Variance Components

Daniel E. Glaser, Karl J. Friston

1. Introduction

The validity of F statistics for classical inference on imaging data depends on the sphericity assumption. This assumption states that the difference between two measurement sets (e.g. those for two levels of a particular variable) has equal variance for all pairs of such sets. In practice this assumption can be violated in several ways, for example, by differences in variance induced by different experimental conditions, and/or by serial correlations within imaging timeseries.

A considerable literature exists in applied statistics that describes and compares various techniques for dealing with sphericity violation in the context of repeated measurements (see e.g. Keselman et al 2001). The analysis techniques exploited by the Statistical Parametrical Mapping (SPM) package also employ a range of strategies for dealing with the variance structure of imaging data, but these have never been explicitly compared with more conventional approaches.

Deductions about what is significant in imaging data depend on a detailed model of what might arise by chance. If you do not know about the structure of random fluctuations in your signal, you will not know what features you should find 'surprising'. A key component of this structure is the covariance of the data. That is, the extent to which different sets of observations within your experiment are dependent upon one another. If this structure is wrongly identified, it can lead to incorrect estimates of the variability of the parameters estimated from the data. This in turn can lead to false inferences.

Classical inference requires the expected likelihood distribution of a test statistic under the null hypothesis. Both the statistic and its distribution depend on hyperparameters controlling different components of the error covariance (this can be just the variance, σ^2, in simple models). Estimates of variance components are used to compute statistics and variability in these estimates determine the statistic's degrees of freedom. Sensitivity depends, in part, upon precise estimates of the hyperparameters (i.e. high degrees of freedom).

In the early years of functional neuroimaging there was debate about whether one could 'pool' (error variance) hyperparameter estimates over voxels. The motivation for this was an enormous increase in the precision of the hyperparameter estimates that rendered the ensuing T statistics normally distributed with very high degrees of freedom (see Chapter 7). The disadvantage was that 'pooling' rested on the assumption that the error variance was the same at all voxels. Although this assumption was highly implausible, the small number of observations in PET renders the voxel-specific hyperparameter estimates highly variable and it was not easy to show significant regional differences in error variance. With the advent of fMRI and more precise hyperparameter estimation this regional heteroscedasticity was established and conventional pooling was precluded. Consequently, most analyses of neuroimaging data now use voxel-specific hyperparameter estimation. This is quite simple to implement, provided there is only one hyperparameter, because its ReML estimate (see section 4 and Chapter 3, 4 and 17) can be obtained non-iteratively and simultaneously through the sum of squared residuals at each voxel. However, there are an increasing number of situations in which the errors have a number of variance components (e.g. serial correlations in fMRI or inhomogeneity of variance in

hierarchical models). The ensuing non-sphericity presents a potential problem for mass univariate tests of the sort implemented by SPM.

Currently, several approaches are taken to this problem. Firstly, departures from a simple distribution of the errors can be modelled using tricks borrowed from the classical statistical literature. This correction procedure is somewhat crude, but can protect to some extent against the tendency towards liberal conclusions. Secondly, a correlation structure can be imposed on the data by smoothing. This runs the risk of masking interesting features of the signal, but can coerce the noise into better behaviour. Finally, the kinds of tests performed can be restricted. For example, tests comparing several measures from each of several subjects from a larger population can be 'forbidden' since they rely more heavily on unjustifiable assumptions about the noise structure.

In this chapter we will describe how this problem has been addressed in various versions of SPM. We first point to a mathematical equivalence between the classical statistical literature and SPM99 in their treatment of violations of assumptions about covariance structure. Classically, the assumed structure is the most liberal and allows a model to be estimated without mathematical iteration. In SPM99, as described in Worsley and Friston (1995), a temporal smoothing stage before the main analysis 'swamps' any intrinsic auto-correlation with an imposed temporal covariance structure. While this structure does not correspond to the assumptions underlying the classical analysis, the same approach is used to take account of this known violation. While it would be possible to estimate and correct directly for the intrinsic covariation structure rather than trying to swamp it, an error in this estimation has been shown to be very costly in terms of the accuracy of the subsequent inference (Friston et al 2000).

Defining sphericity as a quantitative measure of the departure from basic assumptions about the null distribution, we will show how SPM '99 compensates only for sphericity violations associated with serial correlations. It employs a correction to the degrees of freedom that is mathematically identical to that employed by the Greenhouse-Geisser univariate F-test. This correction is applied after a filtering stage which swamps the intrinsic auto-correlation with an imposed structure. It is the known non-sphericity of this imposed structure which is then used to approximate the degrees of freedom.

