final box plot shows the traditional inference approach of ignoring the known fixed effects variance estimating the total mixed effects variance and using OLS to perform inference (labelled OLS). Because this ignores the fixed effects variance, this makes this approach the ''gold standard'' for data set 1, in which $\sigma^2_{\beta_k} \approx 0$. Indeed this is supported by the box plot. However, for data sets 2 and 3, $\sigma^2_{\beta_k} > 0$ and varies over k. For these data sets, OLS will give unbiased statistics, but very inefficient statistics as the $\sigma^2_{\beta_k}$ information is ignored. These box plots illustrate the difference in z statistics between OLS and the ''gold standard'' due to this inefficiency. In data set 4, $\sigma^2_{\beta_k}$ is sufficiently small compared to σ^2_g so that the differences between OLS and MCMC are negligible.

Fig. 3 shows the z statistics obtained for 20 voxels from the three data sets for the inference approaches of UPPER, LOWER, BIDET and OLS. For data set 1, the correspondence of OLS, LOWER and BIDET is reiterated. For data sets 2 and 3, the difference between BIDET and OLS is illustrated, as is the small inaccuracy of the UPPER and LOWER fast approximation approaches compared with BIDET.

Fig. 4 shows the histograms for the four different data sets of the degrees of freedom (DOF) obtained at each voxel from fitting the noncentral t distribution to an MCMC chain of 200,000 samples from the marginal posterior, $p(c^T\boldsymbol{\beta}_g|\boldsymbol{Y})$, as part of BIDET. For data set 1, we know that the OLS solution is the correct one and that the DOF, $v = 7$. In data set 4, σ_{β_k} is sufficiently small compared to σ_g so that the differences between OLS and MCMC are negligible and the range of DOF match those found in data set 1. Fig. 4 shows that BIDET correctly finds the DOF as being seven for the majority of voxels in data set 1. However, for data sets 2 and 3, the OLS DOF will be $v = 7$ and 4, respectively. We should not expect BIDET to have the same DOF values as this. Indeed the histograms show that the DOF obtained from BIDET varies from about these OLS DOF values to values up to about 60 or 70 DOF. Without using BIDET, there would be no way of knowing, for a particular voxel, the required DOF.

Fig. 5 shows box plots of the difference in z statistics between those obtained from a long MCMC chain of 200,000 samples and those obtained from using BIDET on MCMC chains of varying sample sizes. This illustrates the need for an MCMC chain of at least 20,000 samples to achieve accuracies of the order of 0.02 in z statistics.

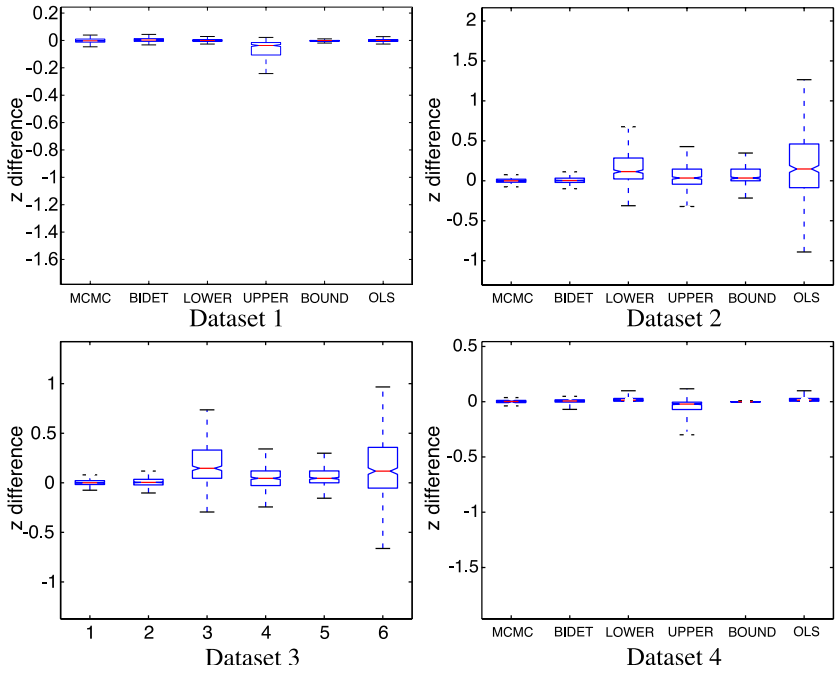

Fig. 2. Box plots over 400 voxels showing the z statistics obtained from a long MCMC chain of 200,000 samples minus the z statistics obtained from the different inference approaches considered. The box has lines at the lower quartile, median and upper quartile values. The length of the whiskers is 1.5 times the interquartile range. The box plots labelled (MCMC) correspond to the difference in z statistics between those obtained from the 200,000 sample MCMC chain and those obtained from another 200,000 sample MCMC chain with a different random seed. The box plots labelled (BOUND) correspond to how far outside the fast approximation bound (described as UPPER and LOWER) the z statistics obtained from the 200,000 sample MCMC chain lie.

