ACADEMIC PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

NeuroImage

NeuroImage 20 (2003) 1052–1063

www.elsevier.com/locate/ynimg

General multilevel linear modeling for group analysis in FMRI

Christian F. Beckmann,[a,b,*,1] Mark Jenkinson,[a,1] and Stephen M. Smith[a]

[a] *Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, University of Oxford, Oxford OX3 9DU, UK*

[b] *Medical Vision Laboratory, Department of Engineering, University of Oxford, Oxford, UK*

Received 18 October 2002; revised 7 July 2003; accepted 14 July 2003

Abstract

This article discusses general modeling of multisubject and/or multisession FMRI data. In particular, we show that a two-level mixed-effects model (where parameters of interest at the group level are estimated from parameter and variance estimates from the single-session level) can be made equivalent to a single complete mixed-effects model (where parameters of interest at the group level are estimated directly from all of the original single sessions' time series data) if the (co-)variance at the second level is set equal to the sum of the (co-)variances in the single-level form, using the BLUE with known covariances. This result has significant implications for group studies in FMRI, since it shows that the group analysis requires only values of the parameter estimates and their (co-)variance from the first level, generalizing the well-established "summary statistics" approach in FMRI. The simple and generalized framework allows different prewhitening and different first-level regressors to be used for each subject. The framework incorporates multiple levels and cases such as repeated measures, paired or unpaired *t* tests and *F* tests at the group level; explicit examples of such models are given in the article. Using numerical simulations based on typical first-level covariance structures from real FMRI data we demonstrate that by taking into account lower-level covariances and heterogeneity a substantial increase in higher-level *Z* score is possible.

Keywords: Mixed effects; FMRI; Group analysis; Fixed and random effects; Multisubject analysis

I. 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 analyzed in such a way that allows for hypothesis tests at the group level (Holmes and Friston, 1998; Worsley et al., 2002), e.g., in order 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 referred to as the General Linear Model (GLM) and much attention to date has been focused on ways of modeling 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).

In order to carry out higher-level analyses it is straightforward to formulate a complete single-level GLM that relates various parameters of interest at the group level to the full set of (time series) data available (Cnaan et al., 1997; Matthews et al., 1990). In FMRI, where the human and computational costs involved in data analysis are relatively high, however, it is desirable to be able to make group-level inferences using the *results* of separate first-level analyses without the need to reanalyse any of the individual subject data; an approach commonly referred to as the "summary statistics" approach to FMRI analysis (Holmes and Friston, 1998). Within such a two-level approach, group parameters of interest can easily be refined as

* Corresponding author. Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB) University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK. Fax: +0044-1865-222717.

E-mail address: beckmann@fmrib.ox.ac.uk (C.F. Beckmann).

[1] These authors contributed equally to this work.

1053-8119/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S1053-8119(03)00435-X

more data become available. The natural question to address is if and how a two-level approach to group analysis restricts the type of hypotheses we are able to infer at the group level.

In order 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 modeling of these additional variance terms (Frison and Pocock, 1992; Holmes and Friston, 1998).

In this article we revisit the single-level mixed-effect GLM for FMRI group analysis which relates the effect of interest at the group level (e.g., difference of patients and controls) to the individual FMRI time-series. This model is well known and encompasses both the familiar single-subject models and a general group model (Cnaan et al., 1997; Everitt, 1995; Holmes and Friston, 1998; Matthews et al., 1990). By using the full GLM at the group level, one obtains a very general and flexible framework which allows examination of more complex relationships beyond the simple tests such as mean group activation. An example of a more general group analysis would be to test whether activation is correlated with some other variable such as drug dosage or disability score.

The main result shows that this single-level GLM can be decomposed into an equivalent two-level version so that group analyses can be performed using only the lower-level parameter *estimates* and their (co-)variances, from the individual subject analyses. The practical consequence of this model is that it is possible to perform valid group analyses in two stages: first the individual subject analyses, and second, a single group analysis performed on the output of the combined estimates from the individual subject analyses. Furthermore, there are very few restrictions placed upon the model and so it can be used in very general conditions, such as when the estimates at the individual subject level are obtained using different prewhitening for each subject or indeed different regressors. We finally give examples of how, in this general framework, group tests of interest, like paired or unpaired *t* tests, can easily be formulated. The model equivalence result applies not only to FMRI studies but to any mixed-effects GLM that can be split into two levels in the same way as presented here.

All models used here are instances of the univariate GLM. This means treating each single, registered voxel separately and assuming Gaussian noise. Parameter estimation is being performed (at all levels) using the prewhitening approach (Bullmore et al., 1996; Woolrich et al., 2001), which is known to be the best linear unbiased estimator (BLUE) for known variances for these models (Searle et al., 1992).

In this article we make a distinction between the problems involved in modeling of FMRI data at the group level and estimation of the relevant (co-)variances. While estimation of variances is an important issue when implementing such models (and is discussed in Section III), we consider it best to treat this as a separate problem. An article on practical issues of estimation, using a novel fully Bayesian method, is currently being prepared and these techniques form the basis of what has been implemented in the latest software release of FSL.

II. Models

To begin with we consider the familiar two-level univariate GLM for FMRI. That is, the model that in the first level deals with 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.

A. Multilevel GLM

Consider an experiment where there are N subjects and that for each subject, k, the preprocessed FMRI data are Y_k (a vector of T time points), the design matrix is X_k, and the parameter estimates are β_k (for $k = 1, \ldots, N$). An individual GLM relates first-level parameters to the k individual data sets,

$$Y_k = X_k\beta_k + \epsilon_k,$$

where ϵ_k specifies the single-subject residuals and $\mathrm{E}(\epsilon_k) = 0$, $\mathrm{Cov}(\epsilon_k) = V_k$. Using the block diagonal forms, i.e., with

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}, \quad X = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & X_N \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix}, \quad \text{and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix},$$

the two-level model for this experiment is

$$Y = X\beta + \epsilon \tag{1}$$

$$\beta = X_G\beta_G + \eta, \tag{2}$$

where X_G is the group-level design matrix (e.g., separating controls from normals), β_G is the final vector of group-level parameters, η specifies the residuals of the group activation

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

(parameter) scores, and $E(\eta) = 0$, $Cov(\eta) = V_G$ and $Cov(\epsilon) = V$ denote the block-diagonal form of first level covariance matrices V_k.

The two-level model written above can be rewritten as a single-level model by substituting Eq. (2) into Eq. (1), so that

$$Y = XX_G\beta_G + \gamma, \tag{3}$$

where

$$\gamma = X\eta + \epsilon,\ E(\gamma) = 0,$$
$$\text{and } Cov(\gamma) = W = XV_GX^T + V.$$

It is easy to see that this *model* is equivalent to the two-level version in Eqs. (1) and (2).

