Published in final edited form as:
Neuroimage. 2012 January 2; 59(1): 490–501. doi:10.1016/j.neuroimage.2011.07.077.

Estimating and Testing Variance Components in a Multi-level GLM

Martin A. Lindquist[1,*], **Julie Spicer**[2], **Iris Asllani**[3], and **Tor D. Wager**[4]

[1]Department of Statistics, Columbia University, USA

[2]Department of Psychology, Columbia University, USA

[3]Program for Imaging and Cognitive Sciences, Columbia University, USA

[4]Department of Psychology, University of Colorado, Boulder, USA

Abstract

Most analysis of multi-subject fMRI data is concerned with determining whether there exists a significant population-wide 'activation' in a comparison between two or more conditions. This is typically assessed by testing the average value of a contrast of parameter estimates (COPE) against zero in a general linear model (GLM) analysis. However, important information can also be obtained by testing whether there exist significant *individual differences* in effect magnitude between subjects, i.e. whether the *variance* of a COPE is significantly different from zero. Intuitively, such a test amounts to testing whether inter-individual differences are larger than would be expected given the within-subject error variance. We compare several methods for estimating variance components, including a) a naïve estimate using ordinary least squares (OLS); b) linear mixed effects in R (LMER); c) a novel Matlab implementation of iterative generalized least squares (IGLS) and its restricted maximum likelihood variant (RIGLS). All methods produced reasonable estimates of within- and between-subject variance components, with IGLS providing an attractive balance between sensitivity and appropriate control of false positives. Finally, we use the IGLS method to estimate inter-subject variance in a perfusion fMRI study (N = 18) of social evaluative threat, and show evidence for significant inter-individual differences in ventromedial prefrontal cortex (VMPFC), amygdala, hippocampus and medial temporal lobes, insula, and brainstem, with predicted inverse coupling between VMPFC and the midbrain periaqueductal gray only when high inter-individual variance was used to define the seed for functional connectivity analyses. In sum, tests of variance provides a way of selecting regions that show significant inter-individual variability for subsequent analyses that attempt to explain those individual differences.

Keywords

fMRI; variance components; multi-level GLM; likelihood ratio tests; iterative generalized least squares; restricted iterative generalized least squares

ADDRESS: Martin Lindquist, Department of Statistics, 1255 Amsterdam Ave, 10th Floor, MC 4409, New York, NY 10027, Phone: (212) 851-2148, Fax: (212) 851-2164, martin@stat.columbia.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

INTRODUCTION

Multi-level models have been a mainstay in the social, behavioral, and agricultural sciences for several decades (Harville, 1977; Raudenbush and Bryk, 2002), and they are now gaining in popularity in the neuroimaging community (Friston, et al., 2002; Beckmann et al., 2003; Woolrich et al., 2004). This class of models includes several forms of 'mixed effects' modeling in which data are conceptualized as coming from two or more levels of analysis (e.g., within-person and between-persons). Multilevel models are particularly appropriate for analyzing hierarchically structured data, in which repeated measures are collected on first-level units, such as individual persons, and the analyst wishes to make population inferences about 1st-level, within-unit effects *and* 2nd-level variables that might predict variation across the units. For example, a longitudinal behavioral study might assess the effects of age on test performance, with each individual person tested across 4 successive years. Researchers might be interested in population inference on effects of age at the 1st level, and whether improvements with age are predicted by 2nd-level variables such as educational interventions. Group fMRI data have a similar structure: Task manipulations influence brain activity within-persons (at the 1st level), and these within-person effects are often predicted by 2nd-level explanatory variables such as individual differences in behavioral performance or [Patient vs. control] differences.

Multi-level models are advantageous when data have a multi-level structure setting because they a) allow for valid population inferences on within-subject effects; b) test within-subject effects controlling for sources of variation across individuals, potentially increasing sensitivity; c) consider potential differences in error variance across individuals, often providing more precise estimates of population-level parameters; and d) provide valid and efficient tests when the designs and error variances are different for different individuals (Raudenbush and Bryk, 2002; Pinheiro and Bates, 2000).

Most analyses of multi-subject fMRI data involve two separate models. A first-level General Linear Model (GLM) analysis is performed on each subject's data, which provides within-subjects contrasts across parameter estimates (COPEs; e.g., activity magnitude estimates for [visual stimulation vs. rest]). A second-level analysis provides population inference on whether COPEs are significantly different from zero and assesses the effects of 2nd-level predictors (e.g., group status, behavioral performance). Mixed-effects implementations exist for popular software packages, including FSL (FSL's Linear Analysis of Mixed Effects; Woolrich et al., 2004) and SPM (spm_mfx.m; Friston et al. 2005). These methods can improve estimation accuracy and increase power in some cases, particularly with dramatically unbalanced designs and/or heterogeneous variances across subjects (Mumford and Nichols, 2009).

These existing methods, and virtually all fMRI results of which we are aware, have been designed primarily to estimate and make inferences about group mean COPEs and thus test regional activation levels. However, a real qualitative strength of multilevel models is that they can provide tests of inter-individual *variances* as well as means, and thus provide tests of whether there are true individual differences in brain activity (or brain connectivity, brain-behavior relationships within-subject, etc.). Tests of inter-individual variance could, for example, allow researchers to determine appropriate regions of interest in which to use as seed regions in a subsequent functional connectivity analysis and to test for brain-behavior correlations or between-group differences in patient studies. Such tests are important because correlating behavioral or other variables with fMRI COPEs has become a mainstay of fMRI analysis, and brain-behavior correlations are often taken as stronger evidence than activation alone that brain activity is related to psychological processes of interest. However, brain-behavior correlations are particularly susceptible to several problems that have led to recent criticism of the approach (Vul et al., 2009; cf. Lindquist and Gelman, 2009 and Lieberman et al. 2009). Such correlational analyses are massively underpowered and likely suffer from high false positive rates (Yarkoni, 2009), due primarily to two factors: a) most studies test correlations across tens or hundreds of thousands of brain voxels, and b) correlations are much more sensitive than group means to outliers, and typically require much larger sample sizes to achieve stability. These factors combine to limit sensitivity: Voxel-selection bias will cause observed effect sizes to be larger than true effect sizes (Vul et al., 2009; Lieberman et al. 2009; Lindquist and Gelman, 2009), and violations of assumptions often result in high false positive rates, even for relatively large samples (Wager et al., 2005; Wager at al., 2007; Loh et al., 2008). Tests of variances in mixed-effects models could be used to provide statistical maps of brain regions in which true inter-individual differences are large and reliable, without biasing voxel selection towards correlation with any particular behavioral measure (Kriegeskorte et al., 2009) as the estimated variances are independent of the estimated regression parameters used to compute the correlations (e.g., Neter et al., 1996). Thus, the number of multiple comparisons tested in brain-behavior correlation analyses could be reduced from thousands of voxel to a few regions of interest [ROIs], reducing both the need for multiple-comparisons correction and effect-size inflation due to voxel selection.

Significant inter-subject variability can alternatively be driven by other sources of between-subjects error, such as individual differences in gray-matter density or inter-subject alignment. The latter might be important particularly at the boundaries between anatomical structures, as even the best nonlinear registration methods cannot perfectly align all these boundaries. Hence, tests of inter-subject variability also have the potential to be an important tool for diagnosing problems with inter-subject alignment and preprocessing in general.

It is important to contrast tests of variance components with the standard approach of testing for significant group differences between conditions. Tests of group means tell us whether the mean is significantly different from 0 in the population, not whether there are individual differences between subjects. In fact, such differences may exist even if the mean is 0 on average in the population. In contrast, tests of the variance components allow us to directly determine whether there exist individual differences in the means and identify regions correlated with individual differences in behavior or showing differences in inter-subject alignment.

This paper describes the statistical development of a multi-level model fit using iterative generalized least squares (IGLS) and its implementation in Matlab software. We first describe the model and several candidate procedures for making inferences about both means and variances. We also describe why, though the implementation of variance tests initially appears straightforward, the problem is more subtle than is apparent at first glance. Next, we use simulations to validate that a) estimates are unbiased, and b) inferential tests control false positive rates at the appropriate level. Third, we compare sensitivity to true effects and false positive rates for two variants of the model (IGLS and its restricted-maximum likelihood cousin, RIGLS), the standard two-stage ordinary least squares (OLS) approach typical in fMRI studies, and Linear Mixed Effects in R (LMER), a model widely considered a leading standard. Finally, we apply the IGLS model to mapping inter-individual differences in a perfusion fMRI study of social evaluative threat (SET), a robust psychological stressor whose effects vary considerably across individuals.

METHODS

Model formulation: A multi-level model for fMRI

We begin by setting up the standard multi-level model used in fMRI data analysis. Suppose we are performing a GLM analysis on *m* separate subjects using the model:

$$\boldsymbol{Y}_i=\boldsymbol{Z}_i\boldsymbol{\beta}_i+\boldsymbol{\varepsilon}_i \quad \boldsymbol{\varepsilon}_i\sim N(\mathbf{0},\mathbf{V}_i) \quad \text{for } i=1,\ldots..m \qquad [1]$$

Here $\mathbf{Y}_i$ is a vector of length *T* representing the fMRI time series data for subject *i* and $\mathbf{Z}_i$ is a subject specific design matrix of size *T* × *p*. The subscript *i* will always refer to subject, and subscript G to group or population effects. For simplicity, we assume that the data have been pre-whitened prior to analysis and that the within-subject covariance matrix can be expressed as $\mathbf{V}_i=\mathbf{I}_T\sigma_i^2$ where $\mathbf{I}_T$ represents the *T* × *T* indicator matrix and σ_i^2 is a subject specific variance; autocorrelation could easily be accommodated, however, without affecting the current results.