In the second part of the chapter we will broadly describe a new approach to the problem. Instead of assuming an arbitrarily restricted covariance structure, we will show how new iterative techniques can be used to simultaneously estimate the actual nature of the errors alongside the estimation of the model. While traditional multivariate techniques also have estimated covariances, here we allow the experimenter to 'build in' knowledge or assumptions about the data, reducing the number of parameters which must be estimated, and restricting the solutions to plausible forms. These techniques are being implemented in new versions of SPM. We will describe briefly the types of previously 'forbidden' models which can be estimated using the new techniques.

More recent approaches that we have developed use a Parametric Empirical Bayesian (PEB) technique to estimate whichever variance components are of interest. This is equivalent to iterative Restricted Maximum Likelihood (ReML). In functional

magnetic resonance imaging (fMRI) time series, for example, these variance components model the white noise component as well as the covariance induced by, for example, an AR(1) component. In a mixed effects analysis the components correspond to the within-subject variance (possibly different for each subject) and the between-subject variance. More generally, when the population of subjects consists of different groups, we may have different residual variance in each group. PEB partitions the overall degrees of freedom (e.g. total number of fMRI scans) in such a way as to ensure that the variance estimates are unbiased. This takes place using a version of an Expectation-Maximisation (EM) procedure where the model coefficients and variance estimates are re-estimated iteratively.

Finally, we will give a mathematical justification of the pooling of covariance estimates underlying this approach.

2. Mathematical equivalences

2.1. Assumptions underlying repeated measures ANOVA

Inference in imaging data under the approach of SPM proceeds by the construction of an F test based on the null distribution. Our inferences are vulnerable to violations of assumptions about the variance structure of the data in just the same way as, for example, in the behavioural sciences:

"Specifically, the conventional univariate method of analysis assumes that the data have been obtained from populations that have the well-known normal (multivariate) form, that the degree of variability (covariance) among the levels of the variable conforms to a spherical pattern, and that the data conform to independence assumptions. Since the data obtained in many areas of psychological inquiry are not likely to conform to these requirements … researchers using the conventional procedure will erroneously claim treatment effects when none are present, thus filling their literatures with false positive claims." – Keselman et al. 2001

It could be argued that limits on the computational power available to researchers have led to a concentration on the limits of models which can be estimated without recourse to iterative algorithms. On this account, sphericity and its associated literature can be considered a historically specific issue. Nevertheless, while the development of methods such as those described in Worsley and Friston (1995) and implemented in SPM99 do not explicitly refer to the repeated measures designs they are in fact mathematically identical, as we will now show.

The assumptions required for both sorts of analysis can be most easily defined by considering the variance-covariance matrix of the observations. Consider a population variance-covariance matrix for a measurement x under k treatments with n subjects. The measurements on each subject can be viewed as a k-element vector with associated covariance matrix

$$\Sigma_x = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & & & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{bmatrix} \qquad (1)$$

This matrix can be estimated on the basis of the data by the sample variance-covariance matrix

$$\hat{\Sigma}_x = S_x = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & & & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \qquad (2)$$

What is the most liberal criterion which we can apply to this matrix without violating the assumptions underlying repeated measures ANOVA? By definition, the following equivalent properties are obeyed by the variance covariance matrix if the covariance structure is **spherical**

$$\begin{aligned} &\forall j \neq j' \\ &\sigma_{jj} + \sigma_{j'j'} - 2\sigma_{jj'} = 2\lambda \\ &\sigma^2_{X_j - X_{j'}} = 2\lambda \\ &\overline{\sigma}_{jj} - \overline{\sigma}_{jj'} = \lambda \end{aligned} \qquad (3)$$

In words, these statements in (3) say that for any pair of levels, the sum of their variances minus twice their covariance is equal to a constant. Equivalently, the variance of the difference between a pair of levels is the same for all pairs. Intuitively it is clear that this assumption is violated, for example, in the case of temporal autocorrelation. In such a case, by definition, pairs of nearby levels (in this case time points) are more highly correlated than those separated by longer times. Another example might be an analysis which took three activations from each of a group member of two groups of subjects. Consider, for example, activation while reading, while writing, and while doing arithmetic. Imagine one wanted to test whether the populations from which two groups were drawn were significantly different, but considering the three types of task together. This would involve an F-test, but it would assume that the covariation between the reading and writing activations was the same as that between the writing and arithmetic. This may or may not be true. If it were not, sphericity would be violated, and the test would be overly liberal.

