



# Testable Hypotheses for Unbalanced Neuroimaging Data

Martyn McFarquhar\*

Neuroscience and Psychiatry Unit, The University of Manchester, Manchester, UK

Unbalanced group-level models are common in neuroimaging. Typically, data for these models come from factorial experiments. As such, analyses typically take the form of an analysis of variance (ANOVA) within the framework of the general linear model (GLM). Although ANOVA theory is well established for the balanced case, in unbalanced designs there are multiple ways of decomposing the sums-of-squares of the data. This leads to several methods of forming test statistics when the model contains multiple factors and interactions. Although the Type I–III sums of squares have a long history of debate in the statistical literature, there has seemingly been no consideration of this aspect of the GLM in neuroimaging. In this paper we present an exposition of these different forms of hypotheses for the neuroimaging researcher, discussing their derivation as estimable functions of ANOVA models, and discussing the relative merits of each. Finally, we demonstrate how the different hypothesis tests can be implemented using contrasts in analysis software, presenting examples in SPM and FSL.

## OPEN ACCESS

### Edited by:

Alexandre Gramfort,  
Université Paris-Saclay, France

### Reviewed by:

Cyril R. Pernet,  
University of Edinburgh, UK  
Phillip M. Alday,  
University of South Australia, Australia

### \*Correspondence:

Martyn McFarquhar  
martyn.mcfarquhar@manchester.ac.uk

### Specialty section:

This article was submitted to  
Brain Imaging Methods,  
a section of the journal  
Frontiers in Neuroscience

**Received:** 26 April 2016

**Accepted:** 30 May 2016

**Published:** 17 June 2016

### Citation:

McFarquhar M (2016) Testable  
Hypotheses for Unbalanced  
Neuroimaging Data.  
Front. Neurosci. 10:270.  
doi: 10.3389/fnins.2016.00270

**Keywords:** GLM, neuroimaging, unbalanced, ANOVA, SPM, FSL, sums of squares

## 1. INTRODUCTION

The general linear model (GLM) is a ubiquitous tool in neuroimaging, forming the basis of most common analysis approaches. Users of neuroimaging software packages are well placed to harness the power of the GLM given that tools such as FSL and SPM provide great flexibility in the forms of models that can be specified and the hypotheses that can be tested. Whilst largely advantageous, one drawback is that such flexibility demands the user understand in detail the underlying theory of the GLM. A part of this understanding is the concept of estimable functions in linear models, particularly in the case of unbalanced data. This is none more relevant than for the different forms of hypotheses that can be tested in unbalanced group-level ANOVA designs containing interactions. In the statistical literature there is a long history of debate around the relative merits of the Type I–III sums of squares. Despite this, there has seemingly been no discussion or clarity on the use of these different forms of hypothesis tests in neuroimaging. In addition, there exists no clear guidance on how such hypothesis tests could be implemented in popular software packages.

In this paper we present a guide for the neuroimaging researcher on the different forms of estimable functions that are possible in the case of unbalanced ANOVA models of neuroimaging data. We also present debate and opinion on the relative merits of each, emphasizing the hypotheses tested by each type in relation to the cell means of the design. To this end we begin with a review of the theory behind balanced ANOVA models in the GLM. This paves the way for the main discussion of unbalanced ANOVA designs, turning to the derivation of the different forms of sums of squares and their implementation using contrasts in neuroimaging software.

## 2. THE BALANCED OVERPARAMETERIZED ANOVA MODEL IN THE GLM

The univariate GLM can be expressed as

$$Y = X\theta + \epsilon$$

where  $Y$  is the  $n \times 1$  vector of observed values,  $X$  is the  $n \times k$  design matrix,  $\theta$  is the  $k \times 1$  vector of parameters, and  $\epsilon$  is the  $n \times 1$  vector of errors. Solving for  $\theta$  is usually achieved using the ordinary least-squares (OLS) estimator,

$$\hat{\theta} = (X'X)^{-1}X'Y \tag{1}$$

assuming that  $X'X$  is invertible. These quantities are guaranteed the best linear unbiased estimates (BLUEs) under the conditions of the Gauss-Markov theorem (Christensen, 2011). Although this theorem does not require specification of a distribution for the data, such assumptions allow for the construction of hypothesis tests. It is therefore usual to assume that the data are drawn from a multivariate normal distribution, denoted  $Y \sim \mathcal{N}(X\theta, \sigma^2I)$ , which is more usefully expressed in terms of the errors as

$$\epsilon \sim \mathcal{N}(0, \sigma^2I)$$

Estimation of the single variance term proceeds from the model residuals

$$\hat{\sigma}^2 = \frac{1}{n-r} \hat{\epsilon}'\hat{\epsilon}$$

where  $r = \text{rank}(X)$  and

$$\hat{\epsilon} = Y - X\hat{\theta}$$

Taken together these results provide the basis for almost all the models typically used for neuroimaging data. Although this framework encompasses both subject-level and group-level neuroimaging models, here we focus solely on the group-level, specifically considering approaches such as the summary-statistic method for group-level modeling of neuroimaging data.

