ACADEMIC
PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

NeuroImage

NeuroImage 20 (2003) 591–600

www.elsevier.com/locate/ynimg

Technical Note

A heuristic for the degrees of freedom of statistics based on multiple variance parameters

Stefan J. Kiebel,[a,*] Daniel E. Glaser,[b] and Karl J. Friston[a]

[a] *Wellcome Department of Imaging Neuroscience, Institute of Neurology, 12 Queen Square, WCIN 3BG, London, UK*
[b] *Institute of Cognitive Neuroscience, 17 Queen Square, London, UK*

Received 7 October 2002; revised 15 April 2003; accepted 8 May 2003

Abstract

In neuroimaging, data are often modeled using general linear models. Here, we focus on GLMs with error covariances which are modeled as a linear combination of multiple variance/covariance components. Each of these components is weighted by one variance parameter. In many analyses variance parameters are estimated using restricted maximum likelihood (ReML). Most classical approaches assume the error covariance matrix can be factorized into a single variance parameter and a nonspherical correlation matrix. In this context, the *F* test based on a single variance parameter, with a suitable correction to the degrees of freedom, is the standard inference tool. This correction can also be adapted to models with multiple variance parameters. However, this extension overlooks the uncertainty about the variance parameter estimates and *P* values tend to be underestimated. Here, we show how one can overcome this problem to render the *F* test more exact. This issue is important, because serial correlations in fMRI time series are generally modeled using multiple variance parameters. Another application is to hierarchical linear models, which are used for modeling multisubject data. To illustrate our approach, we apply it to some typical modeling scenarios in fMRI data analysis.

Introduction

In neuroimaging, data are often analyzed in a voxel-wise fashion. After preprocessing, each voxel provides a univariate data sequence. The most prevalent model for these data is the general linear model (Rao and Toutenburg, 1995). The basic idea is to partition the data into two components. The first can be explained by a linear combination of some user-specified explanatory variables. The second is error. The experimenter is typically interested in the effects modeled by the explantory variables, e.g., how strong a response is. However, the error is also important, because the significance of a response is assessed using its estimated covariance. This depends on the estimated covariance of the error. In other words, a proper assessment of the error is necessary to make inferences about activations.

One could estimate all the elements of the error covariance matrix with a nonparametric method, but because the number of elements is quadratic in the number of scans, this is prohibitive. One generally parameterizes the covariance matrix and estimates a single variance parameter. A well-known example is when the covariance component, scaled by the variance parameter, is the identity matrix. This is called the general linear model, which assumes an identically, independently distributed error. Many statistical tests are based on this model, e.g., the standard *extra sum of squares F* test (Fahrmeir and Tutz, 1994). Positron emission tomography (PET) and single photon emission computed tomography (SPECT) data are examples, where this model is appropriate.

More generally, one can specify a known matrix *Q* as a single covariance component. This is useful, when one knows the correlation structure, but not its variance. This model was proposed for functional magnetic resonance imaging (fMRI) data (Worsley and Friston, 1995; Friston et al., 2002a). Here, the key issue is the correct specification of the correlation matrix. A common approach is to estimate

[*] Corresponding author. Functional Imaging Laboratory, Wellcome Dept of Imaging Neuroscience, 12 Queen Square, WC1N 3BG London, UK. Fax +44-0-20-7813-1420.

E-mail address: skiebel@fil.ion.ucl.ac.uk (S.J. Kiebel).

1053-8119/03/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S1053-8119(03)00308-2

the correlation structure by pooling over many voxels (Friston et al., 2002a; Worsley et al., 2002). When Q is not the identity matrix, the standard F statistic no longer conforms exactly to an F distribution. However, to make inferences, one can approximate the unknown distribution of the test statistic by a F distribution based on the Satterthwaite approximation (Worsley and Friston, 1995). This is formally identical to that used in the Greenhouse-Geisser correction.

The generalization of this approach is to model the covariance matrix by using a mixture of several covariance components. Each of these is scaled by its own variance parameter (Harville, 1977; Friston et al., 2002b). This approach affords the latitude to model unknown covariance structures. Examples, for which multiple covariance component models are useful, include serial correlations in fMRI and hierarchical linear models.

The Satterthwaite approximation as employed in (Worsley and Friston, 1995) is not applicable with multiple covariance components. There are two reasons for this. First, the covariance among different variance parameter estimates is ignored. This results in an increased (false) certainty about these estimates. Therefore, P values are generally too low and give invalid tests. The second reason is that the variability of different contrasts may depend on different variance components (i.e., variance parameters) that may have been estimated with different precisions. This means different contrasts should have different degrees of freedom. The application of the Satterthwaite approximation as described in Worsley and Friston (1995) does not accommodate this.

An appropriate estimation of the null distribution should embody the covariance matrix of the variance parameters estimated. In this note, we propose a general Satterthwaite approximation for multiple covariance components. This is based on a normal approximation to the distribution of the likelihood of the variance parameter estimates. We show how to derive a F ratio for single and multiple variance parameters starting from the log-likelihood ratio normalized by its estimated expectation.

We apply this procedure to three synthetic data sets. We treat some of the typical modeling scenarios encountered in neuroimaging research. For these data, we show that our approach improves P value estimates as compared to the approach adapted from the procedure described in (Worsley and Friston, 1995).

Theory

In this section, we describe linear models that are commonly used in neuroimaging and, in particular, in functional magnetic resonance imaging (fMRI) data analysis. These models belong to the class of the general linear model (Rao and Toutenburg, 1995). We derive a F test for multiple covariance components models. We focus on the F test that obtains when using ordinary least-squares for the regression coefficient or parameter estimates and restricted maximum likelihood (ReML) (Harville, 1977) for the variance parameter estimates of the covariance components.