To illuminate the derivation of the term sphericity, we state without proof an equivalent condition to that in (3). This condition is that there can be found an orthonormal projection matrix M^* which can be used to transform the variables X of the original distribution to a new set of variables Y. This new set of variables has a

covariance matrix Σ_Y which is **spherical** (ie. is a scalar multiple of the identity matrix). This relation will be exploited in the next section.

$$M^{*}M^{*'} = I$$

$$M^{*}\Sigma_x M^{*'} = \Sigma_y = \lambda I = \begin{bmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \lambda \end{bmatrix} \qquad (4)$$

$$\lambda = \frac{\sigma_{jj} + \sigma_{j'j'} - 2\sigma_{jj'}}{2}$$

It is worth mentioning for completeness that while the sphericity condition is necessary it is not necessarily that intuitive nor is it clear by inspection whether a dataset conforms. Historically therefore a more restricted sufficient condition has been adopted, namely compound symmetry. A matrix has compound symmetry if it has the following form:

$$\Sigma_x = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & & & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} \qquad (5)$$

To describe the relation in (5) in words, all the within group variances are assumed equal, and separately all the covariances are assumed equal and this can be assessed directly from the data. There do exist statistical approaches to assess whether a dataset deviates from sphericity such as Mauchly's test (see eg. Winer et al. 1991), but these have very low power.

2.2. A measure of departure from sphericity

Using the notation of the variance-covariance matrix in (1), we can define a measure of departure from sphericity after Box (1954)

$$\varepsilon = \frac{k^2(\overline{\sigma}_{jj} - \overline{\sigma}_{..})^2}{(k-1)\sum\sum(\sigma_{jj'} - \sigma_{j.} - \sigma_{.j} + \sigma_{..})} \qquad (6)$$

where σ_{jj} = mean for diagonal entries, $\sigma_{..}$ = mean of all entries, $\sigma_{j.}$ = mean for row j, $\sigma_{.j}$ = mean for column j. We can rephrase (6) in terms of λ_i, the characteristic roots of the transformed matrix Σ_y from (4)

$$\varepsilon = \frac{\left(\sum \lambda_i\right)^2}{(k-1)\sum \lambda_i^2} \qquad (7)$$

We now informally derive upper and lower bounds for our new measure. If Σ_y is spherical i.e. of form λI then the roots are equal and since Σ_y is of size $(k-1)\times(k-1)$ then

$$\varepsilon = \frac{\left(\sum \lambda\right)^2}{(k-1)\sum \lambda^2} = \frac{((k-1)\lambda)^2}{(k-1)(k-1)\lambda^2} = 1 \qquad (8)$$

At the opposite extreme, it can be shown that for a maximum departure from sphericity,

$$\Sigma_y = \begin{bmatrix} c & c & \cdots & c \\ c & c & & c \\ \vdots & \vdots & & \vdots \\ c & c & \cdots & c \end{bmatrix} \qquad (9)$$

for some constant c. Then the first characteristic root $\lambda_1 = (k-1)c$ and the rest are zeros. From this we see that

$$\varepsilon = \frac{\left(\sum \lambda_i\right)^2}{(k-1)\sum \lambda_i^2} = \frac{\lambda_1^2}{(k-1)\lambda_1^2} = \frac{1}{(k-1)} \qquad (10)$$

Thus we have the following bounds:

$$\frac{1}{(k-1)} \le \varepsilon \le 1 \qquad (11)$$

We have seen that the measure ε can be well defined using basic matrix algebra and expresses the degree to which the standard assumptions underlying the distribution are violated. In the following section we employ this measure to systematically protect ourselves against falsely positive inferences by correcting the parameters of the F distribution.

2.3. Correcting degrees of freedom using ε: The Satterthwaite approximation

Box's motivation for using this measure for the departure from sphericity was in order to harness an approximation due to Satterthwaite. This deals with the fact that the actual distribution of the variance estimator is not χ^2 if the data is not spherical, and thus the F statistic used for hypothesis testing is inaccurate. The solution adopted is to approximate the true distribution with a moment matched scaled χ^2 distribution matching the first and second moments. Using this approximation in the context of repeated measures ANOVA with k measures and n subjects, the overall F statistic will be distributed as $F[(k-1)\varepsilon, (n-1)(k-1)\varepsilon]$. To understand the elegance of this

approach note that, as shown above, when the sphericity assumptions underlying the model are met, $\varepsilon = 1$ and the F distribution is then just $F[(k-1),(n-1)(k-1)]$, the standard degrees of freedom for this model. The correction 'vanishes' when it is not needed.