### 2.1. The 1-Way ANOVA

In the 1-way case there is a single factor variable with  $i$  levels. Letting  $Y_{ij}$  indicate the measurement from the  $i$ th level for the  $j$ th subject ( $i = 1 \dots m, j = 1 \dots n_i$ ) the typical overparameterized 1-way model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{2}$$

where  $\mu$  is the grand mean and  $\alpha_i$  is the effect of group  $i$ . Here, *effect* denotes the difference between the grand mean and the mean of the group. As such, the cell mean for the  $i$ th group is given by

$$\mu_i = \mu + \alpha_i$$

This model is overparameterized because there are more parameters than cell means. In other words, if  $i = 1, 2, 3$  then there are three unique model equations

$$\begin{aligned} \mu_1 &= \mu + \alpha_1 \\ \mu_2 &= \mu + \alpha_2 \\ \mu_3 &= \mu + \alpha_3 \end{aligned}$$

with four unknowns ( $\mu, \alpha_1, \alpha_2, \alpha_3$ ). The consequence of this is that there are no unique solutions for the parameter values. As exemplified by Mumford et al. (2015), this is like trying to find 2 numbers that sum to 10. The choices are infinite.

Moving from the classical formulation of the ANOVA to its implementation within the GLM involves reconceptualizing the model in Equation (2) as a more generic regression model of the form

$$Y_{ij} = \beta_0x_0 + \beta_1x_1 + \dots + \beta_ix_i + \epsilon_{ij}$$

where  $\beta_0 = \mu$  and  $\beta_i = \alpha_i$ . The  $x_i$  predictors are typically indicator variables coding a 1 for group membership and a 0 otherwise. The  $x_0$  variable associated with the intercept is a 1 for all observations. An observation from group 1 would therefore render  $x_1 = 1$ , with all other  $x_i$  set to 0. The model then simplifies to

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 + \epsilon_{ij} \\ &= \mu + \alpha_1 + \epsilon_{ij} \end{aligned}$$

returning us to the classical formulation of the ANOVA. In the GLM, the use of ANOVA models is therefore typified by a design matrix containing only indicator variables of ones and zeroes. If  $i = 1, 2, 3$  the overparameterized 1-way ANOVA model can be expressed in the GLM as

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ \vdots \\ Y_{3n_3} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \epsilon_{31} \\ \vdots \\ \epsilon_{3n_3} \end{bmatrix}$$

Here the inability to solve for the parameters can be demonstrated by considering that  $X$  is rank deficient so that  $(X'X)^{-1}$ , from Equation (1), does not have a unique solution. This can be easily seen by considering that the constant is the sum of the other columns. As such, we have a situation of perfect multicollinearity.

## 2.2. The 2-Way ANOVA

Although the 1-way ANOVA is the most basic and instructive case, it is only when considering models with interaction terms that many of the issues with unbalanced data become apparent. In the 2-way case there are two factors (denoted A and B). In a traditional crossover design the data are sampled from the intersection of the levels of the factors. The additive influences of the factors are termed the *main effects*, with the non-additive influence of the factors termed the *interactions*. These interaction effects can be expressed as  $\gamma_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)$ , where  $\mu_{ij}$  is the cell mean for level  $i$  of factor A and level  $j$  of factor B. Expressed in this form, it is clear that the interaction effect is simply the difference between the actual cell mean value and the expected cell mean value if the model were purely additive.

For subject  $k$  at the  $i$ th level of A and  $j$ th level of B the 2-way model equation is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where  $\mu$  is the grand mean,  $\alpha_i$  is the effect of the  $i$ th level of factor A,  $\beta_j$  is the effect of the  $j$ th level of factor B,  $\gamma_{ij}$  is the interaction effect, and  $\epsilon_{ijk}$  is the error. As with the 1-way case, this model is overparameterized.

## 2.3. Estimable Functions

As indicated earlier, the problem with overparameterized models is that there are no unique solutions for the parameter values. As an example, consider the 2-way model given in Equation (3).

$$\begin{bmatrix} 7 \\ 4 \\ 3 \\ 2 \\ 5 \\ 8 \\ 8 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{bmatrix} + \begin{bmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \end{bmatrix} \quad (3)$$

Four vectors of possible solutions for these parameters are given in **Table 1**. These solutions are, from left to right: solving Equation (1) using a pseudo-inverse of  $\mathbf{X}'\mathbf{X}$ , re-expressing the model using “treatment” coding, re-expressing the model using “sigma-restricted” coding and re-expressing the model using “cell-means” coding (see the Supplementary Materials). Although the existence of alternatives that render  $\mathbf{X}'\mathbf{X}$  invertible may suggest that the overparameterized formulation is overly complex and unnecessary, it is important to realize that non-overparameterized formulations (such as the coding schemes typically used for factors in explicit regression models) often lead to parameters that are more difficult to interpret. Such approaches are in fact equivalent to the overparameterized formulation with appropriate restrictions placed on the parameter values. As such we consider the overparameterized model as the most generic and didactically useful formulation of the ANOVA, despite its mathematical intractability.