Relating fully Bayesian inference to frequentist inference

We have some choices for how we use the posterior distribution $p(c^T\boldsymbol{\beta}_g|\boldsymbol{Y})$. We could simply use the posterior, $p(c^T\boldsymbol{\beta}_g|\boldsymbol{Y})$, to build up posterior probability maps representing the probability of activation at each voxel (Friston and Penny, in press). Another possibility is the use of (spatial) mixture modelling (Everitt and Bullmore, 1999; Hartvig and Jensen, 2000; Woolrich et al., 2003) to classify voxels as activating and nonactivating. We do not attempt to explore or discuss the relative merits of these approaches

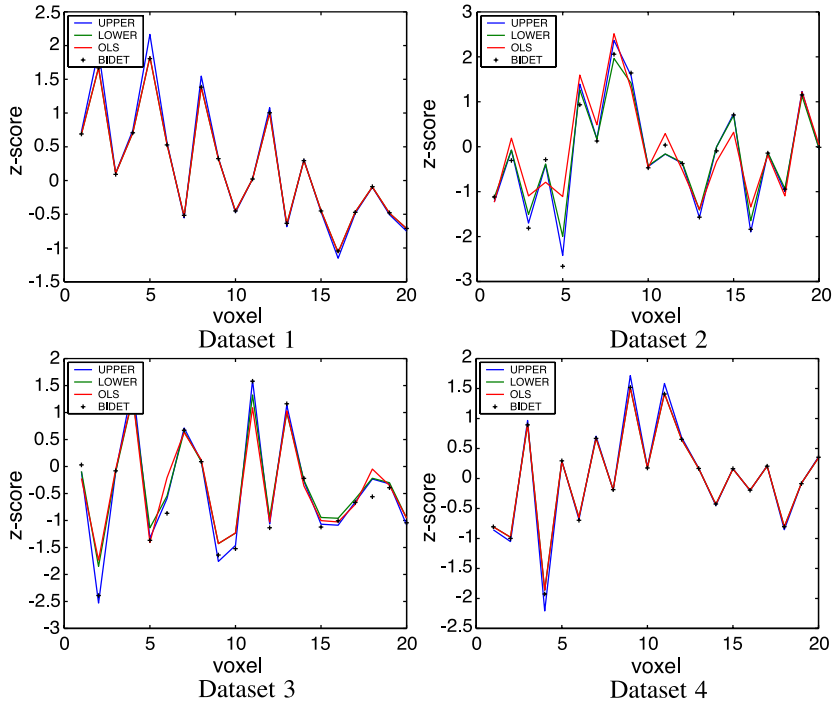

Fig. 3. Plots showing the z statistics for 20 voxels obtained from different inference approaches for the four different artificial data sets.

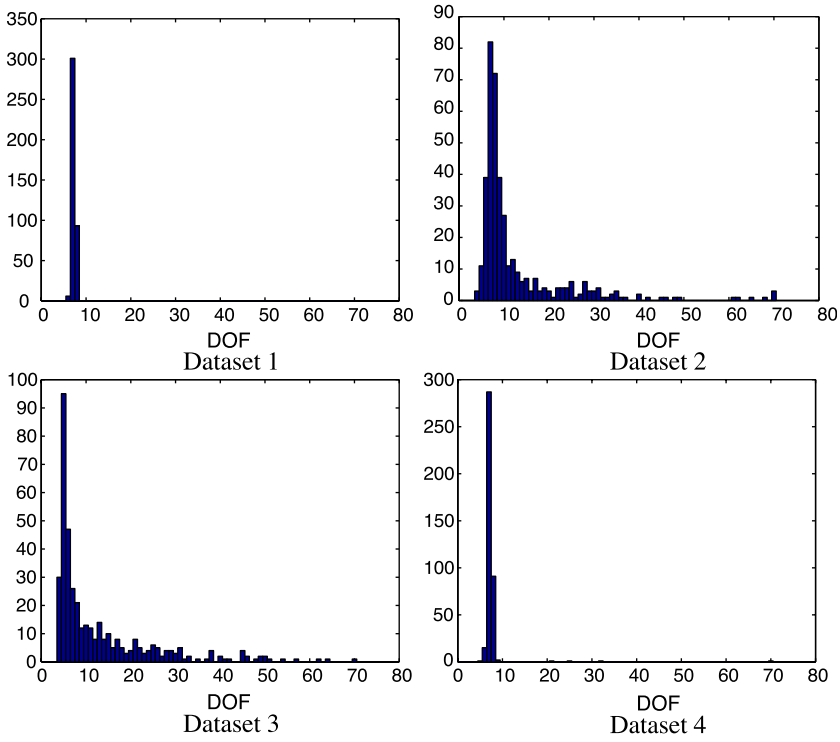

Fig. 4. Histograms over 400 voxels of the DOF estimated by BIDET for the different data sets for the four different artificial data sets.

in this paper. Here, we consider another possibility of the inference produced if we mimic null hypothesis frequentist inference [i.e., controlling a false positive rate (FPR)] by assuming that under the null hypothesis, the z statistics, which the fully Bayesian BIDET approach produces, are standardised (zero mean and standard deviation of one) Normally distributed.

To examine this possibility, Fig. 6 shows the log probability-log probability plots for the four different data sets for BIDET and OLS. These are plots of the nominal or theoretical frequentist FPR against the probabilities obtained empirically from our four null artificial data sets. For all four data sets, OLS does, as expected, produce a log probability plot that matches the nominal or theoretical frequentist FPR. However, this is not true for the BIDET approach.

Data sets 1 and 4 with small σ_{β_k} compared to σ_g give close to the same inference using BIDET as when using OLS. Hence, we would expect the log probability that BIDET produces to match the nominal or theoretical frequentist FPR. Fig. 6 demonstrates that this is true.

However, for data sets 2 and 3 ($\sigma^2_{\beta_k}$ is of the same order as σ_g), BIDET produces different results to OLS. The empirical log probabilities are lower than the nominal or theoretical FPR. Recall from the Relating the MCMC approach to OLS section that we have two ways in which we expect z statistics to change between OLS and MCMC. Firstly, they can increase due to increased efficiency from using lower-level variance heterogeneity. Secondly, they can decrease due to the higher-level variance being constrained to be positive. The first of these effects will introduce no bias into the p-p plots. Hence, only the second of these effects will be visible and the p-p plots for data sets 2 and 3 in Fig. 6 are consistent with this.