The general linear model and the F ratio

Consider the general linear model

$$y = X\beta + \varepsilon, \tag{1}$$

where y is a data vector of length n, X is the $n \times p$ design matrix, β is a p-dimensional parameter vector (regression coefficients), and ε is an error vector of length n. For the general linear model with independent and identically distributed (iid) error assumptions, the error is normally distributed with $\varepsilon \sim N(0, \sigma^2 I_n)$, where I_n is the identity matrix of rank n.

Let

$$\begin{aligned} R &= I_n - XX^- \\ R_0 &= I_n - X_0X_0^- \\ M &= R_0 - R, \end{aligned} \tag{2}$$

where X^- denotes the generalized inverse of X. X_0 is the reduced model; i.e., X_0 represents a subspace of the full model X. The null hypothesis is whether the full model X explains something in addition to the reduced model X_0. Then, M is a projector onto the subspace of X that is not spanned by X_0. $M = R_0 - R$ can be defined either by direct specification of X_0 or, more commonly, through contrast weights c as described in the Appendix.

Classical statistics are generally formed from log-likelihood ratios. In the case of the general linear model, we have

$$\begin{aligned} p(y|X) &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}y^TR^TRy\right) \\ p(y|X_0) &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}y^TR_0^TR_0y\right) \\ l &= \ln\left(\frac{p(y|X)}{p(y|X_0)}\right) \\ &= \frac{1}{2\sigma^2}(y^TR_0^TR_0y - y^TR^TRy) \\ &= \frac{1}{2\sigma^2}y^T My. \end{aligned} \tag{3}$$

Scaling the log-likelihood ratio l [Eq. (3)], so its expectation under the null hypothesis is unity, gives

$$\frac{l}{\langle l_0\rangle} = \frac{y^TMy}{\sigma^2 tr(M)}, \tag{4}$$

where $tr(\cdot)$ denotes the trace operator. $\langle l_0 \rangle$ is the expectation of l under the null hypothesis

$$\langle l_0 \rangle = \left\langle \frac{y^T My}{2\sigma^2} \right\rangle = \left\langle \frac{tr(M\varepsilon\varepsilon^T)}{2\sigma^2} \right\rangle = \frac{tr(M\sigma^2 I)}{2\sigma^2}$$

$$\langle l_0 \rangle = \frac{tr(M)}{2} . \qquad (5)$$

Similarly, $\langle y^T Ry \rangle = \sigma^2 tr(R)$. This equality provides for the unbiased maximum likelihood (ML) estimator $\hat{\sigma}^2 = y^T Ry/tr(R)$. Replacing the unknown error variance variance parameter with this estimator gives the classical F ratio

$$f = \frac{y^T My}{\hat{\sigma}^2 tr(M)} = \frac{y^T My/tr(M)}{y^T Ry/tr(R)}. \qquad (6)$$

Note that the motivation for the F ratio is more commonly portrayed as the ratio of the sum of squares (SSQ) due to the effects of interest and those due to error.

Values of this statistic that are significantly greater than 1 allow rejection of the null hypothesis. The null distribution of f obtains from the ratio of two scaled χ^2 variables. An important interpretation of Eq. (6) is that the F ratio is simply the estimated normalized SSQ due to treatment effects of interest $(y^T My)/\hat{\sigma}^2$ divided by its expectation under the null hypothesis $tr(M)$. The uncertainty in these SSQ estimators enters the inference through the degrees of freedom, where $f \sim F_{\nu_0,\nu}$ with

$$\nu_0 = tr(M) \quad \text{and} \quad \nu = tr(R). \qquad (7)$$

$F_{\nu_0,\nu}$ denotes the F distribution with ν_0 and ν degrees of freedom.

One nonspherical variance component

In many cases, we cannot assume, as in Eq. (6), that the error is independent and identically distributed. If we know the form of the covariance matrix, we can still derive an (approximate) F statistic.

For a general linear model with a single variance parameter and a nonspherical variance component Q, i.e., $\varepsilon \sim N(0, \sigma^2 Q)$, the F statistic and degrees of freedom can be estimated as (Worsley and Friston, 1995)

$$f = \frac{y^T My/tr(MQ)}{y^T Ry/tr(RQ)} \sim F_{\nu_0,\nu}, \qquad (8)$$

where

$$\nu_0 = \frac{tr(MQ)^2}{tr(MQMQ)} \quad \text{and} \quad \nu = \frac{tr(RQ)^2}{tr(RQRQ)}. \qquad (9)$$

There are two approximations here. The first is that the variance parameter estimate

$$\hat{\sigma}^2 = y^T Ry/tr(RQ) \qquad (10)$$

no longer conforms to a scaled χ^2 distribution. However, through the Satterthwaite approximation (e.g., Yandell, 1997, p. 224–225), we can approximate it with one using the method of moments to estimate the effective degrees of freedom given in Eq. (9).

$$\nu = \frac{2\langle \hat{\sigma}^2 \rangle^2}{Var(\hat{\sigma}^2)} , \qquad (11)$$

where $\langle \cdot \rangle$ denotes the expectation and $Var\,(\cdot)$ the variance of a random variable.

The expectation of $\hat{\sigma}^2$ is given by σ^2 itself; i.e.,

$$\langle \hat{\sigma}^2 \rangle = \sigma^2 \qquad (12)$$

and its variance by

$$Var(\hat{\sigma}^2) = \frac{2\sigma^2 tr(RQRQ)}{tr(RQ)^2} . \qquad (13)$$

These results are also derived (in more detail) in (Kiebel and Holmes, 2003). The utility of Eq. (8) lies in being able to model nonsphericity using simple OLS parameter estimates, followed by an adjustment to the degrees of freedom using the Satterthwaite approximation. This is formally identical to the Greenhouse-Geisser correction.