**TABLE 1 | Four possible vectors of solutions for the overparameterized 2-way ANOVA model.**

Parameter	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
$\mu$	2.556	8.5	5.75	0
$\alpha_1$	0.111	-6	-1.75	0
$\alpha_2$	2.444	0	1.75	0
$\beta_1$	1.444	-2	0.25	0
$\beta_2$	1.111	0	-0.25	0
$\gamma_1$	1.389	5	1.25	5.5
$\gamma_2$	-2.778	0	-1.25	2.5
$\gamma_3$	0.056	0	-1.25	6.5
$\gamma_4$	2.389	0	1.25	8.5

Returning to the values in **Table 1**, it is notable that although they are all quite different they are all valid solutions as they all lead to the same fitted values

$$\mathbf{X}\hat{\theta}_1 = \mathbf{X}\hat{\theta}_2 = \mathbf{X}\hat{\theta}_3 = \mathbf{X}\hat{\theta}_4 = \begin{bmatrix} 5.5 \\ 5.5 \\ 2.5 \\ 2.5 \\ 6.5 \\ 6.5 \\ 8.5 \\ 8.5 \end{bmatrix}$$

Because there are multiple solutions that lead to the same estimated values, one may wonder what the worth is of any of the individual estimates given in **Table 1**? In truth, these values tell us nothing because they are dependent on the solution chosen. There are, however, certain linear combinations of the parameters that provide the same result irrespective of the solution. These linear combinations are known as *estimable functions*<sup>1</sup>.

An example of an estimable function for the 2-way model is,

$$\mathbf{L} = [0 \ 1 \ -1 \ 0 \ 0 \ 0.5 \ 0.5 \ -0.5 \ -0.5] \quad (4)$$

as multiplying this vector by any one of the vectors of estimates in **Table 1** produces the same result.

$$\mathbf{L}\hat{\theta}_1 = \mathbf{L}\hat{\theta}_2 = \mathbf{L}\hat{\theta}_3 = \mathbf{L}\hat{\theta}_4 = -3.5$$

An example of a non-estimable function would be,

$$\mathbf{L}^* = [0 \ 1 \ -1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$$

<sup>1</sup>There is an element of subtlety here. The different parameter values given in **Table 1** are not necessarily “meaningless,” rather their meaning changes depending on the solution chosen. As such, linear combinations that are not estimable can be thought of as posing a question that depends on the coding used in the design matrix. Their status as non-estimable simply refers to the fact that they cannot be universally applied to answer the same question. As such, these combinations are legitimate, so long as the parameter restrictions and coding are understood such that the question being posed is clear.

as multiplying this vector by the different solutions in **Table 1** produces different results.

$$\begin{aligned} \mathbf{L}^* \hat{\boldsymbol{\theta}}_1 &= -2.3 \\ \mathbf{L}^* \hat{\boldsymbol{\theta}}_2 &= -6 \\ \mathbf{L}^* \hat{\boldsymbol{\theta}}_3 &= -3.5 \\ \mathbf{L}^* \hat{\boldsymbol{\theta}}_4 &= 0 \end{aligned}$$

As such, the result depends on the method of solving for  $\boldsymbol{\theta}$ . This is problematic, as all methods of finding solutions for the parameters rely on some form of constraint. As argued by Nelder (1994), such constraints should not be considered an intrinsic part of the model as our conclusions should not depend on the constraint. In the case of the GLM, only estimable functions can guarantee this.

More generally, any linear combination of the parameters coded in  $\mathbf{L}$  is estimable if

$$\mathbf{L} = \mathbf{T}\mathbf{X}$$

for some matrix  $\mathbf{T}$  (McCulloch et al., 2008; Christensen, 2011). In words, this result indicates that any linear combination of the rows of  $\mathbf{X}$  produces an estimable function due to the fact that the rows of  $\mathbf{X}$  are estimable functions themselves. This is not surprising given that they dictate the predicted values of the model via  $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$ . However, stating this explicitly leads to a particularly useful result, as any  $\mathbf{L}$  is guaranteed estimable if it is constructed using a linear combination of the rows of  $\mathbf{X}$ .

### 2.4. Some Remarks on the Different Solutions to the ANOVA Model

Throughout our examples, the scheme used to solve for parameters in overparameterized models is the pseudo-inverse of  $\mathbf{X}'\mathbf{X}$ . However, it should be recognized that alternative approaches exist. Indeed, the three other solutions given in **Table 1** are derived from re-specifying  $\mathbf{X}$  such that  $\mathbf{X}'\mathbf{X}$  is invertible. For neuroimaging, knowledge of these alternative approaches is particularly useful. In SPM it is usual to specify ANOVA designs using an  $\mathbf{X}$  of full column rank, with overparameterized designs only occurring through the use of the Flexible Factorial module. In FSL, group-level models using FEAT do not allow a rank-deficient  $\mathbf{X}$ . It is therefore necessary for the user to directly specify the structure of  $\mathbf{X}$ , such that  $\mathbf{X}'\mathbf{X}$  can be inverted. Three common alternative approaches are discussed in the Supplementary Materials.

