The general linear model

S.J. Kiebel A.P. Holmes

Contents

1 Introduction

In the absence of prior anatomical hypotheses regarding the physical location of a particular function, the statistical analysis of functional mapping experiments must proceed by assessing the acquired data for evidence of an experimentally induced effect at every intracerebral voxel individually and simultaneously.

After reconstruction, realignment, spatial normalisation and (possibly) smoothing, the data are ready for statistical analysis. This involves two steps: Firstly, statistics indicating evidence against a null hypothesis of no effect at each voxel are computed. An image of these statistics is then produced. Secondly, this statistical image must be assessed, reliably locating voxels where an effect is exhibited whilst limiting the possibility of false positives. In this chapter we shall address the former topic, the formation of an appropriate statistical image.

Current methods for assessing the data at each voxel are predominantly parametric: Specific forms of probability distribution are assumed for the data, and hypotheses specified in terms of models assumed for the (unknown) parameters of these distributions. The parameters are estimated and a statistic reflecting evidence against the null hypothesis formed. Statistics with a known null distribution are used such that the probability of obtaining a statistic, given that the null hypothesis is true, can be computed. This is hypothesis testing in the classical parametric sense. The majority of the statistical models used are special cases of the General Linear Model.

SPM has become an acronym in common use for the theoretical framework of voxel based analysis of functional imaging data, for the software package implementing this procedure, and for the statistical image (*Statistical Parametric Map*). Here we shall take SPM to refer to (i) the software package in its current version SPM99 and (ii) the conceptual and theoretical framework.

In what follows, we first go through the equations for the general linear model with a spherical error distribution (i.e. we assume an independently and identically distributed error). This theoretical part is presented without any reference to PET or fMRI data and is orientated towards a description as it can be found in a classical statistics textbook. In the third section, we turn to the data in question and illustrate the use of the general linear model on some PET data examples. In the fourth and final section, we introduce the linear model used for fMRI data. This model is a linear model with a normally distributed and non-spherical error.

2 The General Linear Model

Before turning to the specifics of PET and fMRI, we consider the general linear model, which requires some basic matrix algebra and statistical concepts. These will be used to develop an understanding of classical hypothesis testing. Healy (Healy, 1986) presents a brief summary of matrix methods relevant to statistics. Newcomers to statistical methods are directed towards Mould's excellent text "Introductory Medical Statistics" (Mould, 1989), while the more mathematically experienced will find Chatfield's "Statistics for Technology" (Chatfield, 1983) useful. Draper & Smith (Draper and Smith, 1981) give a good exposition of matrix methods for the general linear model, and go on to describe regression analysis in general. The definitive tome for practical statistical experimental design is Winer *et al.* (Winer et al., 1991). An excellent book about experimental design is (Yandell, 1997). A rather advanced, but very useful, text on linear models is (Christensen, 1996).

2.1 The General Linear Model — Introduction

Suppose we are to conduct an experiment during which we will measure a *response variable* (such as rCBF at a particular voxel) Y_j, where $j = 1, \ldots, J$ indexes the observation. Y_j is a random variable, conventionally denoted by a capital letter.[1] Suppose also that for each observation we have a set of L ($L < J$) *explanatory* variables (each measured without error) denoted by x_{jl}, where $l = 1, \ldots, L$ indexes the explanatory variables. The explanatory variables may be continuous (or sometimes discrete) *covariates*, functions of covariates, or they may be *dummy* variables indicating the *levels* of an experimental *factor*.

A *general linear model* explains the response variable Y_j in terms of a linear combination of the explanatory variables plus an error term:

$$Y_j = x_{j1}\beta_1 + \ldots + x_{jl}\beta_l + \ldots + x_{jL}\beta_L + \epsilon_j \tag{1}$$

Here the β_l are (unknown) parameters, corresponding to each of the L explanatory variables x_{jl}. The errors ϵ_j are independent and identically distributed normal random variables with zero mean and variance σ^2, written $\epsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Linear models with other error distributions are *Generalised Linear Models*, for which the acronym GLM is usually reserved.

2.1.1 Examples, dummy variables

Many classical parametric statistical procedures are special cases of the general linear model. We will illustrate this point by going through the equations for two well-known models.

Linear regression

A simple example is linear regression, where only one continuous explanatory variable x_j is measured (without error) for each observation $j = 1, \ldots, J$. The model is usually written as:

$$Y_j = \mu + x_j\beta + \epsilon_j \tag{2}$$

where the unknown parameters are μ, a *constant term* in the model, the regression slope β and $\epsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. This can be re-written as a general linear model by the use of a dummy variable taking the value $x_{j1} = 1$ for all j:

$$Y_j = x_{j1}\mu + x_{j2}\beta_2 + \epsilon_j \tag{3}$$

which is of the form of Eq. 1 on replacing β_1 with μ.

Two-sample t-test

Similarly the two-sample t-test is a special case of a general linear model: Suppose Y_{j1} and Y_{j2} are two independent groups of random variables: The two-sample t-test assumes $Y_{qj} \overset{iid}{\sim}$

[1] We talk of *random variables*, and of observations prior to their measurement, because classical (frequentist) statistics is concerned with what could have occurred in an experiment. Once the observations have been made, they are known, the residuals are known, and there is no randomness.

$\mathcal{N}(\mu_q, \sigma^2)$, for $q = 1, 2$, and assesses the null hypothesis $\mathcal{H}: \mu_1 = \mu_2$. The index j indexes the data points in both groups. The standard statistical way of writing the model is:

$$Y_{qj} = \mu_q + \epsilon_{qj} \tag{4}$$

The q subscript on the μ_q indicates that there are two *levels* to the group *effect*, μ_1 and μ_2. Here, $\epsilon_{qj} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. This can be re-written using two dummy variables x_{qj1} and x_{qj2} as:

$$Y_{qj} = x_{qj1}\mu_1 + x_{qj2}\mu_2 + \epsilon_{qj} \tag{5}$$

which is of the form of Eq. 1 after re-indexing for qj. Here the dummy variables indicate group membership, where x_{qj1} indicates whether observation Y_{qj} is from the first group, in which case it has the value 1 when $q = 1$, and 0 when $q = 2$. Similarly, $x_{qj2} = \begin{cases} 0 & \text{if } q = 1 \\ 1 & \text{if } q = 2 \end{cases}$

2.2 Matrix formulation

In the following few subsections, we use the general linear model in its matrix formulation and derive a least-squares parameter estimation. After this, we describe how one can make inferences based on a contrast of the parameters. This theoretical treatment of the model is useful to derive a set of equations that can be used for the analysis of any data set that can be formulated in terms of the general linear model.

The general linear model can be succinctly expressed using matrix notation. Consider writing out Eq. 1 in full, for each observation j, giving a set of simultaneous equations:

$$\begin{aligned}
Y_1 &= x_{11}\beta_1 + \ldots + x_{1l}\beta_l + \ldots + x_{1L}\beta_L + \epsilon_1 \\
\vdots &= \vdots \\
Y_j &= x_{j1}\beta_1 + \ldots + x_{jl}\beta_l + \ldots + x_{jL}\beta_L + \epsilon_j \\
\vdots &= \vdots \\
Y_J &= x_{J1}\beta_1 + \ldots + x_{Jl}\beta_l + \ldots + x_{JL}\beta_L + \epsilon_J
\end{aligned}$$

This has an equivalent matrix form:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_j \\ \vdots \\ Y_J \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & \cdots & x_{jl} & \cdots & x_{jL} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{J1} & \cdots & x_{Jl} & \cdots & x_{JL} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_l \\ \vdots \\ \beta_L \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_j \\ \vdots \\ \epsilon_J \end{pmatrix}$$

which can be written in matrix notation as

$$Y = X\beta + \epsilon \tag{6}$$

where Y is the column vector of observations, ϵ the column vector of error terms, and β the column vector of parameters; $\beta = [\beta_1, \cdots, \beta_l, \cdots, \beta_L]^T$. The $J \times L$ matrix X, with jl^{th}

element x_{jl} is the *design matrix*. This has one row per observation, and one column (explanatory variable) per model parameter. The important point about the design matrix is that it is a near complete description of our model with the remainder of the model being in the error term. The design matrix is where the experimental knowledge about the expected signal is quantified.

2.3 Parameter estimation

Once an experiment has been completed, we have observations of the random variables Y_j, which we denote by y_j. Usually, the simultaneous equations implied by the general linear model (with $\epsilon = 0$) cannot be solved, because the number of parameters L is typically chosen to be less than the number of observations J. Therefore, some method of estimating parameters that "best fit" the data is required. This is achieved by the method of *ordinary least squares*.

Denote a set of parameter estimates by $\tilde{\beta} = [\tilde{\beta}_1, \ldots, \tilde{\beta}_L]^T$. These parameters lead to *fitted values* $\tilde{Y} = [\tilde{Y}_1, \ldots, \tilde{Y}_J]^T = X\tilde{\beta}$, giving residual errors $e = [e_1, \ldots, e_J]^T = Y - \tilde{Y} = Y - X\tilde{\beta}$. The *residual sum-of-squares* $S = \sum_{j=1}^{J} e_j^2 = e^T e$ is the sum of the square differences between the actual and fitted values, and thus measures the fit of the model with these parameter estimates.[2] The *least squares* estimates are the parameter estimates which minimise the residual sum-of-squares. In full:

$$S = \sum_{j=1}^{J} \left(Y_j - x_{j1}\tilde{\beta}_1 - \ldots - x_{jL}\tilde{\beta}_L \right)^2$$

This is minimized when:

$$\frac{\partial S}{\partial \tilde{\beta}_l} = 2 \sum_{j=1}^{J} (-x_{jl}) \left(Y_j - x_{j1}\tilde{\beta}_1 - \ldots - x_{jL}\tilde{\beta}_L \right) = 0$$

This equation is the l^{th} row of $X^T Y = (X^T X)\tilde{\beta}$. Thus, the least squares estimates, denoted by $\hat{\beta}$, satisfy the *normal equations*:

$$X^T Y = (X^T X)\hat{\beta} \tag{7}$$

For the general linear model, the least squares estimates are the *maximum likelihood estimates*, and are the *Best Linear Unbiased Estimates*[3]. That is, of all linear parameter estimates consisting of linear combinations of the observed data whose expectation is the true value of the parameters, the least squares estimates have the minimum variance.

If $(X^T X)$ is invertible, which it is if and only if the design matrix X is of full rank, then the least squares estimates are:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{8}$$

2.3.1 Overdetermined models

If X has linearly dependent columns, it is *rank deficient*, $(X^T X)$ is singular, and has no inverse. In this case the model is overparameterised: There are infinitely many parameter sets describing

[2] $e^T e$ is the L_2 norm of e — geometrically equivalent to the distance between the model and data

[3] Gauss-Markov theorem

the same model. Correspondingly, there are infinitely many least squares estimates $\hat{\beta}$ satisfying the normal equations. We will illustrate the overdetermined case by an example and discuss the solution that is adopted in SPM.

2.3.2 One way ANOVA Example

A simple example of such a model is the classic Q group one-way analysis of variance (ANOVA) model. Generally, an Anova determines the variability in the measured response which can be attributed to the effects of factor levels. The remaining unexplained variation is used to assess the significance of the effects (Yandell, 1997), page 4 and pages 202ff. The model for a one-way Anova is given by

$$Y_{qj} = \mu + \alpha_q + \epsilon_{qj} \tag{9}$$

where Y_{qj} is the j^{th} observation in group $q = 1, \ldots, Q$. This model clearly does not uniquely specify the parameters: For any given μ and α_q, the parameters $\mu' = \mu + d$ and $\alpha'_q = \alpha_q - d$ give an equivalent model for any constant d. That is, the model is indeterminate up to the level of an additive constant between the constant term μ and the group effects α_q. Similarly for any set of least squares estimates $\hat{\mu}, \hat{\alpha}_q$. Here there is one degree of indeterminacy in the model, resulting in the design matrix having rank Q, which is one less than the number of parameters (the number of columns of X). If the data vector Y has observations arranged by group, then for three groups ($Q = 3$), the design matrix and parameter vectors are

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

Clearly this matrix is rank deficient: The first column is the sum of the others. Therefore, in this model, one cannot test in this model for the effect of one or more groups. However, note that the addition of the constant μ does not effect the relative differences between pairs of group effects. Therefore, *differences* in group effects are uniquely estimated regardless of the particular set of parameter estimates used. In other words, even if the model is overparameterised, there are still useful linear combinations of parameters (i.e. differences between pairs of group effects). This important concept will emerge in many designs, especially for PET and multi-subject data. It will be treated more thoroughly in §2.5.3 (Estimable functions, contrasts).

2.3.3 The pseudoinverse constraint

In the overdetermined case, a set of least squares estimates may be found by imposing constraints on the estimates, or by inverting $(X^T X)$ using a pseudoinverse technique which essentially implies a constraint. In either case it is important to remember that the actual estimates

obtained depend on the particular constraint or pseudoinverse method chosen. This has implications for inference (§2.5): It is only meaningful to consider functions of the parameters that are uninfluenced by the particular constraint chosen.

There are some obvious constraints that are based on removing columns from the design matrix. In the one-way ANOVA example, one can remove the constant term to construct a design matrix which has linearly independent columns. For more complex designs, the form of the design matrix can change a lot such that it becomes difficult to visually recognize the original model. Therefore, in SPM, the overall principle is that each experimentally induced effect is represented by one or more regressors. This excludes the removal of columns as an option to deal with overdetermined models.

Alternatively a pseudoinverse method can be used. Let $(X^TX)^-$ denote the pseudoinverse of (X^TX). Then we can use $(X^TX)^-$ in place of $(X^TX)^{-1}$ in Eq. 8. A set of least squares estimates are given by $\hat{\beta} = (X^TX)^-X^TY = X^-Y$. The pseudoinverse function implemented in MATLAB gives the Moore-Penrose pseudoinverse.[4] This results in the least squares parameter estimates with the minimum sum-of-squares (minimum L_2 norm $||\hat{\beta}||_2$). For example, for the one-way ANOVA model, this can be shown to give parameter estimates $\hat{\mu} = \sum_{j=1}^{Q}(\overline{Y}_{q\bullet})/(1+Q)$ and $\hat{\alpha_q} = \overline{Y}_{q\bullet} - \hat{\mu}$. By $\overline{Y}_{q\bullet}$ we denote the average of Y over the observation index j, i.e. the average of the data in group q.

Using the pseudoinverse for parameter estimation in overdetermined models is the solution adopted in SPM. As mentioned above, this does still not allow to test for those linear combinations of effects for which there exist an infinite number of parameter estimates. (This topic is covered in great detail in chapter 8.) Note that the pseudoinverse constraint leaves us with all columns of X.

2.4 Geometrical Perspective

For some, a geometrical perspective provides an intuitive feel for the procedure. (This section can be omitted without loss of continuity.)

The vector of observed values Y defines a single point in $\Re^J$, J-dimensional Euclidean space. $X\tilde{\beta}$ is a linear combination of the columns of the design matrix X. The columns of X are J-vectors, so $X\tilde{\beta}$ for a given $\tilde{\beta}$ defines a point in $\Re^J$. This point lies in the subspace of $\Re^J$ spanned by the columns of the design matrix, the X-space. The dimension of this subspace is $\text{rank}(X)$. Recall that the space spanned by the columns of X is the set of points Xc for all $c \in \Re^L$. The residual sum-of-squares for parameter estimates $\tilde{\beta}$ is the distance from $X\tilde{\beta}$ to Y. Thus, the least squares estimates $\hat{\beta}$ correspond to the point in the space spanned by the columns of X that is nearest to the data Y. The perpendicular from Y to the X-space meets the X-space at $\hat{Y} = X\hat{\beta}$. It is now clear why there are no unique least squares estimates if X is rank-deficient; for then any point in the X-space can be obtained by infinitely many linear combinations of the columns of X, i.e. the solution exists on a hyperplane and is not a point.

If X is of full rank, then define the projection matrix as $P_X = X\left(X^TX\right)^{-1}X^T$. Then $\hat{Y} = P_XY$, and geometrically P_X is a projection onto the X-space. Similarly, the residual forming matrix is $R = (I_J - P_X)$, where I_J is the identity matrix of rank J. Thus $RY = e$, and R is a projection matrix onto the space orthogonal to the X-space.

As a concrete example, consider a linear regression with only three observations. The observed data $y = [y_1, y_2, y_3]^T$ defines a point in three-dimensional Euclidean space ($\Re^3$). The model

[4] If X is of full rank, then $(X^TX)^-$ is an inefficient way of computing $(X^TX)^{-1}$.

(Eq. 2) leads to a design matrix $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix}$. Provided the x_j's are not all the same, the columns of X span a two dimensional subspace of $\Re^3$, a plane (Fig.1).

[Figure 1 about here.]

2.5 Inference

Here, we derive the t- and F-statistics which are used to test for a linear combination of effects. We will also return to the issue of overdetermined models and determine which linear combinations (contrasts) we can test.

2.5.1 Residual Sum of Squares

The residual variance σ^2 is estimated by the residual sum-of-squares divided by the appropriate degrees of freedom: $\hat{\sigma}^2 = \frac{e^T e}{J-p} \sim \sigma^2 \frac{\chi^2_{J-p}}{J-p}$ where $p = \text{rank}(X)$. See also the Appendix (A2) for a derivation of this result.

2.5.2 Linear Combinations of the Parameter Estimates

It is not too difficult to show that the parameter estimates are normally distributed: if X is full rank then $\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X^T X)^{-1}\right)$. From this it follows that for a column vector c containing L weights (see §2.5.3):

$$c^T \hat{\beta} \sim \mathcal{N}\left(c^T \beta,\ \sigma^2 c^T (X^T X)^{-1} c\right) \tag{10}$$

Furthermore, $\hat{\beta}$ and $\hat{\sigma}^2$ are independent (Fisher's Law). Thus, prespecified hypotheses concerning linear compounds of the model parameters $c^T \beta$ can be assessed using

$$\frac{c^T \hat{\beta} - c^T \beta}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} \sim t_{J-p} \tag{11}$$

where t_{J-p} is a Student's t-distribution with $J - p$ degrees of freedom. For example, the hypothesis $\mathcal{H} : c^T \beta = d$ can be assessed by computing

$$T = \frac{c^T \hat{\beta} - d}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} \tag{12}$$

and computing a p-value by comparing T with a t-distribution having $J - p$ degrees of freedom. In SPM, all tested null hypotheses are of the form $c^T \beta = 0$. Also note that in SPM tests based on this t-value are always one-sided.