If using the OLS approach (often called 'random effects' in the fMRI literature because subject is treated as a random effect), the analyst would then subject the estimates $\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_m$ to a second-level analysis (without covariates, this would be a one-sample t-test). Alternatively, in a multi-level formulation, the data from all *m* subjects is combined into a single GLM as follows:

$$\boldsymbol{Y}=\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon}\sim N(\mathbf{0},\mathbf{V}) \qquad [2]$$

where $\boldsymbol{Y}=(\boldsymbol{Y}_1^T,\ldots,\boldsymbol{Y}_m^T)^T, \boldsymbol{\beta}=(\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_m^T)^T, \boldsymbol{\varepsilon}=(\boldsymbol{\varepsilon}_1^T,\ldots,\boldsymbol{\varepsilon}_m^T)^T$ and $\mathbf{Z}$ and $\mathbf{V}$ are block-diagonal matrices with blocks $\mathbf{Z}_\mathrm{i}$ and $\mathbf{V}_\mathrm{i}$, respectively, on the main diagonals. This arrangement is shown in graphical form in Figure 1.

Next, we assume that the $\boldsymbol{\beta}_i$ values for each subject *i* are a random draw from a distribution centered on $\boldsymbol{\beta}_G$, following a $N(\boldsymbol{\beta}_G, \mathbf{U}_G)$ distribution. The population covariance matrix can be expressed as

$$\mathbf{U}_G=\begin{pmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \cdots & \sigma_{1,p}^2 \\ \sigma_{2,1}^2 & \sigma_{2,2}^2 & & \\ \vdots & & \ddots & \\ \sigma_{p,1}^2 & & & \sigma_{p,p}^2 \end{pmatrix} \qquad [3]$$

Here $\sigma_{j,k}^2$ represents the covariance between the j^th^ and k^th^ element of $\boldsymbol{\beta}_i$(i.e., correlations between first-level parameter estimates). If $\mathbf{U}_\mathrm{G}$ is diagonal, within-subject effects are uncorrelated. Using this notation we can formulate our full multi-level model as follows:

$$\begin{aligned} \boldsymbol{Y}&=\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{\varepsilon} & \boldsymbol{\varepsilon}&\sim N(\mathbf{0},\mathbf{V}) \\ \boldsymbol{\beta}&=\boldsymbol{Z}_G\boldsymbol{\beta}_G+\boldsymbol{\eta} & \boldsymbol{\eta}&\sim N(\mathbf{0},\mathbf{U}) \end{aligned} \qquad [4]$$

The deviation $\boldsymbol{\eta}$ reflects the person-level deviation from the group parameter values $\boldsymbol{\beta}_G$. $\mathbf{U}$ is a block-diagonal matrix representing the between-subject variation in the $\boldsymbol{\beta}$ parameters for each individual, such that $\mathbf{U} = \mathbf{I}_m \otimes \mathbf{U}_G$, where the symbol ⊗ represents the Kronecker product. In other words, $\mathbf{U}$ is the tensor product between an *m* × *m* identity matrix and the group covariance matrix $\mathbf{U}_G$. Because $\mathbf{U}$ is block diagonal, the subjects are assumed to be independent of one another. $\mathbf{Z}_G$ represents the second level design matrix (e.g. separating cases from controls; see Figure 1). In addition, Table 1 provides a short description of all the vectors and matrices defined throughout the paper and their sizes.

As a final step, we can re-express the model described in [4] in single-level format as follows:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{Z}\boldsymbol{Z}_G\boldsymbol{\beta}_G + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}\boldsymbol{\beta}_G + \boldsymbol{\zeta} \end{aligned} \qquad [5]$$

where $\mathbf{X} = \mathbf{Z}\mathbf{Z}_G$, the combined individual and group design matrices, and $\boldsymbol{\zeta} = \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, the combined error vector including within-subject and between-subject components. We denote the covariance matrix of $\boldsymbol{\zeta}$ as

$$\begin{aligned} \sum &= \boldsymbol{Z}Cov(\boldsymbol{\eta})\boldsymbol{Z}^T + Cov(\boldsymbol{\varepsilon}) \\ &= \mathbf{Z}\mathbf{U}\mathbf{Z}^T + \mathbf{V}. \end{aligned} \qquad [6]$$

Here the first term captures the between-subjects covariance, and the second term ($\mathbf{V}$) the within-subjects covariance.

Model Estimation: Iterative Generalized Least Squares (IGLS)

The first goal of the analysis is to obtain maximum likelihood estimates of $\boldsymbol{\beta}_G$, as well as of the unknown variance components contained in $\mathbf{U}$ and $\mathbf{V}$ (the between-subjects and within-subjects variance components, respectively). The IGLS method estimates these, under the assumption the data are multivariate normal (Goldstein, 1986; Browne, 1984), by constructing a second linear model whose unknown parameters are the between-subject ($\sigma^2_{j,k}$) and within-subject (σ^2_i) variances.

To estimate these variances in a linear modeling framework, we begin by expanding Eq. [6], re-expressing the combined covariance matrix Σ as a linear combination of the individual between-subjects variance components $\sigma^2_{j,k}$ and the residual variances σ^2_i, which are embedded in $\mathbf{U}$ and $\mathbf{V}$. This will allow us to subsequently formulate a general linear model with unknown parameters to be estimated for $\sigma^2_{j,k}$ and σ^2_i. We first define $\mathbf{H}$, a matrix that selects the elements of $\mathbf{U}$ and $\mathbf{V}$ corresponding to individual variance parameters (e.g., $\sigma^2_{1,1}$). Let $\mathbf{H}_{jk}$ be a $p \times p$ indicator matrix which is 0 in every element except the $(j, k)^{th}$ where it equals 1 (e.g., $j = 1$ and $k = 1$ for $\sigma^2_{1,1}$). We can now write Σ in the following manner:

$$\sum = \sum_{j=1}^{p}\sum_{k=1}^{p}\sigma^2_{j,k}\boldsymbol{Z}(\mathbf{I}_m \otimes \mathbf{H}_{jk})\boldsymbol{Z}^T + \sum_{i=1}^{m}\sigma^2_i(\mathbf{H}_{ii} \otimes \mathbf{V}_i) \qquad [7]$$

Note that since $\sigma^2_{j,k}=\sigma^2_{k,j}$ the model generally contains a total of $p(p + 1)/2 + m$ unknown variance components that need to be estimated. Figure 2 shows a pictorial representation of Eq. [7] for the special case with two within-subject regressors (i.e., $p = 2$), a random intercept and slope model. Here we define

$$\boldsymbol{\beta}_G^T=\begin{pmatrix}\beta_1 & \beta_2\end{pmatrix},\ \mathbf{U}_G=\begin{pmatrix}\sigma_{1,1}^2 & 0\\ 0 & \sigma_{2,2}^2\end{pmatrix}\text{ and }\mathbf{V}_i=\sigma^2\mathbf{I}_T.$$

This implies that the regression parameters are uncorrelated (off-diagonals of $\mathbf{U}_G$ are 0), the data is pre-whitened, and the within-subject variance is constant across subjects ($\sigma_i \equiv \sigma$ for all i). In this restricted case, there are only $p + 1$ unknown variance parameters ($\sigma_{1,1}^2, \sigma_{2,2}^2$ and σ^2) that combine additively to yield Σ, as shown in Figure 2. If within-subject variances were allowed to differ, as is the default in practical data analysis and in our simulations, the last term in Figure 2 would be expanded into separate, additional terms for each subject's error variance. Similarly, additional terms would typically be included to model covariance between first-level regression parameter estimates.

The next step in IGLS is to formulate the linear model that estimates the variance components. This is done based on Eq. [7], by vectorizing the lower triangular and diagonal elements of the matrix corresponding to each unknown variance term. This can be done using the *vech* operator, which when applied to a matrix stacks its columns after removing all supra-diagonal elements (e.g., Harville, 1997). Using this notation, the summands in Eq. 7 can be written $vech(\mathbf{Z}(\mathbf{I}_m \otimes \mathbf{H}_{1,1})\mathbf{Z}^T)$ and $vech(\mathbf{H}_{ii} \otimes \mathbf{V}_i)$. These become regressors in the new design matrix $\mathbf{X}^*$ (where * will be used to indicate the linear model for variance components, following the notation of Goldstein, 1986). The response variable $\mathbf{Y}^*$ in this model is based on the residual covariance matrix $\mathbf{R} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)^T$, so that $\mathbf{Y}^* = vech(\mathbf{R})$. $\mathbf{Y}^*$ and X^* are used to estimate the variance components $\beta^*=\begin{pmatrix}\sigma_{1,1}^2 & \cdots & \sigma_{p,p}^2 & \sigma_1^2 & \cdots & \sigma_m^2\end{pmatrix}$, and thus predict the combined covariance matrix $\hat{\boldsymbol{\Sigma}}$**.**

Now that we have a model for estimating $\boldsymbol{\beta}_G$ (Eq. [4]) and a model for estimating $\boldsymbol{\Sigma}$ ($\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*$ based on Eq. [7]), the IGLS procedure alternates between estimating $\boldsymbol{\beta}_G$ and $\boldsymbol{\beta}^*$ until convergence. The specific steps are as follows:

1. Start with OLS estimates of the covariance, i.e., $\hat{\boldsymbol{\Sigma}} = \mathbf{I}$.

2. Estimate fixed effects. Use the current estimate of $\hat{\boldsymbol{\Sigma}}$ to calculate $\hat{\boldsymbol{\beta}}_G$ via the standard generalized least squares solution: $\hat{\boldsymbol{\beta}}_G = (\mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}$.