Finally we note that this approximation has been adopted for neuroimaging data in SPM. Consider the expression for the effective degrees of freedom from Worsley and Friston (1995) and applied in SPM99. There

$$\nu = \frac{\text{trace}(RV)^2}{\text{trace}(RVRV)} \qquad (12)$$

Compare (7) above, and see Chapter 7 for a derivation. If we remember that the conventional degrees of freedom for the t statistic is k-1 and consider ε as a correction for the degrees of freedom, then

$$\nu = (k-1)\varepsilon = (k-1)\frac{\left(\sum \lambda_i\right)^2}{(k-1)\sum \lambda_i^2} = \frac{\left(\sum \lambda_i\right)^2}{\sum \lambda_i^2} = \frac{\text{trace}(RV)^2}{\text{trace}(RVRV)} \qquad (13)$$

Thus SPM applies the Satterthwaite approximation to correct the F statistic, implicitly using a measure of sphericity violation. In the next section we will see that this approach is similar to that employed in conventional statistical packages.

2.4. But which covariance matrix is used for the estimation of corrected degrees of freedom?

Returning to the classical approach, in practice of course we do not know Σ_x and so it will be estimated by S_x the sample covariance matrix (2). From this we can generate an $\hat{\varepsilon}$ by substituting $s_{jj'}$ for the $\sigma_{jj'}$ in (6). This correction using the sample covariance is often referred to as 'Greenhouse-Geisser' (e.g. Winer et al 1991). An extensive literature treats the further steps in harnessing this correction and some variants on it in practice. For example, correction can be made more conservative by taking the lower bound on $\hat{\varepsilon}$ as derived in (10). This highly conservative test is (confusingly) also referred to as the 'Greenhouse-Geisser conservative correction'.

The important point to note however, is that the construction of the F statistic is predicated upon a model covariance structure which satisfies the assumptions of sphericity as outlined above, but the degrees of freedom are adjusted based on the **sample** covariance structure. This contrasts with the approach taken in, for example, SPM99 which assumed either IID errors (a covariance matrix which is a scalar multiple of the identity matrix) or a simple autocorrelation structure, but corrected the degrees of freedom only on the basis of the **modelled** covariance structure. In the IID case, no correction was made no matter what the data looked like. In the autocorrelation case, an appropriate correction was made, but ignoring the sample covariance matrix and assuming that the data structure was as modelled. These strategic differences are summarised in a table at the end of the section.

2.5. Estimating the covariance components.

In SPM2, a refinement is made in that the covariance structure can be estimated from the data. This is accomplished by defining a basis set for the covariance matrix and then using an iterative Restricted Maximum Likelihood (ReML) algorithm to estimate parameters controlling these bases. In this way a wide range of sphericity violations can be modelled explicitly. Examples include temporal autocorrelation and more subtle effects of correlations induced by taking several measures on each of several subjects. In all cases however, the modelled covariance structure is used to calculate the appropriate degrees of freedom using the moment-matching procedure described above in Chapter 7. We do not discuss estimation in detail since this is covered in Chapters 13 & 17. We simply state the form of the parameterisation of the variance components, and give illustrations of their typical form. We model the covariance matrix as

$$\Sigma_X = \sum \lambda_j Q_j \qquad (13)$$

where λ_j are some hyperparameters and Q_j represent some basis set for the covariance matrices. Q_j embodies the form of the covariance components at any level and could model different variances for different blocks of data or different forms of correlations within blocks. Estimation takes place using a ReML procedure where the model coefficients and variance estimates are re-estimated iteratively.

As will be discussed in the final section, what we in fact estimate is an intercorrelation matrix or normalised covariance for many voxels at once. This can be multiplied by a scalar variance estimate calculated for each voxel separately. Since this scalar does not affect the correlation structure the corrected degrees of freedom are the same for all voxels.

2.6. Schematic form of covariance constraints.

These can be thought of as 'design matrices' for the second-order behaviour of the response variable and form a basis set for estimating the error covariance, and the hyperparameters scale the contribution of each constraint.

Figure 1 illustrates two possible applications of this technique. One for first level analysis and one for random effects.

(FIGURE 1 ABOUT HERE)

There are three independent choices to be made in dealing with data that may not be distributed according to ones model. We can consider the issues described above separately and could in principle choose any combination of them for an analysis strategy. The following table illustrates the actual combination used in the approaches described in this chapter.

(TABLE 1 ABOUT HERE)

3. Pooling

So far we have discussed the variance structure of our data drawing from a univariate or mass univariate approach. In this section we ask whether we can harness the fact that our voxels come from the same brain. Firstly we will motivate the question by demonstrating that sphericity estimation derives a noisy measure, and that it might therefore be beneficial to pool over voxels. We will then show that under certain assumptions this strategy can be justified, and illustrate an implementation.