Example — Two-sample t-test

For example, consider the two-sample t-test (§2.1.1). The model (Eq. 4) leads to a design matrix X with two columns of dummy variables indicating group membership and parameter vector $\beta = [\mu_1, \mu_2]^T$. Thus, the null hypothesis $\mathcal{H}: \mu_1 = \mu_2$ is equivalent to $\mathcal{H}: c^T\beta = 0$ with $c = [1, -1]^T$.

The first column of the design matrix contains n_1 1's and n_2 0's, indicating the measurements from group one, while the second column contains n_1 0's and n_2 1's for group two. Thus $(X^TX) = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}$, $(X^TX)^{-1} = \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix}$, and $c^T(X^TX)^{-1}c = 1/n_1 + 1/n_2$, giving the t-statistic (by Eq. 11):

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2\,(1/n_1 + 1/n_2)}}$$

which is the standard formula for the two-sample t-statistic, with a Student's t-distribution of $n_1 + n_2 - 2$ degrees of freedom under the null hypothesis.

2.5.3 Estimable functions, contrasts

Recall (§2.3.1) that if the model is overparameterised (i.e. X is rank deficient), then there are infinitely many parameter sets describing the same model. Constraints or the use of a pseudoinverse technique pull out only one set of parameters from infinitely many. Therefore, when examining linear compounds $c^T\beta$ of the parameters it is imperative to consider only compounds that are invariant over the space of possible parameters. Such linear compounds are called *contrasts*. In the following, we will characterize contrasts as linear combinations having two properties, which can be used to determine whether a linear compound is a proper contrast or not.

In detail (Scheffé, 1959), a linear function $c^T\beta$ of the parameters is *estimable* if there is a linear unbiased estimate c'^TY for some constant vector of weights c'. That is $c^T\beta = E(c'^TY)$. ($E(Y)$ is the expectation of the random variable Y.) The natural estimate $c^T\hat{\beta}$ is unique for an estimable function whatever solution, $\hat{\beta}$, of the normal equations is chosen (Gauss-Markov theorem). Further: $c^T\beta = E(c'^TY) = c'^TX\beta \Rightarrow c^T = c'^TX$, so c is a linear combination of the rows of X.

A *contrast* is an estimable function with the additional property $c^T\hat{\beta} = c'^T\hat{Y} = c'^TY$. Now $c'^T\hat{Y} = c'^TY \Leftrightarrow c'^TP_XY = c'^TY \Leftrightarrow c' = P_Xc'$ (since P_X is symmetric), so c' is in the X-space. In summary, a contrast is an estimable function whose c' vector is a linear combination of the columns of X[5].

One can test, whether c is a contrast vector by combining the two properties (i) $c^T = c'^TX$ and (ii) $c' = P_Xc'$ for some vector c'. Combining (i) and (ii), it follows that $c^T = c'^TP_XX$. Because of (i), $c^T = c^T(X^TX)^-X^TX$. In other words, c is a contrast, if it is unchanged by post-multiplication with $(X^TX)^-X^TX$. This test is used in SPM for user-specified contrasts[6].

For a contrast it can be shown that $c^T\hat{\beta} \sim \mathcal{N}\left(c^T\beta,\ \sigma^2c'^Tc'\right)$. Using a pseudoinverse technique, $P_X = X(X^TX)^-X^T$, so $c' = P_Xc' \Rightarrow c'^Tc' = c'^TX(X^TX)^-X^Tc' = c^T(X^TX)^-c$ since $c = c'^TX$ for an estimable function.

[5] In Statistical Parametric Mapping, one usually refers to the vector c as the *vector of contrast weights*. Informally, we will also refer to c as the *contrast*, a slight misuse of the term.

[6] The actual implementation of this test is based on a more efficient algorithm using a singular value decomposition.

This shows that the distributional results given above for unique designs (Eq.10 & Eq.11), apply for contrasts of the parameters of non-unique designs, where $(X^TX)^{-1}$ is replaced by a pseudoinverse.

It remains to characterise which linear compounds of the parameters are contrasts. For most designs, contrasts have weights that sum to zero over the levels of each factor. For example, for the one-way *Anova* with parameter vector $\beta = [\mu, \alpha_1, \ldots, \alpha_Q]^T$, the linear compound $c^T\beta$ with weights vector $c = [c_0, c_1, \ldots, c_Q]^T$ is a contrast if $c_0 = 0$ and $\sum_{q=1}^{Q} c_q = 0$.

2.5.4 Extra Sum of Squares Principle, F-contrasts

The *extra sum-of-squares* principle provides a method of assessing general linear hypotheses, and for comparing models in a hierarchy, where inference is based on a F-statistic. Here, we will describe the classical F-test based on the assumption of an independent identically distributed error. In SPM, both statistics, the t- and the F-statistic, are used for making inferences.

We first describe the classical F-test as found in nearly all introductory statistical texts. After that we will point at two critical limitations of this description and derive a more general and better suited implementation of the F-test for typical models in neuroimaging.

Suppose we have a model with parameter vector β that can be partitioned into two, $\beta = \left[\beta_1^T \vdots \beta_2^T\right]$, and suppose we wish to test $\mathcal{H} : \beta_1 = 0$. The corresponding partitioning of the design matrix X is $X = \left[X_1 \vdots X_2\right]$, and the *full model* is:

$$Y = \left[X_1 \vdots X_2\right] \begin{bmatrix} \beta_1 \\ \cdots \\ \beta_2 \end{bmatrix} + \epsilon$$

which when $\mathcal{H}$ is true reduces to the *reduced model*: $Y = X_2\beta_2 + \epsilon$. Denote the residual sum-of-squares for the full and reduced models by $S(\beta)$ and $S(\beta_2)$ respectively. The *extra sum-of-squares* due to β_1 after β_2 is then defined as $S(\beta_1|\beta_2) = S(\beta_2) - S(\beta)$. Under $\mathcal{H}$, $S(\beta_1|\beta_2) \sim \sigma^2\chi_p^2$ independent of $S(\beta)$, where the degrees of freedom are $p = \text{rank}(X) - \text{rank}(X_2)$. (If $\mathcal{H}$ is not true, then $S(\beta_1|\beta_2)$ has a non-central chi-squared distribution, still independent of $S(\beta)$.) Therefore, the following F-statistic expresses evidence against $\mathcal{H}$:

$$F = \frac{\frac{S(\beta_2) - S(\beta)}{p - p_2}}{\frac{S(\beta)}{J - p}} \sim F_{p-p_2, J-p} \tag{13}$$

where $p = \text{rank}(X)$ and $p_2 = \text{rank}(X_2)$. The larger F gets the more unlikely it is that F was sampled under the null hypothesis H. Significance can then be assessed by comparing this statistic with the appropriate F-distribution. Draper & Smith (Draper and Smith, 1981) give derivations.

This formulation of the F-statistic has two limitations. The first is that two (nested) models, the full and the reduced model, have to be fitted subsequently to the data. In practice, this is implemented by a two-pass procedure on a, typically, large data set. The second limitation is that a partitioning of the design matrix into two blocks of regressors is not the only way one can partition the design matrix space. Essentially, one can partition X into two sets of linear combinations of the regressors. As an example, one might be interested in the difference between

two effects. If each of these two effects is modelled by one regressor, a simple partitioning is not possible and one cannot use Eq. 13 to test for the difference. Rather, one has to re-parameterize the model such that the differential effect is explicitly modelled by a single regressor. As we will show in the following, this re-parameterization is unnecessary.

The key to implement a F-test that avoids these two limitations is the notion of contrast matrices. A contrast matrix is a generalisation of a contrast vector (2.5.3). Each column of a contrast matrix consists of one contrast vector. Importantly, the contrast matrix controls the partitioning of the design matrix X.

A (user-specified) contrast matrix c is used to determine a subspace of the design matrix, i.e. $X_c = Xc$. The orthogonal contrast to c is given by $c_0 = I_p - cc^-$. Then, let $X_0 = Xc_0$ be the design matrix of the reduced model. We wish to compute what effects X_c explain, *after* first fitting the reduced model X_0. The important point to note is that although c and c_0 are orthogonal to each other, X_c and X_0 are possibly not, because the relevant regressors in the design matrix X can be correlated. If the partitions X_0 and X_c are not orthogonal, the temporal sequence of the subsequent fitting procedure attributes their shared variance to X_0. However, the subsequent fitting of two models is unnecessary, because one can construct a projection matrix from the data to the subspace of X_c, which is orthogonal to X_0. We denote this subspace by X_a.

The projection matrix M due to X_a can be derived from the residual forming matrix of the reduced model X_0. This matrix is given by $R_0 = I_J - X_0X_0^-$. The projection matrix is then $M = R_0 - R$, where R is the residual forming matrix of the full model, i.e. $R = I_J - XX^-$.

The F-statistic can then be written as

$$F = \frac{(MY)^T MY}{(RY)^T RY}\frac{J-p}{p_1} = \frac{Y^T MY}{Y^T RY}\frac{J-p}{p_1} \sim F_{p_1,J-p} \tag{14}$$

where p_1 is the rank of X_a. Since M is a projector onto a subspace within X, we can also write

$$F = \frac{\hat{\beta}^T X^T MX\hat{\beta}}{Y^T RY}\frac{J-p}{p_1} \sim F_{p_1,J-p} \tag{15}$$

This equation means that we can conveniently compute a F-statistic for any user-specified contrast without any re-parameterization. In SPM, all F-statistics are based on the full model so that $Y^T RY$ needs only to be estimated once and be stored for subsequent use.

In summary, the formulation of the F-statistic (Eq. 15) is a powerful tool, because by using a contrast matrix c we can test for a subspace spanned by contrasts of the design matrix X. Importantly, we do not need to reparameterise the model and estimate an additional parameter set, but use estimated parameters of the full model. More about F-contrasts and their applications can be found in chapter 8.

Example — one-way ANOVA

For example, consider a one-way ANOVA (§2.3.2, Eq. 9), where we wish to assess the omnibus null hypothesis that all the groups are identical: $\mathcal{H} : \alpha_1 = \alpha_2 = \ldots = \alpha_Q$. Under $\mathcal{H}$ the model reduces to $Y_{qj} = \mu + \epsilon_{qj}$. Since the ANOVA model contains a constant term, μ, $\mathcal{H}$ is equivalent to $\mathcal{H} : \alpha_1 = \alpha_2 = \ldots = \alpha_Q = 0$. Thus, let $\beta_1 = (\alpha_1, \ldots, \alpha_Q)^T$, and $\beta_2 = \mu$. Eq. 13 then gives an F-statistic which is precisely the standard F-statistic for a one-way ANOVA.

Alternatively, we can apply Eq. 15. The contrast matrix c is a diagonal $Q+1$-matrix with Q ones on the upper main diagonal and a zero in the $Q+1$st element on the main diagonal (Fig. 2). This contrast matrix tests, whether there was an effect due to any group after taking into account a constant term across groups. Application of Eq. 15 results in the same F-value as compared to Eq. 13, but without the need to explicitly fit two models.

[Figure 2 about here.]

2.6 Adjusted and fitted data

Adjusted data can be used to illustrate the nature of an effect, some effects having been removed from the raw data Y. For example, when looking at the difference between two groups (two-sample t-test), the effect that is of no interest is the mean over the two groups. Removing this overall mean allows one to have a better chance of visually assessing the difference which is otherwise *hidden away*, because its amplitude is typically only a small fraction of the mean amplitude. This principle of removing effects of no interest to better visually assess the overall effects of interest can be applied to any kind of design.

The question to answer is which effects are of interest and which are not. The partitioning of the design matrix into these two parts is based on the same principles as the F-test developed in the preceding subsection. We can use an F-contrast for the partitioning, which is equivalent to the specification of a full and reduced model. In this context, adjusted data are the residuals of the reduced model, i.e. components that can be explained by the reduced model have been removed from the data. In other words, to compute adjusted data the user needs to tell SPM which part of the design matrix is of no interest (the reduced model). SPM then takes the part of the design matrix, which is orthogonal to the reduced model, as the effects of interest. This process will be illustrated below by an example. Note that the partitioning of the design matrix follows the same logic as the F-test: First, any effect due to the reduced model is removed and only the remaining effects are taken to be of interest. An important point is that any overlap (correlation) between the reduced model and our *partition of interest* is explained by the reduced model. In the context of adjusted data this means that the adjusted data will not contain that component of the effects that can be explained by the reduced model.

Operationally, we compute the adjusted data using the same procedure as used to calculate the F-statistic. A user-specified contrast matrix c induces a partitioning of the design matrix X. The reduced model is given by $X_0 = Xc_0$ and its residual forming matrix $R_0 = I_J - X_0X_0^-$. The adjusted data can then be computed by $\tilde{Y} = R_0Y$. Note that this projection technique makes a re-parameterization redundant.

An alternative way of computing the adjusted data $\tilde{Y}$ is to compute the data explained by the design matrix partition orthogonal to X_0 and add the residuals of the full model, i.e. $\tilde{Y} = Y_f + e$. The residuals are given by $e = RY$, where R is the residual forming matrix of the full model, and $Y_f = MY$, where Y_f is referred to as *fitted data*. The projection matrix M is computed by $M = R_0 - R$ (§2.5.4). In other words, the fitted data is equivalent to the adjusted data minus the estimated error, i.e. $Y_f = \tilde{Y} - e$.

In SPM, both adjusted and fitted data can be plotted for any voxel. For these plots, SPM requires the specification of an F-contrast, which encodes the partitioning of the design matrix into effects of interest and no interest.

2.6.1 Example

As an example, we look at a one-way anova with four groups. The design matrix consists of four columns which indicate group membership. Each group has 12 measurements so that we have altogether 48 measurements. In our example, we are interested in the average of two differences. The first difference is between group 1 and 2 and the second difference between group 2 and 3. If we want to test this difference with a t-statistic, the contrast vector will be $c = [-1,\ 1,\ -1,\ 1]^T$. In Fig. 3 (left), we show what the actual data looks like. It is easy to see that there is a difference between the average of the first two groups compared to the average of the last two groups. (This difference could be tested by using the contrast vector $c = [-1,\ -1,\ 1,\ 1]^T$.) However, by visual inspection, it is hard to tell, whether there is a difference between the average of group 1 and 3 compared to the average of group 2 and 4. This is a situation, where a plot of adjusted and fitted data is helpful. First, we have to specify a reduced model. One way of doing this is to specify a contrast vector or matrix that defines our effect of interest. In our example, the difference is represented by the contrast vector $c = [-1,\ 1,\ -1,\ 1]^T$. The contrast matrix c_0 is given by $c_0 = I_4 - cc^-$. With c_0, we can compute X_0, R_0 and M, all of which are needed to compute the adjusted and fitted data. In Fig. 3 (right), we show the fitted and adjusted data. In this plot, it is obvious that there actually is a difference between group 1 and 2 and between group 3 and 4. This example illustrates that plots of fitted and adjusted data are helpful, when the effect of interest is masked by a comparably large effect of no interest. This is very often the case in neuroimaging, where typically the effect of interest is very small compared to large confounding effects.

Note that a plot of adjusted or fitted data can never substitute for a test of significance. However, for illustration purposes, a plot of the adjusted/fitted data is the closest one can get to the effect that one wishes to test.

[Figure 3 about here.]

2.7 Design matrix images

SPM uses greyscale images of the design matrix to represent linear models. An example for a single subject PET activation study with four scans under each of three conditions is shown in Fig. 4. The first three columns contain indicator variables (consisting of zeros and ones) indicating the condition. The last column contains the (mean corrected) global cerebral blood flow (gCBF) values (see below).

In the greyscale design matrix images, -1 is black, 0 mid-gray, and $+1$ white. Columns containing covariates are scaled by subtracting the mean (zero for centered covariates). For display purposes regressors are divided by their absolute maximum, giving values in [-1, 1]. Design matrix blocks containing factor by covariate interactions (§3.3.5) are scaled such that the covariate values lie in (0,1], thus preserving representation of the padding zeros as mid-grey.

[Figure 4 about here.]

3 PET and basic models

With the details of the general linear model covered, we turn our attention to some actual models used in functional brain mapping, discuss the practicalities of their application, and introduce some terminology used in SPM. As the approach is mass univariate, we must consider a model for each and every voxel. Bear in mind that in the mass univariate approach, the same model is used at every voxel simultaneously, with different parameters for each voxel. We shall concentrate on PET data, with its mature family of standard statistical experimental designs. Models of fMRI data will be presented in the next section.

Although most PET functional mapping experiments are on multiple subjects, many of the key concepts are readily demonstrated using single subject data.

3.1 Heteroscedacity

Heteroscedacity in the context of neuroimaging means that the error variance is allowed to vary between voxels. In PET data, there is substantial evidence against an assumption of constant variance (homoscedasticity) at all points of the brain. This fact is perhaps to be expected, considering the different constituents and activities of grey and white matter. This is unfortunate, as the small sample sizes leave few degrees of freedom for variance estimation. If homoscedasticity can be assumed, variance estimates can legitimately be pooled across all voxels. Provided the image is much greater in extent than its smoothness, this gives an estimate with sufficiently high (effective) degrees of freedom such that its variability is negligible. (Since the images are smooth, neighbouring voxels are correlated and hence the variance estimates at neighbouring voxels are correlated.) t-statistics based on such a variance estimate are approximately normally distributed, the approximation failing only in the extreme tails of the distribution.

3.2 Global normalization

In neuroimaging, one can differentiate between regional and global activity. By regional activity one typically means the activity measured in a single voxel or a small volume of voxels. Global activity refers to a global measure of brain activity. These two informal descriptions of regional and global activity reflect that there may be a number of different definitions. However, the reason why the concept of global activity is important is that there are effects in a single voxel that are caused by global effects. These are usually difficult to model. Typically, we use simple models for global effects. Modelling global effects enhances the sensitivity and accuracy of the subsequent inference step about experimentally induced effects.