The use of different coding schemes provides a direct relationship between the presented ANOVA formulation and modern approaches that use an explicit regression formulation (such as linear and generalized linear mixed effects models). In these approaches, the coding of factors is done using a non-overparameterized scheme, and inference on the individual parameter estimates is often performed. A number of choices exist for the coding used, and each one has an impact on how the parameters, and the subsequent tests on the parameters, are interpreted. Here, a direct link with estimable functions can be made, as the tests on individual parameter estimates in a

regression formulation is akin to the use of a single row of the identity matrix  $\mathbf{I}_k$  as an  $\mathbf{L}$ . For example, in a  $2 \times 2$  treatment coded model, the contrast  $\mathbf{L} = [0\ 1\ 0\ 0]$  cannot be formed from an estimable function of the overparameterized model. It is, however, equivalent to testing the single parameter coding factor  $A$  in an explicit regression using the same coding scheme. As the contrast is not estimable, care must be taken in interpreting the result of any test using it. Within an explicit regression model that contains a  $2 \times 2$  factorial design, such a test on the single parameter for factor  $A$  does not actually produce the equivalent ANOVA main effect test, instead producing something more akin to a *simple* effect. As will be demonstrated later, one possible test to get the ANOVA main effect from a treatment coded design is  $\mathbf{L} = [0\ 1\ 0\ 0.5]$ , which involves a linear combination of parameters, rather than a single test on an estimated value. Though this is clear from the ANOVA perspective, it is not always clear in explicit regression formulations that the coding used directly impacts the questions that are asked by the tests on the individual parameter estimates.

### 2.5. Hypothesis Testing in ANOVA Models

Although solving for parameters is one important aspect of ANOVA models, issues with unbalanced data are most readily seen in terms of performing hypothesis tests. Traditionally, the ANOVA hypothesis testing scheme is presented as a partitioning of the total sums of squares of the data into independent chunks. For a typical 2-way model, this partitioning can be expressed algebraically as

$$\begin{aligned} \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2 &= \sum_i \sum_j \sum_k (\bar{Y}_{i..} - \bar{Y}_{...})^2 \\ &+ \sum_i \sum_j \sum_k (\bar{Y}_{.j.} - \bar{Y}_{...})^2 \\ &+ \sum_i \sum_j \sum_k (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} \\ &+ \bar{Y}_{...})^2 + \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 \end{aligned} \quad (5)$$

where  $\bar{Y}$  denotes a mean, and the dot notation indicates subscripts averaged over. The partitions on the right hand side of this expression are therefore those associated with factor  $A$ , factor  $B$ , the  $A \times B$  interaction, and the error respectively (Searle, 1987). Dividing these quantities by their associated degrees of freedom produces *mean squares*, which are used to construct the ANOVA  $F$  tests.

The GLM hypothesis testing scheme used in neuroimaging software revolves around a slightly more general approach. Here, a sum-of-squares  $Q$  is constructed using,

$$Q = (\mathbf{L}\hat{\boldsymbol{\theta}})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\hat{\boldsymbol{\theta}})$$

for a suitable hypothesis coded as an estimable function in  $\mathbf{L}$ . No matter the coding used for  $\mathbf{X}$ , the sums of squares in Equation (5) can all be constructed using suitable linear combinations of the estimated model parameters in  $\boldsymbol{\theta}$ . The only exception is the error sums of squares, which is constructed from the residuals. An  $F$  test can then be performed using

$$F = \frac{Q}{r\hat{\sigma}^2}$$

where  $r$  is the degrees of freedom for the hypothesis (the rank of  $\mathbf{L}$ ), and  $\hat{\sigma}^2$  is the estimated residual variance of the model (the *mean square error*).

Within this more general testing framework, the hypothesis in question is coded in  $\mathbf{L}$ . As such, understanding the question being put to the data involves interrogation the structure of  $\mathbf{L}$ . As we will demonstrate, it is the form of  $\mathbf{L}$  that dictates the different approaches used for unbalanced data. When using an overparameterized approach, however, it is often most useful to express the hypothesis in  $\mathbf{L}$  in relation to the cell means, rather than in relation to the parameters. Taking an example of the  $\mathbf{L}$  given in Equation (4), the hypothesis test given by  $\mathbf{L}\boldsymbol{\theta}$  can be written as

$$\mathcal{H}_0 : \alpha_1 - \alpha_2 + \frac{1}{2}(\gamma_{11} + \gamma_{12}) - \frac{1}{2}(\gamma_{21} + \gamma_{22}) = 0 \quad (6)$$

Replacing the interaction terms with their expanded form, the expression in Equation (6) can be simplified to

$$\mathcal{H}_0 : \frac{1}{2}(\mu_{11} + \mu_{12}) - \frac{1}{2}(\mu_{21} + \mu_{22}) = 0 \quad (7)$$

This hypothesis therefore equates to the average of the cell means for the first level of factor A, minus the average of the cell means for the second level of factor A. Although somewhat trivial for this example, this process will prove invaluable for understanding more complex estimable functions in unbalanced models later.

## 2.6. An Alternative Perspective on the ANOVA – Model Comparisons and $R()$ Notation

Before moving on to unbalanced designs, it is worth noting that there are in fact two equally useful conceptualizations of hypothesis tests in ANOVA models. Beyond understanding hypothesis testing in terms of the construction of a sum of squares  $Q$ , an alternative approach is via the concept of model comparison. Here the sums of squares for a hypothesis is seen as the difference in the sum of squared residuals for two competing models. In this approach, the sums of squares for a particular model term is seen as the reduction in error gained by the inclusion of that term in the model. Such an understanding then has an intuitive appeal, as the sums of squares can be seen as quantifying the degree of improvement in the model fit when additional terms are added.