B. Parameter estimation

The BLUEs for both the two-level GLM and the single-level GLM, assuming known covariances, can be calculated using the General Least Squares approach (Searle et al., 1992); for the two-level GLM the parameter estimates at the first level are

$$\hat{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}Y,$$
$$Cov(\hat{\beta}) = (X^TV^{-1}X)^{-1}. \tag{4}$$

Similarly, the estimates of the (second-level) group parameters are given by

$$\hat{\beta}_G = (X_G^TV_G^{-1}X_G)^{-1}X_G^TV_G^{-1}\beta,$$
$$Cov(\hat{\beta}_G) = (X_G^TV_G^{-1}X_G)^{-1}.$$

In the summary statistics approach to multilevel GLMs the second-level model uses the *estimates* from the first level as input and not the true (but unobservable) parameters. That is, Eq. (2) is modified so that

$$\hat{\beta} = X_G\beta_G + \eta'. \tag{5}$$

Therefore the two-level model, as used in practice, is specified by Eqs. (1) and (5). This has significant implications, as the two-level version is no longer precisely equivalent to the single-level model with respect to parameter estimates. In particular, the BLUE of the group parameters in the two-level model now is

$$\hat{\beta}_G = (X_G^TV_{G2}^{-1}X_G)^{-1}X_G^TV_{G2}^{-1}\hat{\beta},$$
$$Cov(\hat{\beta}_G) = (X_G^TV_{G2}^{-1}X_G)^{-1}, \tag{6}$$

where $V_{G2} = Cov(\eta')$ represents the potentially different covariance in this modified two-level model.

Substituting Eq. (4) into Eq. (6) gives the two-level group parameter estimates as

$$\hat{\beta}_G = (X_G^TV_{G2}^{-1}X_G)^{-1}X_G^TV_{G2}^{-1}(X^TV^{-1}X)^{-1}X^TV^{-1}Y,$$
$$Cov(\hat{\beta}_G) = (X_G^TV_{G2}^{-1}X_G)^{-1}, \tag{7}$$

where the group parameter of interest, $\hat{\beta}_G$, relates to the original data by means of intermediate estimates $\hat{\beta}$. Now consider the single-level GLM (Eq. (3)) where the BLUE is

$$\hat{\beta}_G = (X_G^TX^TW^{-1}XX_G)^{-1}X_G^TX^TW^{-1}Y,$$
$$Cov(\hat{\beta}_G) = (X_G^TX^TW^{-1}XX_G)^{-1}. \tag{8}$$

This equation directly relates the group parameter estimates of interest, $\hat{\beta}_G$, to the full data vector Y, and as such requires this one GLM to be solved for matrices of greatly increased size. If instead we want to use the computationally more efficient approach of combing the first-level parameter *estimates* $\hat{\beta}$ we need to ensure that the estimated group parameter (Eq. (7)) is equivalent to the single-level estimation (Eq. (8)) by accounting for the covariances of the first-level estimates within the second-level (i.e., setting V_{G2} appropriately).

C. Model equivalence

The two-level model specified by Eqs. (1) and (5) is fully equivalent to the single-level model specified by Eq. (3) in terms of both modeling *and parameter estimation* when

$$V_{G2} = V_G + (X^TV^{-1}X)^{-1}. \tag{9}$$

That is, the second-level covariance is set as the sum of the group covariance from the single-level model and the first-level parameter covariance from the two-level model. This can easily be shown using the *Sherman–Morrison–Woodbury* formula (Golub and van Loan, 1989) to write $W^{-1} = V^{-1} - V^{-1}X(V_G^{-1} + X^TV^{-1}X)^{-1}X^TV^{-1}$ and using some more algebraic manipulations. For this choice of V_{G2}, the group-level parameter estimates can be written as

$$\hat{\beta}_G = (X_G^T(V_G + (X^TV^{-1}X)^{-1})^{-1}X_G)^{-1}X_G^T(V_G + (X^TV^{-1}X)^{-1})^{-1}\hat{\beta};$$

that is, they become a function of the first-level parameter estimates $\hat{\beta}$ and their associated covariances $(X^TV^{-1}X)^{-1}$ only.

Note that by simply applying this equivalence result multiple times, the equivalence extends to any multilevel GLM. For example, one can calculate parameter estimates for groups of groups using a multilevel approach by keeping track of only parameter estimates and associated covariances at each level. Note also that parts of this equivalence can simply be obtained by characterizing the necessary conditions on sufficient statistics for β_G (Searle et al., 1992).

This result assumes that all the first-level parameter estimates would be used in the second-level model. This is often not true, as parameters of no interest (confounds) are often present in the first-level to remove unwanted signals (e.g., motion estimates as regressors to remove motion-related artifacts). However, it is easy to show that this result is equally valid in the presence of confounds: the second-level (group) analysis does *not* need to know about the

confounds defined in the first level and requires only the estimates and covariance estimates of the parameters of interest.

Furthermore it is equally easy to show that this model equivalence result also holds when contrasts are used. Contrasts are defined in the linear model in order to take linear combinations of the given parameters that form more meaningful or useful quantities. For instance, the difference between the responses to two different stimuli is often of greater interest than the individual responses, although it is easier to specify the design matrices, X_k, in terms of the individual stimuli. When contrasts are specified at the first level, only the contrasts of parameter estimates (copes), and their covariances, need to be used in the second level. Both can be shown by reformulating the first-level design matrices into two sets of regressors, where the first set relates to the effects of interests and an orthogonal subspace spanning the remaining (confounds) space (Searle et al., 1992).

D. Limitations

In this section we discuss the limitations of the model framework presented above. The first and most obvious limitation is that the framework is purely linear, as it is an instance of the GLM. Linear modeling, however, is a very common and flexible model for FMRI data. Second, in order for the model to be decomposed into two levels, and hence allow for efficient two-level computation, it is necessary that the stacked first-level covariance matrix, V, does not contain any nonzero off-diagonal blocks. This is equivalent to assuming that the first-level residuals are not correlated between sessions or subjects.

Efficient computation at the second-level requires full access to the first-level parameter estimates and associated covariances. This involves both the variances of the parameter estimates and the covariances *between* different parameters. It is not sufficient to use only the first-level statistical parametric maps (i.e., t scores, z scores, or F scores).

III. Estimation of variance components

In the previous sections it was assumed that all variance terms are known a priori. In practice, these quantities are unknown and will need to be estimated as part of the model fitting. Variance component estimation is a challenging task in itself, having generated a variety of approaches. Any approach to variance estimation (or combination of approaches) can easily be combined with the multilevel GLM to provide a practical multilevel method; this section discusses some of the more popular approaches.

There are no differences between the first level and any other level from the modeling perspective; however, in practice, estimating the variances is substantially different. At the first level there typically exists considerable serial autocorrelation (in FMRI time series data) but with a large number of observations. A considerable amount of literature is devoted to specifying the form for the first-level covariance matrix V and estimating its parameters in the single-session case (Purdon et al., 1998; Woolrich et al., 2001; Worsley et al., 2002). In contrast, higher-level variance component estimation is typically troubled by having very few observations, while serial autocorrelation between these normally is, and often can be, ignored.