3.1. Simulating noise in sphericity measures

To assess the practicality of voxel-wise estimation of the covariance structure we simulated 10000 voxels drawn from a known population with 8 measures of 3 levels of a repeated measure. For each voxel we estimated the variance-covariance matrix using a ReML procedure with a basis set corresponding to the true distribution. We then calculated the ε correction factor and plotted a histogram for the distribution of this over the 10000 voxels (See figure 2). Note the wide distribution even for a uniform underlying variance structure, emphasising the utility of pooling the estimate over many voxels or even the whole brain to generate an intercorrelation matrix. The voxel-wide estimate was 0.65, which is higher (more spherical) than the average of the voxel-wise estimates illustrated below which is 0.56. In this case the ε for the generating distribution was indeed 0.65.

(FIGURE 2 ABOUT HERE)

3.2. Degrees of freedom reprised

As this simulation shows, to make the estimate of effective degrees of freedom valid we require very precise estimates of non-sphericity. However, as mentioned at the start of this chapter 'pooling' is problematic because the true error variance may change from voxel to voxel. We will now expand upon the form described in (13) to describe in detail the strategy used by current fMRI analysis packages like SPM and **multistat** (Worsley *et al* 2002). As stated we hyper-parameterize the error covariance in terms of a single-hyperparameter that is voxel-specific and a series of voxel-independent hyperparameters, that can be estimated with high precision over a large number of voxels. For the *i*th voxel we are then assuming $\varepsilon_i \sim N\{0, \sigma_i^2(\lambda_1 Q_1 +, \ldots, \lambda_n Q_n)\}$. This allows one to use the reduced single-hyperparameter model and the effective degrees of freedom as in (1) while still allowing error variance to vary from voxel to voxel. Here the pooling is over 'similar' voxels (*e.g.* those that activate are spatially close) and that are assumed to express various error variance components in the same proportion but not in the same amounts. In summary, we factorize the error covariance into voxel-specific variance and temporal covariance that is the same for all voxels in the subset. This effectively factorizes the spatiotemporal covariance into non-stationary spatial covariance and

stationary temporal non-sphericity. This enables pooling for, and only for, estimates of temporal covariances.

The problem is that estimating multiple hyperparameters (i) usually requires an iterative procedure that is computationally prohibitive for massive numbers of voxels and (ii) gives imprecise estimates that render the inference less sensitive. The solution, most commonly adopted is to retain the simplicity of the conventional single hyperparameter approach and use a generalized linear model with known non-sphericity (Worsley and Friston 1995). In this approach the different variance components are estimated *a priori* and combined to give the non-sphericity structure in terms of a single error covariance matrix V. This reduces a problem with multiple variance components i.e. $\varepsilon \sim N\{0, \lambda_1 Q_1 + , \ldots , \lambda_n Q_n\}$ into a single component model $\varepsilon \sim N\{0, \sigma^2 V\}$ with a single hyperparameter σ^2. Inference can then proceed using OLS and the appropriate adjustments for the non-sphericity. As described in the previous secion, the hyperparameter is estimated using the sum of squared residuals and the Satterthwaite approximation is used to give the effective degrees of freedom.

The effective degrees of freedom can be thought of as adjusted for known non-sphericity. Here $V \propto \lambda_1 Q_1 + , \ldots , \lambda_n Q_n$ where the constant of proportionality is chosen [arbitrarily] to render $tr\{V\}$ equal to its size, *c.f.* a correlation matrix. As stated above, the Satterthwaite approximation is exactly the same as that employed in the Greenhouse-Geisser (G-G) correlation for non-sphericity in commercial packages. However, there is a fundamental distinction between the SPM adjustment and the G-G correction. This is because the non-sphericity V enters as a known constant (or as an estimate with very high precision). In contradistinction, the non-sphericity in G-G uses the sample covariance matrix or multiple hyperparameter estimates, usually ReML, based on the data themselves to give $\hat{V} \propto \hat{\lambda}_1 Q_1 + , \ldots , \hat{\lambda}_n Q_n$. This gives corrected degrees of freedom that are generally too high, leading to mildly capricious inferences[*]. Compare the following with (12)

$$v_{GG} = tr\{R\}\varepsilon_{GG} = \frac{tr\{R\hat{V}\}^2}{tr\{R\hat{V}R\hat{V}\}} \qquad (14)$$

The reason the degrees of freedom are too high is that G-G fails to take into account the variability in the ReML hyperparameter estimates and ensuing variability in $\hat{V}$. There are simple solutions to this that involve abandoning the single variance component model and forming statistics using multiple hyperparameters directly (Kiebel et al. 2003).