A useful notation for indicating reductions in sums of squares due to model comparisons is the  $R()$  notation described by Searle (1987). Here a slightly simplified version of the  $R()$  notation is used in the interests of clarity. As an example, the reduction in the sum of squared errors found when fitting a model containing  $\mu$  and  $\alpha_i$  compared with a model containing only  $\mu$  can be expressed as

$$R(\alpha_i, \dots, \alpha_k | \mu) = R(\boldsymbol{\alpha} | \mu)$$

Terms to the right of  $|$  are consistent in both models, and terms to the left of  $|$  are only contained in one of the models. Alternatively, this can be read as the effect of  $\boldsymbol{\alpha}$  after correcting for  $\mu$ . When

considering these tests in relation to the reduction in the sum of squared errors between two competing models, the  $R()$  notation can be similarly interpreted as

$$R(\boldsymbol{\alpha} | \mu) = \text{SSE}(\mu) - \text{SSE}(\mu, \boldsymbol{\alpha})$$

where  $\text{SSE}()$  denotes the sum of squared errors for a model containing the terms in brackets. An  $F$ -test can then be understood as taking the form

$$F = \frac{\text{SSE}(\mu) - \text{SSE}(\mu, \boldsymbol{\alpha})}{r\hat{\sigma}^2} = \frac{R(\boldsymbol{\alpha} | \mu)}{r\hat{\sigma}^2}$$

## 3. THE UNBALANCED OVERPARAMETERIZED ANOVA MODEL

The theory behind ANOVA models is well described and understood for balanced data. Unfortunately, greater complexity is found when applying ANOVA models to data where the number of observations per cell differ. For these so-called unbalanced designs, much of the information provided from the balanced case remains relevant. The difference in the unbalanced case is that a clear decomposition of the total sums of squares into the constituent effects of the model is no longer possible. This is due to a loss of orthogonality between the ANOVA effects. This can be demonstrated by considering that the decomposition of sums of squares given in Equation (5) is only true in the balanced case (Searle, 1987). When the data are not balanced, the decompositions on the right hand side will not sum to the total on the left hand side. This indicates that the partitions do not constitute independent elements as the effects now overlap.

An example unbalanced dataset is given in Equation (8). A demonstration of the decomposition of the sums of squares for this dataset, and the preceding balanced dataset, is given in **Table 2**. These sums of squares were obtained using the decomposition in Equation (5). As it turns out, these sums of squares are those associated with tests for each effect as if it were the only effect in the model. For balanced data this is not problematic as the tests are orthogonal. For unbalanced data the lack of orthogonality means it is no longer sensible to consider an effect in isolation given that its sums of squares are no longer independent of the other terms. The only exception to this is the interaction term which, as will be demonstrated, never changes. It is therefore important to not blindly decompose the total sum of squares as would be done in the balanced case. Rather, the focus must be on sensible hypotheses, deriving the appropriate sum of squares and the corresponding estimable functions from there. Within the statistical literature there are 3 generally accepted approaches to deriving estimable functions in unbalanced ANOVA models, known as the Type I–III sums of squares. We note in passing that there is also a Type IV used for data with empty cells, however, given that no neuroimaging software accommodates empty cells we will not discuss them any further.



**TABLE 2 | Example of the sums of squares derived from the traditional decomposition of the total into the constituent effects for the balanced and unbalanced data.**

Sums of squares	Balanced	Unbalanced
Total	47.5	35.43
A	24.5	15.43
B	0.5	4.76
A × B	12.5	8.6
Error	10	9.5
A + B + (A × B) + Error	47.5	38.29

Only in the balanced case does the total also equal the sum of the decomposed elements.

$$\begin{bmatrix} 7 \\ 4 \\ \cdot \\ 3 \\ \cdot \\ 2 \\ \cdot \\ 5 \\ 8 \\ \cdot \\ 8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{bmatrix} + \begin{bmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \end{bmatrix} \quad (8)$$

### 3.1. Type I Sums of Squares

Type I sums of squares, also known as sequential sums of squares, are those associated with the testing of each effect of the model in the order they are specified. In other words, this approach provides tests where each effect is only adjusted for those that precede it in the model equation. As such, the ordering of the model is important. This can be most readily understood using the  $R()$  notation detailed above, so that the Type I ANOVA table for the model  $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$  is as detailed in **Table 3**.

For models with a natural ordering of terms, these tests can provide useful results given that each term added to the model is tested in relation to whether it provides greater predictive power than those terms that are already present. For regression models such tests are particularly useful, providing complimentary tests to those on the individual coefficients. For neuroimaging, these tests also provide the opportunity for a similar approach to stepwise regression modeling, without the necessity of fitting multiple models. However, it should be emphasized that better model building strategies exist, such as the use of Akaike Information Criterion (Akaike, 1998), and that the Type I tests simply provide a workable approach within the confines of the GLM testing procedures implemented in neuroimaging software.

In terms of the computation of the Type I tests, there are two approaches consistent with either the model comparison or estimable functions perspective. In relation to the interpretation of these tests, their sums of squares can be readily computed using the model comparisons detailed in **Table 3**. However, for neuroimaging applications, where both models must be fit at every voxel, this is a more demanding proposition. As such, it is also possible to specify these tests using an estimable function **L**. Though there are many ways to derive the weights of **L**, here we present a generic procedure based on an LU factorization

**TABLE 3 | The model comparisons that form the Type I sums of squares.**

Effect	Sum of squares
Constant	$R(\mu)$
A	$R(\alpha \mu)$
B	$R(\beta \mu, \alpha)$
A × B	$R(\gamma \mu, \alpha, \beta)$

of  $X'X$  using the Doolittle algorithm (Gaylor et al., 1970; Goodnight, 1979). See the Supplementary Materials for example MATLAB code implementing this approach for the purpose of calculating contrast weights. As an example, the coefficients for the unbalanced data presented above are given in **Table 4**.