As an example, consider a simple single subject PET experiment. The subject is scanned repeatedly under both *baseline* (control) and *activation* (experimental) conditions. Inspection of regional activity, (used as a measure of regional cerebral blood flow (rCBF)), alone at a single voxel may not indicate an experimentally induced effect. However, the additional consideration of global activity (the global cerebral blood flow (gCBF)) for the respective scans may clearly differentiate between the two conditions (Fig. 5).

[Figure 5 about here.]

In Statistical Parametric Mapping, the precise definition of global activity is user-dependent. The *default* definition is that global activity is the global average of image intensities of intrac-

erebral tissue. If Y_j^k is the image intensity at voxel $k = 1, \ldots, K$ of scan j, then denote the estimated global activity by $g_j = \overline{Y}_j^{\bullet} = \sum_{k=1}^{K} Y_j^k / K$.

Having estimated the global activity for each scan, a decision must be made about what model of global activity should be used. In SPM, there are basically two alternatives or various mixtures between them. The first is *proportional scaling* and the second is an AnCova approach.

3.2.1 Proportional scaling

One way to account for global changes is to adjust the data by scaling each scan by its estimated global activity. This approach is based on the assumption that the measurement process introduces a (global) scaling of the image intensities at each voxel, a gain factor. This has the advantage of converting the raw data into a physiological range to give parameters in interpretable scale. The mean global value, is usually chosen to be the canonical normal gCBF of 50ml/min/dl. The scaling factor is thus $\frac{50}{\overline{g}_{\bullet}}$. We shall assume that the count rate recorded in the scanner (counts data) has been scaled into a physiologically meaningful scale. The normalised data are $Y_j'^k = \frac{50}{g_j} Y_j^k$. The model is then

$$Y_j^k = \frac{g_j}{50} (X\beta^k)_j + \epsilon_j'^k \tag{16}$$

where $\epsilon'^k \sim \mathcal{N}(0, \sigma_k^2 \times diag((g_j/50)^2))$. The $diag()$ operator transforms a column vector to a diagonal matrix with the vector on its main diagonal and zero elsewhere. This is a weighted regression, i.e. the shape of the error covariance matrix is no longer I_J, but a function of the estimated global activity. Also note that the jth row of X is weighted by g_j.

The adjustment of data, from Y to Y' is illustrated in Fig. 6a.

[Figure 6 about here.]

3.2.2 Ancova approach

Another approach is to include the mean corrected global activity vector g as an additional regressor into the model. In this case the model (Eq. 6) becomes

$$Y_j^k = (X\beta)_j + \zeta^k (g_j - \overline{g}_{\bullet}) + \epsilon_j^k \tag{17}$$

where $\epsilon^k \sim \mathcal{N}(0, \sigma_k^2 I_J)$ and ζ_k is the slope parameter for the global activity vector. In this model, the data is explained as the sum of experimentally induced regional activity and some global activity which varies over scans. Note that the model of Eq. 17 can be considerably extended by allowing for different slopes between replications, conditions, subjects and groups.

3.2.3 Proportional scaling versus ancova

Clearly a decision has to be made which global normalization approach shall be used for a given data set. One cannot apply both, because proportional scaling will normalize the global mean activity such that the mean corrected g in the AnCova approach will consists of a zero vector. The proportional scaling approach is most appropriate for any data set for which there is a gain (multiplicative) factor that varies over scans. This is a useful assumption for fMRI data (see

next section). In contrast to this, an AnCova approach is appropriate, if the gain factor does not change over scans. This is the case for PET scans acquired on modern scanners using protocols which control for the administered dose rate. This means that a change in estimated global activity reflects a change in a subject's global activity and not a change in a global (machine specific) gain factor. Moreover, the AnCova approach assumes that regional experimentally induced effects are independent of changes in global activity. Note that the AnCova approach should not be used for PET data, where the administered dose is not controlled and varies over scans. In this case, the true underlying gCBF might be constant over scans, but the global gain factor varies. Similarly, for SPECT scans, it is known that the global gain factor can vary over scans, so that it is recommended to prefer proportional scaling over the AnCova approach for SPECT data.

Special considerations apply if there are condition dependent changes in global activity.

Implicit in allowing for changes in gCBF (either by proportional scaling or ANCOVA) when assessing condition specific changes in rCBF, is the assumption that gCBF represents the underlying background flow, above which regional differences are assessed. That is, gCBF is independent of condition. Clearly, since gCBF is calculated as the mean intracerebral rCBF, an increase of rCBF in a particular brain region must cause an increase of gCBF unless there is a corresponding decrease of rCBF elsewhere in the brain. Similar problems can arise when comparing a group of subjects with a group of patients with brain atrophy, or when comparing pre and post-operative rCBF.

If gCBF actually varies considerably between conditions, as in pharmacological activation studies, then testing for an activation after allowing for global changes involves extrapolating the relationship between regional and global flow outside the range of the data. This extrapolation might not be valid, as illustrated in figure 7a.

If gCBF is increased by a large activation that is not associated with a corresponding deactivation, then comparison at a common gCBF will make non-activated regions (whose rCBF remained constant) appear falsely de-activated, and the magnitude of the activation will be similarly decreased. (Figure 7b illustrates the scenario for a simple single subject activation experiment using ANCOVA.) In such circumstances a better measure of the underlying background flow should be sought, for instance by examining the flow in brain regions known to be unaffected by the stimulus.

[Figure 7 about here.]

3.2.4 Grand Mean Scaling

Grand mean scaling multiplies all scans by some factor such that the resulting estimated mean global activity is a (user specified) constant over scans. [7] Note that this common factor has no effect on the inference, because in the t- and F-statistic (Eqs. 12 and 14) such a factor cancels out. It will also not change relative interpretations of the fitted or adjusted data. The default behaviour of SPM with respect to PET and fMRI data is described in §3 and §4.

3.2.5 Mixtures of scaling and AnCova

For PET and SPECT data, the user can choose from a wide range of global normalization models that lie in between proportional scaling and an AnCova approach. An intermediate approach

[7] Clearly grand mean scaling is redundant when followed by proportional scaling.

is possible by scaling groups of scans. These groups can be scaling by replication, condition, subject and group grand mean. This can be then applied together with an AnCova approach that estimates different slopes for such a grouping. For example, one can apply proportional scaling by the grand mean within-subject and combine this with a (within-) subject AnCova approach.

3.3 PET models

In the following subsections, the flexibility of the general linear model is demonstrated using models for various PET functional mapping experiments. For generality, ANCOVA style models are used, with gCBF included as a confounding covariate. The corresponding ANCOVA models for data adjusted by proportional scaling can be obtained by omitting the global terms. Voxel level models are presented in the usual statistical notation, alongside the SPM description and design matrix images. The form of contrasts for each design are indicated, and some practicalities of the SPM interface are discussed.

Single subject models

3.3.1 Single subject activation design

The simplest experimental paradigm is the single subject activation experiment. Suppose there are Q conditions, with M_q scans under condition q. Let Y_{qj}^k denote the rCBF at voxel k in scan $j = 1, \ldots, M_q$ under condition $q = 1, \ldots, Q$. The model is:

$$Y_{qj}^k = \alpha_q^k + \mu^k + \zeta^k (g_{qj} - \overline{g}_{\bullet\bullet}) + \epsilon_{qj}^k \tag{18}$$

There are $Q + 2$ parameters for the model at each voxel: The Q condition effects, the constant term μ^k, and the global regression effect, giving parameter vector $\beta^k = (\alpha_1^k, \ldots, \alpha_Q^k, \mu^k, \zeta^k)^T$ at each voxel. In this model, replications of the same condition are modelled with a single effect. The model is overparameterised, having only $Q + 1$ degrees of freedom, leaving $N - Q - 1$ residual degrees of freedom, where $N = \sum M_q$ is the total number of scans.

[Figure 8 about here.]

Contrasts are linear compounds $c^T \beta^k$ for which the weights sum to zero over the condition effects, and give zero weight to the constant term, i.e. $\sum_{q=1}^{Q} c_q = 0$ (Fig. 8). Therefore, linear compounds that test for a simple group effect or for an average effect over groups cannot be contrasts. However, one can test for differences between groups. For example, to test the null hypothesis $\mathcal{H}^k : \alpha_1^k = (\alpha_2^k + \alpha_3^k)/2$ against the one sided alternative $\overline{\mathcal{H}}^k : \alpha_1^k > (\alpha_2^k + \alpha_3^k)/2$, the appropriate contrast weights would be $c = [1, -\frac{1}{2}, -\frac{1}{2}, 0, \ldots, 0]^T$. In words, one tests for a (positive) difference between the effect of group 1 compared to the average of groups 2 and 3? Large positive values of the t-statistic express evidence against the null hypothesis, in favour of the alternative hypothesis.

3.3.2 Single subject parametric design

Consider the single subject parametric experiment where a single covariate of interest, or "score", is measured. For instance, the covariate may be a physiological variable, a task difficulty rating, or a performance score. It is desired to find regions where the rCBF values are

highly correlated with the covariate, taking into account the effect of global changes. Figure 9a depicts the situation. If Y_j^k is the rCBF at voxel k of scan $j = 1, \ldots, J$ and s_j is the independent covariate, then a simple ANCOVA style model is a multiple regression with two covariates:

$$Y_j^k = \varrho^k(s_j - \overline{s}_\bullet) + \mu^k + \zeta^k(g_j - \overline{g}_\bullet) + \epsilon_j^k \tag{19}$$

Here, ϱ is the slope of the regression plane in the direction of increasing score, fitted separately for each voxel.

There are three model parameters, leaving $J-3$ residual degrees of freedom. The design matrix (Fig. 9b) has three columns, a column containing the (centered) score covariate, a column of dummy 1's corresponding to μ^k, and a column containing the (centered) global values.

In SPM this is a "Single subject: Covariates only" design. The design is uniquely specified, so any linear combination of the three parameters is a contrast. The null hypothesis of no score effect at voxel k, $\mathcal{H}^k : \varrho^k = 0$, can be assessed against the one sided alternative hypotheses $\overline{\mathcal{H}}^k : \varrho^k > 0$ (rCBF increasing with score) with contrast weight for the effect of interest $c_1 = +1$, and against $\overline{\mathcal{H}}^k : \varrho^k < 0$ (rCBF decreasing as score increases) with contrast weight $c_1 = -1$.

[Figure 9 about here.]

This simple model assumes a linear relationship between rCBF and the covariate (and other explanatory variables). More general relationships may be modelled by including other functions of the covariate. These functions are essentially new explanatory variables, which if linearly combined still fit in the framework of the general linear model. For instance, if an exponential relationship is expected, the logarithm of s_j, i.e. $\ln(s_j)$, would be used in place of s_j. Fitting powers of covariates as additional explanatory variables leads to *polynomial regression*. More generally, a set of *basis functions* can be used to expand the covariate to allow flexible modelling. This theme will be developed later in this chapter (for fMRI), and in other chapters.

3.3.3 Simple single subject activation revisited

As discussed in the general linear model section (§2), it is often possible to reparameterise the same model in many ways. As an example, consider a two condition ($Q = 2$) single subject experiment, discussed above (§3.3.1). The model (Eq. 18) is:

$$Y_{qj}^k = \alpha_q^k + \mu^k + \zeta^k(g_{qj} - \overline{g}_{\bullet\bullet}) + \epsilon_{qj}^k$$

The model is over-determined, so consider a sum-to-zero constraint on the condition effects. For two conditions this implies $\alpha_1^k = -\alpha_2^k$. Substituting for α_2^k the resulting design matrix has a column containing $+1$'s and -1's indicating the condition $q = 1$ or $q = 2$ respectively, a column of 1's for the overall mean, and a column containing the (centered) gCBF (Fig.10). The corresponding parameter vector is $\beta^k = [\alpha_1^k, \mu^k, \zeta^k]^T$. Clearly this is the same design matrix as that for a parametric design with (non-centered) "score" covariate indicating the condition as active or baseline with $+1$ or -1 respectively. The hypothesis of no activation at voxel k, $\mathcal{H}^k : \alpha_1^k = 0$ can be tested against the one sided alternatives $\overline{\mathcal{H}}^k : \alpha_1^k > 0$ (activation) and $\overline{\mathcal{H}}^k : \alpha_1^k < 0$ with contrast weights for the effects of interest $c_1 = 1$ and $c_1 = -1$ respectively. This example illustrates how the SPM interface may be used to enter "hand-built" blocks of design matrix as non-centered covariates.

[Figure 10 about here.]

3.3.4 Single subject: conditions and covariates

Frequently there are other confounding covariates in addition to gCBF that can be added into the model. For example, a linear time component could be modeled simply by entering the scan number as covariate. In SPM these appear in the design matrix as additional covariate columns adjacent to the global flow column.

3.3.5 Factor by covariate interactions

A more interesting experimental scenario is when a parametric design is repeated under multiple conditions in the same subject(s). A specific example would be a PET language experiment in which, during each of twelve scans, lists of words are presented. Two types of word list (the two conditions) are presented at each of six rates (the parametric component). Interest may lie in locating regions where there is a difference in rCBF between conditions (accounting for changes in presentation rate), the *main* effect of condition; locating regions where rCBF increases with rate (accounting for condition), the main effect of rate; and possibly assessing evidence for condition specific responses in rCBF to changes in rate, an interaction effect.[8] Let Y^k_{qrj} denote the rCBF at voxel k for the j-th measurement under rate $r = 1, \ldots, R$ and condition $q = 1, \ldots, Q$, with s_{qr} the rate covariate (some function of the rates). A suitable model is:

$$Y^k_{qrj} = \alpha^k_q + \varrho^k_q(s_{qr} - \overline{s}_{\bullet\bullet}) + \mu^k + \zeta^k(g_{qrj} - \overline{g}_{\bullet\bullet\bullet}) + \epsilon^k_{qrj} \tag{20}$$

Note the q subscript on the parameter ϱ^k_q, indicating different slopes for each condition. Ignoring for the moment the global flow, the model describes two simple regressions with common error variance (Fig. 11a). The SPM interface describes such factor by covariate interactions as "factor specific covariate fits". The interaction between condition and covariate effects is manifest as different regression slopes for each condition. There are $2Q + 2$ parameters for the model at each voxel, $\beta^k = [\alpha^k_1, \ldots, \alpha^k_Q, \varrho^k_1, \ldots, \varrho^k_Q, \mu^k, \zeta^k]^T$, with $2Q + 1$ degrees of freedom. A design matrix image for the two condition example is shown in figure 11b. The factor by covariate interaction takes up the third and fourth columns, corresponding to the parameters ϱ^k_1 and ϱ^k_2, the covariate being split between the columns according to condition, the remaining cells filled with zeros.

Only the constant term and global slope are designated confounding, giving $2Q$ effects of interest to specify contrast weights for, $\beta^k_1 = [\alpha^k_1, \ldots, \alpha^k_Q, \varrho^k_1, \ldots, \varrho^k_Q]^T$. As with the activation study model, contrasts have weights which sum to zero over the condition effects. For the 2 condition word presentation example, contrast weights $c_1 = [0, 0, 1, 0]^T$ for the effects of interest express evidence against the null hypothesis that there is is no covariate effect in condition one, with large values indicating evidence of a positive covariate effect. Weights $c_1 = [0, 0, \frac{1}{2}, \frac{1}{2}]^T$ address the hypothesis that there is no average covariate effect across conditions, against the one sided alternative that the average covariate effect is positive. Weights $c_1 = [0, 0, -1, +1]^T$ address the hypothesis that there is no condition by covariate interaction, that is, that the regression slopes are the same, against the alternative that the condition 2 regression is steeper.

Conceptually, contrast weights $c_1 = [-1, +1, 0, 0]^T$ and $c_1 = [+1, -1, 0, 0]^T$ for the effects of interest assess the hypothesis of no condition effect against appropriate one-sided alternatives.

[8] Two experimental factors *interact* if the level of one affects the expression of the other.

However, the comparison of main effects is confounded in the presence of an interaction: In the above model, both gCBF and the rate covariate were centered, so the condition effects α_q^k are the relative heights of the respective regression lines (relative to μ^k) at the mean gCBF and mean rate covariate. Clearly if there is an interaction, then difference in the condition effects (the separation of the two regression lines) depends on where you look at them. Were the rate covariate not centered, the comparison would be at mean gCBF and zero rate, possibly yielding a different result.

Thus main effects of condition in such a design must be interpreted with caution. If there is little evidence for a condition dependent covariate effect then there is no problem. Otherwise, the relationship between rCBF and other design factors should be examined graphically to assess whether the perceived condition effect is sensitive to the level of the covariate.

[Figure 11 about here.]

Multi-subject designs

Frequently, experimentally induced changes of rCBF are subtle, such that analyses must be pooled across subjects to find statistically significant evidence of an experimentally induced effect. In this chapter, we will discuss some fixed effects models. Random or mixed effects models are covered in chapter 12.

The single subject designs presented above must be extended to account for subject to subject differences. The simplest type of subject effect is an additive effect, otherwise referred to as a *block* effect. This implies that all subjects respond in the same way, save for an overall shift in rCBF (at each voxel). We extend our notation by adding subscript i for subjects, so Y_{iqj}^k is the rCBF at voxel k of scan j under condition q on subject $i = 1, \ldots, N$.

3.3.6 Multi subject activation (replications)

For instance, the single subject activation model (Eq.18) is extended by adding subject effects γ_i^k giving the model:

$$Y_{iqj}^k = \alpha_q^k + \gamma_i^k + \zeta^k (g_{iqj} - \overline{g}_{\bullet\bullet\bullet}) + \epsilon_{iqj}^k \tag{21}$$

A schematic plot of rCBF vs. gCBF for this model is shown in figure 12a. In SPM terminology, this is a "multi-subject: replication of conditions" design. The parameter vector at voxel k is $\beta^k = [\alpha_1^k, \ldots, \alpha_Q^k, \gamma_1^k, \ldots, \gamma_N^k, \zeta^k]^T$. The design matrix (Fig.12b) has N columns of dummy variables corresponding to the subject effects. (Similarly a multi-subject parametric design could be derived from the single subject case (§3.3.2) by including appropriate additive subject effects.)