This means that whilst we produce more accurate estimates of the total mixed effects variance, it also means that the z statistics resulting from BIDET are not standardised normally distributed under the null hypothesis. This is not a problem if we just report posterior probability maps or use mixture modelling.

However, if we do choose to proceed assuming that the z statistics from BIDET are standardised normally distributed since the empirical log probabilities are lower than the nominal or theoretical frequentist FPR, then the validity of our statistics will not be violated. In other words, the z statistics from BIDET are on average conservative. The disadvantage of this is that we will lose some sensitivity when compared with using the unknown, correct null distribution. The advantage is that we can utilise cluster-based inference techniques on the z statistic maps,

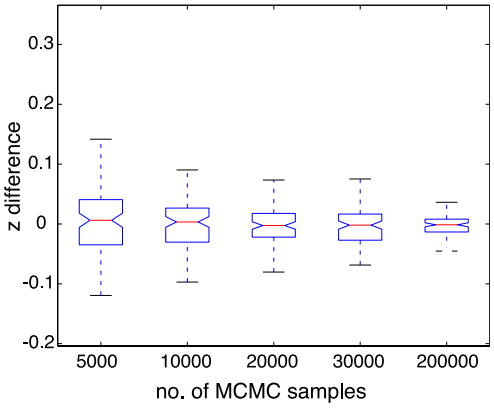

Fig. 5. Box plots over 400 voxels showing the difference in z statistics between those obtained from a long MCMC chain of 200,000 samples and those obtained from using BIDET on MCMC chains of varying sample sizes on data set 1. The box has lines at the lower quartile, median and upper quartile values. The length of the whiskers is 1.5 times the interquartile range.

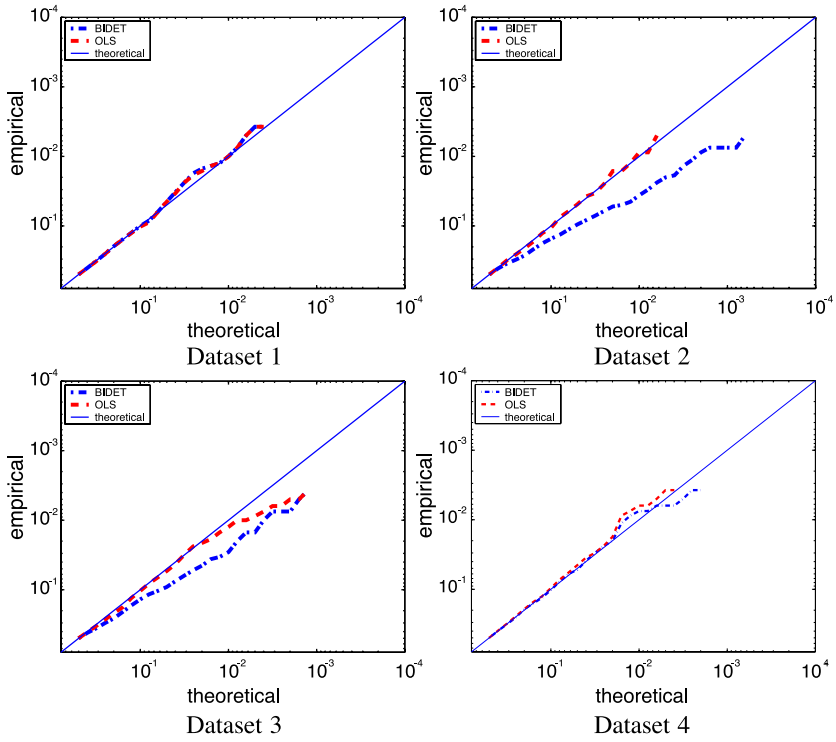

Fig. 6. Log probability-log probability plots over 400 voxels for the four different data sets for BIDET and OLS. These show plots of (nominal or theoretical) FPR against that obtained experimentally from our four null artificial data sets. The straight diagonal line shows the result for what would be a perfect match.

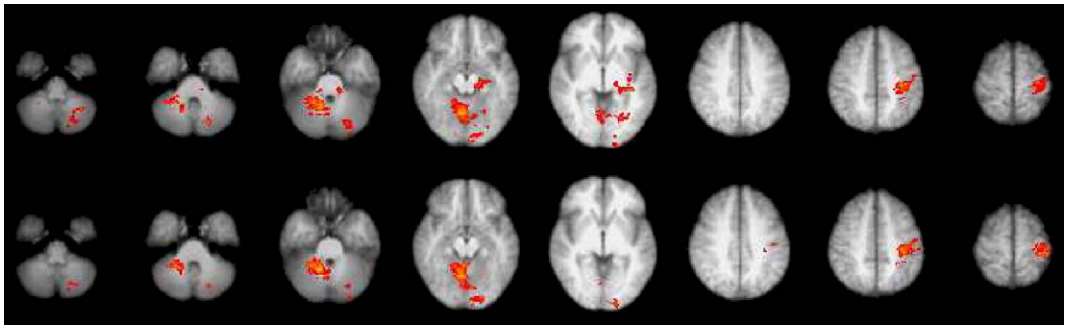

Fig. 7. Cluster thresholded ($z > 2.3$, $P < 0.01$) group activation from the INDEX data set. (Top) The OLS approach, and (bottom) the HYBRID approach.