When the number of observations is very low, this imposes restrictions in the types of model which are practically estimable. For instance, while it is possible to formulate a model where the variance about the group mean is different for each session/subject, such a model is not estimable because there is only a single measurement per session/subject. Several approaches to estimation within the multilevel GLM currently exist, of which the parametric techniques can, most easily, be split into Bayesian and non-Bayesian approaches. Classically, variance components tend to be estimated separately using iterative estimation schemes employing Ordinary Least Squares (OLS), Expectation Maximization (EM), or Restricted Maximum Likelihood (ReML); see Searle et al. (1992) for details.

The specific choice of the variance component estimator will also determine the "effective" degrees of freedom of the variance estimate. For OLS estimates, the degrees of freedom simply is $n - 1$. When more sophisticated estimation techniques and/or variance structures are being used, this changes significantly (Jezzard et al., 2001; Worsley et al., 2002); e.g., when enforcing mixed-effects variance to be greater than the fixed-effects variance, this can increase the effective degrees of freedom, a quantity typically estimated using a Satterthwaite (1946) approximation.

As an example of a non-Bayesian approach (Worsley et al., 2002) estimates variance components at each level of the split-level model separately. First-level estimation incorporates autoregressive AR(p) noise estimated from the lag-l-autocovariance matrices and utilizes the prewhitening approach to generate BLUEs for the first-level parameters. At higher levels, they propose EM for estimation of the random-effects variance contribution, in order to reduce bias in the variance estimation—a potential problem in higher-level analyses if simple OLS were used (note that Marchini and Smith (2003) show that this is not necessary at first-level). Positivity of the random-effects variance, avoiding what is known as the "negative variance problem" (Leibovici, 2001) (where mixed-effects variance estimates are lower than fixed-effects variances implying negative random-effects variance), is partially addressed but not strictly enforced. As a separate stage, in order to boost effective degrees of freedom for later mixed-effects inference, the authors propose to post hoc spatially regularize the random-effects variance via smoothing a random variance ratio image. As a consequence, the resulting analysis does not produce a mixed-effects analysis. It should be noted that this specific form of spatial regularization (or indeed any) is

not a necessary ingredient of an EM approach to variance estimation.

In contradistinction Friston et al. (2002) have proposed an empirical–Bayesian approach for estimation of the full single-level model. Unlike the previous case and the techniques advocated in this article, this relates the parameters of interest to the full set of original data, i.e., does not utilize the summary statistics approach. Parameter and variance component estimation are no longer separated. Conditional posterior point estimates are generated using EM which gives rise to posterior probability maps.

More recently, Behrens et al. (2003) have placed the higher-level GLM estimation in a *fully* Bayesian framework. Using appropriate reference priors, the method is based on Markov–Chain–Monte-Carlo sampling from the full posterior distribution of $p(\beta_G|\hat{\beta})$. Under the parametric model assumptions, the posterior has a noncentral *t*-distribution which the method fits to the MCMC samples. This is done in order to both estimate the appropriate degrees of freedom for the mixed-effects inference, and in order to avoid having to sample densely far into the tail of the posterior. As in the empirical–Bayesian case, parameters and their relevant variance components are estimated together. Also, as the number of degrees of freedom is estimated as part of the *t*-distribution fit, there is no need to separately approximate this quantity, e.g., via the Satterthwaite (1946) approximation. This technique provides for an unbiased and efficient estimation of multilevel GLMs within the summary statistics approach, also strictly enforcing positivity of variance components at all levels. This technique, combined with a prewhitening approach to first-level estimation (Woolrich et al., 2001), has been implemented as part of FSL (Behrens et al., 2003; FSL).

IV. Examples

In this section we show how various group-level parameters of interest can easily be calculated within the GLM framework. This amounts to specifying a suitable group design matrix X_G, a covariance structure V_G, and possibly a contrast vector C_G. Note that unlike the case of first-level designs, the mean parameter value is often of interest and hence the design matrix, X_G, must always model the group mean activation; that is, the unit vector must always be included in the span of X_G.

For several of the examples it is easy to show the added benefit of the proposed framework using numerical simulations. These principally contrast the heteroscedastic model (allowing for different individual first-level variances) with the homoscedastic model (where first-level variances are assumed to be identical). These comparisons show substantial increases in *Z*-statistics over a wide range of realistic scenarios.

A. Average group activation

In order to calculate the average group activation, we model the individual subject activation as being normally distributed according to

$$\beta_k \sim \mathcal{N}(\beta_G, \sigma_s^2),$$

where β_G represents the average group activation and is usually estimated as

$$\hat{\beta}_G = \frac{1}{N}\sum_k \hat{\beta}_k,$$

and where σ_s^2 denotes the between-subject variance.

We will model the first-level within-subject covariances to be subject-specific and model the between-subject variances (from the group mean) as equal across the group. That is,

$$V = \begin{bmatrix} V_1 & & 0 \\ & \ddots & \\ 0 & & V_N \end{bmatrix}, \quad V_G = \sigma_s^2 I,$$

$$\text{and } X_G = (1 \cdots 1)^T.$$

Then the adjusted second-level covariance matrix is

$$V_{G2} = V_G + (X^T V^{-1} X)^{-1} = \begin{bmatrix} (X_1^T V_1^{-1} X_1)^{-1} & & 0 \\ & \ddots & \\ 0 & & (X_N^T V_N^{-1} X_N)^{-1} \end{bmatrix} + \sigma_s^2 \mathbf{I}.$$

Define $u_k = (X_k^T V_k^{-1} X_k)^{-1} + \sigma_s^2$, which is the sum of the within- and between-subject covariances. Then the estimate of the group parameter is

$$\hat{\beta}_G = (X_G^T V_{G2}^{-1} X_G)^{-1} X_G^T V_{G2}^{-1} \hat{\beta} = (\sum_i u_i^{-1})^{-1} (u_1^{-1} \cdots u_N^{-1}) \hat{\beta}, \tag{10}$$

where the inverse sum over u_k^{-1} is the associated variance.

Hence we see that in the general framework, the mean group activation parameter is a weighted average of the combined subject-specific activations, where the weights are inversely proportional to the subject-specific variances. This adjustment is advantageous in the case where the individual time series model does not fit well for a particular subject, *j*, generating an unusual value for $\hat{\beta}_j$ (an outlier) but also a large $\mathrm{Var}(\hat{\beta}_j)$. If no correction for the first-level variance is done, then this outlier can significantly affect the estimation of the group (between-subject) variance. If, however, this first-level correction is performed, the increased variance in this parameter will effectively de-weight the contribution of this outlier to the group variance estimate, since we use Generalised Least Squares estimation.