Again, the model is overparameterised, though this time we have omitted the explicit constant term from the confounds, since the subject effects can model an overall level. Adding a constant to each of the condition effects and subtracting it from each of the subject effects gives the same model. Bearing this in mind, it is clear that contrasts must have weights that sum to zero over both the subject effects and the condition effects.

[Figure 12 about here.]

3.3.7 Condition by replication interactions

The above model assumes that (accounting for global and subject effects) replications of the same condition give the same (expected) response. There are many reasons why this assumption may be inappropriate, such as learning effects or more generally effects that change as a function of time. For example, some time effects can be modelled by including appropriate functions of the scan number as confounding covariates. With multi-subject designs we have sufficient degrees of freedom available to enable the consideration of replication by condition interactions. Such interactions imply that the (expected) response to each condition is different between replications (having accounted for other effects in the model). Usually in statistical models, interaction terms are added to a model containing main effects. However, such a model is so overparameterised that the main effects may be omitted, leaving just the interaction terms. The model is:

$$Y^k_{iqj} = \alpha\vartheta^k_{(qj)} + \gamma^k_i + \zeta^k(g_{iqj} - \overline{g}_{\bullet\bullet\bullet}) + \epsilon^k_{iqj} \tag{22}$$

where $\alpha\vartheta^k_{(qj)}$ is the interaction effect for replication j of condition q, the condition-by-replication effect. As with the previous model, this model is overparameterised (by one degree of freedom), and contrasts must have weights which sum to zero over the condition-by-replication effects. There are as many of these condition-by-replication terms as there are scans per subject. (An identical model is arrived at by considering each replication of each experimental condition as a separate condition.) If the scans are reordered such that the j-th scan corresponds to the same replication of the same condition in each subject, then the condition-by-replication corresponds to the scan number. An example design matrix for 5 subjects scanned twelve times is shown in figure 13a, where the scans have been reordered. In SPM this is termed a "Multi-subject: conditions only" design.

This is the "classic" SPM ANCOVA described by Friston *et al.* (Friston et al., 1990), and implemented in the original SPM software.[9] It offers great latitude for specification of contrasts. Appropriate contrasts can be used to assess main effects, specific forms of interaction, and even parametric effects. For instance, consider the verbal fluency data-set described by Friston *et al.* (Friston et al., 1995)[10]: Five subjects were scanned twelve times, six times under each of two conditions, word shadowing (condition A) and intrinsic word generation (condition B). The scans were reordered to ABABABABABAB for all subjects. Then a contrast with weights (for the condition-by-replication effects) of $c_1 = [-1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1]^T$ assesses the hypothesis of no main effect of word generation (against the one-sided alternative of activation). A contrast with weights of $c_1 = [5\frac{1}{2}, 4\frac{1}{2}, 3\frac{1}{2}, 2\frac{1}{2}, 1\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -1\frac{1}{2}, -2\frac{1}{2}, -3\frac{1}{2}, -4\frac{1}{2}, -5\frac{1}{2}]^T$ is sensitive to linear decreases in rCBF over time, independent of condition, and accounting for subject effects and changes in gCBF. A contrast with weights of $c_1 = [1, -1, 1, -1, 1, -1, -1, 1, -1, 1, -1, 1]^T$ assesses the interaction of time and condition, subtracting the activation in the first half of the experiment from that in the latter half.

[Figure 13 about here.]

3.3.8 Interactions with subject

While it is (usually) reasonable to use ANCOVA style models to account for global flow, with regression parameters constant across conditions, the multi-subject models considered thus

[9] The original SPM software is now fondly remembered as SPM*classic*.

[10] This data set is available via http://www.fil.ion.ucl.ac.uk/spm/data/

far assume additionally that this regression parameter is constant across subjects. It is quite possible that rCBF at the same location for different subjects will respond differentially to changes in gCBF – a subject by gCBF covariate interaction. The gCBF regression parameter can be allowed to vary from subject to subject. Extending the multi-subject activation (replication) model (Eq. 21) in this way gives:

$$Y^k_{iqj} = \alpha^k_q + \gamma^k_i + \zeta^k_i (g_{iqj} - \overline{g}_{\bullet\bullet\bullet}) + \epsilon^k_{iqj} \tag{23}$$

Note the i subscript on the global slope term, ζ^k_i, indicating a separate parameter for each subject. A schematic plot of rCBF vs. gCBF for this model and an example design matrix image are shown in figure 14. In the terminology of the SPM interface, this is an "ANCOVA by subject". The additional parameters are of no interest, and contrasts are as before.

[Figure 14 about here.]

Similarly, the SPM interface allows subject by covariate interactions, termed "subject specific fits". Subject by condition interactions can be entered by using "Multi-subject: Conditions × subject interaction and covariates".

3.4 Multi-study designs

The last class of SPM models for PET we consider are the "multi-study" models. In these models, subjects are grouped into two or more *studies*. The "multi-study" designs fit separate condition effects for each study. In statistical terms this is a *split plot* design. As an example consider two multi-subject activation studies, the first with five subjects scanned twelve times under two conditions (as described above in section 3.3.6), the second with three subjects scanned six times under three conditions. An example design matrix image for a model containing study specific condition effects, subject effects and study specific global regression (termed "ANCOVA by group" in SPM) is shown in figure 15. The first two columns of the design matrix correspond to the condition effects for the first study, the next two to the condition effects for the second study, the next eight to the subject effects, and the last to the gCBF regression parameter. (The corresponding scans are assumed to be ordered by study, by subject within study, and by condition within subject.)

Contrasts for multi-study designs in SPM have weights, when considered for each of the studies individually, would define a contrast for the study. Thus, contrasts must have weights which sum to zero over the condition effects within each study. There remain three types of useful comparison available. The first is a comparison of condition effects within a single study, carried out in the context of a multi-study design; the contrast weights appropriate for the condition effects of the study of interest is entered, padded with zeros for the other study, e.g. $c_1 = [1, -1, 0, 0, 0]^T$ for the first study in our example. This may have additional power when compared to an analysis of this study in isolation, since the second study observations change the variance estimates. The second is an average effect across studies; contrasts for a particular effect in each of the studies are concatenated, the combined contrast assessing a mean effect across studies. For example, if the second study in our example has the same conditions as the first, plus an additional condition, then such a contrast would have weights for the effects of interest $c_1 = [-1, 1, -1, 1, 0]^T$. Lastly, differences of contrasts across studies can be assessed, such as differences in activation. The contrasts weights for the appropriate main effect in each study are concatenated, with some studies contrasts negated. In our example,

$c_1 = [-1, 1, 1, -1, 0]^T$ would be appropriate for locating regions where the first study activated more than the second, or where the second deactivated more than the first.

[Figure 15 about here.]

Assumption of model fit in this case includes the assumption that the error terms have equal variance (at each voxel) across studies. For very different study populations, or studies from different scanners or protocols (possibly showing large differences in the measured global activity between studies), this assumption may not be tenable and the different variances should be modelled (chapter 9).

3.5 Basic models

In this section, we will discuss some of the models that are referred to in Statistical Parametric Mapping as *Basic models*. Typically, basic models are used for analyses at the second level to implement mixed effects models (chapter 12). For example, basic models include the one-sample t-test, the two-sample t-test, the paired t-test and a one-way AnCova, all of which are described in the following. For clarity, we shall drop the voxel index superscript k.

3.5.1 One-sample t-test

The one-sample t-test can be used to test the null hypothesis that the mean of J scans equals zero. This is the simplest model available in SPM and the design matrix consists of just a constant regressor. The model is

$$Y = x_1\beta_1 + \epsilon \tag{24}$$

where x_1 is a constant vector of ones and $\epsilon \sim N(0, \sigma^2 I_J)$. The null hypothesis is $\mathcal{H} : \beta_1 = 0$ and the alternative hypothesis is $\overline{\mathcal{H}} : \beta_1 > 0$. The t-value is computed using Eq. 12 as

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/J}} \sim t_{J-1} \tag{25}$$

where $\hat{\sigma}^2 = Y^T RY/(J-1)$, where R is the residual forming matrix (see above). In other words, $Y^T RY$ is the sum of squares of the residuals. This could also be expressed as $Y^T RY = \sum_{j=1}^{J}(Y_j - \hat{Y}_j)^2$, where $\hat{Y}_j = (x_1\hat{\beta}_1)_j = \hat{\beta}_1$.

3.5.2 Two-sample t-test

The two-sample t-test allows one to test the null hypothesis that the means of two groups are equal. The resulting design matrix (in SPM) consists of three columns, the first two encode the group membership of each scan and the third models a common constant across scans of both groups. This model is overdetermined by one degree of freedom, i.e. the sum of the first two regressors equals the third regressor. Notice the difference in parameterization compared to the two-sample t-test example in §2.5. As it turns out, the resulting t-value is nevertheless the same for a differential contrast. Let the number of scans in the first and second group be

J_1 and J_2, where $J = J_1 + J_2$. The three regressors consists of ones and zeros, where the first regressor consists of J_1 ones, followed by J_2 zeros. The second regressor consists of J_1 zeros, followed by J_2 ones. The third regressor contains ones only.

Let the contrast vector be $c = [-1, 1, 0]^T$, i.e. the alternative hypothesis is $\overline{\mathcal{H}} : \beta_1 < \beta_2$. Then $(X^T X) = \begin{pmatrix} J_1 & 0 & J_1 \\ 0 & J_2 & J_2 \\ J_1 & J_2 & J \end{pmatrix}$. This matrix is rank deficient so we use the pseudo-inverse $(X^T X)^-$ to compute the t-statistic. We sandwich $(X^T X)^-$ with the contrast and get $c^T (X^T X)^- c = 1/J_1 + 1/J_2$. The t-statistic is then given by

$$T = \frac{\hat{\beta}_2 - \hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / (1/J_1 + 1/J_2)}} \sim t_{J-2} \tag{26}$$

and $\hat{\sigma}^2 = Y^T R Y / (J - 2)$. Note that one assumption for the two-sample t-test is that $J_1 = J_2$, i.e. the number of scans in both groups is the same. However, it turns out that the two-sample t-test is rather robust against a violation of this assumption. Another assumption which we implicitly made in Eq. 26 is that we have equal variance in both groups. This assumption may not be tenable (e.g. when comparing normal subjects with patients) and we potentially have to take this non-sphericity into account (chapter 9).

3.5.3 Paired t-test

The model underlying the paired t-test is an extension to the model underlying the two-sample t-test. It is assumed that the scans come in pairs, i.e. one scan of each pair is in the first group and the other is in the second group. The extension is that the means over pairs are not assumed to be equal, i.e. the mean of each pair has to be modelled separately. For instance, let the number of pairs be $N_{pairs} = 5$, i.e. the number of scans is $J = 10$. The design matrix consists of 7 regressors. The first two model the deviation from the pair-wise mean within group and the last five model the pair-specific means. The model has degrees of freedom one less than the number of regressors.

Let the contrast vector be $c = [-1, 1, 0, 0, 0, 0, 0]^T$, i.e. the alternative hypothesis is $\overline{\mathcal{H}} : \beta_1 < \beta_2$. This leads to

$$T = \frac{\hat{\beta}_2 - \hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / (1/J_1 + 1/J_2)}} \sim t_{J-J/2-1} \tag{27}$$

The difference to the two-sample t-test lies in the degrees of freedom $J - J/2 - 1$. The two-sample t-test and the paired t-test are an example of compromising when selecting a model. The paired t-test can be a more appropriate model for a given data set, but more effects are modelled, i.e. there are less error degrees of freedom. This might come at the price of a decrease in sensitivity so that the two-sample t-test can be less appropriate, but more sensitive. This compromise is increasingly harder to make with a smaller number of scans J.

3.5.4 one-way AnCova

A one-way AnCova allows one to model group effects, i.e. the mean of each of Q groups. This model includes the one-sample and two-sample t-tests, i.e. the cases, when $1 \leq Q \leq 2$.

In our example, let the number of groups be $Q = 3$, where there are 5 scans within each group, i.e. $J_q = 5$ for $q = 1, \ldots, Q$. There are a range of different contrasts available. For instance, we could test the null hypothesis that the group means are all equal using the F-contrast as described in example §2.5.4. Here, we wish to test the null hypothesis, whether the mean of the first two groups is equal to the mean of the third group, i.e. $\mathcal{H} : (\beta_1 + \beta_2)/2 - \beta_3 = 0$ and our alternative hypothesis is $\overline{\mathcal{H}} : (\beta_1 + \beta_2)/2 < \beta_3$. This can be tested based on a t-statistic, where we use the contrast $c = [-1/2, -1/2, 1, 0]^T$. The resulting t-statistic and its distribution is

$$T = \frac{(\hat{\beta}_1 + \hat{\beta}_2)/2 - \hat{\beta}_3}{\sqrt{\hat{\sigma}^2/(1/J_1 + 1/J_2 + 1/J_3)}} \sim t_{J-Q} \tag{28}$$

4 fMRI models

In this section, we describe the analysis of functional magnetic resonance imaging (fMRI) data. For PET, we showed that we can use the general linear model to analyze the data. The models used to interpret fMRI data are modified due to differences in the character of fMRI data compared to PET. These differences include (i) serial temporal correlations, (ii) fast event-related designs, (iii) the large number of observations. A linear model can still be used, however, the normally distributed error term is non-spherical[11].

Historically, SPM was developed for and applied to PET data and therefore it is not a surprise that SPM for fMRI data was initially based on the understanding that SPM would just need some extensions to cope with the new kind of data. In this section, we therefore not only describe these extensions, but also describe the model from scratch. This has the benefits (i) that the modelling issues in fMRI analysis are described without the need to refer to PET issues and (ii) that one can skip most of the PET section if trying to learn about fMRI analysis.

The topics of this section are a linear time series model for fMRI data, temporal serial correlations and their estimation, temporal filtering, parameter estimation and inference.

4.1 A linear time series model

One of the main stays of SPM is that we use the same temporal model at each voxel, i.e. we use a mass-univariate model and perform the same analysis at each voxel. Therefore, we can describe the complete temporal model for fMRI data by looking at how the data from a single voxel (a time series) is modelled. A time series consists of the sequential measures of fMRI signal intensities over the period of the experiment. Usually, fMRI data is acquired for the whole brain with a sample time of roughly 2 to 4 seconds using an echo planar imaging (EPI) sequence. This means that a time series at a single voxel is acquired with a sample time of 2 to 4 seconds.

Multi-subject data is acquired in sessions, there being one or more sessions for each subject[12]. Here, we only talk about a model for one of these sessions, i.e. a single subject analysis. Multi-subject studies are based on multiple single-subject models and are described in chapter 12.

The process which we are going to describe in the following is at the heart of SPM. We take as an input a single time-series and transform it to a single statistical value. This statistic can then be used to derive a p-value. This is done simultaneously at all voxels so that a Statistical Parametric Map is formed with one statistic at each voxel.

Suppose we have a time series of N observations $Y_1, \ldots, Y_s, \ldots, Y_N$, acquired at one voxel at times t_s, where $s = 1, \ldots, N$ is the *scan number*. The approach is to model at each voxel the observed time series as a linear combination of explanatory functions, plus an error term:

$$Y_s = \beta_1 f^1(t_s) + \ldots + \beta_l f^l(t_s) + \ldots + \beta_L f^L(t_s) + \epsilon_s \tag{29}$$

Here the L functions $f^1(.), \ldots, f^L(.)$ are a suitable set of *regressors*, designed such that linear combinations of them span the space of possible fMRI responses for this experiment, up to the

[11] Non-sphericity refers to the deviation of the error covariance matrix from a diagonal shape or a shape that can be transformed into a diagonal shape. See also chapter 9

[12] The term session will be defined below in 4.3.1.

level of error. Consider writing out the above Equation 29 for all time points t_s, to give a set of equations:

$$\begin{aligned} Y_1 &= \beta_1 f^1(t_1) + \ldots + \beta_l f^l(t_1) + \ldots + f^L(t_1)\beta_L + \epsilon_1 \\ \vdots &= \vdots \\ Y_s &= \beta_1 f^1(t_s) + \ldots + \beta_l f^l(t_s) + \ldots + f^L(t_s)\beta_L + \epsilon_s \\ \vdots &= \vdots \\ Y_N &= \beta_1 f^1(t_N) + \ldots + \beta_l f^l(t_N) + \ldots + f^L(t_N)\beta_L + \epsilon_N \end{aligned}$$

swhich in matrix form is:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_s \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} f^1(t_1) & \ldots & f^l(t_1) & \ldots & f^L(t_1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ f^1(t_s) & \ldots & f^l(t_s) & \ldots & f^L(t_s) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ f^1(t_N) & \ldots & f^l(t_N) & \ldots & f^L(t_N) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_l \\ \vdots \\ \beta_L \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_s \\ \vdots \\ \epsilon_N \end{pmatrix} \tag{30}$$

or in matrix notation

$$Y = X\beta + \epsilon \tag{31}$$

Here each column of the design matrix X contains the values of one of the continuous regressors evaluated at each time point t_s of fMRI time series. That is, the columns of the design matrix are the discretised regressors.

The regressors must be chosen to span the space of all possible fMRI responses for the experiment in question, such that the error vector ϵ is normally distributed with zero mean. As will be discussed later, ϵ is not assumed to be spherically distributed. Rather, we will consider other forms for the covariance matrix of ϵ. This leads us out of the realm of the General Linear Model to the much broader class, the Generalized Linear Model (GLM). However, because we are modelling the data with a normally distributed error term and do not consider other error distributions or so called *link functions*, we look at a rather constrained class of GLMs.

4.2 Proportional and grand mean scaling

Before we proceed to the description of how the regressors in the design matrix are generated, we want to mention the issue of global normalization. fMRI data is known to be subject to various processes that cause globally distributed confounding effects, e.g. (Andersson et al., 2001). A rather simple global confounding source is the scanner gain. This volume-wise gain is a factor that scales the whole image and is known to slowly vary during a session. A simple way to remove the effect of such a varying gain is to estimate this gain per image and multiply all image intensities by this gain estimate. This method is known as *proportional scaling*.