The second approximation is that the F value [Eq. (8)], although valid, is no longer a log-likelihood ratio. To construct a log-likelihood ratio we would have to use projector matrices based on the maximum-likelihood estimates and restricted maximum likelihood estimates of the variance parameters. The classical F ratio [Eq. (8)] is based on OLS projectors.

The ensuing loss of efficiency, when using classical F statistics, in estimating the variance parameter is reflected in the fact that the effective degrees of freedom ν are always less than $tr(R)$.

As noted above, a better estimator of the variance parameter σ^2 is the ReML estimator. The ReML estimate uses decorrelated residuals about the maximum likelihood fit, see Appendix 2 of Friston et al., 2002b. The ReML estimate is given by

$$\hat{\sigma}^2 = \frac{y^T R_{ML}^T Q^{-1} R_{ML} y}{tr(R_{ML})} , \qquad (14)$$

where

$$R_{ML} = I_n - X(X^T C_\varepsilon^{-1} X)^{-1} X^T C_\varepsilon^{-1} \qquad (15)$$

and the error covariance matrix $C_\varepsilon = \sigma^2 Q$ so that

$$R_{ML} = I_n - X(X^T Q^{-1} X)^{-1} X^T Q^{-1}. \qquad (16)$$

$\nu = tr(R_{ML}) = tr(R)$; i.e., the effective degrees of freedom for the variance parameter revert to those in Eq. (7). We will pursue the construction of F ratios that use OLS projectors but incorporate ReML variance parameter estimates.

Multiple variance parameters

A generalization of the single covariance component case is when one assumes that the error covariance matrix can be characterized as a linear combination of several components.

For general linear models with m variance parameters, the errors are distributed with $\varepsilon \sim N(0, C_\varepsilon)$ and $C_\varepsilon = \lambda_1 Q_1 + \cdots + \lambda_m Q_m = \Sigma_{i=1}^{m} \lambda_i Q_i$ has multiple components. ReML variance parameters estimates $\hat{\lambda} = [\hat{\lambda}_1, \ldots, \hat{\lambda}_m]^T$ are usually obtained iteratively and do not generally have an easily derived covariance matrix. This covariance matrix is needed to specify an appropriate null distribution for F. In the following two subsections we describe how one can use the ReML variance parameter estimates to find an approximating F null distribution using the Satterthwaite and a normal approximation.

As in the single variance parameter case [Eq. (8)], one can use the estimated normalized sum of squares due to treatment divided by its estimated expectation under the null hypothesis, where $\langle y^T My \rangle = \langle \varepsilon^T M \varepsilon \rangle = tr(MC_\varepsilon)$, giving

$$f = \frac{y^T My}{\hat{\lambda}_1 tr(MQ_1) + \cdots + \hat{\lambda}_m tr(MQ_m)} = \frac{y^T My}{\hat{\lambda}^T T}, \quad (17)$$

where the elements of vector T are

$$T_i = tr(MQ_i). \quad (18)$$

The degrees of freedom of the numerator is $\nu_0 = tr(M\hat{C}_\varepsilon)^2/tr(M\hat{C}_\varepsilon M\hat{C}_\varepsilon)$, [cf. Eq. (9)]. The degrees of freedom of the denominator (ν), whose distribution is approximately a mixture of scaled χ^2 variables, depend on the distribution of each of the estimated variance parameters. Again, we can use the Satterthwaite approximation to compute the effective degrees of freedom

$$\nu = \frac{2\langle \hat{\lambda}^T T \rangle^2}{Var(\hat{\lambda}^T T)} = \frac{2\langle \hat{\lambda}^T T \rangle^2}{T^T Cov(\hat{\lambda}) T}, \quad (19)$$

where $Cov(\cdot)$ denotes the covariance. The expectation of the variance parameter estimates is simply $\langle \hat{\lambda} \rangle = \lambda$. However, the covariance matrix of the estimated variance parameters can, in many cases, only be approximated. We will use a normal approximation to $Cov(\hat{\lambda})$.

Normal approximation

Assuming we have ReML estimators of the variance parameters λ, we can use a Taylor series expansion for the log-likelihood $\log(p(y|\lambda))$. The expansion gives

$$\log p(y|\lambda) = \log p(y|\hat{\lambda}) + \frac{1}{2}(\lambda - \hat{\lambda})^T \frac{\partial^2}{\partial \lambda^2} \log p(y|\hat{\lambda})(\lambda - \hat{\lambda}) + \cdots. \quad (20)$$

Note that the first-order term in the Taylor series expansion is zero, because the partial derivative is zero at $\hat{\lambda}$. A second order approximation implies

$$\log p(y|\lambda) \approx \frac{1}{2}(\lambda - \hat{\lambda})^T \frac{\partial^2}{\partial \lambda^2} \log p(y|\hat{\lambda})(\lambda - \hat{\lambda}) \quad (21)$$

and therefore

$$p(y|\lambda) \approx N(\hat{\lambda}, I(\hat{\lambda})^{-1}), \quad (22)$$

where $I(\hat{\lambda})$ is the observed information matrix with

$$I(\hat{\lambda}) = -\frac{\partial^2}{\partial \lambda^2} \log p(y|\hat{\lambda}). \quad (23)$$

Here, we use the expectation of $I(\hat{\lambda})$ over the data y, the expected information matrix

$$I_E(\hat{\lambda}) = -\left\langle \frac{\partial^2 \log p(y|\hat{\lambda})}{\partial \lambda_i \partial \lambda_j} \right\rangle_y. \quad (24)$$

From (Harville, 1977), the elements of this matrix are given by

$$I_E(\hat{\lambda})_{ij} = \frac{1}{2} tr(PQ_i PQ_j), \quad (25)$$

where

$$P = C_\varepsilon^{-1} - C_\varepsilon^{-1} X (X^T C_\varepsilon^{-1} X)^{-1} X^T C_\varepsilon^{-1}. \quad (26)$$

When we substitute C_ε by its estimate $\hat{C}_\varepsilon$, the normal approximation to $p(y|\hat{\lambda})$ is $N(\hat{\lambda}, \hat{I}_E(\hat{\lambda})^{-1})$.