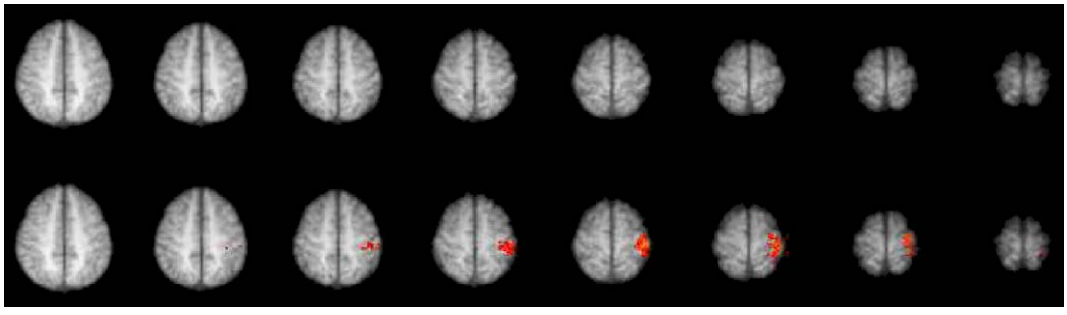

Fig. 8. Cluster thresholded ($z > 2.3$, $P < 0.01$) group activation from the SEQUENTIAL data set. (Top) The OLS approach, and (bottom) the HYBRID approach.

such as Gaussian Random Field Theory (Poline et al., 1997; Worsley, 2001).

FMRI data

Methods

Here, we consider two different FMRI data sets, both of which are simple motor tasks:

- INDEX: index finger vs. rest tapping task.
- SEQUENTIAL: sequential finger tapping vs. index finger tapping.

Each data set consists of single sessions for eight different subjects. In both data sets, the overall aim is to infer the group means at the top level. For each subject, echo planar images (EPI) were acquired using a 3-T system with TR = 3 s, time to echo (TE) = 30 ms, in-plane resolution 4 mm and slice thickness 7 mm. The first four scans were removed and the data were motion corrected using MCFLIRT (Jenkinson et al., 2002) and high-pass filtered as described in (Woolrich et al., 2001).

The overall model for this group experiment consists of two levels. The first level is a standard FMRI GLM with a design matrix for subject k, $\mathbf{X}_k$, containing regressors modelling the response to the task within each subject's data set. The second level is a GLM, which models the group mean of the individual subject's responses to the tasks, via a design matrix $\mathbf{X}_g = [1, 1, 1, 1, 1, 1, 1, 1]^T$.

To infer on this two-level model, we utilise the summary statistic approach we have laid out in this paper. To do this, we firstly produce the multivariate noncentral t distribution summary statistics of Eq. (16) using a first-level analyses of standard generalised least squares (GLS). This GLS analysis was performed using FEAT (FSL, n.d.). FEAT performs voxel-wise time series statistical analysis using local autocorrelation estimation to pre-whiten the data (Woolrich et al., 2001).

To infer the group mean, we now need to infer on the marginal posterior, $p(\beta_g|\boldsymbol{Y})$, using the multivariate noncentral t distribution summary statistics obtained from these first-level analyses (Eq. (14)).

To do this, we use two different approaches. Firstly, the OLS approach as described in Artificial data. Secondly, a hybrid approach that provides a compromise between the fast posterior approximation approach and the slower but more accurate approach of using Markov Chain Monte Carlo (MCMC) sampling and the fitting of a noncentral multivariate t distribution BIDET. The HYBRID approach is now described in detail. It is this which is implemented as the FMRIB's local analysis of mixed effects (FLAME) C++ program used for higher level analyses in FEAT (part of FSL v3.1).

Hybrid inference approach

Firstly, we can determine bounds on the accuracy of the fast approximation's z statistic bounds by using artificial data with "worst case scenario" variance components by comparing the LOWER and UPPER inference approaches with BIDET (as described in Artificial data). For the design matrices we are using here, the corresponding artificial data set we need to use is data set 1 from Artificial data.

We can then run the fast approximation approach on our real FMRI data first and subsequently only run the computationally expensive MCMC sampling (with 30,000 samples and a burn-in of 1000 samples) and the fitting of a noncentral multivariate t distribution BIDET on voxels at which the desired z threshold lies within the estimated bounds.

This hybrid approach takes approximately 1 h (for the data sets considered here) on a 2-GHz Intel PC on a full volume.

Thresholding

Using HYBRID, we obtain the marginal posterior, $p(\beta_g|\boldsymbol{Y})$, as a multivariate noncentral t distribution. We can then use a contrast $c = 1$ to produce $p(c^T\beta_g|\boldsymbol{Y})$. As discussed in Relating fully Bayesian inference to frequentist inference, we have some choices as to how we infer on this posterior distribution. Here we take the option of performing a t to p to z transform and mimicking a null hypothesis frequentist inference (i.e., controlling a FPR) by assuming that under the null hypothesis, the z statistics produced are standardised normally distributed (see Relating fully Bayesian inference to frequentist inference). One advantage of doing this is that we can utilise Gaussian Random Field Theory (GRFT) (Poline et al., 1997; Worsley et al., 1992). Here we use GRFT to threshold the z statistic maps and generate activation clusters determined by $z > 2.3$ with a significance threshold of $P = 0.01$.