If one does not use proportional scaling, SPM performs by default a session-specific scaling. This type of scaling divides each volume by a session-specific gain. This is known in SPM as

grand mean scaling. Session-specific grand mean scaling is highly recommended, because the session-specific gains can strongly vary between sessions, masking any activations.

To estimate the gain factors, SPM uses a rough estimate of the volume-wise intracerebral mean intensity. Note that both kinds of scaling also scale the mean global activity (either of a volume or of a session) to 100. The data and a signal change can then be conveniently interpreted as percent with respect to the estimated global intracerebral mean.

4.3 Generation of regressors

In the following, we will describe how the regressors in Eq. 30 are generated and what the underlying model of the BOLD response is. This process consists of several stages. Although these are mostly hidden from the user, it can be helpful to know about intermediate processing steps and their temporal sequence.

The overall aim of regressor generation is to come up with a design matrix that models the expected fMRI response at any voxel as a linear combination of its columns. Basically, there are two things SPM needs to know to construct the design matrix. The first is the timings of the experiment and the second is the expected shape of the BOLD response due to stimulus presentation. Given this information, SPM computes the design matrix. In the following, we will go through the stages of this process.

4.3.1 Timings

We describe how a design matrix for one session of functional data is generated. Let the number of scans in a session be N_{scans}. Furthermore, it is important that the data is ordered according to acquisition order.

In SPM a session starts at session time zero. This time point is given when the first slice of the first scan was started to be acquired by the scanner. Session time can be measured both in scans or in seconds. In both cases the session starts at time zero whatever the units. The duration of a session is the number of scans multiplied by the volume repetition time (RT) which is the time spent from the beginning of the acquisition of one scan to the beginning of the acquisition of the next scan. We assume that RT stays constant throughout a session. The RT and the number of scans of a given session completely define the start and the end of a session. Moreover, because we assume that RT stays constant throughout the experiment, one also knows the onset of each scan.

The design of the experiment is described as a series of trials or events, where each trial is associated with a trial type. Let N^m_{trials} be the number of trials of trial type m and N_{types} the number of trial types. For each trial j of trial type m, one needs to specify its onset and duration. Note that we do not need to make a distinction between event-related or blocked designs so that a trial can be either a short event or an epoch. Let the onset vector of trial type m be O^m so that O^m_j is the onset of trial j of trial type m. For example, the onset of a trial that started at the beginning of scan 4 is at session time 3 (in scans) or at session time $3 \cdot RT$ (in seconds).

Let vector D^m contain the user-specified stimulus durations of each trial for trial type m.

Given all onsets O^m and durations D^m, SPM generates an internal representation of the session and the experiment. This representation consists of the discretized stimulus function S^m for each trial type m. All time bins of a session are covered such that the vectors S^m represent a

contiguous series of time bins. These time bins typically do not cover a time period of length RT, but a fraction of it to provide a well sampled discretized version of the stimulus functions S^m, i.e. they are over-sampled.

The occurrence of a stimulus is binarily represented in the stimulus functions. The elements of the stimulus function can also contain other values. An important application of this lies in the concept of *parametric modulation*, see §4.3.3.

Note that the degree of discretization of the stimulus functions is controlled by the user. Time bin size is specified in number of time bins per RT[13].

For example, assume the RT is 3.2 seconds. Then each time bin given the default of 16 bins/RT covers 200 milliseconds. The length of the vector S^m is $16N_{scans}$. Note that choosing a smaller time bin size does not necessarily provide a higher temporal precision for the resulting regressors in the design matrix. This is because the expected BOLD response is located in a rather low frequency band. Therefore, responses to trials being only a few milliseconds apart from each other are virtually indistinguishable.

4.3.2 High-resolution basis functions

After the generation of the stimulus functions, we need to describe the shape of the expected response. This is done using temporal basis functions. During the development of Statistical Parametric Mapping over the last few years, some effort has gone into designing sets of basis functions which appropriately model the expected blood oxygen level dependent (BOLD) response. The underlying model is that the BOLD response for a given trial type m is generated by feeding the stimulus function through a linear finite impulse response (FIR) system, whose output is the observed data Y. This is expressed by the model

$$Y = d(\sum_{m=1}^{N_{types}} h^m \otimes S^m) + \epsilon \tag{32}$$

where h^m is the impulse response function for trial type m. The $\otimes$ operator denotes the convolution of two vectors (Bracewell, 1986). $d(\cdot)$ denotes the down-sampling operation which is needed to sample the convolved stimulus functions at each sampled time point. In other words, the observed data Y is modelled by summing the output of N_{types} different linear systems. Additionally, we add some (measurement) noise ϵ. The input to the mth linear system is the stimulus function of trial type m.

The impulse response functions h^m are not known, but we assume that they can be modelled as linear combinations of some basis functions b_i:

$$Y = \sum_{m=1}^{N_{types}} \sum_{i=1}^{N_{bf}} d(b_i \beta_i^m \otimes S^m) + \epsilon \tag{33}$$

where β_i^m is the ith coefficient for trial type m and N_{bf} is the number of basis functions b_i.

We can move the coefficients outside the sampling operator so that we get

$$Y = d(\left[(b \otimes S^1)\beta^1 + \cdots + (b \otimes S^{N_{types}})\beta^{N_{types}}\right]) + \epsilon \tag{34}$$

[13]The effective time bin size is accessible in SPM as variable fMRI_T. Its default value is 16.

where $b = \left[b_1, \ldots, b_{N_{bf}}\right]$ and $\beta^m = \left[\beta_1^{m^T}, \ldots, \beta_{N_{bf}}^{m^T}\right]^T$. Note that we define the convolution to operate on the columns of matrix b. If we let $X = \left[(b \otimes S^1) \vdots \cdots \vdots (b \otimes S^{N_{types}})\right]$ and $\beta = \left[\beta^{1^T}, \ldots, \beta^{N_{types}{}^T}\right]^T$, we see that Eq. 34 is a linear model like Eq. 31. The columns of the design matrix X are given by the discretely sampled convolution of each of the N_{types} stimulus functions with each of the N_{bf} basis functions. Note that although we assumed different impulse response functions for each trial type m, our parameterization leads to the same basis functions b_i for each trial type, but different parameter vectors $\beta_1^m, \ldots, \beta_{N_{bf}}^m$.

In summary, when we choose a specific basis function set b_i, we express our belief that a linear combination of the convolved basis functions is able to model an experimentally induced effect. The question remains which basis function set is appropriate for fMRI data. In SPM, the default choice is a parameterised model of the expected impulse response function. This function is a superposition of two gamma functions. To form an appropriate basis function set, one usually complements this function with its first partial derivatives with respect to some generating parameters. In SPM, the default choice is to add partial derivatives with respect to two generating parameters, the onset and dispersion. This gives a basis function set with three basis functions; b_1 is the expected response function, b_2 its partial derivative with respect to onset (time) and b_3 its partial derivative with respect to dispersion. In SPM, this set is usually referred to as the 'haemodynamic response function (HRF) with derivatives'. In practice, this set can model a BOLD response that (i) can be slightly shifted in time with respect to the expected delay or (ii) has a different width than the HRF model b_1. This issue is dealt within more detail in chapter 10.

4.3.3 Parametric Modulation

When we first introduced the stimulus functions S^m they were described as vectors consisting of ones and zeros. However, one can also assign numbers other than 1 to the S^m. More interestingly, one can assign different values to different individual trials. As one can see from Eq. 34, after convolution of the S^m with the basis functions b_i, different weights in S^m essentially control the relative height of the expected response of all trials. This weighting allows models where one can parametrically modulate the relative response height over trials. There is a wide range of applications for parametric modulations. For instance, one can weight events by a linear function of time, which models a linear change in the individual responses over time. Another application is the weighting of S^m with some external measure that was acquired trial-wise, e.g. reaction times. Such a modulated regressor would allow one to test for a linear dependence between reaction times and height of response while taking into account all other modelled effects. Higher order modulations can be modelled by polynomial expansions of the modulation, which give us multiple parametrically modulated regressors per trial type.

4.3.4 Low-resolution basis functions

In Eq. 34, a down-sampling operator d was applied to sample the high-resolution (continuous) regressors to the low-resolution space of the data Y. Here, one has to be aware of a slight limitation of the SPM model for event-related data that arises due to the use of the same temporal model at each voxel.

fMRI data is typically acquired slice-wise so that a small amount of time elapses from the acquisition of one slice to the next. Given standard EPI sequences, acquisition of one slice

takes roughly 100 ms. Therefore, an optimal sampling of the high-resolution basis functions does not exist, because any chosen sampling will only be optimal for one slice, but not for all the others. The largest timing error is given for a slice that lies in acquisition order $\lfloor N_{slices}/2 \rfloor$ slices away from the slice for which the temporal model is exact[14]. This sampling issue is only relevant for event-related designs, where one typically uses short stimulus durations that elicit BOLD responses lasting only some seconds. For these transient responses, an appropriate temporal model is critical. Any difference in expected and actual onset may decrease the sensitivity of the analysis, if one uses a naive HRF model (e.g. only the HRF model without its derivatives). For blocked designs, timing errors are small compared to epoch length so that the potential loss in sensitivity is negligible.

In SPM, there are two ways to solve this timing issue and take the different slice acquisition times into account. The first is to choose one time point within volume acquisition time and temporally interpolate all slices at this time. This is called *slice timing correction*. However, note that this interpolation requires rather short RT (< 3 seconds), because the sampling should be dense enough in relation to the width of the BOLD response to capture its interesting peak. The second option is to model latency differences with the temporal derivative of the HRF set. As discussed above, the temporal derivative can model a temporal shift of the expected BOLD response. This temporal shift can not only capture onset timing differences due to different slice times, but also differences due to, for example, a different vascular response onset. However, due to the linear nature of the model, the temporal derivative can only model small shifts (forwards or backwards in time). With the HRF basis functions set, the temporal derivative can accommodate a shift backwards or forwards of slightly more than one second. The slice timing interpolation is recommended if one looks for voxel-specific timing differences between conditions. Independently of this, we recommend the use of the temporal derivative as part of the model to capture any potential latency differences.

One also needs to specify at what time bin, in scan time, SPM samples the regressors to generate the design matrix. The SPM default is 1, i.e. the first time bin after the start of a scan.[15]

Finally, the down-sampled basis functions are mean corrected and entered column-wise into the design matrix X (Eq. 30). A baseline is modelled by adding a constant regressor to the design matrix.

4.3.5 Additional regressors

It is possible to use additional regressors in the model without going through the process described above. For instance, consider the case that an additional physiological measurement was acquired during the session at a high temporal resolution. These measurements can be added to the design matrix after suitable down-sampling. Another important example for user-specified regressors is the modelling of movement correlated effects. These can be taken into account to a first order by adding the estimated movement parameters as regressors (see chapter 2). Note that all user-specified regressors are automatically mean-corrected by SPM.

4.4 Serial correlations

fMRI data exhibits short range serial temporal correlations. By this we mean that the error ϵ_s at a given scan s is correlated with its temporal neighbours. This has to be modelled,

[14] $\lfloor x \rfloor$ denotes the nearest integer less or equal to x

[15] This sampling point is accessible in SPM as variable $fMRI_{T0}$ and lies between 1 and $fMRI_T$, the number of time bins.

because correlations play an important role when assessing the significance of a test statistic. Ignoring correlations leads to an inappropriate estimate of the error covariance matrix, which is propagated to the estimated parameter covariance matrix. In other words, when forming a t- or F-statistic, we have a biased estimate of the variability of a contrast. Additionally, when using ordinary least squares estimates, the null statistic, which is used when computing p-values, is also dependent on the error covariance matrix. This dependency enters when estimating the effective degrees of freedom of a null distribution[16]. With serial correlations present and modelled, the effective degrees of freedom are lower than in the independent case. The overall picture is that ignoring serial correlations leads generally to too lenient and therefore invalid tests. To derive correct tests we have to appropriately estimate the error covariance matrix by assuming some kind of non-sphericity (chapter 9). Then, we use this estimate in the computation of the statistic and the effective degrees of freedom.

Note that we are only concerned about the serial correlations of the error component ϵ of the time series (Eq. 31). The correlations induced by the experimental design should be modelled by the design matrix X.

Serial correlations in fMRI data are caused by various sources including cardiac, respiratory and vasomotor sources (Mitra et al., 1997).

There are two issues that need to be resolved. The first is how to estimate the error covariance matrix and the second is how to incorporate this estimate into our modelling framework and derive a valid statistical test. In what follows we describe the statistics that are based on ordinary least squares (OLS) parameter estimates.

One model that seems to capture the observed form of serial correlations in fMRI data is the autoregressive (order 1) plus white noise model ($AR(1)$+wn) (Purdon and Weisskoff, 1998)[17]. This model accounts for short range correlations. Note that the order of the model (one) means that the form and amount of correlations can be modelled by one (AR) coefficient. The model order does not refer to the range of the serial correlations in time, which can be up to 6 to 8 scans. We only need to model short range correlations, because we also apply a highpass filter to the data (s. 4.5). The highpass filter removes any low frequency components and thus long range correlations from the data. We refer the interested reader to the Appendix for a mathematical description of the $AR(1)$+wn model.

4.4.1 Estimation of the error covariance matrix

Having decided that the $AR(1)$+wn is an appropriate model for the fMRI error covariance matrix, we need to estimate its three hyperparameters (s. Appendix) at each voxel. The hyperparameterised model gives an autocovariance matrix at each voxel (Eq. 45), which we want to estimate. In SPM, an additional assumption is made to estimate this matrix more efficiently, which is described in the following.

Mathematically, the error covariance matrix can be partitioned into two components. The first component is the correlation matrix and the second component is the variance. The assumption made by SPM is that the correlation matrix is the same at all voxels of interest (see chapter 9 for further details). The variance is assumed to be different between voxels. In other words, SPM assumes that the pattern of serial correlations is the same over all interesting voxels, but its amplitude is different at each voxel. This assumption seems to be quite a sensible one, because

[16] Effective degrees of freedom refer to the degrees of freedom of an approximation to the underlying null distribution (Worsley and Friston, 1995).

[17] The $AR(1)$+wn is also known as the autoregressive moving-average model of order (1,1) (ARMA(1,1))

we observed that the serial correlations over voxels within tissue types are very similar. The estimate of the serial correlations is therefore extremely precise because of the large number of voxels involved in the estimation. Therefore, the correlation matrix at each voxel can be assumed to be known.

In the following, we describe the model and estimation of the error covariance matrix. Let us start with the linear model for voxel k

$$Y^k = X\beta^k + \epsilon^k \tag{35}$$

where Y^k is a $N \times 1$ observed time series vector at voxel k, X is a $N \times L$ design matrix, β^k is the parameter vector and ϵ^k is the error at voxel k. The error ϵ^k is normally distributed with $\epsilon \sim N(0, {\sigma^k}^2 V)$. The critical difference to Eq. 6 is the distribution of the error term where the identity matrix I is replaced by the correlation matrix V. Note that V does not depend on the voxel position k, i.e. we make the above mentioned assumption that the correlation matrix V is the same for all voxels $k = 1, \ldots, K$. However, the variance ${\sigma^k}^2$ is assumed to be different for each voxel.

How can the correlation matrix V be estimated over all voxels? Since we made the assumption that V is the same at each voxel, we can either estimate V at each voxel and then pool our estimate, or we can pool data from all voxels and then estimate V on this pooled data. We use the second method, because it is computationally much more efficient. The pooled data is given by summing the sampled covariance matrix of all interesting voxels k, i.e. $V_Y = 1/K \sum_k Y^k {Y^k}^T$. Note that the pooled V_Y is a mixture of two variance components, the experimentally induced variance and the error variance component:

$$V_Y = \sum_k X\beta^k {\beta^k}^T X^T + \epsilon^k {\epsilon^k}^T \tag{36}$$

One way of estimating the error covariance matrix $Cov(\epsilon^k) = {\sigma^k}^2 V$ is to use the Restricted Maximum Likelihood (ReML) method (Harville, 1977; Friston et al., 2002). ReML takes the space spanned by the design matrix into account and is an unbiased estimator of the hyperparameters. ReML works with linear covariance constraints, i.e. the estimated covariance matrix is modelled as a linear combination of some covariance constraints. The concept of covariance constraints is a very general concept that can be used to model all kinds of non-sphericity (see chapter 9). The model described in the Appendix (Eq. 44) is nonlinear in the hyperparameters so ReML cannot be used directly. But if we linearize the covariance constraints

$$V = \sum_l \lambda_l Q_l \tag{37}$$

where Q_l are $N \times N$ constraint matrices and the λ_l are the hyperparameters, ReML can be applied. We are interested in specifying the Q_l such that they form an appropriate model for serial correlations of fMRI data when using standard EPI sequences. The default model in SPM is to use two constraints Q_1 and Q_2. These are $Q_1 = I_N$ and

$$Q_{2_{ij}} = \begin{cases} e^{-|i-j|} & : \quad i \neq j \\ 0 & : \quad i = j \end{cases} \tag{38}$$

Fig. 16 shows the shape of Q_1 and Q_2.

[Figure 16 about here.]

A voxel-wide estimate of V is then derived by rescaling V such that V is a correlation matrix.

This method of estimating the covariance matrix at each voxel uses the two voxel-wide (global) hyperparameters λ_1 and λ_2. A third voxel-wise (local) hyperparameter (the *variance* σ^2) is estimated at each voxel using the usual estimator (Worsley and Friston, 1995)

$$\sigma^{2^k} = \frac{{Y^k}^T R Y^k}{trace(RV)} \tag{39}$$

where R is the residual forming matrix. This completes the estimation of the serial correlations at each voxel k. Before we can use these estimates to derive statistical tests, we still need to describe the highpass filter and what role it plays in modelling fMRI data.