3. Estimate variance components. Use the residuals to form $\mathbf{Y}^*$ and update $\hat{\boldsymbol{\beta}}^*$, estimates for the within- and between-subjects variance components, again using generalized least squares. Following Goldstein (1986), the covariance of the variance components can be shown to be: $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}$. This result, which follows from the theory of the matrix normal distribution, holds if and only if the data are multivariate normal or if the sample variance matrix is Wishart distributed (Bilodeau & Brenner, 1999). Thus, the GLS solution is given by $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*T}\boldsymbol{\Sigma}^{*-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\boldsymbol{\Sigma}^{*-1}\mathbf{Y}^*$. If negative estimates are obtained for the variance components $\sigma_{j,j}^2$ and σ_i^2, they are truncated at 0. However, covariance terms ($\sigma_{j,k}^2$, $j \neq k$) are allowed to be negative. Re-form $\hat{\boldsymbol{\Sigma}}$ from $\hat{\boldsymbol{\beta}}^*$.

4. Repeat steps 2–3 until convergence.

Model Estimation: Restricted Iterative Generalized Least Squares (RIGLS)

As with all MLEs of variance components, the results will be biased due to the fact that they are estimated with reduced degrees of freedom because they are conditioned on $\hat{\boldsymbol{\beta}}_G$. This problem will be more severe when dealing with small sample sizes. By making a simple modification to the IGLS algorithm, we can instead obtain restricted maximum likelihood

(ReML) estimates (Goldstein, 1989) using a procedure called restricted iterative generalized least squares (RIGLS). These estimates are unbiased since they take into consideration the loss in degrees of freedom resulting from the estimation of the regression parameters and are the preferred estimates for variance components. The modification of the algorithm involves altering the term on the left hand side of Eq. [7] as follows:

$$\sum - \mathbf{X}\left(\mathbf{X}^T \sum^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T = \sum_{j=1}^{p}\sum_{k=1}^{p}\sigma_{j,k}^2\mathbf{Z}(\mathbf{I}_m \otimes \mathbf{H}_{jk})\mathbf{Z}^T + \sum_{i=1}^{m}\sigma_i^2(\mathbf{H}_{ii}\otimes \mathbf{V}_i) \qquad [8]$$

Proceeding in an analogous manner as in the IGLS algorithm, while using this modified equation, gives us the unbiased RIGLS estimates.

Statistical inference: Hypothesis tests for variance components

Once model parameters are estimated, it is desirable to make inferences about the likely population values of estimated variance components. Inference on fixed effects (i.e., within-subject effects of time or condition in the examples above) can be performed in the traditional manner within the GLM framework, by calculating either t or F statistics on linear combinations of the elements in the vector $\boldsymbol{\beta}_G$ (Lindquist, 2008). However, inferential procedures for testing the statistical significance of the variance components contained in the vector $\boldsymbol{\beta}^*$ are less straightforward. There are several types of hypothesis we may be interested in testing. For example, in the case of a single between-subject variance component $\sigma_{1,1}^2$ we may want to test the hypothesis $H_0{:}\sigma_{1,1}^2=0$ against the alternative that it is a non-negative scalar. That is, we may want to test whether the variance is significantly different from zero. Alternatively, in the case where multiple variance components are included in the model, we may want to test the hypothesis:

$$H_0{:}\mathbf{U}_G=\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} \qquad [9]$$

where $\mathbf{D}$ is a $(p-1)\times(p-1)$ positive definite matrix, against the alternative hypothesis that $\mathbf{U}_G$ is a general $p \times p$ positive definite matrix. That is, we may want to test whether a single between-subject variance component is statistically different from zero, conditional on estimates of the other variance components. For example, suppose that

$$\boldsymbol{\beta}_G^T=\begin{pmatrix} \beta_1 & \beta_2 \end{pmatrix} \text{ and } \mathbf{U}_G=\begin{pmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 \\ \sigma_{2,1}^2 & \sigma_{2,2}^2 \end{pmatrix}.$$

In order to test whether there is significant variation attributable to the parameter β_2, we can use the hypothesis stated in Eq. [9] with $\mathbf{D}=\sigma_{1,1}^2$.

The Likelihood Ratio Test (LRT) and restricted LRT (RLRT)

The Likelihood Ratio Test (LRT) and restricted LRT (RLRT)—In each of the situations outlined above we are interested in testing a null hypothesis H_0, with parameters $(\hat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}_0^*})$, against an alternative H_A that involves fitting additional parameters, $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}^*)$. That is, the null model is nested within the alternative (full) model. A likelihood ratio test can be derived for comparing nested models with different covariance structures. The likelihood ratio test statistic can be written:

$$\begin{aligned} LRT &= -2\log\left(\frac{L_{ML}(\widehat{\boldsymbol{\beta}}_0,\widehat{\boldsymbol{\beta}}_0^*)}{L_{ML}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}^*)}\right) \\ &= -2(\log(L_{ML}(\widehat{\boldsymbol{\beta}}_0,\widehat{\boldsymbol{\beta}}_0^*)) - \log(L_{ML}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}^*))) \end{aligned} \qquad [10]$$

Here L_{ML} denotes the likelihood function where

$$\log(L_{ML}(\boldsymbol{\beta},\boldsymbol{\beta}^*)) \propto -\frac{1}{2}\log\left|\sum(\boldsymbol{\beta}^*)\right| - \frac{1}{2}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta})^T\sum(\boldsymbol{\beta}^*)^{-1}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}) \qquad [11]$$

and the parameters $(\hat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_0^*)$ and $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}^*)$ are the MLEs obtained from maximizing L_{ML} under the null and full model, respectively. These values can be obtained using the IGLS algorithm.

If using RIGLS, a valid restricted likelihood ratio test (RLRT) for the variance components can be performed using ReML estimation. Here the general format of the test statistic remains the same, but the term L_{ML} is replaced by the ReML likelihood function (Lindstrom and Bates, 1988) $L_{\mathrm{Re}\,ML}$ where:

$$\log(L_{\mathrm{Re}ML}(\boldsymbol{\beta},\boldsymbol{\beta}^*)) \propto -\frac{1}{2}\log\left|\mathbf{X}^T\sum(\boldsymbol{\beta}^*)^{-1}\mathbf{X}\right| - \frac{1}{2}\log\left|\sum(\boldsymbol{\beta}^*)\right| - \frac{1}{2}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta})^T\sum(\boldsymbol{\beta}^*)^{-1}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}) \qquad [12]$$

In order to perform the RLRT one must also assume that the fixed effects do not vary across the two models and only allow the variance components to differ. The appropriate estimates of the variance components can be estimated using the RIGLS algorithm.

Distributions for the LRT and RLRT: chi-square and chi-square mixtures— Standard statistical theory states that the asymptotic distribution of the likelihood ratio statistic is distributed as χ_k^2 where k is the difference in the number of parameters included in the full and reduced models. However, this result only holds under certain regularity assumptions, one being that H_0 does not lie on the boundary of parameter space. Since variances are by definition non-negative, the H_0 value of zero will indeed lie on the boundary, invalidating the distributional assumptions of the LRT. It can be shown that under certain conditions (described below) the asymptotic null distribution for the likelihood ratio test is instead a mixture of chi-square distributions (Self & Liang, 1987; Stram & Lee, 1994). When testing the hypothesis $H_0{:}\sigma_{j,j}^2=0$ against the alternative that it is a non-negative scalar, they suggest defining the limiting distribution as a mixture of a χ_1^2 distribution with a χ_0^2 distribution with equal weight 0.5. Here a χ_0^2 distribution is defined as having probability mass 1 at the value 0. The intuition for this result becomes clear if we assume that $\sigma_{j,j}^2=0$. In this situation the probability is 0.5 that this expression would have been negative had this been allowed and therefore truncated at 0. The null distribution is therefore a mixture with probability 0.5 taking the value 0 and 0.5 taking the standard χ_1^2 distribution. When testing the hypothesis in Eq. [9], the asymptotic variance will instead be a mixture of a χ_p^2 distribution with a χ_{p-1}^2 distribution with equal weights of 0.5. While both these results were originally derived for the maximum likelihood case, they have also been shown to hold for the ReML case (Morrell, 1998).

One caveat is that these results are asymptotic large-sample approximations (i.e., the number of subjects are assumed to approach infinity). Crainiceanu and Ruppert (Crainiceanu & Ruppert, 2004) provide an alternative, simulation-based method that produces more accurate results for a single variance component. However, similar results are not easily obtainable for models with multiple variance components, which is the scenario most likely to occur while analyzing fMRI data. Thus, we assessed bias, power, and false positive rates for the chi-square mixture with multiple variance components, to assess its validity and usefulness as an inferential test.

Simulations

We assessed the validity and efficiency of our approach for estimating and testing the significance of variance components in a multi-level GLM. Simulation 1 compares the bias and efficiency of OLS, IGLS, RIGLS, and LMER (in the R software) for estimating both fixed effects and variance components. Simulation 2 assesses the use of the mixture of chi-square statistics as the limiting distribution of the likelihood ratio test under the null hypothesis. Finally, Simulation 3 studies the efficiency and power of the likelihood ratio test for determining the significance of variance components under the alternative hypothesis.