Results

Figs. 7 and 8 show cluster-thresholded ($z > 2.3$; $P < 0.01$) group activation for the two motor tasks. Fig. 9 shows the number of suprathreshold voxels and the maximum z statistics for the two tasks. Fig. 7 shows the results from index finger tapping against rest (INDEX data set). There is a general decrease in z statistics in potentially activating voxels. This demonstrates the dominance of one of the two possible effects of incorporating first-level variances into the second level estimation process—that is, we get an increase in estimated group variance, σ_g, due to it being constrained to be positive. Fig. 8 shows the results of a contrast of sequential finger tapping vs. index finger tapping (SEQUENTIAL data set). There is a general increase in z statistics in potentially activating voxels. This demonstrates the dominance of the other possible effect of incorporating first-level variances into the second level estimation process—that is, we get increased efficiency in parameter estimation due to the use of lower level variance heterogeneity.

No. suprathreshold voxels

	[INDEX]	[SEQUENTIAL]
[OLS]	5206	0
[HYBRID]	3861	657

Max. z

	[INDEX]	[SEQUENTIAL]
[OLS]	5.20	4.02
[HYBRID]	4.69	4.22

Fig. 9. (Left) Number of suprathreshold voxels and (right) maximum z statistic from the INDEX and SEQUENTIAL FMRI data sets using the two different inference techniques OLS and HYBRID.

Conclusions

We have shown how multilevel hierarchical GLM inference can be split into different levels with the summary statistics of a multivariate noncentral t distribution being passed between the levels. This was achieved by formulating the model in a fully Bayesian framework and using reference analysis to drive our crucial choice of priors (see First level and Two-level). Using this framework, we have proposed two approaches to inferring at the top level. A fast approximation to the marginal posterior and a slower approach utilising Markov Chain Monte Carlo (MCMC) followed by a multivariate noncentral t distribution fit to the MCMC chains. These inference approaches are applicable whether we are attempting to infer using the all-in-one approach or the summary statistic split-model approach. We have validated the crucial assumption of the marginal distribution of the GLM regressions parameters being a multivariate noncentral t distribution at levels higher than the first using artificial data. The artificial data also demonstrates the difference between a standard OLS approach and the approach proposed in this paper. We have also shown results on FMRI data.

Discussion

When we attempt to infer on mixed effects models, we need to deal with the fact that the variance components are unknown. Classically, variance components tend to be estimated separately using iterative estimation schemes employing ordinary least squares (OLS), expectation maximisation (EM) or restricted maximum likelihood (ReML), see Searle et al. (1992) for details. As an example of a non-Bayesian approach, Worsley (2001) estimates variance components at each split level of the model separately. At higher than first levels, they propose EM for estimation of the random effects variance contribution to reduce bias in the variance estimation—a potential problem in higher level analyses if simple OLS were used. Positivity of the random-effects variance, avoiding what is known as the 'negative variance problem' (where mixed-effects variance estimates are lower than fixed-effects variances implying negative random-effects variance Leibovici and Smith, 2001), is partially addressed but not strictly enforced.

However, only in certain special cases (not including the model presented here) is it possible to derive analytical forms for the null distributions required by frequentist statistics. In the absence of analytical forms, frequentist solutions rely on null distributions derived from the data using such techniques as permutation tests (Nichols and Holmes, 2001). However, these lose the statistical power gained from educated assumptions about, for example, the distribution of the noise and limit inference to the number of available points in the empirical null distribution. Bayesian statistics gives us a tool for inferring on any model we choose and guarantees that uncertainty will be handled correctly.

Friston et al. (2002) have proposed a parametric empirical-Bayesian (PEB) approach for estimation of the all-in-one multilevel model. Unlike Worsley (2001), they relate the parameters of interest to the full set of original data, that is, they do not utilise the 'summary statistics' approach. Conditional posterior point estimates are generated using EM, which causes posterior probability maps.

Working in a fully Bayesian reference analysis framework, we have the capacity to infer either using the summary statistic split-level (Worsley) approach or the all-in-one (Friston et al., 2002) approach. However, all-in-one inference is not part of this paper and is an area of future work. The difference between an all-in-one inference based on the work described in this paper and the PEB work of (Friston et al., 2002) is that they assume a multivariate Gaussian marginal posterior for the regression parameters (and then heuristically convert it to a t statistic), whereas we work in a fully Bayesian framework using reference priors that we can validate as giving a multivariate t distribution with certain degrees as freedom using MCMC. Without reference priors, Friston et al. (2002) have nothing principled to drive the important choice of prior at the top level and as a result assume flat priors.

Importantly, one of the results demonstrated in this paper is that the inference we would obtain at the top level will be approximately the same regardless of whether we infer using the summary statistic split level (Worsley, 2001) or the all-in-one approaches (Friston et al., 2002) (assuming that first-level temporal autocorrelations are effectively known). However, it is very important to realise that there will be a difference if we look to infer at intermediate levels in the model. This is because in the all-in-one approach, the regression parameters at these intermediate levels will be regularised by the levels above in the hierarchy, whereas in the split-level approach they will not. Whether or not an experimenter would like to infer on, for example, a subject in isolation or on a subject in the context of the group of which it is a member, is a choice for the experimenter to make.

Acknowledgments

The authors would like to acknowledge support from the UK MRC, EPSRC and GSK.

Appendix A. Gamma distribution

x has a two-parameter gamma distribution, denoted by $Ga(a, b)$, with parameters a and b, if its density is given by:

$$\Gamma(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \tag{22}$$

where $\Gamma(a)$ is the single-parameter Gamma function. Note that a two-parameter gamma distribution has mean = a/b and variance = a/b^2.