In the much simpler case, where the within-subject covariances are $V_k = \sigma_w^2 \mathbf{I}$, and the X_k are normalized, such that $X_k^T X_k = 1$ for all *k*, then

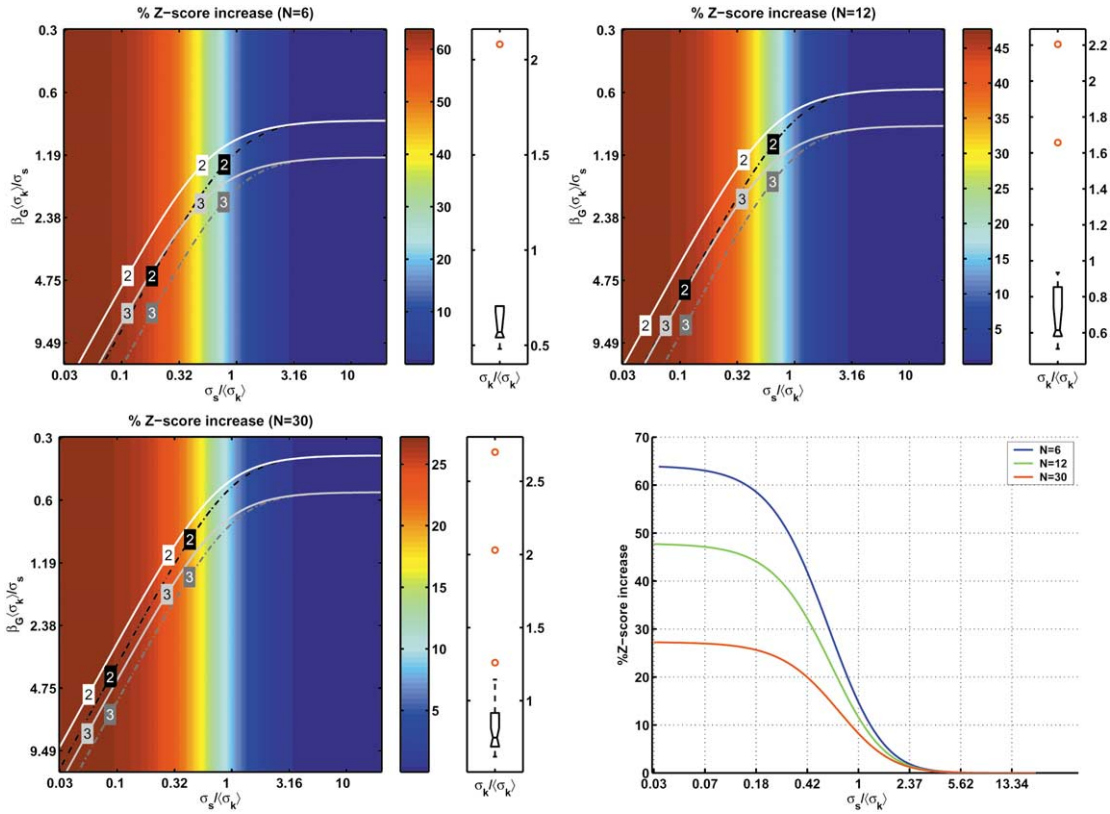

Fig. 1. Expected percentage *Z*-score increase for the mean group activation for the case of typical first level covariance containing a few large outliers with different group sizes ($N = 6$, 12, and 30). The first three images show the increase as a function of the ratio of group level and mean first-level standard deviations ($\sigma_s/\langle\sigma_x\rangle$ on the *x* axis) vs relative group-level activation (*y* axis). Contour lines are shown for group-level $Z = 2.0$ and $Z = 3.0$ in the case of the heteroscedastic model (Eq. (10) solid lines) and the homoscedastic model (OLS; Eq. (11); dashed–dotted). The difference between the solid and dashed–dotted lines corresponds exactly to the color-coded percentage increase. The boxplots show the individual first-level standard deviation compared to their mean ($\sigma_k/\langle\sigma_k\rangle$) as an indicator of heteroscedacity. The percentage changes, which are independent of β_G, are summarized in the fourth plot for all three variance configurations.

$$u_k = \sigma_w^2 + \sigma_s^2 \text{ and}$$

$$\hat{\beta}_G = \frac{1}{N}\sum_k \hat{\beta}_k \text{ with associated variance}$$

$$\mathrm{Var}(\hat{\beta}_G) = \frac{\sigma_w^2 + \sigma_s^2}{N}. \tag{11}$$

The test for significance is then carried out in a *t*-test where

$$T = \hat{\beta}_G / \sqrt{\mathrm{Var}(\hat{\beta}_G)}$$

has an approximate *t*-distribution.

Numerical simulation

In this section we illustrate the potential benefit of a heteroscedastic variance model (i.e., by allowing for separate first-level variances) compared to the homoscedastic variance model (otherwise known as OLS for random-effects analysis of FMRI data (Holmes and Friston, 1998)). Both the heteroscedastic model (Eq. (10)) and the homoscedastic model (Eq. (11)) provide an unbiased estimate of the group-level parameter of interest β_G. They differ, however, in their associated variance, $\mathrm{Var}(\hat{\beta}_G)$, and therefore will give different *Z*- or *T*-statistic. The associated variance for the heteroscedastic model is always less than or equal to that of the homoscedastic model, as can be shown using Jensen's inequality. This will result in an increase in the expected statistics values for the model in Eq. (10):

$$\frac{\langle \Delta Z \rangle}{\langle Z_{\mathrm{homo}} \rangle} = \sqrt{\frac{\langle \mathrm{Var}(\hat{\beta}_G) \rangle_{\mathrm{homo}}}{\langle \mathrm{Var}(\hat{\beta}_G) \rangle_{\mathrm{hetero}}}} - 1.$$

As a quantitative example of this increase we have generated a simulated group FMRI study, with a known set of first-level variances $V_k = \sigma_k^2 \mathbf{I}$, together with a second-level variance σ_s^2. From these known values we calculate the expected percentage increase in *Z*-statistic (see above) and show this versus changes in the ratio of the second-level variance to the mean first-level variance (the assumed variance in the homoscedastic model). First-level variances were taken from a real FMRI group study (simple motor paradigm) estimated using the GLS prewhitening approach as implemented in FILM (part of FSL; (FSL; Woolrich et al., 2001). These estimates of first-level variances are real-