The critical difference between conventional G-G corrections and the SPM-adjustment lies in the fact that SPM is a mass univariate approach that can pool non-sphericity estimates $\hat{V}$ over subsets of voxels to give a highly precise estimate V. Conventional univariate packages cannot do this because there is only one data sequence.

[*] This is only a problem if the variance components interact (*e.g.* as with serial correlations in fMRI).

3.3. Separable errors

The final issue addressed in this chapter is how relative values of the voxel-independent hyperparameters are estimated and how precise these estimates are. There are many situations in which the hyperparameters of mass-univariate observations factorize. In the present context we can regard fMRI time-series as having both spatial and temporal correlations among the errors that factorize in a Kronecker tensor product. Consider the data matrix $Y = [y_i, \ldots, y_n]$ with one column, over time, for each of *n* voxels. The spatio-temporal correlations can be expressed as the error covariance matrix in a vectored GLM

$$Y = X\beta + \varepsilon$$

$$vec\{Y\} = \begin{bmatrix} y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X & & 0 \\ & \ddots & \vdots \\ 0 & \cdots & X \end{bmatrix} \begin{bmatrix} \beta_i \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad (15)$$

$$\mathrm{cov}\{vec\{\varepsilon\}\} = \Sigma \otimes V = \begin{bmatrix} \Sigma_1 V & & \Sigma_{1n} V \\ & \ddots & \vdots \\ \Sigma_{n1} V & \cdots & \Sigma_n V \end{bmatrix}$$

Note that (15) assumes a separable form for the errors. This is the key assumption underlying the pooling procedure. Here V embodies the temporal non-sphericity and Σ the spatial non-sphericity. Notice that the elements of Σ are voxel specific whereas the elements of V are the same for all voxels. We could now enter the vectored data into a ReML scheme, directly, to estimate the spatial and temporal hyperparameters. However, we can capitalize on the assumed separable form of the non-sphericity over time and space by only estimating the hyperparameters of V and then use the usual estimator (Worsley et al. 1995) to compute a single hyperparameter $\hat{\Sigma}_i$ for each voxel according to (15).

The hyperparameters of V can be estimated with the algorithm presented in Friston *et al* (2003, Appendix 1). This uses a Fisher scoring scheme to maximize the log likelihood $\ln p(Y \mid \lambda, \Sigma)$ (i.e. the ReML objective function) to find the ReML estimates. In the current context this scheme is

$$\begin{aligned} &\lambda \leftarrow \lambda + W^{-1} g \\ &g_i = \frac{\partial \ln p(Y \mid \lambda, \Sigma)}{\partial \lambda_i} = -\tfrac{n}{2} tr\{PQ_i\} + \tfrac{1}{2} tr\{P^T Q_i P Y \Sigma^{-1} Y^T\} \\ &W_{ij} = E\left\{ -\frac{\partial^2 \ln p(Y \mid \lambda, \Sigma)}{\partial \lambda_{ij}^2} \right\} = \tfrac{n}{2} tr\{PQ_i PQ_j\} \end{aligned} \qquad (16)$$

$$\begin{aligned} &P = V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} \\ &V = \lambda_1 Q_1 + , \ldots, \lambda_n Q_n \end{aligned}$$

where E{} is the expectation operator. Notice that the Kronecker tensor products and vectorized forms disappear. Critically *W,* the precision of the hyperparameter estimates, increases linearly with the number of voxels. With sufficient voxels this allows us to enter the resulting estimates, through *V*, into (15) as known variables, because they are so precise. The nice thing about (16) is that the data enter only as $Y\Sigma^{-1}Y^T$ whose size is determined by the number of scans as opposed to the massive number of voxels. $Y\Sigma^{-1}Y^T$ is effectively the sample temporal covariance matrix, sampling over voxels (after spatial whitening) and can be assembled voxel by voxel in a memory efficient fashion. (16) assumes that we know the spatial covariances. In practice, $Y\Sigma^{-1}Y^T$ is approximated by selecting voxels that are spatially dispersed (so that $\Sigma_{ij} = 0$) and scaling the data by a ReML estimate of Σ_i^{-1} obtained non-iteratively assuming temporal sphericity. This is the approach used in SPM and is described further in chapter 17.

4. Conclusions

We have shown that classical and recent approaches did not explicitly estimate the covariance structure of the noise in their data but instead assumed it had a tractable form, and then corrected for any deviations from the assumptions by an approximation. This approximation could be based on the actual data, or on a defined structure which was imposed on the data. More modern approaches explicitly model those types of covariation which the experimenter expects to find in the data. This estimation can be noisy, and is therefore best conducted over pooled collections of voxels.