4.5 Temporal filtering

The concept of filtering is based on the observation that certain frequency bands in the data contain more noise than others. In an ideal world, our experimentally induced effects would live in one frequency band and all the noise in another. Applying a filter that removes the noise frequency range from the data would then give us increased sensitivity. However, the data is a mixture of activation and noise that can share some frequency bands. One of the experimenter's tasks is therefore to make sure that the interesting effects do not lie in a frequency range which is especially exposed to noise processes. In fMRI, the low frequencies (say less than half a cycle per minute, i.e. 1/120 Hz) are known to contain scanner drifts and possibly cardiac/respiratory artifacts. Any activations that lie within this frequency range are virtually undistinguishable from these noise processes. This is why (i) fMRI data should be highpass filtered to remove noise and (ii) the experimenter should take care to construct a design that puts the interesting contrasts into higher frequencies than 1/120 Hz. This issue is especially important for event-related designs and is dealt with in chapter 10. Here, we describe how the highpass filter is implemented.

The *highpass filter* is implemented using a set of discrete cosine transform (DCT) basis functions. These are part of the design matrix. To the user of SPM, they are *invisible* in the sense that the DCT regressors are never plotted. This is simply to save space on the display. In practice, the parameters of the DCT part of the design matrix are not estimated, but the residual forming matrix of the DCT regressors are applied to the data. Only after this step, the other (visible) part of the design matrix is fitted to the resulting residuals. This procedure is equivalent to estimating all model parameters simultaneously, where all tests of hypotheses automatically take low frequency noise components into account.

Mathematically, for time points $t = 1, \ldots, N$, the discrete cosine set functions are $f_r(t) = \sqrt{2/N}\left(cos\left(r\pi\frac{t}{N}\right)\right)$. See Fig. 17 for an example. The integer index r ranges from 1 (giving half a cosine cycle over the N time points), to a user-specified maximum R. Note that SPM asks for a highpass cutoff d_{cut} in seconds. R is then chosen as $R = \lfloor 2NRT/d_{cut} + 1 \rfloor$.

[Figure 17 about here.]

To summarize, the following picture emerges. The regressors in the design matrix X must account for all components in the fMRI time series up to the level of residual noise. The high-pass filter is part of the design matrix and removes unwanted low-frequency components from

the data. The estimation of the error covariance matrix is based on a model similar to the $AR(1)$+wn model and uses the ReML-method for estimation of the hyperparameters. In the next section, we describe how the model parameter estimates are used to form a t- or F-statistic at each voxel.

4.6 Parameter estimates and distributional results

In this section, we describe the equations that lead to a t- or F-statistic. These statistics can be used to make inferences about the data by computing a p-value at each voxel. For clarity, we shall drop the voxel index superscript k.

Ordinary least-squares parameter estimates $\hat{\beta}$ are given by

$$\hat{\beta} = (X^T X)^- X^T Y = X^- Y \tag{40}$$

As described above, we estimate the error correlation matrix V using the ReML method. The error covariance matrix is then given by $\hat{\sigma}^2 V$ (Eq. 39). The covariance matrix of the parameter estimate is

$$Var(\hat{\beta}) = \sigma^2 X^- V X^{-T} \tag{41}$$

A t-statistic can then be formed by dividing a contrast of the estimated parameters $c^T\hat{\beta}$ by its estimated standard deviation:

$$T = \frac{c^T\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T X^- V X^{-T} c}} \tag{42}$$

where σ^2 is estimated using Eq. 39.

The key difference to the spherical case, i.e. when the error is i.i.d. is that the correlation matrix V enters into the denominator of the t-value. This gives us a more accurate t-statistic. However, because of V the denominator of Eq. 42 is not the square root of a χ^2-distribution. (The denominator would be exactly χ^2 distributed, when V describes a spherical distribution.) This means that Eq. 42 is not t-distributed and we cannot simply make inferences by comparing with a t- null distribution with $trace(RV)$ degrees of freedom.

Instead, one approximates the denominator with a χ^2-distribution (Eq. 42). Consequently, T is then approximated by a t-distribution. The approximation proposed in (Worsley and Friston, 1995) is the Satterthwaite approximation (see also (Yandell, 1997)) which is based on fitting the first two moments of the denominator distribution with a χ^2 distribution. The degrees of freedom of the approximating χ^2-distribution are called the *effective* degrees of freedom and are given by

$$\nu = \frac{2E(\hat{\sigma}^2)^2}{Var(\hat{\sigma}^2)} = \frac{\mathrm{trace}(RV)^2}{\mathrm{trace}(RVRV)} \tag{43}$$

See the Appendix for a derivation of this Satterthwaite approximation.

Similarly, the null distribution of an F-statistic in the presence of serial correlations can be approximated. In this case, both the numerator and denominator of the F-value are approximated by a χ^2-distribution.

4.7 Summary

After reconstruction, realignment, spatial normalisation and smoothing, functional imaging data are ready for statistical analysis. This involves two steps: Firstly, statistics indicating evidence against a null hypothesis of no effect at each voxel are computed. An image of these statistics is then produced. Secondly, this statistical image must be assessed, reliably locating voxels where an effect is exhibited whilst limiting the possibility of false positives. These two steps are referred to as (1) Modelling and (2) Inference and they are covered separately in sections 2 and 3 of this book.

As models are designed with inference in mind it is often difficult to separate the two issues. However, the inference section, section 3, in this book is largely concerned with the multiple comparison, that is, how to correctly make inferences from large volumes of statistic images. A distinction can be made between such 'image-level' inference and statistical inference at a single voxel. This second sort of inference has been covered in this chapter and will be dealt with further in the remainder of section 2.

We have shown how the general linear model, the workhorse of functional imaging analysis, provides a single framework for many statistical tests and models, giving great flexibility for experimental design and analysis. The use of such models will be further highlighted in the following chapters, especially Chapters 8 and 9. Additionally, to incorporate non-spherical error distributions, SPM uses covariance constraints and the ReML estimator. This is described further in Chapter 9.

In Chapters 10 and 11 we focus on modelling issues specific to fMRI and in chapters 12 and 13 consider making inferences from multiple subject fMRI and PET studies. In Chapter 13 we take up recent developments in the field which make used of hierarchical models. This introduction to the area paves the way for further development in section 2, in particular Chapter 17.

Appendix

A1 — The autoregressive model of order 1 plus white noise

Mathematically, the $AR(1)$+wn model at voxel k can be written in state-space form:

$$\begin{aligned} \epsilon(s) &= z(s) + \delta_\epsilon(s) \\ z(s) &= az(s-1) + \delta_z(s) \end{aligned} \tag{44}$$

where $\delta_\epsilon(s) \sim N(0, {\sigma_\epsilon}^2)$, $\delta_z(s) \sim N(0, {\sigma_z}^2)$ and a is the $AR(1)$ coefficient. This model describes the error component $\epsilon(s)$ at time point s and at voxel k as the sum of an autoregressive component $z(s)$ plus white noise $\delta_\epsilon(s)$. We have three hyperparameters[18] at each voxel k, the variances of the two error components δ_ϵ and δ_z and the autoregressive coefficient a. The resulting error covariance matrix is then given by

[18] We call these parameters hyperparameters to distinguish them from the parameter vector β

$$E(\epsilon\epsilon^T) = {\sigma_z}^2(I_N - A)^{-1}(I_N - A)^{-T} + {\sigma_\epsilon}^2 \tag{45}$$

where A is a matrix with all elements of the first lower off-diagonal set to a and zero elsewhere. I_N is the identity matrix of dimension N.

A2 — The Satterthwaite approximation

The unbiased estimator for σ^2 is given by dividing the sum of the squared residuals by its expectation (Worsley and Friston, 1995). Let e be the residuals $e = RY$, where R is the residual forming matrix.

$$\begin{aligned} E(e^T e) &= E(trace(ee^T)) \\ &= E(trace(RYY^T R^T)) \\ &= trace(R\sigma^2 V R^T) \\ &= \sigma^2 trace(RV) \end{aligned}$$

An unbiased estimator of σ^2 is given by $\hat{\sigma}^2 = \frac{e^T e}{trace(RV)}$. If V is a diagonal matrix with identical non-zero elements, $trace(RV) = trace(R) = J - p$, where J is the number of observations and p the number of parameters.

In what follows, we derive the Satterthwaite approximation to a χ^2-distribution given a non-spherical error covariance matrix.

We approximate the distribution of the squared denominator of the t-value (Eq. 42) $d = \hat{\sigma}^2 c^T (X^T X)^- X^T V X (X^T X)^- c\ \hat{\sigma}^2$ with a scaled χ^2-variate, i.e.

$$d \sim p(ay) \tag{46}$$

where $p(y) \sim \chi^2(\nu)$. We want to estimate the effective degrees of freedom ν. Note that, for a $\chi^2(\nu)$ distribution, $E(y) = \nu$ and $Var(y) = 2\nu$. The approximation is made by matching the first two moments of d to the first two moments of ay:

$$E(d) = a\nu \tag{47}$$

$$Var(d) = a^2 2\nu \tag{48}$$

If the correlation matrix V (Eq. 42) is assumed to be known, it follows that

$$\nu = \frac{2E(\hat{\sigma}^2)^2}{Var(\hat{\sigma}^2)} \tag{49}$$

With $E(\hat{\sigma}^2) = \sigma^2$ and

$$\begin{aligned} E(e^T e e^T e) &= E(2trace((e_i e_i^T)^2) + trace(e_i e_i^T)^2) \\ &= \sigma^4(2trace(RVRV) + trace(RV)^2) \end{aligned}$$

we have

$$\begin{aligned} Var(\hat{\sigma}^2) &= E(\hat{\sigma}^4) - E(\hat{\sigma}^2)^2 \\ &= \frac{\sigma^4(2trace(RVRV) + trace(RV)^2)}{trace(RV)^2} - \sigma^4 \\ &= \frac{2\sigma^4 trace(RVRV)}{trace(RV)^2} \end{aligned}$$

Using Eq. 49, we get

$$\nu = \frac{trace(RV)^2}{trace(RVRV)} \tag{50}$$

References

Andersson, J. L., Huttton, C., Ashburner, J., Turner, R., and Friston, K. 2001. Modeling geometric deformations in EPI time series. *Neuroimage*, 13:903–919.

Bracewell, R. 1986. *The Fourier transform and its applications.* McGraw-Hill International Editions, 2nd edition.

Chatfield, C. 1983. *Statistics for Technology.* Chapman & Hall, London.

Christensen, R. 1996. *Plane Answers to Complex Questions: The Theory of Linear Models.* Springer-Verlag, Berlin.

Draper, N. and Smith, H. 1981. *Applied Regression Analysis.* John Wiley & Sons, New York, 2nd edition.

Friston, K., Frith, C., Liddle, P., Dolan, R., Lammertsma, A., and Frackowiak, R. 1990. The relationship between global and local changes in PET scans. *Journal of Cerebral Blood Flow and Metabolism*, 10:458–466.

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., and Frackowiak, R. 1995. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210.

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

Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the Am. Stat. Assoc.*, 72:320–338.

Healy, M. 1986. *Matrices for Statistics.* Oxford University Press, Oxford.

Mitra, P., Ogawa, S., Hu, X., and Ugurbil, K. 1997. The nature of spatiotemporal changes in cerebral hemodynamics as manifested in functional magnetic resonance imaging. *Magnetric Resonance Imaging in Medicine*, 37:511–518.

Mould, R. 1989. *Introductory Medical Statistics.* Institute of Physics Publishing, London, 2nd edition.

Purdon, P. and Weisskoff, R. 1998. Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human Brain Mapping*, 6:239–249.

Scheffé, H. 1959. *The Analysis of Variance.* Wiley, New York.

Winer, B., Brown, D., and Michels, K. 1991. *Statistical Principles in Experimental Design.* McGraw Hill, New York, 3rd edition.

Worsley, K. and Friston, K. 1995. Analysis of fMRI time-series revisited — again. *Neuroimage*, 2:173–181.

Yandell, B. S. 1997. *Practical data analysis for designed experiments.* Chapman & Hall, first edition.

List of Figures

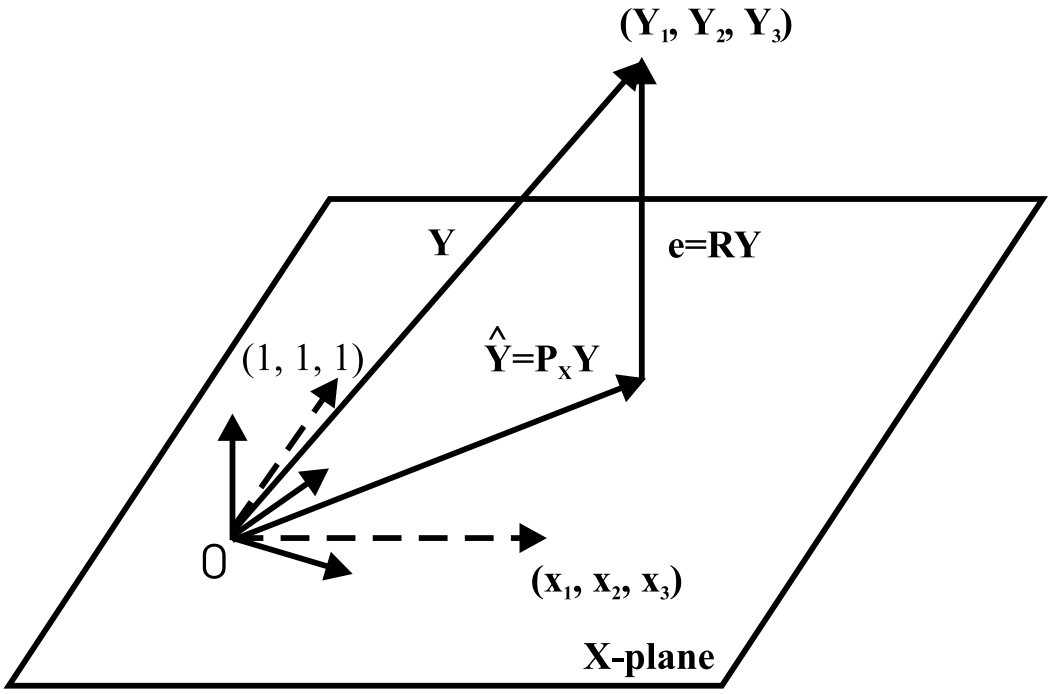


Figure 1: Geometrical perspective on linear regression: The three-dimensional data Y lies in a three-dimensional space. In this observation space, the (two-column) design matrix spans a subspace. Note that the axes of the design space are not aligned with the axes of the observation space. The least-squares estimate is the point in the space spanned by the design matrix that has minimal distance to the data point.

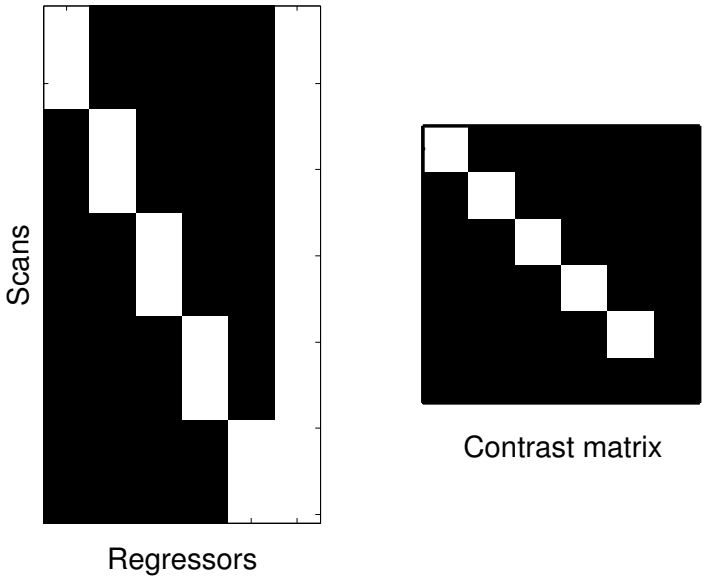


Figure 2: Example of ANOVA design and contrast matrix. Both matrices are displayed as images, where 0s are coded by black and 1s by white, cf. 2.7. (Left:) Design matrix, where five groups are modelled by their mean and overall mean. The model is overdetermined by one degree of freedom. (Right:) F-contrast matrix which tests for any group-specific deviation from the overall mean.

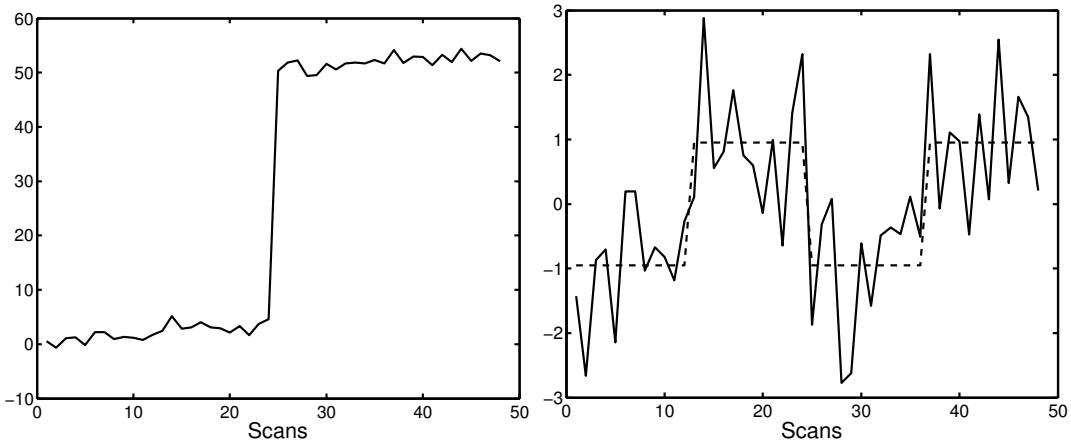


Figure 3: Adjusted and fitted data. Left: Plot of raw data. Right: (Solid line:) adjusted data, (Dashed line:) fitted data

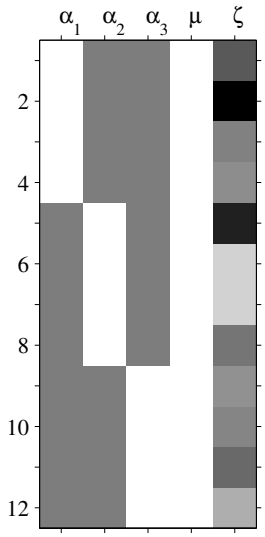


Figure 4: Single subject activation experiment, ANCOVA design (§3.3.1). Illustrations for a three-condition experiment with four scans in each of three conditions, ANCOVA design. Design matrix image, with columns labelled by their respective parameters. The scans are ordered by condition.