Each row in **Table 4** can be taken as an **L** matrix for testing the model effects in the order they appear in **X**. For example, the second row provides an **L** for testing the Type I main effect of factor A, where the effect of the constant has been removed. Similarly, the third row provides an **L** for testing the Type I main effect of factor B, where the effect of both factor A and the constant are removed. The final row contains the **L** for testing the interaction term. Notably this has not changed from its familiar form. If any effect spans more than 2 columns of **X** (such as factor with > 2 levels) then there will be more than a single row containing weights for the effect. In these cases, **L** will be a matrix of weights consisting of the  $k - 1$  relevant rows from the Doolittle factorization (where  $k$  is the number of levels of the factor).

Although somewhat less intuitive than model comparison, the calculation of **L** allows a greater insight into the hypotheses being tested by the Type I approach. Consideration of only model comparison can lead one to conclude that such tests need only be considered on the basis of whether a comparison such as  $R(\beta|\mu, \alpha)$  is sensible in the context of the model. Although true in part, the weights used to calculate an expression such as  $R(\beta|\mu, \alpha)$  reveals one of the key disadvantages of this approach for unbalanced data, as the hypotheses are dependent on the cell frequencies. To see this, consider the **L** matrix for  $R(\alpha|\mu)$  presented in the second row of **Table 4**. Using the method detailed in Equations (6 and 7), this specification can be simplified to a more intuitive form, as shown in Equation (9).

$$\mathcal{H}_0 : \frac{1}{2}\mu_{11} + \frac{1}{2}\mu_{12} - \frac{2}{3}\mu_{21} - \frac{1}{3}\mu_{22} = 0 \quad (9)$$

Here it is clear that the weights associated with each cell mean depend on the cell frequency. In this case, each weight is derived from the number of subjects in the cell divided by the total number of subjects at that level of factor A. Intuitively this can be understood by considering that this **L** estimates a sum of squares equivalent to comparing two models that contains neither  $\beta_j$  nor  $\gamma_{ij}$ . To get such a test from a model that *does* contain these terms, it is necessary to adjust the parameter estimates. As will be shown later, the procedure to do so is unavoidably dependent on the number of observations in each cell. The classical objection to this approach is therefore that the hypothesis can no longer be considered a testable statement about population parameters, unless the frequency in the sample

**TABLE 4 | Type I contrast weights derived from Doolittle factorization of X'X in the example unbalanced dataset.**

	$\mu$	$\alpha_1$	$\alpha_2$	$\beta_1$	$\beta_2$	$\gamma_{11}$	$\gamma_{12}$	$\gamma_{21}$	$\gamma_{22}$
Constant	1	0.571	0.429	0.571	0.429	0.286	0.286	0.286	0.143
A	0	1	-1	-0.167	0.167	0.5	0.5	-0.667	-0.333
B	0	0	0	1	-1	0.6	-0.6	0.4	-0.4
A × B	0	0	0	0	0	1	-1	-1	1

is comparable to the frequency in the population (Searle, 1987). Furthermore, given its dependence on the order of the model terms, it is debatable how useful the hypothesis in Equation (9) actually is for a traditional ANOVA model. Although these insights appear damning of Type I tests, it is worth reiterating that this approach is legitimate and useful in cases of ordered regression models.

### 3.2. Type II Sums of Squares

Type II sums of squares are those associated with the testing of model terms under the assumption that higher-order effects containing those terms are zero. For example, when testing the main effect of A the A × B interaction is assumed zero. Unlike the Type I tests, however, the ordering of the model does not matter. As such the main effect of A is adjusted for B, and the main effect of B is adjusted for A. To make this clear, in a 3-way ANOVA model with effects A, B, and C the Type II main effect of A would be adjusted for B, C and B × C, but not A × B, A × C, or A × B × C as these effects contain A. This approach is shown from the model comparison perspective in Table 5.

Like the Type I sums of squares, the model comparison approach is relatively easy to implement via the model comparisons depicted in Table 5. From the estimable function perspective, Doolittle factorization of X'X can again be used, recognizing that this would need to be conducted multiple times with different Xs. In each case, the effect of interest is added *after* terms for which it should be adjusted. If the model contains continuous covariates care must be taken to place the covariates before the factor of interest in X so that the sums of squares are also adjusted for the covariates. As such, there can be a large amount of model re-ordering in order to calculate the Type II weights correctly. A more generic procedure that does not require model re-ordering is detailed in the SAS algorithms (<https://support.sas.com/documentation/>). As an example, the Type II weights for the unbalanced data given earlier are shown in Table 6.