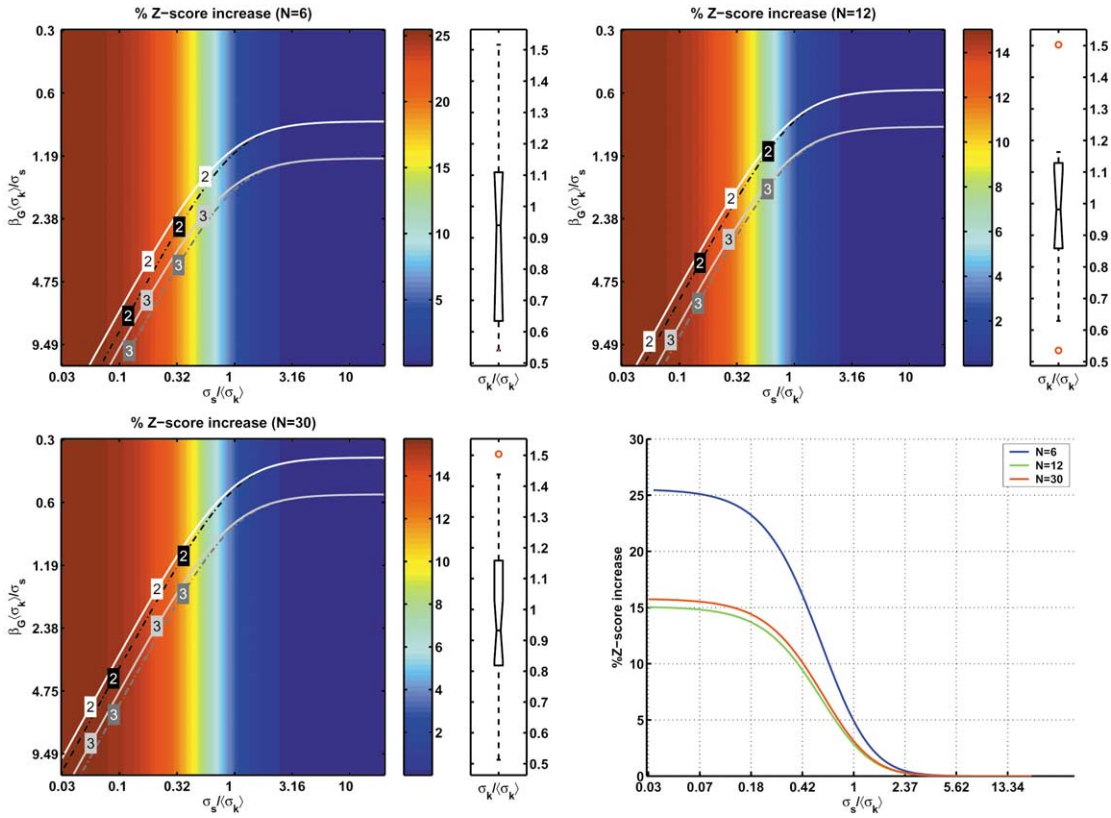

Fig. 2. Expected percentage *Z*-score increase for the mean group activation for the case of typical first level covariance (no large outliers) with different group sizes ($N = 6$, 12, and 30). See Fig. 1 for details.

istic representatives of typical first-level variance structures and are defined to be ground-truth in the following simulation.

Fig. 1 shows the expected percentage increase in *Z* scores for three different sets of first-level variances, each of them containing one or more "outliers" (see boxplots). As can be seen from the first three images, the expected percentage increase, for a given first-level variance structure, depends only on the ratio of second-level variance to mean first-level variance ($\sigma_s/\langle\sigma_k\rangle$) and is independent of the group-level effect size β_G. As the second-level variance becomes considerably larger than the mean first-level variance, the expected increase in *Z*-score tends to zero. When these variances are approximately equal, the heteroscedastic model allows for an ~7–14% increase in *Z*-score, increasing to much larger values as the second-level variance decreases relative to the mean first-level variance. The group-level effect size determines the actual *Z*-level, as shown by the overlaid contour plots (solid lines show *Z*-score levels for the heteroscedastic model while dashed–dotted lines show the *Z*-score levels for the homoscedastic model). Over a large range of variance ratios the improvements gained by the heteroscedastic model have substantial implications on post-thresholded results; e.g., typical subthreshold values of 2.0 increase to superthreshold values of ~3.0.

Fig. 2 shows similar plots for cases where the first-level variances do not contain large outliers; that is, they are close to the homoscedastic model. In this case, the expected percentage *Z*-score increase is smaller, but still potentially important.

It is clear from these figures that the exact configuration of first-level variances has a major impact on the improvements gained by using the heteroscedastic model. This dependency is further investigated in Fig. 3, where we show the expected (log-) increase in *Z*-score for a set of variance

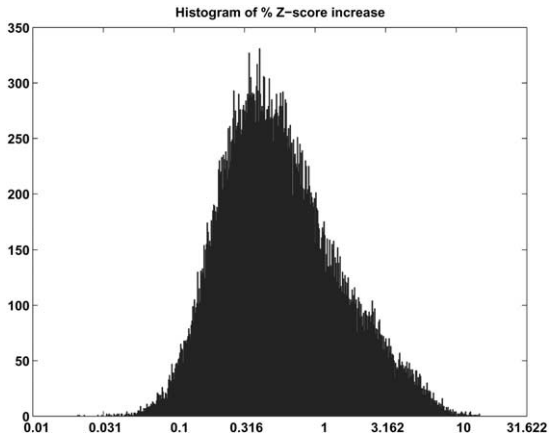

Fig. 3. Histogram of log-percent *Z*-score increase (50,200 voxels, $N = 12$ subjects, $\sigma_s/\langle\sigma_k\rangle = 1$): ~2% of voxels (about 1000) have an increase of >5% in *Z*-score.

configurations estimated from 50,200 voxels in a real group study, assuming equal second-level and mean first-level variances. This histogram shows that approximately 1000 voxels have a >5% increase in expected *Z*-score.

B. Unpaired group difference

We assume that the individual subjects are grouped in two groups and that within each group the first-level parameters are normally distributed around a group-specific mean. That is,

$$\beta_k \sim \mathcal{N}(\mu_1,\sigma_{s_1}^2),\ k \in G_1 \quad \text{and}$$

$$\beta_k \sim \mathcal{N}(\mu_2,\sigma_{s_2}^2),\ k \in G_2.$$

In order to simplify further notation and without loss of generality we assume that the subjects $1,\ldots,r$ belong to the first group and subjects $r+1,\ldots,N$ belong to the second group. We do not make any assumption about the first-level covariance structure and simply set

$$V = \begin{bmatrix} V_1 & & 0 \\ & \ddots & \\ 0 & & V_N \end{bmatrix},\quad V_G = \begin{bmatrix} \sigma_{s_1}^2\mathbf{I}_r & 0 \\ 0 & \sigma_{s_2}^2\mathbf{I}_{N-r} \end{bmatrix},$$

$$X_G = \begin{pmatrix} 1 \cdots 1 & & -1 \cdots -1 \\ 1 & \cdots & 1 \end{pmatrix}^T$$

and $C_G = (2\ 0)^T$.