Effective degrees of freedom

The normal approximation to $p(y|\hat{\lambda})$ allows us to compute the effective degrees of freedom using the Satterthwaite approximation [Eq. (19)]. Using the estimated information matrix, we have

$$\hat{\nu} = \frac{(\hat{\lambda}^T T)^2}{T^T W^{-1} T}, \quad (27)$$

where T is defined in Eq. (18) and

$$W_{ij} = tr(\hat{P} Q_i \hat{P} Q_j), \quad (28)$$

i.e., W is the information matrix multiplied by two. In many instances, the variance parameters in the numerator and denominator in Eq. (27) cancel and the expression is exact, i.e., $\hat{\nu} = \nu$.

Summary

To use our approach for multiple variance parameters, one estimates the parameters β using ordinary least-squares (OLS) while estimating the variance parameters λ using restricted maximum likelihood. One then computes an F ratio according to Eq. (17). The inference is made by computing the P value of this F ratio by using a null F distribution with $\nu_0 = tr(M\hat{C}_\varepsilon)^2/tr(M\hat{C}_\varepsilon M\hat{C}_\varepsilon)$ and $\hat{\nu} = ((T^T\hat{\lambda})^2)/(T^TW^{-1}T)$ [Eq. (27)] degrees of freedom.

There are two critical points about this approach. The first is that different contrasts have different degrees of freedom. The second is that the covariance among variance parameters is taken into account when estimating the effective degrees of freedom. Both these considerations should ensure reasonably valid tests.

An alternative approach

Below, we will compare our approach to that based on (Worsley and Friston, 1995). Their approach assumes that the error covariance matrix is known and employs a Satterthwaite approximation to compute the effective degrees of freedom for a t or F distribution. One can use this approach for multiple variance parameter models by first estimating the correlation matrix and then treating it as known. This estimate V obtains by scaling $\hat{C}_\varepsilon$ so that $V = (n\hat{C}_\varepsilon)/(trace(\hat{C}_\varepsilon))$. In other words, one reduces the multiple variance parameter model to a single variance parameter one ($C_\varepsilon = \lambda V$), for which distributional results are provided in (Worsley and Friston, 1995).[1] In the remainder of the note, we will refer to this as the *WF procedure*, which is applied in this deliberately inappropriate way to illustrate the improvement conferred by the current approach.

Some important cases

In the following, we will discuss three models, which are special cases of the multiple variance parameter model. For these, we derive the effective degrees of freedom using Eq. (27).

Generally, if there is only one variance parameter, T [Eq. (18)] becomes a scalar and disappears from Eq. (27). In other words, the effective degrees of freedom of the denominator do not depend on the contrast of effects tested. In contradistinction, when there are multiple variance parameters, they do.

The general linear model

Here, $\varepsilon \sim N(0, \lambda I_n)$, i.e., $Q = I_n$, and $C_\varepsilon^{-1} = I_n/\lambda$ and $P = (I_n - X(X^TX)^{-1}X^T)/\lambda = R/\lambda$. This gives the classical result for the degrees of freedom

$$T = tr(M)$$

$$W = \frac{tr(R)}{\hat{\lambda}^2}$$

$$\nu = tr(R). \qquad (29)$$

General linear model with a single variance parameter

Let $\varepsilon \sim N(0, \lambda Q)$. Here, $C_\varepsilon^{-1} = Q^{-1}/\lambda$ and therefore $P = (Q^{-1} - Q^{-1}X(X^TQ^{-1}X)^{-1}X^TQ^{-1})/\lambda = Q^{-1}R_{ML}/\lambda$, where R_{ML} as defined in Eq. (15). This again gives the classical degrees of freedom

$$T = tr(MQ)$$

$$W = \frac{tr(\hat{R}_{ML})}{\hat{\lambda}^2}$$

$$\nu = tr(\hat{R}_{ML}), \qquad (30)$$

where $tr(\hat{R}_{ML}) = tr(R)$. (These results can be derived using $tr(A + B) = tr(A) + tr(B)$ and the cyclical property of the trace operator, i.e. $tr(ABC) = tr(BCA)$.)

The Satterthwaite approximation used by Worsley and Friston (1995) reduces to Eq. (30) when prewhitening is used for the variance parameter estimation [Eq. (10)]. Prewhitening means the multiplication of the model $Y = X\beta + \varepsilon$ by $\hat{C}_\varepsilon^{-1/2}$. Using the notation in (Worsley and Friston, 1995), matrix V becomes the identity matrix, the variance parameter estimate becomes the ReML estimate, and the effective degrees of freedom $\nu = tr(R)$.

General linear model with spherical variance components

If the Q_i are leading diagonal and nonoverlapping such that $W_{ij} = 0$ for $i \neq j$ we have

$$\hat{\nu} = \frac{\sum_i \hat{\lambda}_i^2 tr(MQ_i)^2}{\sum_i \hat{\lambda}_i^2 tr(MQ_i)^2 tr(R_{ML}Q_iR_{ML}Q_i)^{-1}}. \qquad (31)$$

If we impose the additional constraint on the F contrast that renders only one $tr(MQ_\mathrm{i})$ nonzero, we have

$$\hat{\nu} = tr(\hat{R}_{ML}Q_i\hat{R}_{ML}Q_i). \qquad (32)$$

Critically, the degrees of freedom ν now depend on M. This means different subspaces tested (by different contrasts) will have different degrees of freedom. We will demonstrate this below.