In all simulations, the response for each subject was assumed to be $y_i = \beta_{0i} + \beta_{1i}x_i + \varepsilon_i$ where x_i is an indicator function of length 200 time units, taking the values 1 at time points [1, 41, 81, 121, 161] and 0 otherwise, convolved with a canonical hemodynamic response function (Friston et al., 1998). Both the intercept and slope were random draws from populations with distributions $\beta_{0i} \sim N(1.5, \sigma^2_{\beta_0})$ and $\beta_{1i} \sim N(3, \sigma^2_{\beta_1})$, respectively. Finally the error term consisted of independent and identically distributed normal random variables, i.e. $\varepsilon_i \sim N(0, \sigma_i^2)$. Though we chose particular values for these design and variance parameters, the conclusions from these simulations are not expected to depend on these particular analysis choices.

Simulation 1—Simulated data were generated for 20 subjects ($m = 20$) for two different event-related fMRI designs (though these results apply equally to multi-level designs from any field). In Simulation 1a, the true within-subject standard deviation was $\sigma_i = 1$ for all subjects. In Simulation 1b, it was a random draw from a population with a chi-square distribution with 1 degree of freedom. In both simulations the value of $\sigma^2_{\beta_0}$ was set to either 0 or 0.4, and the value of $\sigma^2_{\beta_1}$ was set to either 0 or 0.5, with all combinations of these values tested. These values were chosen to correspond with results observed in real data (Mumford and Nichols, 2008).

We fit the data using OLS (the ordinary least squares approach typically employed by the neuroimaging community), IGLS, RIGLS, and LMER, from the statistical software package R (R v.2.11.1, GNU General Public License). LMER is a generic function for fitting a linear mixed-effects model using the EM-algorithm, following the framework of Lindstrom and Bates (1988). The OLS procedure begins by fitting individual regression coefficients for each subject using Eq. [1]. Thereafter group estimates of the slope and intercept are obtained by averaging across subjects. Group-level variance components are obtained by computing the variance of the estimates across subjects. Since this analysis is performed on the estimated regression coefficients, the variance will contain contributions from both the standard error of the estimates and the between-subject variance components. Using the standard errors estimated in the first-level of the analysis, the within-subject variance components can be computed by taking the difference between the variance of the estimates and the mean standard error. If the values are negative, they are set to zero. This method is included as it is the most popular method in the neuroimaging community for estimating the parameters of a mixed-effects model. Though it has been shown to be effective for estimating the slope and intercept (Mumford and Nichols, 2010), its efficacy in estimating variance components has not yet been explored to date.

1,000 iterations of dataset generation and fitting all models were performed for each of the four unique values for σ_β^2 for each of Simulations 1a and 1b. We assessed whether each model yielded estimates equal to the true values, on average (i.e., was unbiased) and compared the method-related variance by comparing the variances in parameter estimates across models. As a final step, to access the methods ability to handle model misspecification we repeated the analysis of the data from Simulation 1a using an incorrect first-level design matrix. In Simulation 1c we incorrectly assumed that x_i was an indicator function of length 200 time units, taking the values 1 at time points [1, 81, 161] and 0 otherwise, convolved with a canonical hemodynamic response function. This was done to study how robust the methods are to potential model misspecification.

Simulation 2—In Simulation 2 we sought to study the distributions of the LRT and RLRT statistics under two different types of null hypotheses. In Simulation 2a both variance components $\sigma_{\beta_0}^2$ (corresponding to the intercept) and $\sigma_{\beta_1}^2$ (corresponding to the slope) were set to 0, and all other values are set as outlined above. We tested a random slope model, in which the intercepts are assumed to be fixed (the same for all subjects) and we are interested in testing whether the value of the slope differed across subjects (i.e. $H_0{:}\sigma_{\beta_1}^2=0$). For IGLS analyses, the LRT was computed for each of 10,000 iterations (more iterations were included to obtain stable P-values in the tails of the distribution) using the standard χ_1^2 distribution, the 50:50 mixture of χ_1^2 and χ_0^2 distributions, and the simulation-based approximation of Crainiceanu and Ruppert (Crainiceanu & Ruppert, 2004). For RIGLS analyses, the same tests were performed, yielding RLRT tests.

In Simulation 2b, $\sigma_{\beta_0}^2$ was set to 0.4 and $\sigma_{\beta_1}^2$ was set to 0. This simulation gives us null-hypothesis data for the case when the intercept is random and we are interested in testing for a significant random slope. This corresponds to a random intercept, random slope model, in which both intercept and slope are modeled as random effects and variances are estimated. As above, we test the null hypothesis that the between-subject variance of the slope is zero:

$$H_0{:}\mathbf{U}_G=\begin{pmatrix} \sigma_{\beta_0}^2 & 0 \\ 0 & 0 \end{pmatrix}$$

The alternative hypothesis is that $\mathbf{U}_G$ is a general positive definite matrix. The likelihood ratio statistic for testing this hypothesis is calculated at each of 10,000 iterations. In standard likelihood ratio testing the commonly used limiting distribution for the statistic would be the χ_2^2 distribution, as there are two additional parameters in the full model ($\sigma_{1,2}^2$ and $\sigma_{2,2}^2$). The resulting empirical distribution of the statistic is compared with a χ_2^2 distribution, a χ_1^2 distribution, and a 50:50 mixture of the two. Note that in this situation, an approximation of the finite null distribution of the LRT does not exist. The simulations are repeated using RIGLS, to yield RLRT tests.

Simulation 3—In the third simulation the efficiency and power of the LRT for determining the significance of variance components was studied in a variety of situations. The general outline for the simulations was equivalent to those described above with slight alterations. In Simulation 3a, the between-subject variance components ($\sigma^2_{\beta_0}, \sigma^2_{\beta_1}$) are set to the fixed values (0, 0) and the value of the within-subject variance σ was allowed to vary from 0.2 to 4. This corresponds to within-subject signal-to-noise ratios (SNR, equivalent to Cohen's *d*) ranging from 0.25 to 5. For each value, 1,000 repetitions are simulated. For each repetition both IGLS (and RIGLS were fit and LRT/RLRTs were performed (using the 50:50 mixture and the Crainiceanu and Ruppert method) on the inter-subject variance of the slope. For each value of σ we estimated the false positive rate at the nominal α =0.05. In Simulation 3b, the simulation was repeated using a fixed value of $\sigma = 1$ and allowing the number of "subjects" included in the study to vary between 2 and 36. In Simulation 3c, we assessed efficiency under the alternative hypothesis by letting $\sigma^2_{\beta_1}$ vary between 0.05 to 0.55 in steps of 0.10 (m = 20 subjects and $\sigma = 1$, 1000 iterations).

In Simulations 3d–3f, we generated and fit data according to the random intercept, random slope model (with $\sigma^2_{\beta_0}=0.5$) and tested the hypothesis that $\sigma^2_{\beta_1}=0$ conditional on estimates of the intercept. In Simulations 3d–3e, we tested false positive rates ($\sigma^2_{\beta_1}=0$, with SNR varying as in Simulation 3a and *m* varying as in Simulation 3b), and in Simulation 3f, we tested efficiency ($\sigma^2_{\beta_1}$ varied as in Simulation 3c).

Experimental Data

The method was applied to cerebral blood flow (CBF) measures collected with continuous arterial spin labeling (CASL) fMRI while participants (n=18) performed a social evaluative threat (SET) task, allowing us to detect regions exhibiting significant individual differences. CBF images were computed as described in detail in Asllani et al. (Asllani et al. 2009). Briefly, preprocessing was implemented using SPM software (Wellcome Department of Cognitive Neurology) and other in-house code written in MATLAB (Mathworks, Natick MA). For each subject, images were preprocessed as follows: (1) all EPI images were realigned to the first acquired. (2) GM, WM, and CSF posterior probability images were obtained from SPGR image using SPM99's segmentation algorithm. (3) The SPGR and posterior probability maps were co-registered to the first acquired EPI using the mutual information co-registration algorithm. (4) An analysis mask was made for each subject by summing subject's posterior probability images; only voxels within this mask were included in the analysis. (5) SPGR and average CBF images were transformed into the Talairach standard. The spatially normalized control/label pairs were used to calculate percent change maps, which were subsequently used to compute CBF using the two-compartment formula derived by Alsop & Detre (Alsop and Detre, 1996) and later modified by Wang et al. (Wang et al., 2005).

The SET task was administered using a variant of the Trier paradigm (Kirschbaum et al., 1993) and 8 blocks of CASL images were acquired in the following order: Pre-stress Baseline (5.3 minutes), Practice Math (5.3 min), Fun Math (5.3 min), Anticipation (2.7 min) Speech Preparation (2.4 min), Stress Math (5.3 min), Recovery (2.7 min) and Post-stress Baseline (5.3 min). Following Anticipation and prior to the Speech Preparation block, participants were introduced to two confederates posing as professors over what they were told was a live video feed (they remained in the same position in the scanner). The confederates were paid actors, and they instructed participants to silently prepare an 8 minute speech in a period of 3 minutes, a speech that would be delivered to the professors after the scanning portion of the study. Participants were then told that the speech topic would be their strengths and weaknesses as a candidate for their dream job, and the Speech Preparation period began immediately thereafter. Before the Recovery block, participants were told that they were randomly selected not to deliver the speech, so speeches were never actually delivered. Data shown are for the contrast [Speech Preparation – Pre-stress Baseline].