Then V_{G2} is block diagonal with elements $u_k = (X_k^T V_k^{-1} X_k)^{-1} + \sigma_{sk}^2$. If we define $s_1 = \Sigma_{G1}\, u_k^{-1}$ and $s_2 = \Sigma_{G2}\, u_k^{-1}$, the group parameter estimate writes as

$$\hat{b}_G = C_G^T \left[\begin{pmatrix} \frac{1}{u_1} \cdots \frac{1}{u_r} & \frac{-1}{u_{r+1}} \cdots \frac{-1}{u_N} \\ \frac{1}{u_1} \cdots \frac{1}{u_r} & \frac{1}{u_{r+1}} \cdots \frac{1}{u_N} \end{pmatrix} X_G \right]^{-1} X_G^T V_{G2}^{-1} \hat{\beta}$$

$$= \frac{C_G^T}{4_{s_1 s_2}} \begin{bmatrix} s_1 + s_2 & s_2 - s_1 \\ s_2 - s_1 & s_1 + s_2 \end{bmatrix} \begin{bmatrix} \sum_{G_1} \frac{\hat{\beta}_k}{u_k} - \sum_{G_2} \frac{\hat{\beta}_k}{u_k} \\ \sum_{G_1} \frac{\hat{\beta}_k}{u_k} + \sum_{G_2} \frac{\hat{\beta}_k}{u_k} \end{bmatrix}$$

$$= \left(\sum_{G_1} \frac{\hat{\beta}_k}{s_1 u_k} - \sum_{G_2} \frac{\hat{\beta}_k}{s_2 u_k} \right),$$

where the variance, as usual, is calculated from the first term as

$$\mathrm{Var}(\hat{b}_G) = C_G^T\, (X_G^T V_{G2}^{-1} X_G)^{-1}\, C_G = \frac{1}{s_1} + \frac{1}{s_2}.$$

Under the same assumptions as before, of equal covariance at the first level and normalized designs (i.e., homoscedastic model), these equation simplify to

$$u_k = \sigma_w^2 + \sigma_{sk}^2,\ s_1 = \frac{r}{\sigma_w^2 + \sigma_{s_1}^2},$$

$$s_2 = \frac{N-r}{\sigma_w^2 + \sigma_{s_2}^2},\ \text{and thus}\ \hat{b}_G = \frac{1}{r}\sum_{G_1} \hat{\beta}_k - \frac{1}{N-r}\sum_{G_2} \hat{\beta}_k$$

with

$$\mathrm{Var}(\hat{b}_G) = \frac{\sigma_{s_1}^2}{r} + \frac{\sigma_{s_2}^2}{N-r} + \frac{N\sigma_w^2}{r(N-r)}.$$

Note that the second-level contrast includes an appropriate scaling constant. This factor becomes irrelevant once the group parameter of interest is combined with its variance to form a test statistic.

Numerical simulation

Fig. 4 shows numerical simulation of the expected *Z*-score increase for different first-level variance configurations. As before, the expected increase is independent of the second-level effect size but will depend on the first-level variance configuration for group 1 and group 2 as well as the different second-level variances σ_{s1} and σ_{s2}. The added flexibility of the heteroscedastic model is important for a variety of real FMRI experiments where the two groups naturally will have different variance configurations, e.g., studies of patients vs nonpatients. Once again, significant changes in *Z* score (e.g., >10%) can be seen over a large set of configurations.

C. Repeated measures

In the case of repeated measures, where some or all N subjects were scanned more than once, we might want to account for a possible difference in the between-subject variance and the between-session variance of the same subject. Thus, we simply extend the previous model to write

$$\beta_{kj} \sim \mathcal{N}(\bar{\beta}_k,\sigma_b^2) \quad \text{and} \quad \bar{\beta}_k \sim \mathcal{N}(\beta_G,\sigma_s^2),$$

where β_{kj} denotes the first-level parameter for subject k obtained at session j, β_G is the population mean, $\bar{\beta}_k$ is the subject-specific mean, σ_s^2 is the between-subject variance, and σ_b^2 is the between-session variance. Note that this can be extended to the case of repeated measures under different conditions by introducing more variance terms. The existence of these additional variance terms will be reflected in the covariance matrix V_G simply by allowing for off-diagonal elements. As an example, consider the case of three subjects each measured twice, then with $\sigma_c^2 = \sigma_b^2 + \sigma_s^2$

$$V_G = \begin{bmatrix} \sigma_c^2 & \sigma_s^2 & & & & \\ \sigma_s^2 & \sigma_c^2 & & & & \\ & & \sigma_c^2 & \sigma_s^2 & & \\ & & \sigma_s^2 & \sigma_c^2 & & \\ & & & & \sigma_c^2 & \sigma_s^2 \\ & & & & \sigma_s^2 & \sigma_c^2 \end{bmatrix},$$

where, for simplicity, we order the subjects in pairs appropriately. Hence, we can simply adjust any covariance structure V_G at the group level to take account of the existence of repeated measurements. Alternatively, note that because our previous results extend to multilevel GLMs, we can simply combine the different sessions per subject on a second level and calculate group parameters of interest in a third level.

D. Paired t tests

Let us assume that for each N subject there exist two measurements obtained under different conditions s_1, s_2 and that we are interested in the significance of the mean group difference $b_G = \Sigma_k b_k/N = \Sigma_k(\beta_{ks_1} - \beta_{ks_2})/N$. We will assume that the between-subject variance and the between-session variance is equal across subjects and conditions. Note that this is a notational simplification within this framework which might or might not become a necessary condition once we try to estimate the associated group-level parameters. Similar to the previous sections, we model this as

$$\beta_{ks_i} \sim \mathcal{N}(\mu_i,\ \sigma_b^2 + \sigma_s^2), \text{ and } \mathrm{Cov}(\beta_{ks_1},\ \beta_{ks_2}) = \sigma_s^2.$$

Let

$$V = \begin{bmatrix} V_1 & & 0 \\ & \ddots & \\ 0 & & V_{2N} \end{bmatrix},$$

$$V_G = \begin{bmatrix} U & & 0 \\ & \ddots & \\ 0 & & U \end{bmatrix} \text{ with } U = \begin{bmatrix} \sigma_c^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_c^2 \end{bmatrix} \text{ and}$$

$$X_G = \begin{bmatrix} 1 & 1 & & & \\ -1 & 1 & & & \\ & & 1 & & \\ \vdots & & 1 & & \\ & & & \ddots & \\ 1 & & & & 1 \\ -1 & & & & 1 \end{bmatrix},$$

where, again, $\sigma_c^2 = \sigma_b^2 + \sigma_s^2$ and where the group design matrix, X_G, de-means the first-level estimates for each subject. Assume for simplicity that $V_k = \sigma_w^2\mathbf{I}$, $X_k^T X_k = 1$ and define $u^2 = \sigma_w^2 + \sigma_c^2$. Then V_{G2}^{-1} will be block diagonal with blocks of

$$\tilde{U}^{-1} = \frac{1}{u^4 - \sigma_s^4}\begin{bmatrix} u^2 & -\sigma_s^2 \\ -\sigma_s^2 & u^2 \end{bmatrix},$$

$$\text{where } \tilde{U} = \begin{bmatrix} u^2 & \sigma_s^2 \\ \sigma_s^2 & u^2 \end{bmatrix}.$$