Simulations

Synthetic data

In the following, we apply the current and the WF procedure to synthetic data. To assess the validity of our approach, we evaluated the approximation to the null distribution. Therefore, it was sufficient to generate synthetic data under the null hypothesis that does not include signal, i.e., activation.

We first deal with a simple but classic (Behrens-Fisher) problem of comparing the means of two samples with different error variances. This is a somewhat trivial example that sets the scene for two further examples of nonsphericity with multiple variance components. These are (1) components induced by the hierarchical nature of observation models and (2) by serial correlations in a single level model.

[1] It should be noted that this approach is adapted by SPM2. However, the estimate of V obtains by pooling over all responsive voxels. The ensuing precision is so high that V can be treated as known (Glaser et al., 2002).

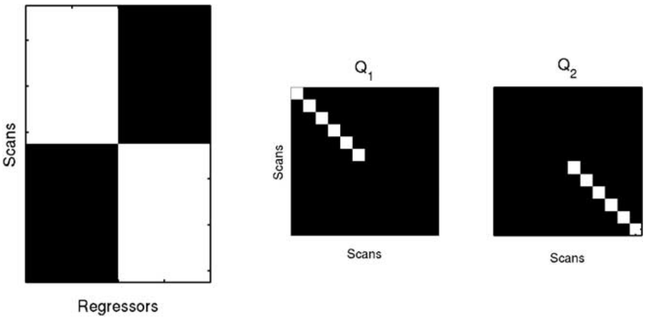

Fig. 1. Graphical display of a model for two groups with different means and variances. Left, design matrix, right, two covariance components.

Two groups with different variances

In the first experiment, we model two groups with different means and variances. We do not assume any covariance between groups. We used six observations in each group. We want to test for the difference between group means while allowing for different variances. The design matrix consists of two columns with dummy variables indicating that an observation belongs to either group 1 or group 2 (Fig. 1, left). The two variance components have ones and zeros on the diagonal, indicating group 1 or 2 (Fig. 1, right).

The interesting contrasts are the vectors $c = [-1\ 1]^T$ and $[1\ -1]^T$ that test for a difference between group means. For these contrasts, WF gives the same effective degrees of freedom as our procedure. However, for a contrast that is zero for one of the columns of the design matrix, e.g., $[0\ 1]^T$, the effective degrees of freedom diverge. This means the null distributions and subsequently the P values are different. The $[0\ 1]^T$ contrast might appear nonsensical at first, because one can test this hypothesis simply by using a one-sample t test. However, note that our formulation has the advantage that one model can be used to test multiple hypotheses without refitting different models.

To see how the diversion of the effective degrees of freedom effects the computed P values, we repeated this experiment 10^4 times by drawing a Gaussian identically independently distributed (iid) data vector with variance 1 and applied both approaches. The resulting cumulative P value distributions are shown in a PP plot (Fig. 2). In this plot, lines above the identity represent invalid, or capricious, performance, and regions below the identity represent conservative performance. For small P values, the WF approach slightly underestimates P values. Our approach estimates P values that indicate a more valid test. The degrees of freedom of the associated F statistic were ranging between 5.48 and 10.00 (mean 8.68) for WF and 5 for the current approach.

Hierarchical model

Another model that is often used in neuroimaging is a hierarchical model with two levels (Friston et al., 2002a). In our example, the first level models scans within subjects and the second level models effects over subjects. We modeled three scans for each of 12 subjects. At the second level, we model two groups of 6 subjects each. We assume different variances between groups at the second level. See Fig. 3 and 4 for a graphical display of the two design matrices and the three covariance components at both levels. Hierarchical models induce multiple variance components in the following way

$$y^{(1)} = X^{(1)}\beta^{(1)} + \varepsilon^{(1)} \tag{33}$$

$$\beta^{(1)} = X^{(2)}\beta^{(2)} + \varepsilon^{(2)} \tag{34}$$

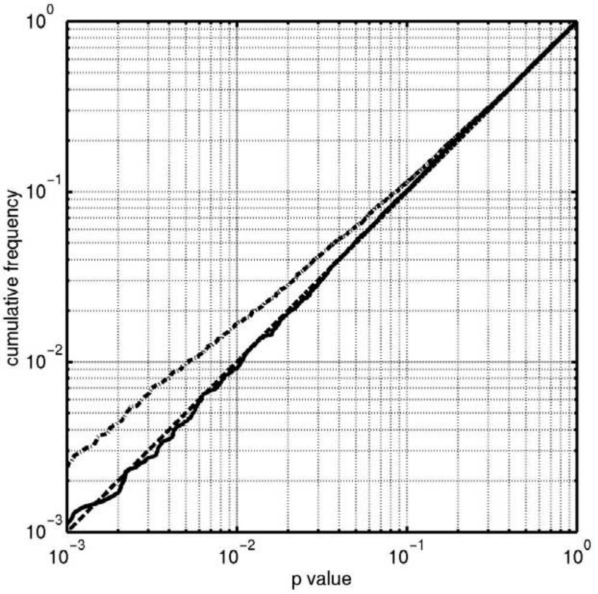

Fig. 2. Comparison of cumulative P values using our approach and WF. We modeled two groups with different means and different variances and tested for the mean of group 2. Results are displayed on a log-log plot. Dashed line, P values required for an exact test; dot-dashed line, WF approach; Solid line, current approach.