Given the definition of the Type II tests, the only row that has changed in Table 6 compared to Table 4 is the row associated with factor A. The test coded in this row now produces a sum of squares that is adjusted for the intercept *and* factor B. As with before, the new hypothesis for factor A can be expressed in terms of the cells means, as shown in Equation (10).

$$\mathcal{H}_0 : \frac{3}{5}\mu_{11} + \frac{2}{5}\mu_{12} - \frac{3}{5}\mu_{21} - \frac{2}{5}\mu_{22} = 0 \quad (10)$$

In this form, one would be forgiven for thinking that the hypothesis does not look intuitive, let alone useful. However, as

**TABLE 5 | The model comparisons that form the Type II sums of squares.**

Effect	Sum of squares
Constant	$R(\mu)$
A	$R(\alpha \mu, \beta)$
B	$R(\beta \mu, \alpha)$
A × B	$R(\gamma \mu, \alpha, \beta)$

we will argue, framing a hypothesis test that reflects a model without interaction terms within a model that does contain these terms can be misleading. The hypothesis in Equation (10) is actually identical to

$$\mathcal{H}_0 : \alpha_1 - \alpha_2 = 0 \quad (11)$$

in the model  $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$ . From the model comparison perspective, we are therefore comparing,

$$\begin{aligned} Y_{ijk} &= \mu + \beta_j + \epsilon_{ijk} \\ Y_{ijk} &= \mu + \alpha_i + \beta_j + \epsilon_{ijk} \end{aligned} \quad (12)$$

as indicated in Table 5. Despite the seemingly unintuitive nature of this test when viewed from a cell means perspective, it is clear from both Equations (11 and 12) that such a test is readily understandable. As such, the slightly more involved arithmetic necessitated by the cell means model should not distract from the question that the test is posing. We shall return to this issue later.

#### 3.2.1. The Principle of Marginality

Though we have now covered the interpretation and calculation of the Type II tests, one may still wonder why such tests are of any interest at all. This is particularly as, given the weights in Table 6, similar caveats with the Type I approach exist as the tests appear dependent on the cell frequencies. There is, however, an appealing logic to these tests, known as the *principle of marginality* (Nelder, 1977, 1994; Nelder and Lane, 1995). In brief, this principle is based on the idea that interpreting a main effect in the presence of an interaction is uninformative. Indeed, some authors have gone as far as suggesting that any attempt to do so is “...an exercise in fatuity” (Kempthorne, 1952). Based on this logic, assuming that the interaction effect is zero is the only way to produce a meaningful test of a main effect. As such, main effects should only be adjusted for each other, as well as any interaction that does not contain the main effect in question. If it turns out that there is a significant interaction effect, the main effects should be ignored as a matter of course. As such, the fact that

**TABLE 6 | Type II contrast weights derived from Doolittle factorization of  $X'X$  in the example unbalanced dataset.**

	$\mu$	$\alpha_1$	$\alpha_2$	$\beta_1$	$\beta_2$	$\gamma_{11}$	$\gamma_{12}$	$\gamma_{21}$	$\gamma_{22}$
Constant	1	0.571	0.429	0.571	0.429	0.286	0.286	0.286	0.143
A	0	1	-1	0	0	0.6	0.4	-0.6	-0.4
B	0	0	0	1	-1	0.6	-0.6	0.4	-0.4
A $\times$ B	0	0	0	0	0	1	-1	-1	1

In order to correctly calculate the weights for the main effect of factor A, the order of factor A and factor B must be swapped before performing the factorization of  $X'X$ .

the model comparison in Equation (12) assumes no interaction is moot. As will be explained shortly, the Type III main effects tests can be interpreted in terms of adjustments of main effects for all other model terms. From the model comparison perspective, this involves testing models containing interactions *without* all their corresponding main effects. As such, the Type III tests implicitly entertain models that are arguably unrealistic. It is for this reason that the Type II tests are sometimes regarded more favorably than the Type III (Nelder, 1977, 1994; Nelder and Lane, 1995; Langsrud, 2003; Fox, 2008; Fox and Weisberg, 2011).

### 3.3. Type III Sums of Squares

Type III sums of squares are those associated with model comparisons in sigma-restricted models where only single terms are removed at a time. These are therefore tests where each effect is adjusted for *all* other model terms, thus violating the principle of marginality. Because of this, the model comparison perspective on the Type III tests is where much of the controversy surrounding the approach is found. However, the logic of the tests from the perspective of hypotheses about cell means makes the situation much clearer because these tests are the only ones that do not depend on the cell frequencies. These tests are also equivalent to Yates' weighted square-of-means approach (Yates, 1934), are often used by default in statistical software packages (e.g., SAS, SPSS, STATA), and correspond to the contrasts that researchers are taught to use within the GLM in neuroimaging.

The Type III tests are shown in Table 7 from the model comparison perspective. Here the notation from Searle (1987) is adopted to indicate two key points about these tests. Firstly, as stated above, the Type III tests based on model comparisons can only be considered for models using sigma-restricted coding (see the Supplementary Materials), a technicality discussed in Searle (1987). It is this seemingly arbitrary aspect of the Type III tests that has been used as an argument against their use (Venables, 1998), particularly given our earlier discussion of the fact that the model constraints should not influence the answers gained from the data. Here, both the Type I and Type II tests have an advantage, as their values do not depend on the constraint chosen for the model. Secondly, as indicated earlier, from the model comparison perspective the Type III tests of main effects involve comparing models with and without main effects, but maintaining all interaction terms. In comparison to the Type II approach, these tests do not treat the interaction effect as zero, rather they average over them. In the balanced case these two approaches are equivalent due to the orthogonality between the main effect and interaction tests. For unbalanced data this is not