RESULTS

Simulations

Simulation 1—Figures 3 and 4 show results of 1,000 repetitions of the simulated event-related designs assuming constant and random within-subject variance, respectively. Figure 5 shows equivalent results assuming constant within-subject variance and a misspecified first-level design matrix. The different rows of the figures correspond to the estimates of the slope and intercept parameters (β_0, β_1) and their between-subject variances ($\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_1}$). The different columns correspond to the 4 different pair-wise combinations of ground truth values for the parameters ($\sigma^2_{\beta_0}, \sigma^2_{\beta_1}$). For each combination, box plots of the estimates over the 1,000 repetitions are shown for LEMR, OLS, IGLS and RIGLS.

It is clear that all four methods provide almost equivalent estimates of the slope and intercept parameters, and that these estimates are unbiased (i.e., they center around the true values shown by dashed horizontal lines). In addition, the estimates of the variance components are similar for the four methods except for OLS which appears to have a negative bias. The bias of the OLS makes theoretical sense, and often n/(n − p) is used as a correction factor (Hinkley, 1977; Mumford & Nichols, 2009). However, it does not appear that this correction factor alone is sufficient to remove the amount of bias observed in the simulations. Of note, the estimates obtained using IGLS and RIGLS are comparable to those obtained using LMER, which has been extensively validated in the statistical community. The estimates of the variance components obtained using IGLS show a slight bias, but interestingly appear to have slightly smaller variance that those obtained using RIGLS. In addition, the estimated between-subject variance appears to be robust to model misspecification even when the sample size is relatively small. The within-subject variance is however effected and the introduction of robust estimation techniques (e.g., Waldorp, 2009) is an important topic for future research.

Simulation 2—Figure 6A shows the empirical distribution of the likelihood ratio statistic for testing the variance component associated with the slope assuming the intercept is fixed (i.e. the random slope model). It is clear that the χ^2_1 distribution used in standard likelihood ratio testing is overly conservative. In addition the 50:50 mixture distribution suggested by Stram and Lee (1994) also appears to be somewhat conservative. However, the approximation of the finite sampling distribution (denoted LTR$_{CR}$) appears to accurately reflect the behavior in the tail of the distribution as it more or less lies on top of the empirical LRT. Figure 6B shows the equivalent results for the RLRT. Here the difference between the 50:50 mixture distribution and the approximation of the finite sampling distribution is less pronounced, with both distributions taking roughly the same shape.

Figure 6C–D shows the empirical distribution of the LRT, and RLRT, statistic computed for testing the variance components associated with the slope in a random intercept, random slope model. From the plot it is clear that the χ^2_2 distribution used in standard likelihood ratio testing is overly conservative while the χ^2_1 distribution is anti-conservative. It would therefore appear that a 50:50 mixture of these distributions would accurately reflect the behavior in the tail of both distributions.

In sum, these simulations together these simulations indicate that using a mixture of χ^2 distributions give a sensible approximation of the null distribution for both tests even for relatively small sample sizes.

Simulation 3—Figure 7A shows the performance of the LRT and RLRT, for testing the significance of $\sigma^2_{\beta_1}$ in a random slope model, as a function of the within-subject variance. It shows the proportion of false positives when the likelihood ratio test is thresholded at the standard $p < 0.05$ level. Though both are close to the nominal level of 0.05, the results based on the LRT are slightly more conservative than those based on the RLRT, which shows a slightly inflated false positive rate. This finding is consistent with results in the literature (Morrell, 1998). In general, it appears that the size of the within-subject SNR (x-axis in Figure 7A) does not significantly impact the results. Figure 7B shows equivalent results when the sample size is allowed to vary (x-axis in Figure 7B). Again, both IGLS (LRT) and RIGLS (RLRT) give values close to the nominal values. The RLRT again shows a slightly inflated false positive rate, whereas the LRT is particularly conservative for small samples (false positive rates are lower than the nominal accepted rate), but are accurate at about N = 20 and above. Figure 7C shows the true positive rate plotted as a function of σ_1. The RLRT gives a marginal increase in power compared to the LRT, but this is likely a function of its increase in false-positive rate, so there is no clear advantage over the computationally similar LRT.

Figure 7D–F shows equivalent results for testing $\sigma^2_{\beta_1}$ in a random slope, random intercept model. The results are similar to those presented in Figure 7A–C. Again, there is a slightly inflated false positive rate for the RLRT (Figure 7D), though less pronounced than before and thus there is no clear advantage in power for the RLRT vs. the LRT (Figure 7E). In addition, for small sample sizes (N < 10), there appears to be a large increase in the number of false positives with both models.

For completeness we also performed t-tests on the regression parameters (fixed effects) for a variety of settings. The results (not presented here) show that while both IGLS and RIGLS adequately control for false positives, the latter tends to give slightly more powerful results.

Experimental Data

We focused on the inter-subject variance in the [Speech – Preparation] contrast, which has identified areas of increased activity related to social evaluative threat (SET) in previous studies (Wager et al., 2009a; 2009b). We focus on inter-subject variance parameter testing as it is the focus of interest in the current paper. Activation data (i.e., COPE estimates) will be discussed in a separate report.

Regions with significant variation across subjects (p < .005) are shown on axial brain slices in Figure 8. These regions include circumscribed parts of the midbrain, parahippocampal cortex, amygdala, ventromedial prefrontal cortex, insula, lateral prefrontal cortex, and pons. The significance of inter-individual variation is unbiased with respect to other predictors of individual variation (e.g., perceived threat or anxiety), and so these regions can be used to define ROIs for tests of these other effects. Such ROIs can also be used to assess functional connectivity across the brain. For example, ROI can be used as seed regions and a voxel-wise search employed to find regions that correlate with the seed.

Here, a functional connectivity analysis was performed by identifying a significant region in the left ventromedial prefrontal cortex (vMPFC; see Fig. 9A) and using it as a seed region in a correlation study. For each subject, time series data was extracted from this region and used to compute correlations with other voxel’s time series in a whole brain seed analysis.

Using robust regression techniques (Wager et al., 2005) a between-subject t-test was performed on the Fisher transformed correlations at each voxel. Fig. 9B shows regions of significant positive correlation in the right vMPFC and negative correlation in the periaqueductal gray (PAG). We selected PAG as a region of particular *a priori* interest because it is a region with strong anatomical connections with vMPFC and an extensive animal literature relates vMPFC-PAG function to the generation and regulation of threat (e.g., Amat et al., 2005; Bandler et al., 2000). In our previous work and others' (Mobbs et al., 2007; Wager et al., 2009a), vMPFC-PAG relationships have been thought to be key mediators of experienced threat states such as those evoked in this experiment, with inverse coupling between the vMPFC and PAG as found here. The procedure was repeated using a nearby seed region based on significant COPE estimates, the current standard way of generating regions of interest (see Fig. 9C). No correlation was found with PAG activity; as shown in Fig. 9D, among midline regions, only vMPFC and hypothalamus were positively correlated with the seed.

We next turn to the issue of using regions of high inter-subject variability to diagnose other, potentially artifactual sources of between-subjects error, such as individual differences in gray-matter density or inter-subject alignment. The latter might be important particularly at the boundaries between anatomical structures, as even the best nonlinear registration methods cannot perfectly align all these boundaries. Figure 10 shows detail for several brain regions (top panels) compared with maps of the inter-subject variability in gray-matter intensity as assessed from anatomical T1-weighted images (bottom panels). Anatomical variability is expressed in terms of the coefficient of variation (CV), the standard deviation in T1-intensity after nonlinear registration divided by the mean intensity across subjects. Variation in the midbrain and parahippocampal cortex are co-localized with areas showing high anatomical CV, shown by lighter blue and orange regions in the bottom panels of Figure 10. This suggests that inter-subject anatomical alignment may play a large role in the functional COPE inter-subject variability in these areas. However, other areas of theoretical interest in SET show large COPE variability without markedly large inter-subject anatomical variability. These include the hippocampus, amygdala, and ventromedial prefrontal cortex. All of these have been linked to SET in previous human studies (Critchley, 2005; Eisenberger et al., 2007; Gianaros et al., 2008, 2009; Preussner et al., 2010; Wager et al., 2009a, b), and so are promising ROIs for investigating correlations between brain activity and psychological correlates of SET. In sum, the inter-subject variance maps can be useful for both diagnosing problems with alignment and for identifying potential sources of individual differences in behavior.

DISCUSSION

In the current fMRI literature, multi-level models have primarily been used to estimate and make inferences about group mean COPEs and thus test regional activation levels. In this paper we introduce the concept of testing inter-individual variances as well as means, thus providing tests of whether there are true individual differences in brain activity. We believe that inter-subject variance maps can be useful for diagnosing problems with alignment in the preprocessing stage of the analysis, identifying sources of individual differences in behavior and choosing seed regions for functional connectivity studies.