Appendix B. Multivariate normal distribution

x is a $P \times 1$ random vector and has a multivariate normal distribution, denoted by $N(\mu,\sigma^2\Sigma)$, if its density is given by:

$$\mathcal{N}(x; \mu, \sigma^2\Sigma) = \frac{1}{(2\pi)^{P/2} \mid \sigma^2\Sigma \mid^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{23}$$

The multivariate normal distribution has mean = μ and covariance = $\sigma^2\Sigma$.

Appendix C. Multivariate noncentral *t* distribution

x is a $P \times 1$ random vector and has a multivariate noncentral t distribution, denoted by $t(\mu,\sigma^2\Sigma,v)$, if its density is given by:

$$\mathcal{T}(x;\mu,\sigma^2\Sigma,v) = \frac{\Gamma[(v+P)/2]}{(\pi v)^{P/2} \mid \sigma^2\Sigma \mid^{1/2} \Gamma[v/2]} \times \left(1 + \frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{\sigma^2 v}\right)^{-(v+P)/2} \quad (24)$$

where $\Gamma(a)$ is the single-parameter Gamma function. The noncentral t distribution has mean = μ and covariance = $\sigma^2\Sigma v / (v-2)$ for $v > 2$.

We can represent a multivariate noncentral t distribution using a two-parameter gamma distribution and a multivariate normal distribution in a Bayesian framework. If we introduce a variable τ and specify a joint posterior over x and τ as:

$$p(\tau,x \mid \mu,\sigma^2\Sigma,v) \propto p(x \mid \tau,\mu,\sigma^2\Sigma)p(\tau \mid v)$$
$$x \mid \tau,\mu,\sigma^2\Sigma \sim N(\mu,(\sigma^2\Sigma/\tau)) \quad (25)$$
$$\tau \mid v \sim Ga(v/2,v/2)$$

then the marginal posterior for x is a multivariate noncentral t distribution, that is,

$$p(x \mid \mu,\sigma^2\Sigma,v) = \int p(\tau,x \mid \mu,\sigma^2\Sigma,v)d\tau$$
$$x \mid \mu,\sigma^2\Sigma,v \sim t(\mu,\sigma^2\Sigma,v) \quad (26)$$

Appendix D. Multivariate noncentral *t* distribution fit

In this section, we describe how the multivariate noncentral t distribution fit is performed in BIDET.

Assume that we have $P \times N_J$ matrix, x, with elements, (x_{jp}), where $j = 1 \ldots N_J$ indexes samples and $P = 1 \ldots P$ indexes parameters. The task is to fit to these samples a multivariate noncentral t distribution, $t(\mu,\sigma^2\Sigma,v)$ (as described in Appendix C).

In BIDET, we constrain the mean of the multivariate noncentral t distribution, μ_{β_g}, to be equal to that from the fast posterior approximation for μ_{β_g} described in Fast posterior approximation. If we are not using this constraint, then we can set the mean μ to the sample mean, that is,

$$\mu_p = \frac{1}{N_J}\sum_j x_{jp} \quad (27)$$

We can also directly estimate the normalised covariance Σ using the sample covariance, $\hat{\Sigma}$:

$$\tilde{\Sigma} = \hat{\Sigma} / \mid \hat{\Sigma} \mid^{1/P}$$
$$\hat{\Sigma} = (x-M)(x-M)^T/(N_J-1) \quad (28)$$

where $M = \{\mu, \mu, \ldots, \mu\}^T$.

We still need to estimate σ^2 and v. Fortunately, we can represent a multivariate noncentral t distribution using a two-parameter gamma distribution and a multivariate normal distribution in a Bayesian framework by introducing hidden variables τ_i (see Appendix C). With hidden variables, we can use the Expection–Maximisation (EM) algorithm. In the E-step, we obtain the expected value of the hidden variables, τ_j:

$$E_{\tau_j \mid v_{(t)},\sigma^2_{(t)},x}[\tau_j] = \frac{\sigma^2_{(t)}(v_{(t)}+P)}{v_{(t)}\sigma^2_{(t)} + s_j} \quad (29)$$

where:

$$s_j = (x_j - \mu_j)^T\tilde{\Sigma}^{-1}(x_j - \mu_j) \quad (30)$$

and then in the M-step, we can minimise the joint posterior over v, σ^2 given $\tau_j(t) = E_{\tau_j \mid v(t),\sigma^2_{(t)},x}[\tau_j]$ to get updates for v, σ^2 as:

$$\sigma^2_{(t+1)} = \frac{1}{N_J P}\sum_j \tau_{j(t)}s_j$$
$$v_{(t+1)} = \frac{2}{1 - \sigma^2_{(t)}/\left(\frac{1}{N_J-1}\sum_j s_j\right)} \quad (31)$$

Convergence normally occurs after about 10 iterations. To be conservative, we therefore use 50 iterations.

Appendix E. Determining reference priors

Here we show how we determine the reference prior for a vector of parameters θ for a model with likelihood $p(y|\theta)$. This is taken from Section 5.4.5. of Bernardo and Smith (2000):

The Fisher information matrix, $\mathbf{H}(\theta)$, is given by:

$$\mathbf{H}(\theta) = -E_{y|\theta}\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(y \mid \theta)\right\} \quad (32)$$

For the models in this paper, the Fisher information matrix, $\mathbf{H}(\theta)$, is block diagonal:

$$\mathbf{H}(\theta) = \begin{bmatrix} h_{11}(\theta) & 0 & \cdots & 0 \\ 0 & h_{22}(\theta) & 0 & \vdots \\ \vdots & 0 & O & 0 \\ 0 & \cdots & 0 & h_{mm}(\theta) \end{bmatrix} \quad (33)$$

and we can separate out the block $h_{jj}(\theta)$ as being the product:

$$\{h_{jj}(\theta)\}^{1/2} = f_j(\theta_j)g_j(\theta_{-j}) \quad (34)$$

where $f_j(\theta_j)$ is a function depending only on θ_j and $g_j(\theta_{-j})$ does not depend on θ_j. The Berger–Bernardo reference prior is then given by:

$$\pi(\theta) \propto \prod_j^m f_j(\theta_j) \quad (35)$$

Note that this approach yields the Jeffreys prior in one-dimensional problems.