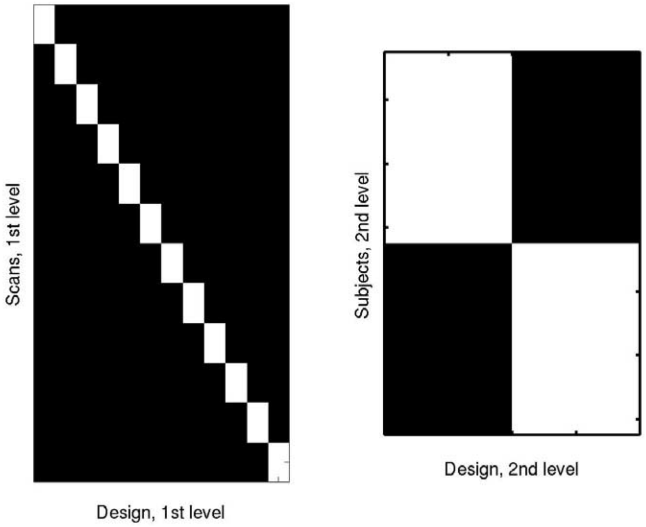

Fig. 3. First and second level design matrix for a hierarchical model.

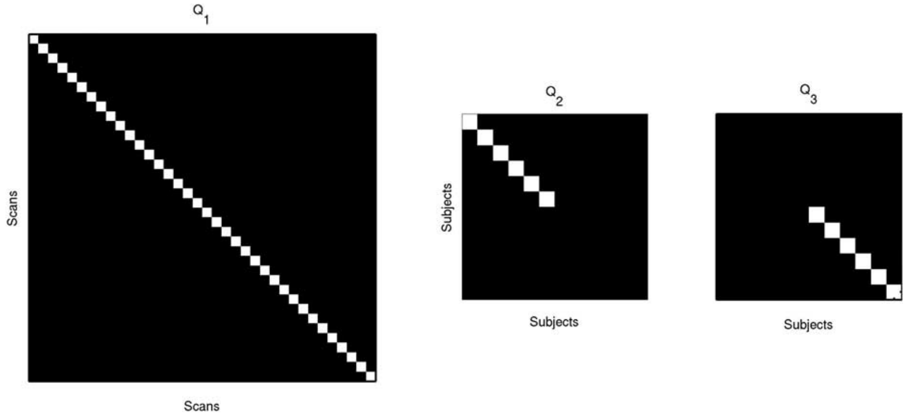

Fig. 4. First and second level covariance components for a hierarchical model.

giving

$$y^{(1)} = X^{(1)}X^{(2)}\beta^{(2)} + X^{(1)}\varepsilon^{(2)} + \varepsilon^{(1)}. \tag{35}$$

The ensuing error has three components (Fig. 4) corresponding to $\lambda_1 Q_1$ (first level), $\lambda_2 Q_2$, and $\lambda_3 Q_3$ (second level). Note that the two variance parameters at the second level are hyperparameters of the first level parameters $\beta^{(1)}$.

Here, we use a $[0\ 1]^T$ contrast. We repeated this experiment 10^4 times by drawing a data vector with variance 1 from a iid Gaussian distribution and applied both approaches. We observed discrepancies in the *P* value estimation between the two approaches (Fig. 5). For low *P* values, the WF approach underestimates the *P* values. Our approach gives higher *P* values, which provide a nearly valid test. The degrees of freedom of the associated *t* statistic were 10 for WF and 5 for the current approach.

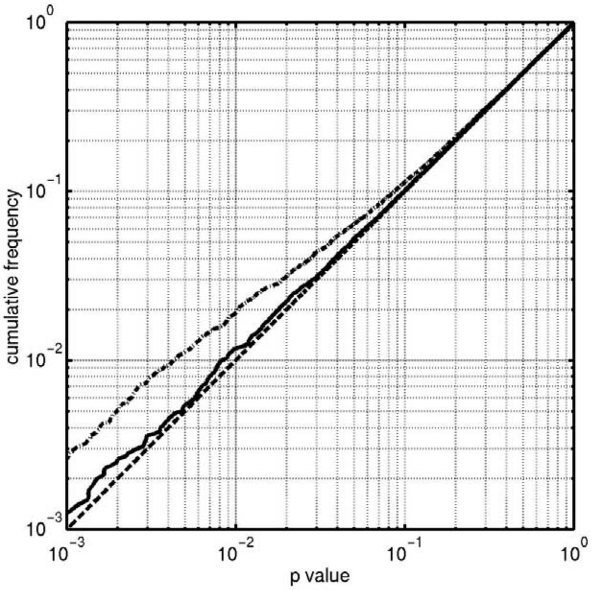

Fig. 5. Comparison of *P* values using our approach and WF. We modeled two groups with 12 subjects with three observations each. We tested for the effect of group 2. Results are displayed on a log-log plot. Dashed line, *P* values required for an exact test; dot-dashed line, WF approach; solid line, presented approach.

Serial correlations

Our third simulation embodies simulated serial correlations in fMRI time series. These can be modeled by two error covariance components. The choice of the covariance components are based on an AR(1) plus white noise model and are shown in Fig. 6. The AR(1) coefficient is fixed to be 1/*e*, which is typical for data acquired at 1.5 or 2 T with a standard echo planar imaging (EPI) sequence.