The use of this technique will allow the experimenter to perform types of analysis which were previously 'forbidden' under the less sophisticated approaches. These are of real interest to many researchers and include: better estimation of autocorrelation structure for fMRI data; the ability to take more than one scan per subject to the second level and thus conduct F-tests to draw conclusions about populations. In event-related studies where the exact form of the haemodynamic response can be critical, more than one aspect of this response can be analysed in a random effects context. For example a canonical form and a measure of latency or spread in time can jointly express a wide range of real responses. Alternatively, a more general basis set (e.g. Fourier or finite impulse response) can be used. There is of course a trade-off between better detectability by making assumptions about the form of the haemodynamic response, or greater robustness to variable forms. Finally, comparisons can be made between fMRI and PET data at the second level.

5. Key references

Box GEP (1954), Some theorems on quadratic forms applied in the study of analysis of variance problems. The Annals of Mathematical Statistics, 25, 290-302.

Friston KJ, Josephs O, Zarahn E, Holmes AP, Rouquette S, Poline J. (2000) To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. Neuroimage. Aug;12(2):196-208.

Friston KJ, Glaser DE, Henson RN, Kiebel S, Phillips C, Ashburner J. (2002) Classical and Bayesian inference in neuroimaging: applications. Neuroimage. 2002 Jun;16(2):484-512.

Keselman HJ, Algina J, Kowalchuk RK. (2001), The analysis of repeated measures designs: a review. Br J Math Stat Psychol. 54(Pt 1):1-20.

Kiebel S.J., Glaser, D.E. and Friston, K.J. (2003) A heuristic for the degrees of freedom of statistics based on multiple hyperparameters. Submitted.

Ollinger JM & McAvoy MP, A Homogeneity Correction for Post-hoc ANOVAs in fMRI, HBM2000.

Winer BJ et al (1991), Statistical Principles in Experimental Design, McGraw-Hill.

Worsley KJ & Friston KJ (1995), Analysis of fMRI Time-Series Revisited - Again, Neuroimage 2:173-181.

Table 1

	Classical approach Greenhouse-Geisser	SPM99	SPM 2
Choice of model	Assume sphericity	Assume IID or AR(1)	Use ReML to estimate covariance structure parameterised with a basis set
Corrected degrees of freedom based on covariance structure of …	Actual data	Model	Model
Estimation of degrees of freedom is voxel-wise or for whole brain		Whole brain	Whole brain