Researchers are often interested in finding brain regions that are highly correlated with some behavioral measure of interest and with other brain regions. However, often these regions are chosen by searching through the brain for voxels that correlate with the same measure of interest, and only those that lie above a certain threshold are reported. This selection process will lead to the creation of radically inflated and spurious correlations due to the relationship between the selection process and the result of interest. In this work we suggest an analysis technique that allows one to determine important regions for which to calculate the correlation of interest. Our technique is based on detecting regions which show significant individual differences between subjects. Hence our approach towards defining regions of interest is independent of the measure in which we want to calculate correlations. This is increasingly important as studies that use the behavioral measure both to determine regions and thereafter calculate correlations may give severely inflated correlations (Vul et al., 2009; Lieberman et al. 2009; Lindquist and Gelman, 2009).

In this work, we suggest a complete framework for determining voxels that have significant variance components. Our framework consists both of an estimation and inference step. Our estimation technique is base on the use of IGLS/RIGLS. Both are intuitive and easily implemented techniques which we show in a series of simulations give accurate results. In addition, IGLS can be shown to be equivalent to the Fisher scoring technique used in SPM (Friston et al., 2002). The choice between using IGLS or RIGLS ultimately depends on the goal of the analysis. If one is interested in performing inference on COPEs then RIGLS is preferred as it provides unbiased estimates of the variance components. However, if the primary interest is performing inference on variance components, it appears that IGLS though biased, provides more powerful analysis.

Our inference technique consists of using likelihood ratio tests. We show that while calculating the test statistic itself is straightforward, finding the appropriate limiting distribution to calculate p-values is subtle. In this work we present asymptotic results that appear to give reasonable results. In addition, we discuss an approximation of the finite limiting distribution that is valid when the model contains a single between-subject variance component. As this is a rather limited case, we often rely on the asymptotic results in practice. Choosing the appropriate mixture of chi-square distributions can be tricky and some results point to that it should depend on the number of subjects. Though we use a 50:50 mixture in our simulations it may be more appropriate to weight the 0 degrees of freedom distribution somewhat higher, for example a 60:40 mixture. As an alternative one could use resampling techniques (e.g. the Bootstrap) to find the appropriate p-values, but we find that this approach is too time consuming to be a serious alternative in neuroimaging studies.

References

Alsop DC, Detre JA. Reduced transit-time sensitivity in noninvasive magnetic resonance imaging of human cerebral blood flow. J Cereb Blood Flow Metab. 1996; 16:1236–49. [PubMed: 8898697]

Amat J, Baratta MV, Paul E, Bland ST, Watkins LR, Maier SF. Medial prefrontal cortex determines how stressor controllability affects behavior and dorsal raphe nucleus. Nat Neurosci. 2005; 8(3): 365–371. [PubMed: 15696163]

Asllani I, Borogovac A, Wright C, Sacco R, Brown TR, Zarahn E. An investigation of statistical power for continuous arterial spin labeling imaging at 1.5 T. Neuroimage. 2008; 39:1246–56. [PubMed: 18036834]

Bandler R, Keay KA, Floyd N, Price J. Central circuits mediating patterned autonomic activity during active vs. passive emotional coping. Brain Res Bull. 2000; 53(1):95–104. [PubMed: 11033213]

Beckmann CF, Jenkinson M, Smith SM. General multilevel linear modeling for group analysis in fMRI. Neuroimage. 2003; 20:1052–1063. [PubMed: 14568475]

Bilodeau, M.; Brenner, D. Theory of Multivariate Statistics. Springer-Verlag; New York: 1999.

Browne M. Generalized least squares estimators in the analysis of covariance structures. South African Statistical Journal. 1974; 8:1–24.

Crainiceanu CM, Ruppert D. Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society Series B. 2004; 66(1):165–185.

Critchley HD. Neural mechanisms of autonomic, affective, and cognitive integration. J Comp Neurol. 2005; 493(1):154–166. [PubMed: 16254997]

Eisenberger NI, Taylor SE, Gable SL, Hilmert CJ, Lieberman MD. Neural pathways link social support to attenuated neuroendocrine stress responses. Neuroimage. 2007; 35(4):1601–1612. [PubMed: 17395493]

Friston KJ, Fletcher P, Josephs O, Holmes AP, Rugg MD, Turner R. Event-related fMRI: characterising differential responses. NeuroImage. 1998; 7:30–40. [PubMed: 9500830]

Friston KJ, Penny W, Phillips C, Kiebel S, Hinton G, Ashburner J. Classical and Bayesian inference in neuroimaging: theory. Neuroimage. 2002; 16(2):465–483. [PubMed: 12030832]

Friston KJ, Stephan KE, Lund TE, Morcom A, Kiebel SJ. Mixed-effects and fMRI studies. NeuroImage. 2005; 24:244–252. [PubMed: 15588616]

Gianaros PJ, Sheu LK, Matthews KA, Jennings JR, Manuck SB, Hariri AR. Individual differences in stressor-evoked blood pressure reactivity vary with activation, volume, and functional connectivity of the amygdala. J Neurosci. 2008; 28(4):990–999. [PubMed: 18216206]

Gianaros PJ, Sheu LK. A review of neuroimaging studies of stressor-evoked blood pressure reactivity: emerging evidence for a brain-body pathway to coronary heart disease risk. Neuroimage. 2009; 47(3):922–936. [PubMed: 19410652]

Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika. 1986; 73(1):43–56.

Goldstein H. Restricted unbiased iterative generalised least squares estimation. Biometrika. 1989; 76:622–623.

Harville DA. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association. 1977; 72:320–338.

Harville, DA. Matrix Algebra from a Statisticians's Perspective. Springer-Verlag; New York: 1997.

Hinkley D. Jackknifing in unbalanced situations. Technometrics. 1977; 19:285–292.

Kirschbaum C, Pirke KM, Hellhammer DH. The ‘Trier Social Stress Test’--a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology. 1993; 28(1–2):76–81. [PubMed: 8255414]

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience. 2009; 12:535–540.

Kriegeskorte N, Lindquist MA, Nichols TE, Poldrack RA, Vul E. Everything you never wanted to know about circular analysis, but were afraid to ask. Journal of Cerebral Blood Flow & Metabolism. 2010; 30:1551–1557. [PubMed: 20571517]

Lieberman MD, Berkman ET, Wager TD. Correlations in Social Neuroscience Aren't Voodoo: Commentary on Vul et al. (2009). Perspectives on Psychological Science. 2009; 4(3):299–307.

Lindstrom MJ, Bates DM. Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. Journal of the American Statistical Association. 1988; 83:1014–1022.

Lindquist MA. The Statistical Analysis of fMRI Data. Statistical Science. 2008; 23(4):439–464.

Lindquist MA, Gelman A. Correlations and Multiple Comparisons in Functional Imaging: A Statistical Perspective - Commentary on Vul et al., (2009). Perspectives on Psychological Science. 2009; 4(3):310–313.

Loh JM, Lindquist MA, Wager TD. Residual Analysis for Detecting Mis-modeling in fMRI. Statistica Sinica. 2008; 18:1421–1448.

Mobbs D, Petrovic P, Marchant J, Hassabis D, Weiskopf N, Seymour B, Dolan RJ, Frith CD. When fear is near: threat imminence elicits prefrontal-periaqueductal gray shifts in humans. Science. 2007; 317:1079–83. [PubMed: 17717184]

Morrell CH. Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. Biometrics. 1998; 54(4):1560–1568. [PubMed: 9883552]

Mumford JA, Nichols TE. Simple Group fMRI Modeling and Inference. NeuroImage. 2009; 47(4): 1469–1475. [PubMed: 19463958]

Mumford JA, Nichols TE. Power Calculation for Group fMRI Studies Accounting for Arbitrary Design and Temporal Autocorrelation. NeuroImage. 2008; 39(1):261–268. [PubMed: 17919925]

Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. Applied Linear Statistical Models. 4. McGraw Hill; 1996.

Pinheiro, JC.; Bates, DM. Mixed Effects Models in S and S-PLUS. Springer Verlag; New York: 2000.

Pruessner JC, Dedovic K, Pruessner M, Lord C, Buss C, Collins L, et al. Stress regulation in the central nervous system: evidence from structural and functional neuroimaging studies in human populations - 2008 Curt Richter Award Winner. Psychoneuroendocrinology. 2010; 35(1):179–191. [PubMed: 19362426]

Raudenbush, SW.; Bryk, AS. Hierarchical Linear Models. 2. Thousand Oaks: Sage Publications; 2002.

Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association. 1987; 82:605–610.

Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. Biometrics. 1994; 50(4):1171–1177. [PubMed: 7786999]

Vul E, Harris C, Winkielman P, Pashler H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. Perspectives on Psychological Science. 2009; 4:274–290.

Wager TD, Keller MC, Lacey SC, Jonides J. Increased sensitivity in neuroimaging analyses using robust regression. Neuroimage. 2005; 26(1):99–113. [PubMed: 15862210]

Wager TD, Lindquist MA, Kaplan L. Meta-analysis of functional neuroimaging data: current and future directions. Social Cognitive Affective Neuroscience. 2007; 2:150–8.

Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN. Prefrontal-subcortical pathways mediating successful emotion regulation. Neuron. 2008; 59:1037–50. [PubMed: 18817740]

Wager TD, Waugh CE, Lindquist M, Noll DC, Fredrickson BL, Taylor SF. Brain mediators of cardiovascular responses to social threat, Part I: Reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. Neuroimage. 2009a; 47:821–835. [PubMed: 19465137]

Wager TD, van Ast VA, Hughes BL, Davidson ML, Lindquist MA, Ochsner KN. Brain mediators of cardiovascular responses to social threat, Part II: Prefrontal-subcortical pathways and relationship with anxiety. Neuroimage. 2009b; 47:836–851. [PubMed: 19465135]

Waldorp L. Robust and Unbiased Variance of GLM Coefficients for Misspecified Autocorrelation and Hemodynamic Response Models in fMRI. International Journal of Biomedical Imaging. 2009:723912. [PubMed: 19746181]

Wang J, Zhang Y, Wolf RL, Roc AC, Alsop DC, Detre JA. Amplitude-modulated continuous arterial spin-labeling 3.0-T perfusion MR imaging with a single coil: feasibility study. Radiology. 2005; 235:218–28. [PubMed: 15716390]

Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM. Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage. 2004; 21:1732–47. [PubMed: 15050594]

Yarkoni T. Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). Perspectives on Psychological Science. 2009; 4(3):294–298.

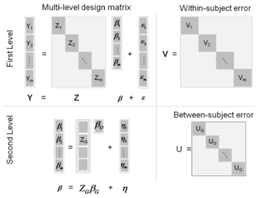

Figure 1.

A pictorial representation of the multi-level general linear model typically used in functional neuroimaging; mathematically described in Eq. [4]. The top panel shows the first-level (subject-specific) parameters and the bottom panel the second-level (group) parameters.

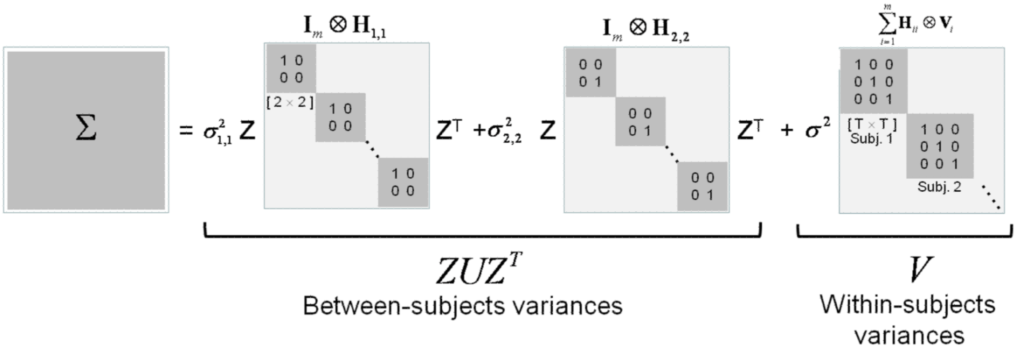
Σ
$= \sigma^2_{1,1}$ Z
$\mathbf{I}_m \otimes \mathbf{H}_{1,1}$
1 0
0 0
[2 × 2]
$Z^T + \sigma^2_{2,2}$ Z
$\mathbf{I}_m \otimes \mathbf{H}_{2,2}$
0 0
0 1
Z^T + σ^2
$\sum_{i=1}^{m} \mathbf{H}_{ii} \otimes \mathbf{V}_i$
1 0 0
0 1 0
0 0 1
[T × T]
Subj. 1
Subj. 2
ZUZ^T
Between-subjects variances
V
Within-subjects variances

Figure 2.

A pictorial representation of Eq. [7] for the special case when $p = 2$ (e.g., a random intercept and slope model), where the slope and intercept are uncorrelated, the data is pre-whitened and the within-subject variance is constant across subjects.

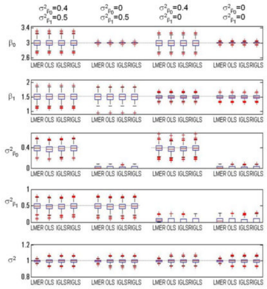

Figure 3.

Box plots of the parameter estimates from 1,000 repetitions of the simulation study performed assuming constant within-subject variance are shown for the LEMR, OLS, IGLS and RIGLS methods. The different rows correspond to the different parameters β_0, β_1, $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_1}$, and the different columns to the 4 different pair-wise combinations of values for the parameters ($\sigma^2_{\beta_0}, \sigma^2_{\beta_1}$). The results show that while all four algorithms give equivalent results for the fixed effects β_0 and β_1, the estimates of $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_1}$ differ with the OLS method giving negatively biased results.

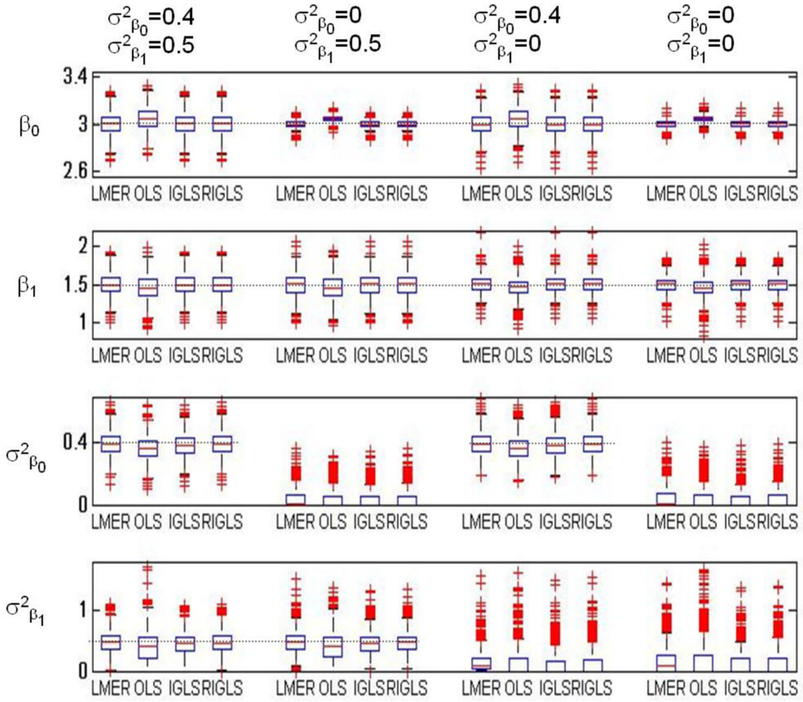

Figure 4.

Box plots of the parameter estimates from 1,000 repetitions of the simulation study performed assuming randomly varying within-subject variances are shown for the LEMR, OLS, IGLS and RIGLS methods. The different rows correspond to the parameters β_0, β_0, $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_1}$, and the different columns to the 4 different pair-wise combinations of values for the parameters ($\sigma^2_{\beta_0}, \sigma^2_{\beta_1}$). The results coincide with those shown in Fig. 3 for the constant variance case.

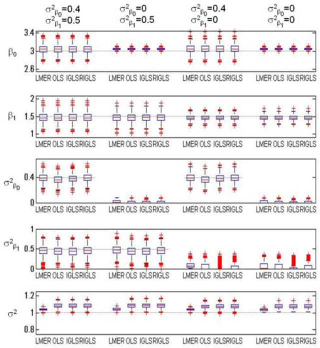

Figure 5.

Box plots of the parameter estimates from 1,000 repetitions of the simulation study performed assuming constant within-subject variance but misspecified first-level design matrix, are shown for the LEMR, OLS, IGLS and RIGLS methods. The different rows correspond to the different parameters β_0, β_1, $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_1}$, and the different columns to the 4 different pair-wise combinations of values for the parameters ($\sigma^2_{\beta_0}, \sigma^2_{\beta_1}$). The results are consistent with those shown in Figures 3 and 4.

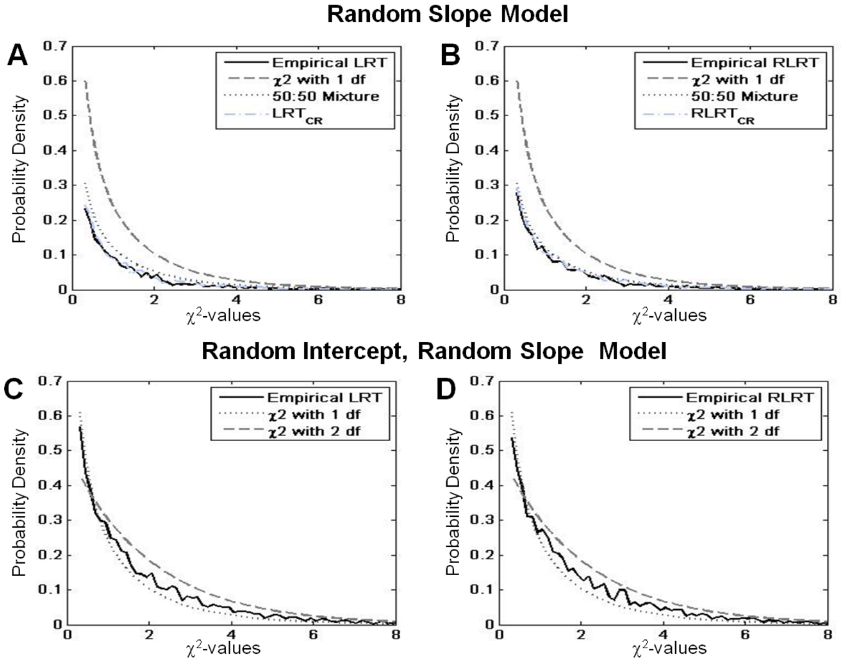

Figure 6.