We sampled 10^4 time series of length 48 from an AR(1) process with a Gaussian iid error with variance 1. The AR(1) coefficient was 1/*e*. The design matrix *X* was simply a constant regressor. We tested for the mean effect with a contrast of [1]. Any difference between *P* value estimates is due to treating the relative sizes of the covariance components as known. In reality there are two unknown covariance components. Our method takes this into account, the WF approach does not. The resulting estimates of the *P* values using the WF and our approach are shown in Fig. 7. Additionally, we estimated *P* values using a model for which the error was assumed to be iid (the current and the WF procedures are the same here). Clearly, this assumption leads to highly underestimated *P* values. For small *P* values,

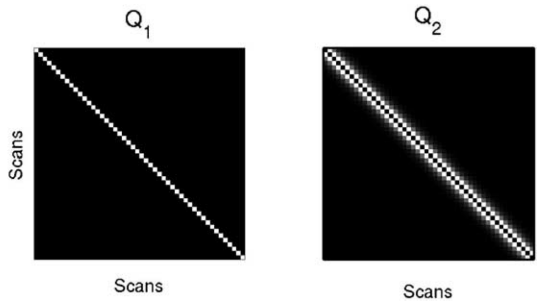

Fig. 6. Covariance components which model serial correlations in a fMRI time series.

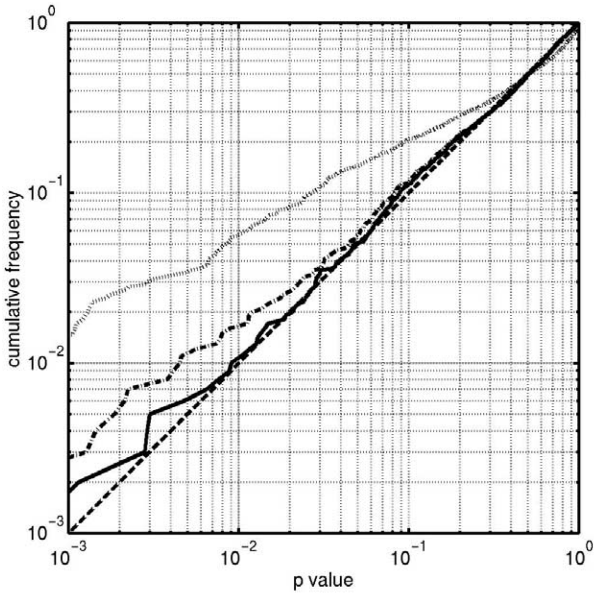

Fig. 7. Comparison of *P* values using our approach and WF. We modeled a fMRI time series using two covariance components. Results are displayed on a log-log plot. Dashed line, *P* values required for an exact test; dotted line, conventional *F* test (not modeling serial correlations); dot-dashed line, WF approach (modeling serial correlations); solid line, presented approach.

the WF approach still underestimates *P* values, although to a much lesser extent. Our approach gives *P* values closer to the *P* values required for a valid test. The degrees of freedom of the associated *F* statistic were between 24.51 and 46.99 (mean 36.40) for WF (modeling serial correlations), 47 when not modeling serial correlations, and between 3.98 and 28.12 (mean 15.44) for the current approach.

Discussion

We have developed a general estimator for the effective degrees of freedom of a *F* distribution when using multiple variance parameters in general linear models. The variance parameters are coefficients of covariance components which are linearly mixed to model the error covariance. The estimate of the effective degrees of freedom is based on the Satterthwaite approximation and a normal approximation to the likelihood of the variance parameters. We construct a *F* statistic from ordinary least-squares estimates of the parameters and restricted maximum likelihood estimates of the variance parameters. Inference can be made by approximating the distribution of the *F* statistic by a null *F* distribution. Only in the case of the general linear model with iid error is this null distribution exact. In all other cases the null *F* distribution approximates the underlying null distribution.

We derived a Satterthwaite approximation to the distribution of the estimated expectation of the treatment effect under the null hypothesis [denominator of Eq. (17)] when using multiple variance parameters. As we have shown with synthetic data the resulting effective degrees of freedom indicate reasonably valid tests. When using another method (Worsley and Friston, 1995) that was developed for a known covariance matrix structure with a single variance parameter, the resulting *P* values tended to be underestimated. The WF procedure is formally identical to the Greenhouse-Geisser correction.

There are two reasons for discrepancy between the WF and our approach when using multiple variance parameters. The first is that the effective degrees of freedom should be a function of the contrast defining *M*. With the WF approach, this is not the case. This is because one assumes a known correlation matrix and the uncertainty in the single variance parameter estimate effects all contrasts equally. The resulting underestimation is clearly illustrated in the first two synthetic data examples. With our approach, the contrast enters the estimation of the effective degrees of freedom through *M*.

Another reason for a difference between both approaches is that the covariance between estimated variance parameters should be taken into account during inference. Clearly, this cannot be done with the WF approach, because there is only one variance parameter. However, the covariance between estimated variance parameters is important, if one deals with overlapping covariance components [see Eq. (27)]. An example of an overlap is the model for serial correlations in fMRI.

In practice, for many studies, the differences between using the current and the WF approach will be small. This is because the null distribution does not differ much once both estimates of the effective degrees of freedom are above (say) 32. To see this, we show three different null *t* distributions (the cumulative distribution function (CDF)) for 9, 32, and 128 degrees of freedom (Fig. 8). The interesting part of the CDFs is the upper tail, which is associated with small *P* values. The difference between 9 and 32 degrees of freedom can make a difference. For 32 and 128 degrees of freedom, this difference in the resulting *P* value (given the same *t* value) is small.

For which data does a proper distributional approximation make a difference? Basically, one will observe differences when using multiple covariance components in hierarchical models, where one has only few data points at the subject level. A relevant example is a two-level model, where the second level models effects over subjects. If one uses multiple covariance components at the subject level (to model different variances), one may encounter differences for small *P* values for group-specific contrasts and contrasts testing for differences between groups.