Appendix F. Marginalising over $(\beta_K;\sigma^2_K)$ in the two-level model

From the two-level model the full joint posterior distribution is (Eq. (12)):

$$p(\beta_g,\sigma^2_g,\beta_K,\sigma^2_K|Y)\propto\prod_k\{p(Y_k|\beta_k,\sigma^2_k)\}$$

$$\times p(\beta_K|\beta_g,\sigma^2_g)p(\beta_g,\sigma^2_g,\sigma^2_K), \tag{36}$$

where the prior is the reference prior for this full two-level model (Eq. (13)):

$$p(\beta_g,\sigma^2_g,\sigma^2_K)=\frac{1}{\sigma^2_g}\prod_k\frac{1}{\sigma^2_k}. \tag{37}$$

If we marginalise out σ^2_K, then we get:

$$p(\beta_g,\sigma^2_g,\beta_K|Y)\propto\prod_k\left\{\int p(Y_k|\beta_k,\sigma^2_k)/\sigma^2_k d\sigma^2_k\right\}$$

$$\times\mathcal{N}(\beta_K;X_g\beta_g,\sigma^2_g I)1/\sigma^2_g \tag{38}$$

and then substitute in the summary result of the first-level model in isolation (Eq. (10)):

$$p(\beta_g,\sigma^2_g,\beta_K|Y)\propto\prod_k\{\mathcal{T}(\beta_k;\mu_{\beta_k},\sigma^2_{\beta_k}\Sigma_{\beta_k},\nu_{\beta_k})\}$$

$$\times\mathcal{N}(\beta_K;X_g\beta_g,\sigma^2_g I)1/\sigma^2_g. \tag{39}$$

We can represent a multivariate noncentral t distribution using a two-parameter Gamma distribution and a multivariate normal distribution (see Appendix C). This is achieved by introducing a parameter τ_k for each vector β_k:

$$p(\beta_g,\sigma^2_g,\beta_K,\tau_K|Y)\propto\prod_k\{\mathcal{N}(\beta_k;\mu_{\beta_k},(\sigma^2_{\beta_k}\Sigma_{\beta_k}/\tau_k))$$

$$\times\Gamma(\tau_k;\nu_{\beta_k}/2,\nu_{\beta_k}/2)\}\mathcal{N}(\beta_K;X_g\beta_g,\sigma^2_g I)1/\sigma^2_g. \tag{40}$$

Writing $\mathcal{N}(\beta_K;X_g\beta_g,\sigma^2_g I)=\Pi_k\mathcal{N}(\beta_k;X_{gk}\beta_g,\sigma^2_G I)$, where X_{gk} is the kth row vector of the second-level design matrix X_g, we can now easily integrate out β_k for all k to give:

$$p(\beta_g,\sigma^2_g,\tau_K\mid Y)\propto\prod_k\{\mathcal{N}(\mu_{\beta_k};X_{gk}\beta_g,(\sigma^2_{\beta_k}\Sigma_{\beta_k}/\tau_k)+\sigma^2_g I)$$

$$\times\Gamma(\tau_k;\nu_{\beta_k}/2,\nu_{\beta_k}/2)\}1/\sigma^2_g \tag{41}$$

where τ_K is a $(K\times 1)$ vector of the variables τ_k for $k=1\ldots,K$.

Appendix G. Fast approximation point estimates

Here we describe how we obtain the point estimates of σ^2_g and β_g for use in the fast approximation approach described in Fast posterior approximation. We start by rewriting Eq. (14) as:

$$p(\beta_g,\sigma^2_g,\tau_K\mid Y)=N(\mu_{\beta_K};X_G\beta_g,U)1/\sigma^2_g$$

$$U=\begin{bmatrix}S_1 & 0 & \cdots & 0\\ 0 & S_2 & & 0\\ \vdots & & O & \vdots\\ 0 & \cdots & 0 & S_N\end{bmatrix}$$

$$S_k=(\sigma^2_{\beta_k}\Sigma_{\beta_k}/\tau_k)+\sigma^2_g I \tag{42}$$

Point estimate of σ^2_g

We get a point estimate of σ^2_g by finding the maximum a posterior (MAP) over the marginal posterior distribution $p(\sigma^2_g,\tau_K|Y)$. If we marginalise out β_g, then the marginal posterior is:

$$p(\sigma^2_g,\tau_K\mid Y)=\mid U\mid^{-1/2}\mid X^T_G U^{-1}X_G\mid^{-1/2}$$

$$\times\exp\left\{-\frac{1}{2}(\mu^T_{\beta_K}U^{-1}\mu_{\beta_K}-\tilde{\beta}^T_g X^T_G U^{-1}X_G\tilde{\beta}_g)\right\}$$

$$\times 1/\sigma^2_g \tag{43}$$

where

$$\tilde{\beta}_g=(X^T_G U^{-1}X_G)^{-1}X^T_G U^{-1}\mu_{\beta_K} \tag{44}$$

We then assume $\tau_k=1$ and look to find the MAP for σ^2_g. However, there is a question of parameterisation. The mode we get will depend on the parameterisation we use. For example, we could look to maximise with respect to σ^2_g, σ_g, $\log(\sigma^2_g)$ or $\phi_g=1/\sigma^2_g$, etc., all of which will give us different MAPs. Note that as we reparameterise, the reference prior might change but the reference posterior always stays the same, see Bernardo and Smith (2000). Hence, a natural way to reparameterise such that the parameter we use gives us a uniform reference prior.