(A) The empirical null distribution of the LRT statistic computed for testing the significance of the slope in a random slope model. This distribution acts as our ground truth in this simulation. The distribution used in standard likelihood ratio testing (χ^2_1) is overly conservative. The 50:50 mixture distribution also appears to be somewhat conservative. However, the approximation of the finite sampling distribution (LTR_{CR}) appears to accurately reflect the behavior in the tail of the distribution. (B) Equivalent results for the RLRT. Here the difference between the 50:50 mixture distribution and $RLTR_{CR}$ is less pronounced. (C–D) Equivalent results for testing the slope in a random slope, random intercept model (see Eq. [9]). The distribution used in standard likelihood ratio testing is overly conservative, while a 50:50 mixture accurately reflects the behavior in the tail of both distributions. Note that LTR_{CR} and $RLTR_{CR}$ are not defined when there are more than one between-subject variance component.

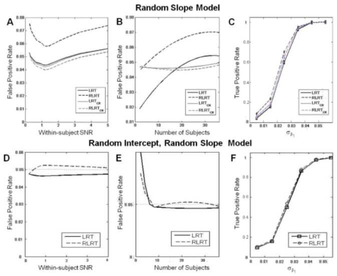

Figure 7.

(A) The proportion of false positives, when testing the slope in a random slope model, for the LRT and RLRT when thresholded at the 5% level, as a function of within-subject SNR. Though both are close to the nominal level of 0.05, the results based on the LRT are slightly more conservative than those based on the RLRT. In general, the size of the within-subject SNR does not significantly impact the results. (B) Equivalent results for varying number of subjects. Both the LRT and RLRT give values close to the nominal values. (C) The true positive rate plotted as a function of σ_1. The RLRT gives a marginal increase in power compared to the LRT. Hence, it appears ReML does not offer much benefit; a slight power increase, but also an increase in false positives. (D–F) Similar results for testing the slope in a random slope, random intercept model. For small number of subjects (<10) there appear to be a large increase in the number of false positives.

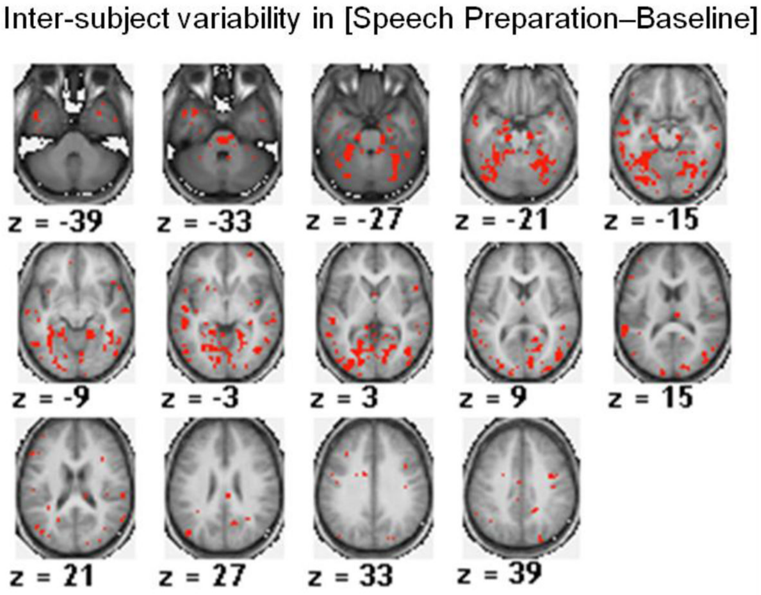

Figure 8.

Inter-subject variance maps for the [Speech Preparation - Baseline] contrast estimated using IGLS. Regions with significant variation across subjects (p < .005) are shown on axial brain slices.

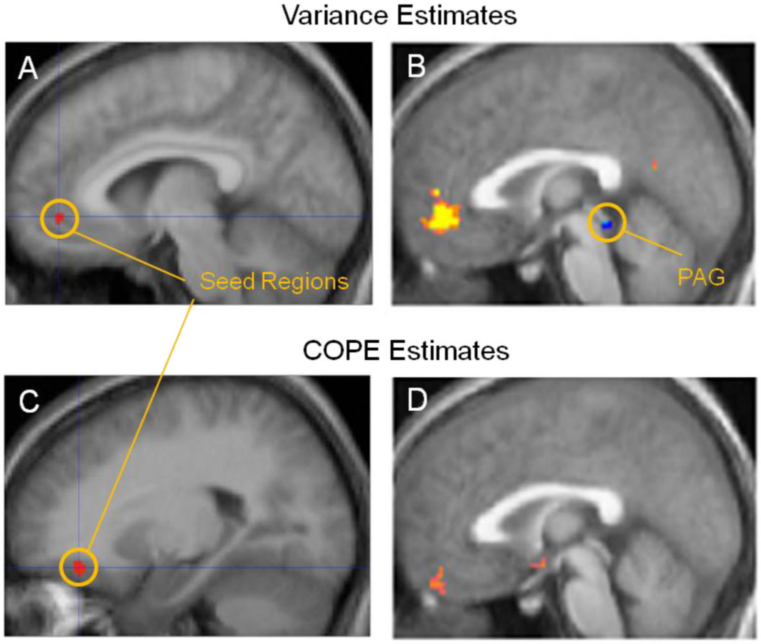

Figure 9.

(A) A seed region in the left vMPFC chosen because its variance was significantly different from 0. (B) Results of the seed analyses shown for a single sagittal slice. (C) A seed region in the left vMPFC chosen because the COPE was significantly different from 0 and it lies in close proximity to the region shown in (A). (D) Results of the seed analysis shown for the same slice as in (C). Notably, the seed with significant variance was significantly correlated with the PAG while the seed with significant COPE was not.

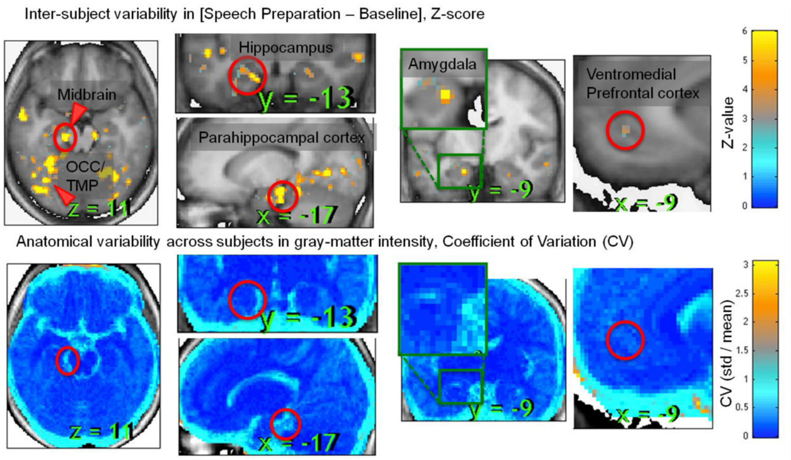

Figure 10.

Inter-subject variance maps for the [Speech Preparation - Baseline] contrast estimated using IGLS. Details of some significant regions (top panels) compared with inter-subject variability in gray-matter intensity as assessed from anatomical T1-weighted images (bottom). Variation in the midbrain and parahippocampal cortex may be caused by variation in inter-subject image registration (nonlinear warping), because the areas with significant inter-subject variance shown on the top are co-localized with areas showing a high coefficient of variation (CV) across subjects in gray-matter intensities. Inter-subject variability in the hippocampus, amygdala, and ventromedial prefrontal cortex do not show unusually large anatomical variability, and these results are more likely to reflect individual differences in functional brain processes.

Table 1

	Symbol	Description	Dimensions
Subject *i*	$\mathbf{Y}_\mathrm{i}$	fMRI time series data	T × 1
	$\mathbf{Z}_\mathrm{i}$	Design matrix	T × p
	β_i	Regression parameters	p × 1
	$\mathbf{V}_\mathrm{i}$	Covariance matrix	T × T
	σ_i^2	Subject-specific variance after pre-whitening	1 × 1
First-level All subjects	$\mathbf{Y}$	fMRI time series data – concatenated across subjects	mT × 1
	$\mathbf{Z}$	First-level design matrix - Block diagonal matrix with $\mathbf{Z}_\mathrm{i}$ in the blocks.	mT × mp
	β	Regression parameters – concatenated across subjects	mp × 1
	$\mathbf{V}$	Covariance matrix - Block diagonal matrix with $\mathbf{V}_\mathrm{i}$ in the blocks.	mT × mT
Second-level	$\mathbf{Z}_\mathrm{G}$	Group-level design matrix	mp × p
	β_G	Group-level regression parameters	p × 1
	$\mathbf{U}_\mathrm{G}$	Group-level covariance matrix	p × p
	$\sigma_{j,k}^2$	Covariance between the jth and kth element of $\boldsymbol{\beta}_{G.}$	1 × 1
Single-level	$\mathbf{X}$	Design matrix for single-level model - Can also be expressed as $\mathbf{ZZ}_\mathrm{G}$.	mT × p
	$\mathbf{\Sigma}$	Covariance matrix	mT × mT
VC Estimation	$\mathbf{R}$	Residual covariance matrix	mT × mT
	$\mathbf{Y}^*$	Vectorized version of $\mathbf{R}$.	$(mT)^2 \times 1$
	$\mathbf{X}^*$	Design matrix for variance components estimation	$(mT)^2 \times (p^2+m)$
	β^*	Vector containing all variance components	$(p^2+m) \times 1$
	$\mathbf{\Sigma}^*$	Covariance matrix of Y*	$(mT)^2 \times (mT)^2$

Note. The table contains a list of all vectors and matrices defined in the paper, together with a short description and information about their dimensionality.