Finally, we would like to point out the differences between the approach described here and the one adapted by statistical parametric mapping (SPM). SPM uses a spatial, hierarchical model to estimate the correlation matrix structure for all voxels. In a first step, SPM estimates from the sample covariance matrix (over a subset of voxels) the variance parameters of the specified covariance components using ReML. Note that ReML estimates the variance parameters in the null space of the design matrix. The estimates are based on many voxels; i.e., the certainty or precision about the estimates is extremely high. Therefore, at

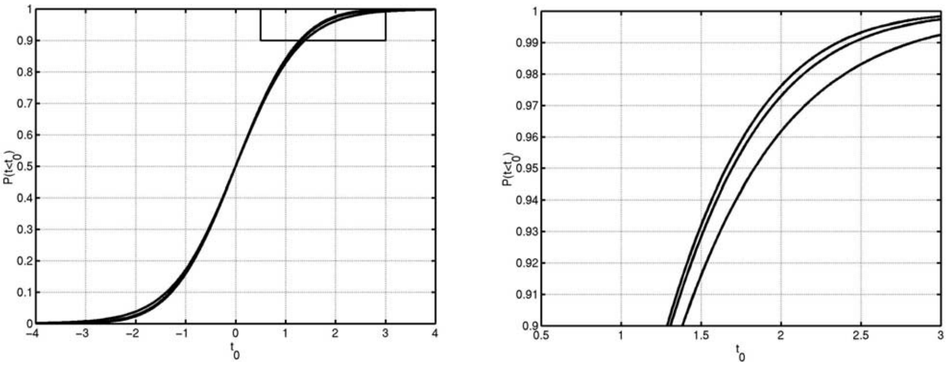

Fig. 8. Cumulative density distributions (CDF) for *t* distributions with 9, 32, and 128 degrees of freedom. Left, CDFs plotted for *t* values between −4 and 4; right, Zoomed display of (interesting) part of CDFs.

each voxel, one can treat the estimates and therefore the correlation matrix as known. Then, SPM uses the Worsley and Friston (1995) procedure. In other words, SPM uses knowledge derived from a global estimate to finesse local variance parameter estimates.

Conclusion

We have derived an estimator for the effective degrees of freedom when using multiple variance parameters in a regression model with normally distributed error. Our approach is based on the Satterthwaite and a normal approximation. Inference is made using *F* tests, using ordinary least-squares estimates for the parameters and restricted maximum likelihood for the variance parameters. We have shown that our approach gives valid tests, whereas classical approaches based on a single variance parameter underestimate *P* values and yield invalid tests.

Acknowledgments

S.J.K. and K.J.F. were supported by the Wellcome Trust. D.E.G. was supported by the Medical Research Council.

Appendix

In this appendix, we show how to compute the projection matrix *M*, which is necessary to form *F* statistics, based on OLS parameter estimates.

The general linear model is given by

$$y = X\beta + \varepsilon, \tag{36}$$

where y is a data vector of length *n*, *X* is the $n \times p$ design matrix, β is a *p*-dimensional parameter vector (regression coefficients), and ε is an error vector of length *n*. If one has a reduced model X_0, one can test whether the full model *X* explains anything in addition to X_0. X_0 spans a subspace of the space spanned by *X*. One often talks, in this context, of a *nested* model.

The test based on the *F* statistic [Eq. (6)] is a function of the data *y* and two projection matrices *R* and *M*. $R = I_\text{n} - XX^-$ is the residual forming matrix of the full model *X*, i.e., a projection matrix from the measurement to the residual space. *M* projects the data to a subspace of *X* that is the null space of X_0. This subspace spans the effects of interest after taking into account the reduced model.

Computing *M* requires the reduced model X_0. One can specify X_0 directly for simple models, but for more complex models, this can be difficult. The most commonly used approach is to specify contrast weights *c* which define the effects one wishes to test for *c* can be a vector or a matrix.

To get X_0, one first finds the null space of *c*.

$$c_0 = I_p - cc^- I_p. \tag{37}$$

The reduced model is $X_0 = X\text{c}_0$. Note that although the contrast weights c_0 and *c* are orthogonal to each other, the resulting partitions of the design matrix (Xc and Xc_0) are not necessarily orthogonal. The residual forming matrices of *X* and X_0 are then

$$\begin{aligned} R &= I_n - XX^- \\ R_0 &= I_n - X_0 X_0{}^- \end{aligned} \tag{38}$$

and $M = R_0 - R$. The important point to note is that an overlap of the subspaces spanned by X_0 and Xc is resolved by assigning it to X_0. The space orthogonal to X_0 is

$$X_1 = Xc - X_0 X_0^- Xc. \tag{39}$$

X_1 is never needed in the computation of the *F* statistic, but it is the space that one tests for.

References

Fahrmeir, L., Tutz, G., 1994. Multivariate Statistical Modelling Based on Generlized Linear Models. Springer-Verlag, Berlin.

Friston, K.J., Glaser, D.E., Henson, R.N., Kiebel, S., Phillips, C., Ashburner, J., 2002a. Classical and bayesian inference in neuroimaging: Applications. NeuroImage 16, 484–512.

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

Glaser, D.E., Kiebel, S.J., Friston, K., 2003. Pooling and covariance component estimation in statistical parametric mapping. Submitted for publication.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Statis. Assoc. 72, 320–338.

Kiebel, S., Holmes, A., 2003. Human Brain Function, Part II, Imaging Neuroscience Theory and Analysis, second ed. Elsevier Science, San Diego.

Rao, C.R., Toutenburg, H., 1995. Linear Models, Least Squares and Alternatives. Springer Series in Statistics.

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

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

Yandell, B.S., 1997. Practical Data Analysis for Designed Experiments. Chapman & Hall, New York.