Furthermore, let $c_1 = u^2 + \sigma_s^2 = \sigma_w^2 + \sigma_b^2 + 2\sigma_s^2$ and $c_2 = u^2 - \sigma_s^2 = \sigma_w^2 + \sigma_b^2$. Then

$$X_G^T V_{G2}^{-1} = \begin{bmatrix} c_2^{-1} & -c_2^{-1} & \cdots & c_2^{-1} & -c_2^{-1} \\ c_1^{-1} & c_1^{-1} & & & \\ & & \ddots & & \\ & & & c_1^{-1} & c_1^{-1} \end{bmatrix},$$

$$\text{and } (X_G^T V_{G2}^{-1} X_G)^{-1} = \frac{1}{2}\begin{bmatrix} c_2/N & & & \\ & c_1 & & \\ & & \ddots & \\ & & & c_1 \end{bmatrix},$$

so that

$$\hat{\beta}_G = \frac{1}{2}\left(\sum_k \frac{\hat{\beta}_{ks_1} - \hat{\beta}_{ks_2}}{N}, \hat{\beta}_{1s_1} + \hat{\beta}_{1s_2}, \ldots, \hat{\beta}_{Ns_1} + \hat{\beta}_{Ns_2}\right)^T.$$

Using $C_G = (1, 0, \ldots, 0)^{\mathrm{T}}$, the group parameter estimate writes as

$$\hat{b}_G = C_G^T \hat{\beta}_G = \sum_k \frac{\hat{\beta}_{ks_1} - \hat{\beta}_{ks_2}}{N} \quad \text{and} \quad \mathrm{Var}(\hat{b}_G) = \frac{c_2}{2N}.$$

As expected, the variance of the group-level estimate no longer depends on the between-subject variance σ_s^2. Note that this approach is equivalent to using a three-level approach with an unpaired t test of de-meaned repeated measures.

Numerical simulation

Similar to before, Fig. 5 shows the expected percentage increase in Z score for the heteroscedastic vs the homoscedastic model. Again, depending on the explicit configurations of first-level variances, a substantial increase in Z score can be observed.

E. F tests

Assume that for a set of N subjects we model their hemodynamic response functions using the same set of M basis functions for each subject and consider the case where we wish to test for the population mean activation. Thus, we implicitly assume that there exists a single hemodynamic response function that is representative for the group activation. We will not restrict the choice of basis functions (i.e., we do not require the basis functions to be orthogonal) and therefore allow for general correlations between the individual basis function fits, but assume that the covariance structure is the same for each individual. That is, we model the subject-specific vector of fits as distributed according to a multivariate normal distribution

$$\beta_k \sim \mathcal{N}(\beta_G, V_B),$$

and let

$$V_G = \begin{bmatrix} V_B & & 0 \\ & \ddots & \\ 0 & & V_B \end{bmatrix},$$

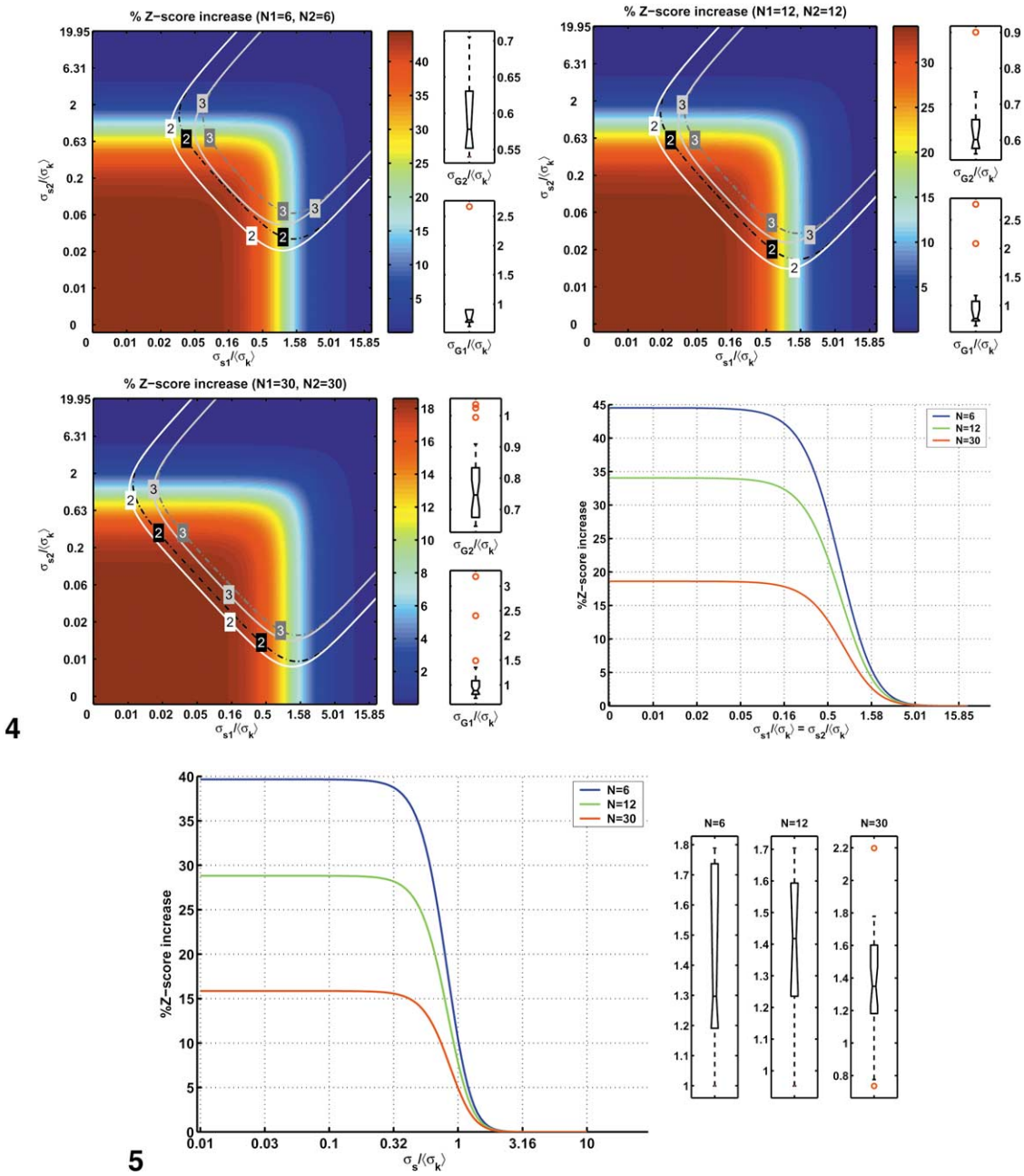

Fig. 4. Expected percentage *Z*-score increase for the unpaired group difference at different group sizes ($N = 6$, 12, and 30). The first three images show the increase as a function of the standard deviation of group 1 and mean first-level standard deviation ($\sigma_{s1}/\langle\sigma_k\rangle$ on the *x* axis) vs the standard deviation of group 2 relative to the mean first-level standard deviation ($\sigma_{s2}/\langle\sigma_k\rangle$ on the *y* axis). Contour lines are shown for the $Z = 2.0$ and $Z = 3.0$ group-level thresholds in the case of the heteroscedastic model (solid lines) and homoscedastic model (OLS; dashed–dotted). The boxplots show the individual first level standard deviation compared to the mean first-level standard deviation ($\sigma_k/\langle\sigma_k\rangle$) as an indicator of heteroscedacity. The expected increases in *Z*-score (assuming $\sigma_{s_1} = \sigma_{s_2}$) are shown in the fourth plot.