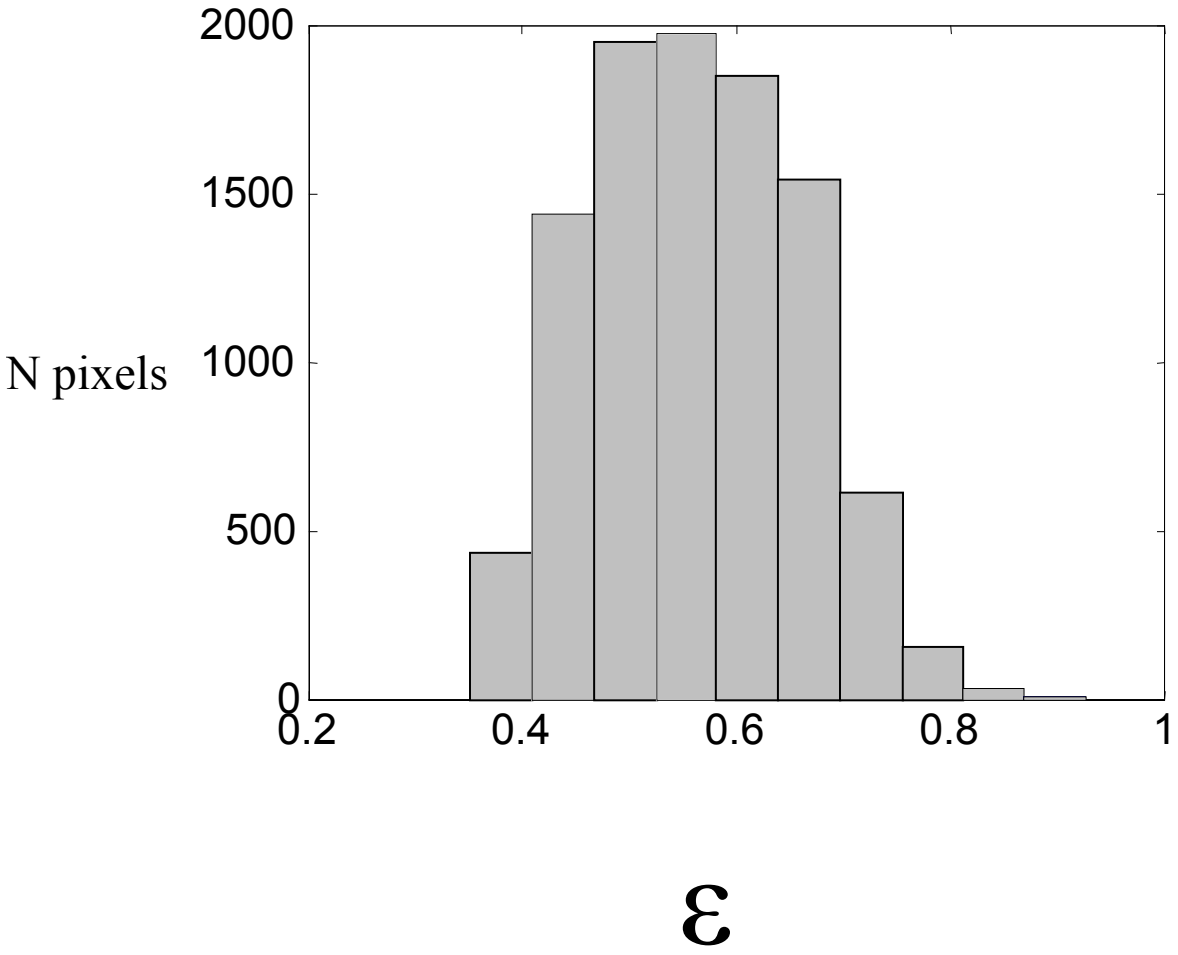


Figure 2

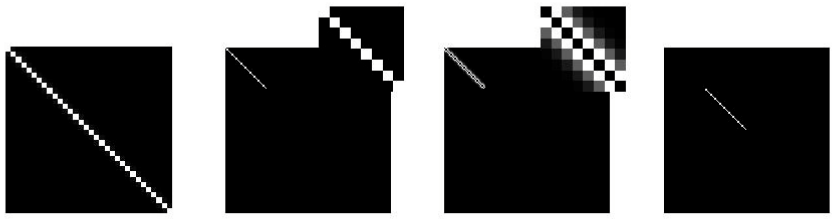
Histogram illustrating voxel-wise sphericity measure, ε, for 10000 simulated voxels drawn from a known population with 8 measures of 3 levels of a repeated measure. Average of the illustrated voxel-wise estimates is 0.56. The voxel-wide estimate was 0.65, and the ε for the generating distribution was indeed 0.65.

Figure 1

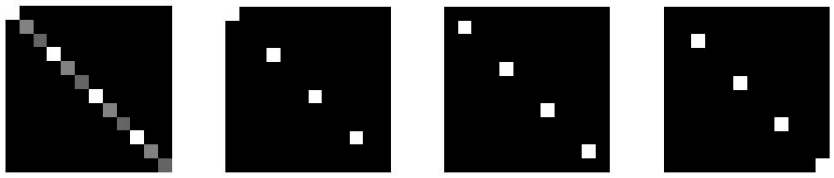
Two examples:

Top row: Here we imagine that we have a number of observations over time and a number of subjects. We decide to model the autocorrelation structure by a sum of a simple autocorrelation AR(1) component and a white noise component. A separately scaled combination of these two can approximate a wide range of actual structures, since the white noise component affects the 'peakiness' of the overall autocorrelation. For this purpose we generate two bases for each subject, and here we illustrate the first three overall. The first is an identity matrix (no correlation) restricted to the observations from the first subject, the second is the same but blurred in time and with the diagonal removed. The third illustrated component is the white noise for the second subject and so on.

Second row: In this case we imagine that we have three measures for each of several subjects. For example, as suggested above, consider a second level analysis in which we have a scan while reading, while writing and while doing arithmetic for several members of a population. We would like to make an inference about the population from which the subjects are drawn. We want to estimate what the covariation structure of the three measures is, but we assume that this structure is the same for each of the individuals. Here we generate three bases in total, one for all the reading scores, one for all the writing, and one for all the arithmetic. We then iteratively estimate the hyperparameters controlling each basis, and hence the covariance structure.



$$C_\varepsilon = \lambda_1 Q_1 + \lambda_2 Q_2 + \lambda_3 Q_3 + \ldots$$



$$C_\varepsilon = \lambda_1 Q_1 + \lambda_2 Q_2 + \lambda_3 Q_3$$