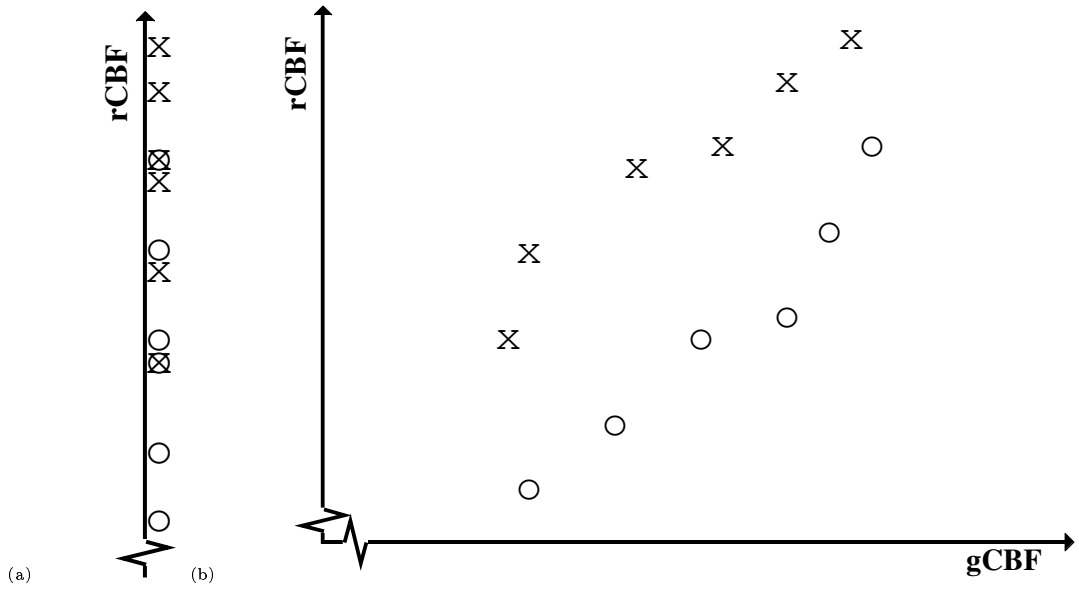
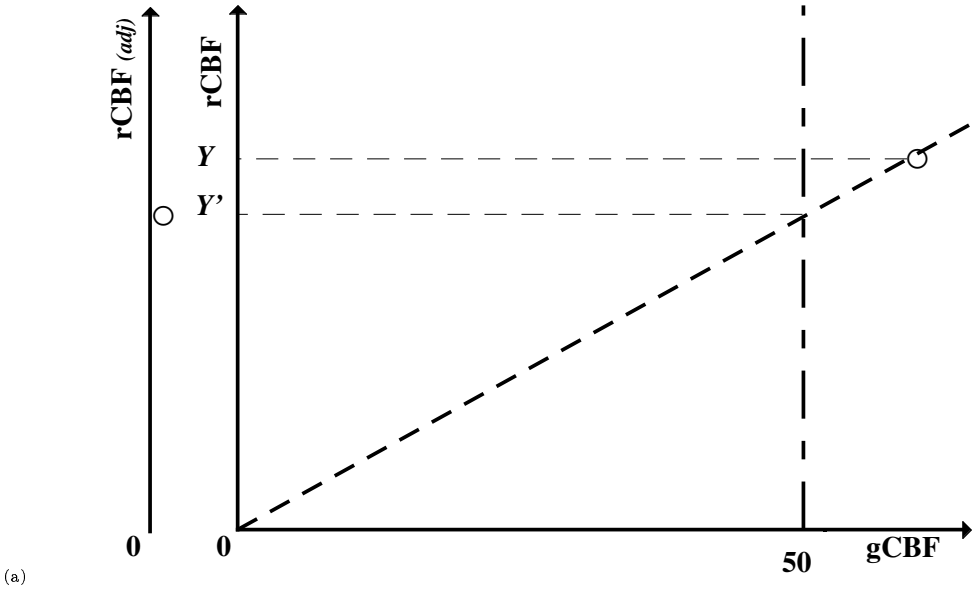
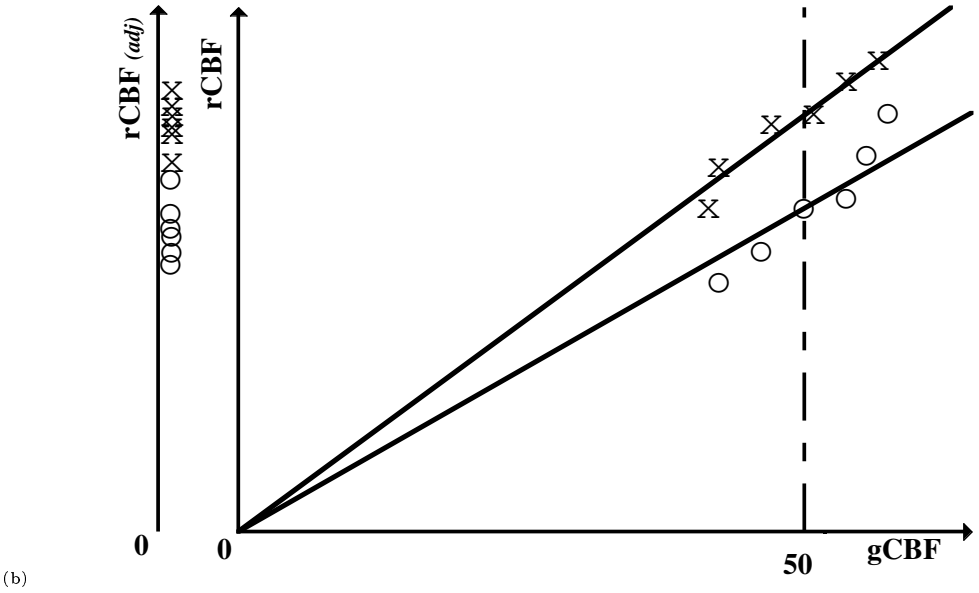


Figure 5: Single subject PET experiment, illustrative plots of rCBF at a single voxel: (a) Dot-plots of rCBF(b) Plot of rCBF vs. gCBF. Both plots indexed by condition: $\circ$ for baseline, $\times$ for active.



(a)



(b)

Figure 6: (a) Adjustment by proportional scaling (b) Simple single subject activation as a t-test on adjusted rCBF: Weighted proportional regression

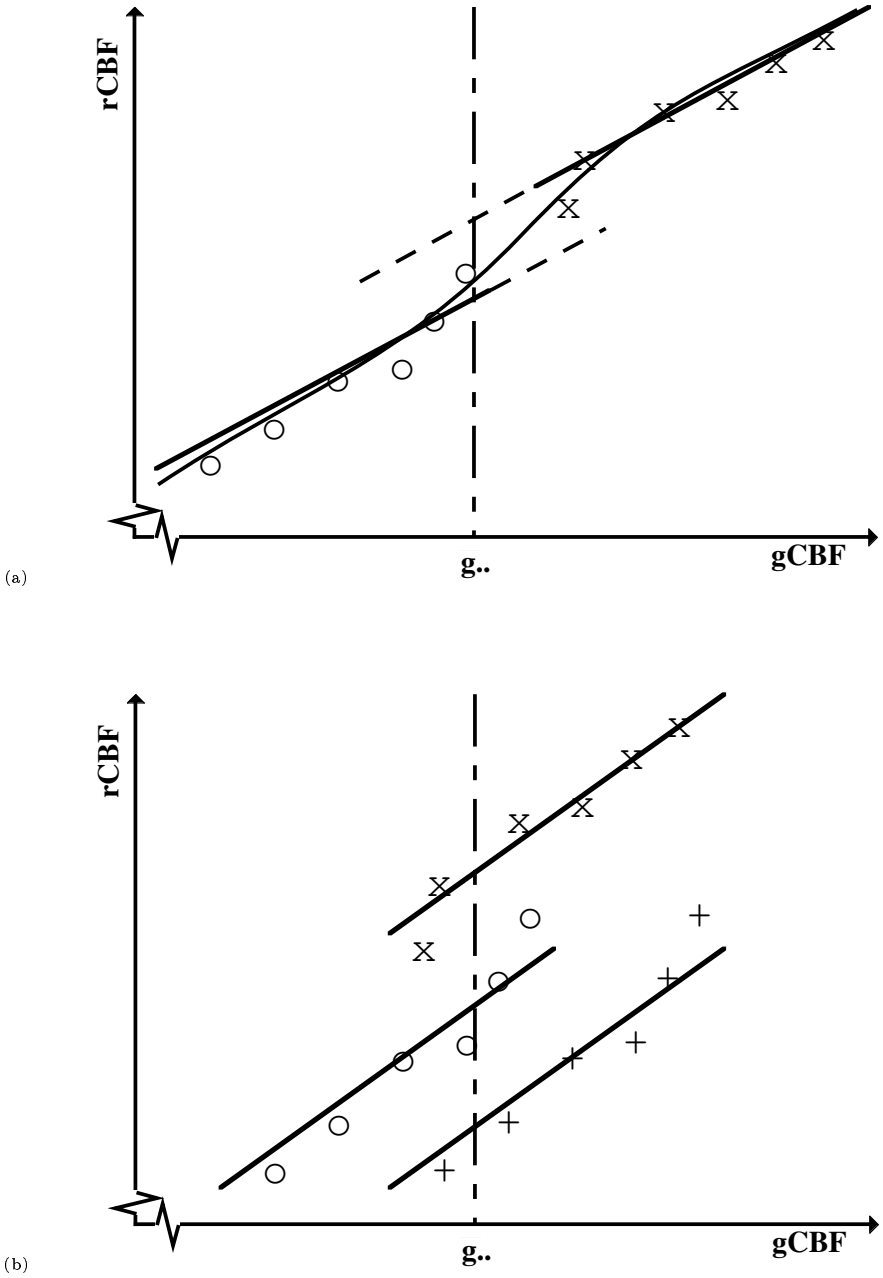
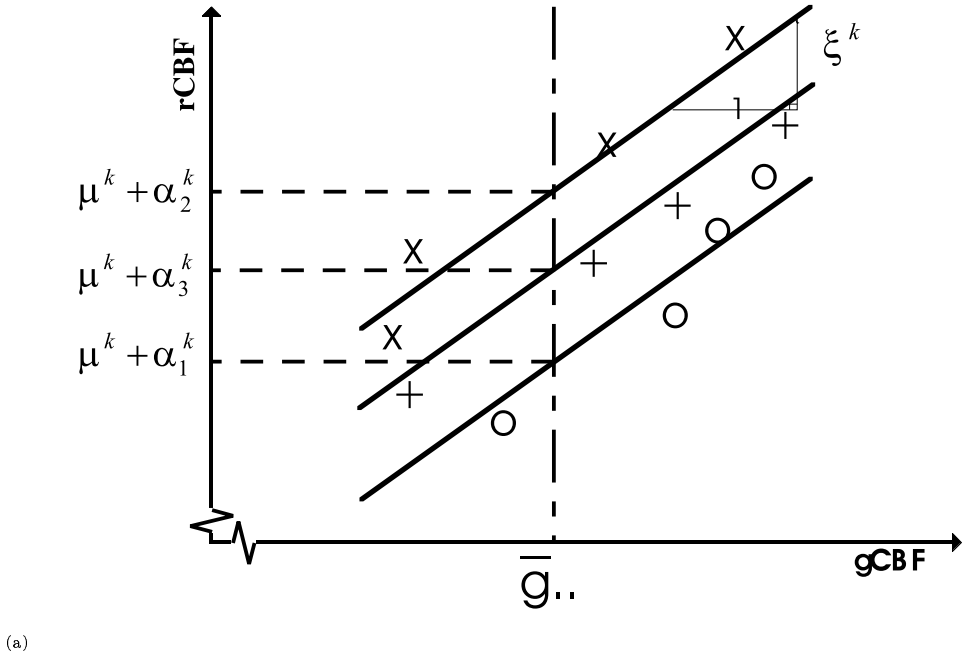
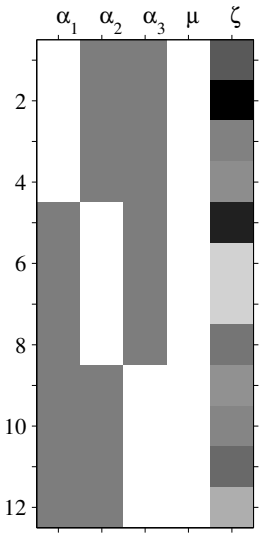


Figure 7: Single subject data, illustrative (ANCOVA) plots of rCBF vs. gCBF at a single voxel showing potential problems with global changes: (a) Large change in gCBF between conditions. The apparent activation relies on linear extrapolation of the baseline and active condition regressions (assumed to have the same slope) beyond the range of the data. The actual relationship between regional and global for no activation may be given by the curve, in which case there is no activation effect. (b) Large activation inducing increase in gCBF measured as brain mean rCBF. Symbol $\circ$ denotes rest, $\times$ denotes active condition values if this is a truly activated voxel (in which case the activation is underestimated), while $+$ denotes active condition values where this voxel is not activated (in which case an apparent deactivation is seen).



(a)



(b)

Figure 8: Single subject study, ANCOVA design (§3.3.1). Illustration of a three-condition experiment with four scans in each of three conditions, ANCOVA design. (a) Illustrative plot of rCBF vs. gCBF. (b) Design matrix image with columns labelled by their respective parameters. The scans are ordered by condition.

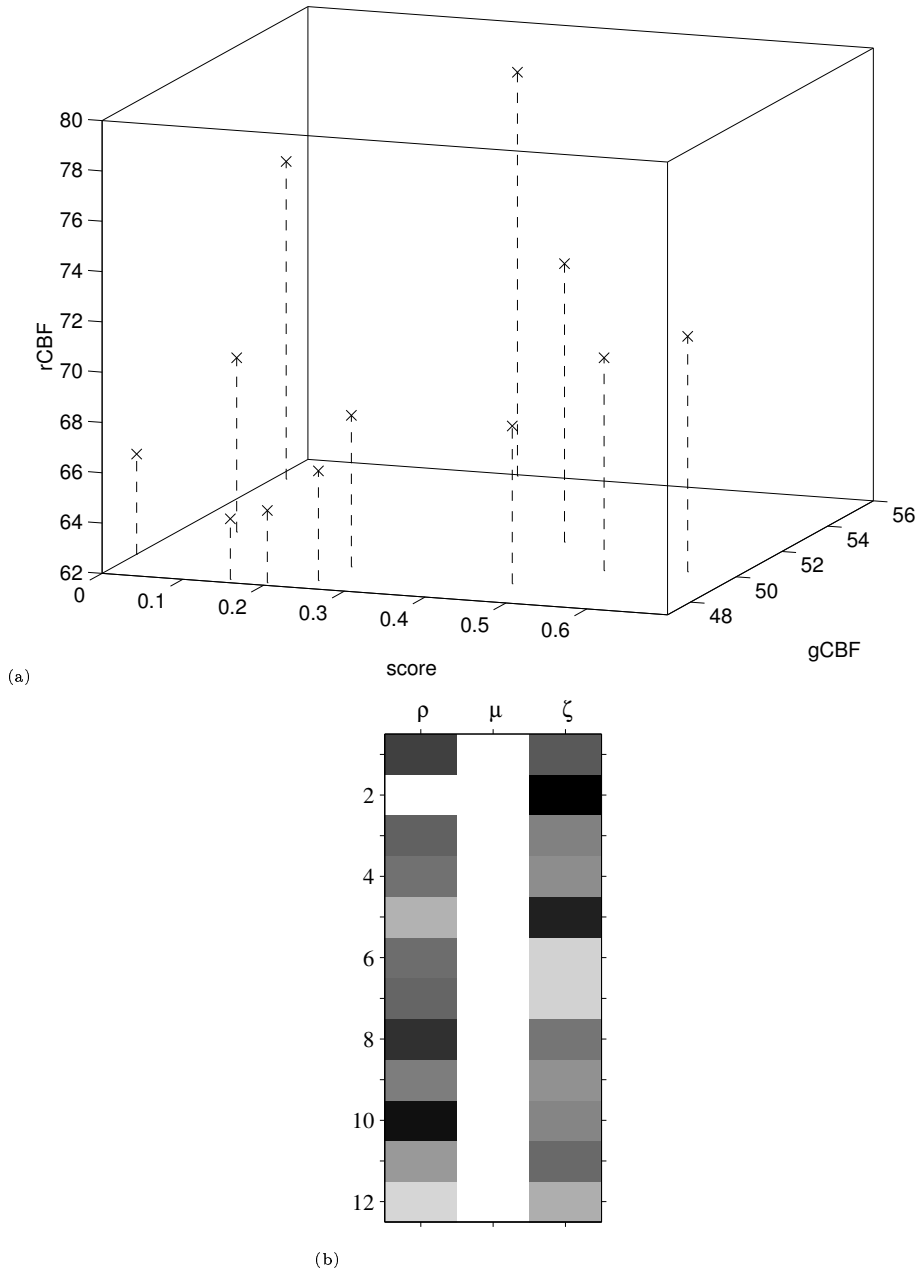


Figure 9: Single subject parametric experiment (§3.3.2): (a) Plot of rCBF vs. score and gCBF. (b) Design matrix image for Eq.19, illustrated for a 12 scan experiment. Scans are ordered in the order of acquisition.

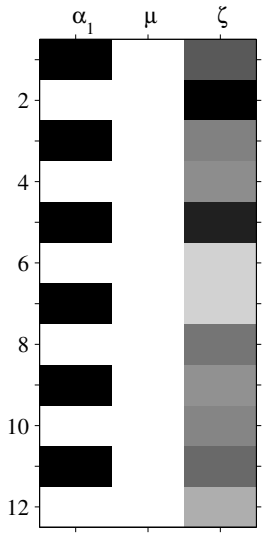


Figure 10: Example design matrix image for single subject activation study, with six scans in each of two conditions, formulated as a parametric design (§3.3.3). The twelve scans are ordered alternating between baseline and activation conditions, as might have been the order of acquisition.