Fig. 5. Expected percentage increase in *Z* score for the paired *t* test at different group sizes ($N = 6$, 12, and 30). Boxplots show the standard deviation of the paired differences.

where V_B is the covariance matrix of the M basis function fits. Then the group-level design matrix

$$X_G = [\mathbf{I}_M, \ldots, \mathbf{I}_M]^T$$

combines the M individual basis functions across subjects such that

$$\hat{\beta}_G = (\hat{\beta}_{G_1}, \ldots, \hat{\beta}_{G_M})^T,$$

where the individual values are the M mean basis coefficients. In order to assess the average population activation, we need to test if any of the basis function coefficients are significantly nonzero. This can be achieved by calculating

$$F = \frac{1}{M}\beta_G^T C_G(\widehat{\mathrm{Var}}(C_G\beta_G))^{-1}C_G\beta_G, \text{ with } C_G = \mathbf{I}_M,$$

which approximately follows an F-distribution. If, instead, we wish to assess if the final p basis functions contribute significantly to the mean fit, we simply set

$$C_G = [\ 0 \quad \mathbf{I}_p\ ]$$

and change M to p.

V. Conclusion

In this article we have shown that it is possible to efficiently test, using only first-level parameter and (co-)variance estimates, general hypotheses for a mixed-effects[3] group analysis model within the framework of the multilevel GLM under the BLUE for known variances. In particular, we have demonstrated the equivalence between a multilevel GLM and a single-level version if (and only if) the (co-)variance structure is modified appropriately. The result has a natural interpretation in that the variance used in the second-level analysis must be the sum of the (co-)variances from the individual and group levels. This equivalence allows the second-level analysis to be carried out efficiently, using only first-level results, without any need to revisit the FMRI time series data.

A further consequence of the model equivalence is that we can switch freely between the two forms of the model, depending on which one is more convenient at any one time. Also, the equivalence relation naturally extends to multilevel GLMs, so this framework encompasses more complex analysis scenarios, e.g., where data are grouped first across sessions and later across subjects.

This article deliberately restricts itself to the issues involved in modeling and does not directly address the problem of varience component estimation. The main problem in estimation for group studies in FMRI is the generally small number of subjects/sessions and the associated difficulty in obtaining accurate (co-)variance estimates. This problem will occur irrespective of the specific model used. Naturally, any advanced estimation technique can be used for this multilevel model. These problems of estimation have been investigated separately using a fully Bayesian approach (Behrens et al., 2003). This is implemented within the current release version of FSL and will be the topic of a forthcoming article.

We have provided examples of how various group designs can be formulated within this general framework, including commonly used designs such as paired and unpaired t tests or F tests at the group level. The explicit equations in each of these cases can be derived easily from Eqs. (6) and (9), such that the differences in the exact formulation of these tests and their assumptions become explicit. For example, all results in Section IV have been derived without the need for balanced designs, i.e., they allow for different first-level regressors. Balanced designs, therefore, are not a requirement of the multilevel GLM but potentially are a constraint introduced by a specific estimation technique. This becomes relevant when the first-level regressors need to be subject-specific, e.g., are related to the subject's performance in a cognitive task.

To demonstrate some of the benefits of more flexible mixed-effects modeling some numerical simulations have been generated which contrasted the heteroscedastic model versus the standard homoscedastic (OLS) random-effects model. Using typical first-level variance structures from real FMRI data, these simulations demonstrate that by taking into account the relevant lower-level variances and heterogeneity among them, a substantial increase in higher-level Z-scores is possible.

[3] This naturally includes fixed-effects tests simply by assuming that the group-level (co-)variance matrix V_G is zero.

Acknowledgments

The authors thank Mark Woolrich, Didier Leibovici, Timothy Behrens, and Joe Devlin for the contribution they made by taking part in many helpful discussions and thinking through examples and counterexamples which helped to greatly improve the manuscript. Funding from the UK MRC and EPSRC and GSK is gratefully acknowledged.

References

Behrens, T. Woolrich, M.W. Smith, S.M. 2003. Multi-subject null hypothesis testing using a fully Bayesian framework: theory, in: Ninth Int. Conf. on Functional Mapping of the Human Brain.

Bullmore, E., Brammer, M., Williams, S.C.R., 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.

Cnaan, A, Laird, N.M., Slasor, P., 1997. Tutorial in biostatistics: using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Statist. Med. 16, 2349–2380.

Everitt, B.S., 1995. The analysis of Repeated Measures: a practical review with examples. Statistician 44 (1), 113–135.

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

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

Golub, G.H., Van Loan, C.F., 1989. Matrix Computations, second ed. Johns Hopkins University Press, Baltimore.

Holmes, A.P. Friston, K.J. 1998. Generalisability, random effects & population inference, in: Fourth Int. Conf. on Functional Mapping of the Human Brain, NeuroImage, Vol. 7, p. S754.

Jezzard, P., Matthews, P.M., Smith, S.M. (Eds.), 2001. Functional MRI: An Introduction to Methods. OUP, Oxford.

Leibovici, D.G. 2001. Negative variance problem in fMRI, in: Seventh Int. Conf. on Functional Mapping of the Human Brain.

Marchini, J.L., Smith, S.M., 2003. On bias in the estimation of autocorrelations for fMRI voxel time-series analysis. NeuroImage 18 (1), 83–90.

Matthews, J.N.S., Altman, D.G., Campbell, M.J., Royston, P., 1990. Analysis of serial measurements in medical research. Brit. Med. J. 300, 230–235.

Purdon, P.L. Solo, V. Brown, E. Buckner, R. Rotte, M. Weisskoff, R.M. 1998. fMRI noise variability across subject and trials: Insights for noise estimation methods, in: Fourth Int. Conf. on Functional Mapping of the Human Brain, NeuroImage, Vol. 7, p. S617.

Satterthwaite, F.E., 1946. An approximate distribution of estimates of variance components. Biometr. Bull. 2, 110–114.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

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

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

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