The parameterisation that gives us a uniform reference prior is $\theta=\log(\sigma^2_g)$. Hence, we need to solve:

$$\hat{\theta}=\arg\max_\theta p(\sigma^2_g\mid Y,\tau_K=1) \tag{45}$$

where $p(\sigma^2_g|Y,\tau_K=1)$ is the marginal in Eq. (43) with $\tau_K=1$. To solve for $\hat{\theta}$ using this equation, we use Brent's (1973) algorithm. We can then easily convert from $\hat{\theta}$ to $\widehat{\sigma^2_g}$.

Point estimate of β^2_g

We approximate β_g using the point estimate $\widehat{\sigma^2_g}$ and $\tau_K=1$:

$$\hat{\beta}=\arg\max_{\beta_g}p(\beta_g\mid Y,\sigma^2_g=\widehat{\sigma^2_g},\tau_K=1) \tag{46}$$

where $p(\beta_g|Y,\sigma^2_g=\widehat{\sigma^2_g},\tau_K=1)$ is Eq. (15) with $\sigma^2_g=\widehat{\sigma^2_g}$ and $\tau_K=1$. The solution to this is:

$$\hat{\beta}_g=(X^T_G U^{-1}X_G)^{-1}X^T_G U^{-1}\mu_{\beta_K} \tag{47}$$

with U as in Eq. (15), but with $S_k=(\sigma^2_{\beta_k}\Sigma_{\beta_k})+\widehat{\sigma^2_g}I$.

References

Beckmann, C., Jenkinson, M., Smith, S., 2003. General multi-level linear modelling for group analysis in FMRI. NeuroImage 20, 1052–1063 (first two authors contributed equally).

Bernardo, J., Smith, A., 2000. Bayesian Theory. Wiley.

Brent, R., 1973. Algorithms for Minimization without Derivatives. Prentice Hall.

Bullmore, E., Brammer, M., Williams, S., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P., 1996. Statistical methods of estimation and inference for functional MR image analysis. Magn. Reson. Med. 35 (2), 261–277.

Cox, R., 1946. Probability, frequency and reasonable expectation. Am. J. Phys. 14, 1–13.

Everitt, B., Bullmore, E., 1999. Mixture model mapping of brain activation in functional magnetic resonance images. Hum. Brain Mapp. 7, 1–14.

Frison, L., Pocock, S., 1992. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. Stat. Med. 11, 1685–1704.

Friston, K.J., 2002. Bayesian estimation of dynamical systems: an application to fMRI. NeuroImage 16, 513–530.

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

Friston, K., Josephs, O., Zarahn, E., Holmes, A., Rouquette, S., Poline, J.-B., 2000. To smooth or not to smooth? NeuroImage 12, 196–208.

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

FSL. Available: http://www.fmrib.ox.ac.uk/fsl

Gamerman, D., 1997. Markov Chain Monte Carlo. Chapman & Hall, London.

Gilks, W., Richardson, S., Spiegalhalter, D., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

Hartvig, N., Jensen, J., 2000. Spatial mixture modelling of fMRI data. Hum. Brain Mapp. 11 (4), 233–248.

Holmes, A., Friston, K., 1998. Generalisability, random effects and population inference. Fourth International Conference on Functional Mapping of the Human Brain, NeuroImage, vol. 7, p. S754.

Jenkinson, M., Bannister, P., Brady, J., Smith, S., 2002. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17 (2), 825–841.

Kass, R., Wasserman, L., 1996. Formal rules for selecting prior distributions: a review and annotated bibliography. JASA 91, 1343–1370.

Lee, P., 1997. Bayesian Statistics. Arnold.

Leibovici, D., Smith, S., 2001. Min-max filter for multi-subject analysis. Seventh International Conference on Functional Mapping of the Human Brain.

Nichols, T.E., Holmes, A.P., 2001. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

Poline, J.-B., Worsley, K., Evans, A., Friston, K., 1997. Combining spatial extent and peak intensity to test for activations in functional imaging. NeuroImage 5, 83–96.

Searle, S., Casella, G., McCulloch, C., 1992. Variance Components. Wiley.

Woolrich, M., Ripley, B., Brady, J., Smith, S., 2001. Temporal autocorrelation in univariate linear modelling of FMRI data. NeuroImage 14 (6), 1370–1386.

Woolrich, M., Behrens, T., Beckmann, C., Smith, S., 2004a. Mixture models with adaptive spatial regularisation for segmentation with an application to FMRI data. Technical Report TR04MW1, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Oxford University, Oxford, UK. Available at http://www.fmrib.ox.ac.uk/analysis/techrep for downloading (submitted for publication).

Woolrich, M., Jenkinson, M., Brady, J., Smith, S., 2004b. Fully Bayesian spatio-temporal modelling of FMRI data. IEEE Trans. Med. Imag. 22 (2) (in press).

Worsley, K., 2001. in Functional MRI: An Introduction to Methods. In: Jezzard, P., Matthews, P.M., Smith, S.M. (Eds.), OUP, Oxford, pp. 251–270 (Chap. 14).

Worsley, K., Friston, K., 1995. Analysis of fMRI time series revisited—Again. NeuroImage 2, 173–181.

Worsley, K., Evans, A., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., Evans, A., 2002. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.