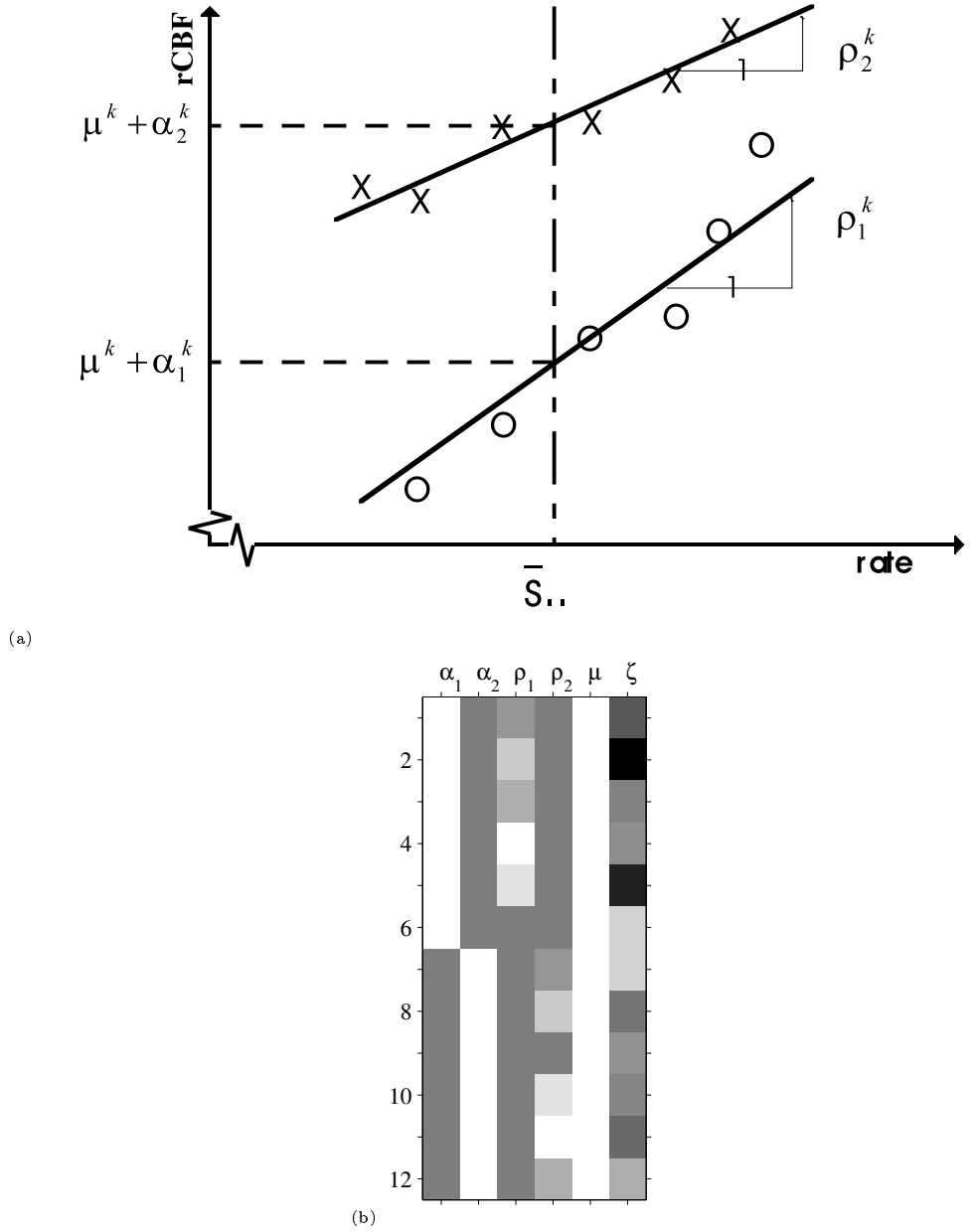


Figure 11: Single subject experiment with conditions, covariate, and condition by covariate interaction. (§3.3.5): (a) Illustrative plot of rCBF vs. rate. (b) Design matrix image for Eq. 20. Both illustrated for the two condition 12 scan experiment described in the text. The scans have been ordered by condition.

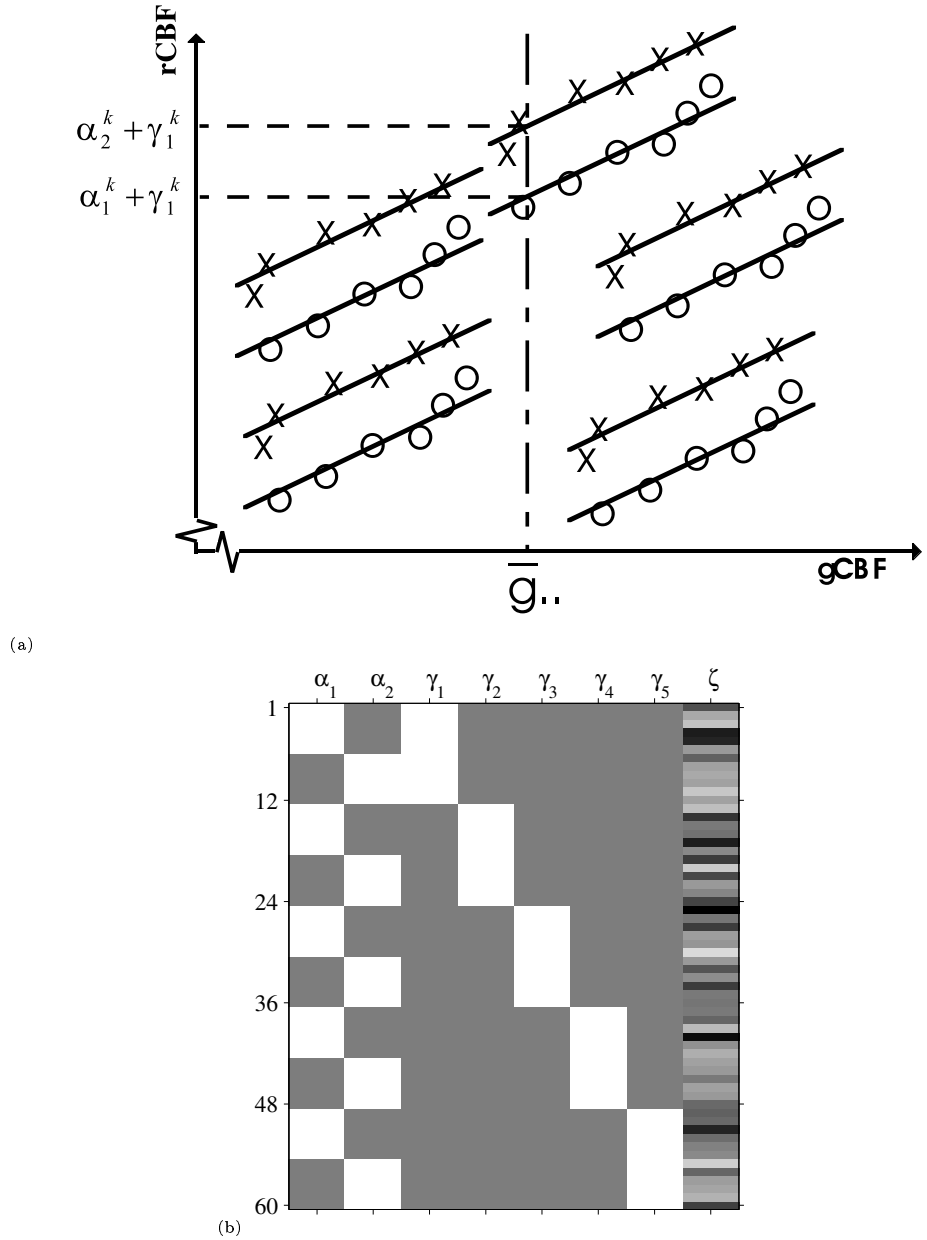


Figure 12: Multi-subject activation experiment, replication of conditions (§3.3.6), model Eq.21. Illustrations for a 5 subject study, with six replications of each of two conditions per subject: (a) Illustrative plot of rCBF vs. gCBF. (b) Design matrix image: The first two columns correspond to the condition effects, the next five to the subject effects, the last to the gCBF regression parameter. The design matrix corresponds to scans ordered by subject, and by condition within subjects.

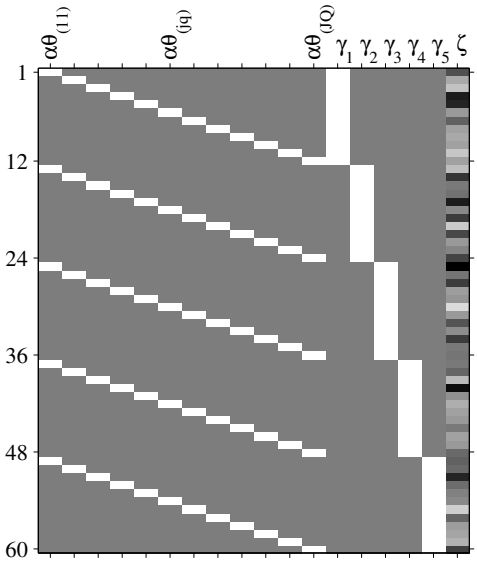


Figure 13: Multi-subject activation experiment, "classic" SPM design, where each replication of each experimental condition is considered as a separate condition (Eq.22). Illustrative design matrix image for five subjects, each having 12 scans, the scans having been ordered by subject, and by condition and replication within subject. The columns are labelled with the corresponding parameter. The first twelve columns correspond to the "condition" effects, the next five to the subject effects, the last to the gCBF regression parameter.

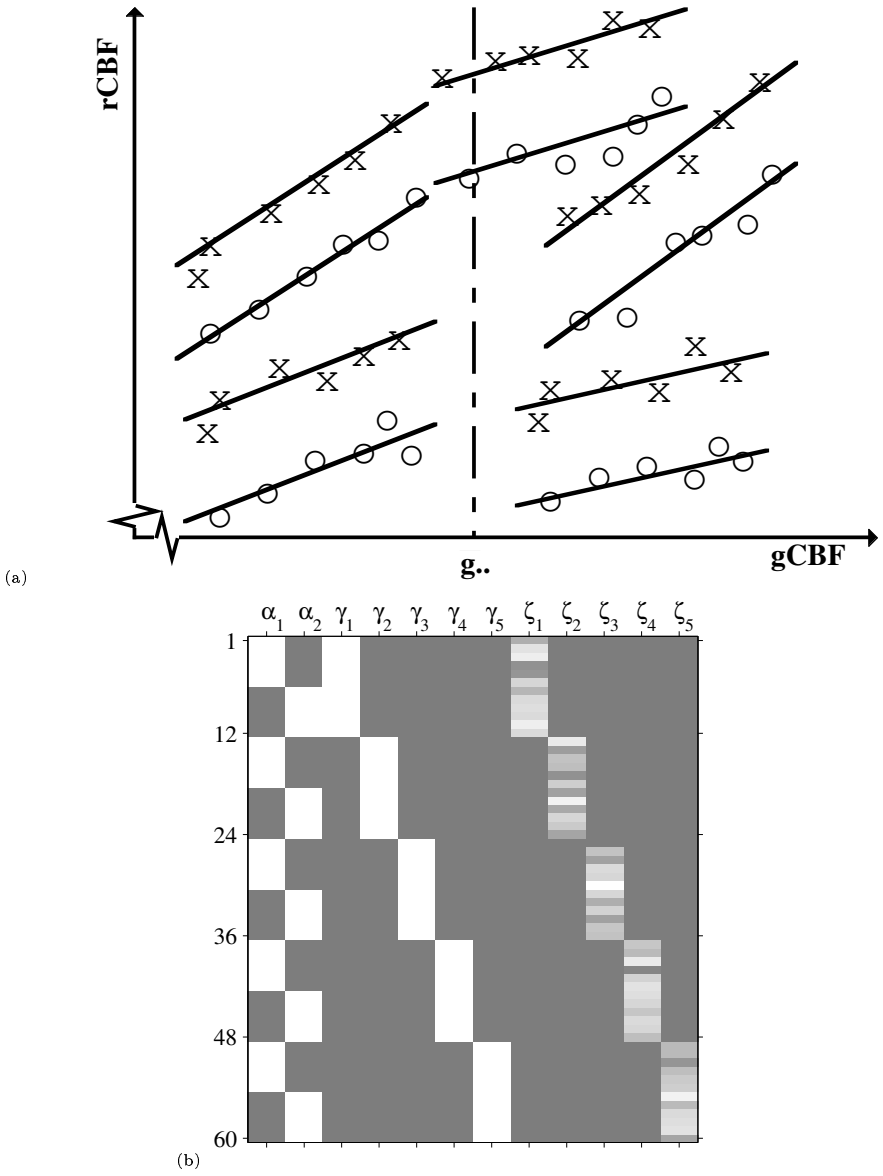


Figure 14: Multi-subject activation experiment, replication of conditions, ANCOVA by subject. Model Eq.23. Illustrations for a 5 subject study, with six replications of each of two conditions per subject: (a) Illustrative plot of rCBF vs. gCBF. (b) Design matrix image: The first two columns correspond to the condition effects, the next five to the subject effects, the last five the gCBF regression parameters for each subject. The design matrix corresponds to scans ordered by subject, and by condition within subjects.

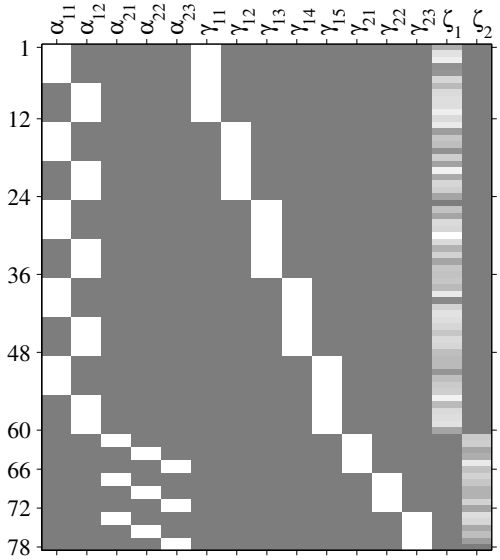


Figure 15: Design matrix image for the example multi-study activation experiment described in section 3.4.

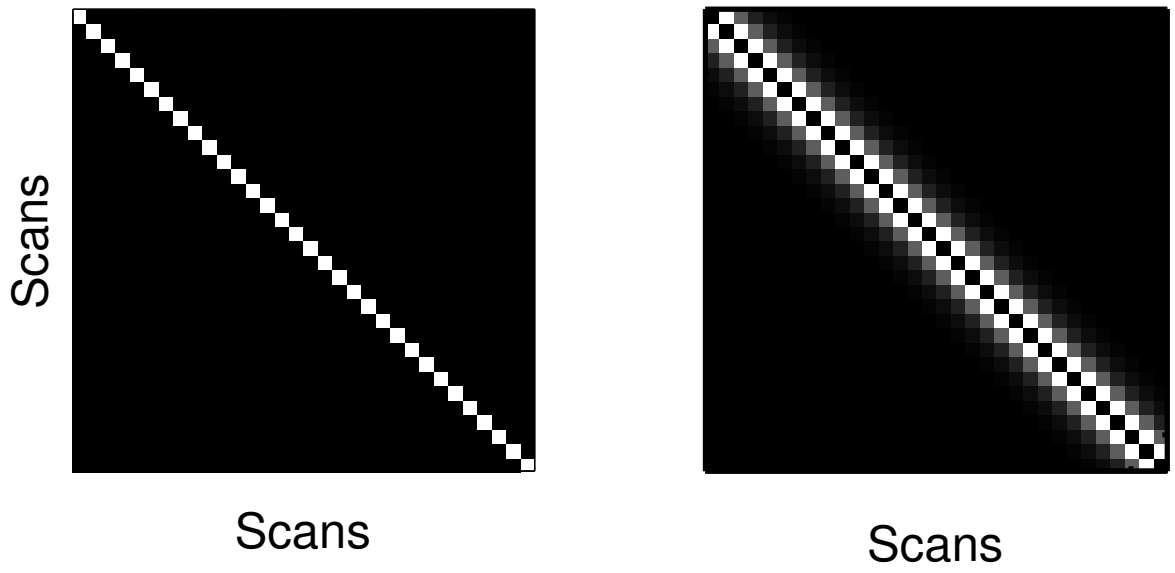


Figure 16: Graphical illustration of the two covariance constraints which are used for estimating the error correlation matrix. (Left:) Constraint Q_1 that imposes a stationary variance onto the estimate, (Right:) Constraint Q_2 that implements the $AR(1)$ model part with an autoregressive coefficient of $1/e$.

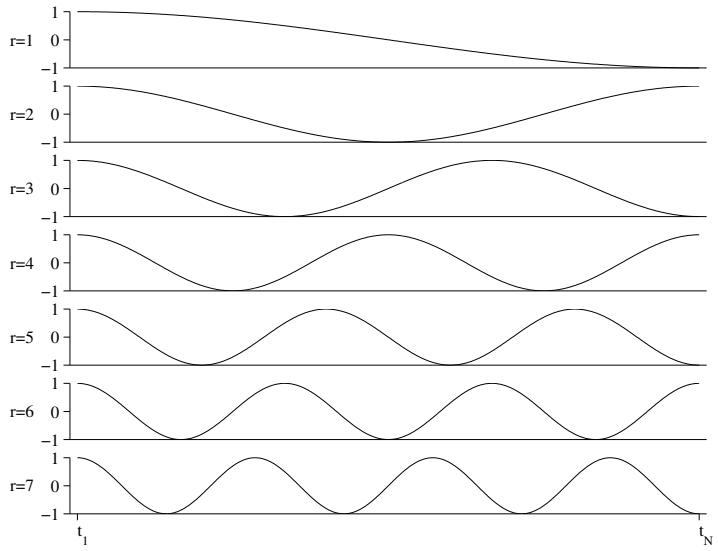


Figure 17: A discrete cosine transform set.