**TABLE 7 | The model comparisons that form the Type III sums of squares.**

Effect	Sum of squares
Constant	$R(\mu)$
A	$R(\hat{\alpha} \mu, \hat{\beta}, \hat{\gamma})$
B	$R(\hat{\beta} \mu, \hat{\alpha}, \hat{\gamma})$
A $\times$ B	$R(\hat{\gamma} \mu, \hat{\alpha}, \hat{\beta})$

The dot notation indicates that the parameters adhere to sigma-restrictions.

so. As an example, consider that the Type III main effect of factor A involves the following model comparison,

$$Y_{ijk} = \mu + \hat{\beta}_j + \hat{\gamma}_{ij} + \epsilon_{ijk}$$

$$Y_{ijk} = \mu + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} + \epsilon_{ijk}$$

and thus implicitly entertains a model that contains an interaction term with only one of the corresponding main effects. As argued by a number of authors, it is debatable how sensible this is (Venables, 1998; Langsrud, 2003; Fox, 2008; Fox and Weisberg, 2011).

As with all the other approach discussed so far, an  $L$  matrix can also be used to develop the Type III sums of squares. For the Type III tests this is particularly advantageous, as this approach is applicable to any coding of  $X$ . Again, the  $L$  weights can be derived using Doolittle factorization of  $X'X$ , but only after  $X$  has been reduced to its unique rows. This is therefore equivalent to calculation of these effects in the balanced case. Here we see the argument for using Type III tests emerging, as the  $L$  matrix used for Type III tests does not depend on the cell frequencies. Using this approach, the Type III weights for the example unbalanced dataset are shown in Table 8.

Looking at the main effect of factor A, the Type III test provides cell means comparison identical to Equation (6). As such, the hypotheses on the cell means do not differ between the balanced and unbalanced cases. As such the Type III tests can be interpreted as testing hypotheses that are generalizable, as they do not depend on the size of the sample. From this perspective, it has been argued that these tests are therefore the most appropriate for unbalanced data (Searle, 1987). However, this is not necessarily reason enough to dismiss the fact that this approach tests hypotheses that are arguably not sensible when considered from a model comparison perspective. Indeed, Nelder (1977) insists that the forms of models that the Type III tests actually compare are of "...no practical interest." Such divergences in opinion make it clear why the difference between the Type II



**TABLE 8 | Type III contrast weights derived from Doolittle factorization of  $X'X$  in the example unbalanced dataset.**

	$\mu$	$\alpha_1$	$\alpha_2$	$\beta_1$	$\beta_2$	$\gamma_{11}$	$\gamma_{12}$	$\gamma_{21}$	$\gamma_{22}$
Constant	1	0.5	0.5	0.5	0.5	0.25	0.25	0.25	0.25
A	0	1	-1	0	0	0.5	0.5	-0.5	-0.5
B	0	0	0	1	-1	0.5	-0.5	0.5	-0.5
A $\times$ B	0	0	0	0	0	1	-1	-1	1

Here the weights are derived after reduction of  $X$  to unique rows.

**TABLE 9 | Demonstration of the differing approaches to producing marginal means in the presence of interaction effects using either equally weighted or frequency weighted cell means.**

Approach	Cell 1	Cell 2	Result
Averaged – ignoring cells	7	8,9	8
Equally weighted means (Type III)	$\frac{1}{2} \times 7$	$\frac{1}{2} \times 8.5$	7.75
Frequency weighted means (Type I-II)	$\frac{1}{3} \times 7$	$\frac{2}{3} \times 8.5$	8

and Type III sum of squares remains such a contentious topic. This is particularly true given that the perspective one takes on such tests can lead to equally valid arguments for and against their use. As such, it is important to realize that none of the tests are “wrong” *per-se*, rather they are simply asking different questions.

### 3.4. Overview and Merits of the Type I–III Tests

Now that the differences between the Type I–III tests has been covered, we turn to more general debate on their individual merits for hypothesis testing in classical ANOVA designs. Before doing so, we first present a concrete example to help elucidate the differences between how the means for the varying hypotheses are calculated in the Type I–III cases. Given that it is in the main effects tests that the various methods differ, it should be no surprise that it is the calculation of the marginal means that holds the key to understanding the different approaches.

#### 3.4.1. Calculation of Marginal Means in the Type I–III Tests

In Table 9 there are three data points from two hypothetical cells of data containing different numbers of observations. In row 1 we simply average all the data ignoring the cells, in row 2 we use equally weighted cell means, and in row 3 we use cell means weighted by the cell frequency. When the two cells are ignored, and the data are treated as coming from the same source, the results differ from when the cell means are averaged over. If instead the cell means are weighted by the cell frequency, the original row mean can be recovered. Herein lies the key conceptual differences between the Type III (equally weighted means) and the Type I–II (cell frequency weighted means) approaches. Here it is clear that the very process of correcting the cell means to recover the original row mean necessitates the use of the cell frequencies.

#### 3.4.2. The Type II vs. Type III Debate

As covered earlier, the Type I tests have limited utility beyond ordered regression models, where they should be considered as complimentary to the standard *t*-tests on the coefficient values. As such, much of the debate in the ANOVA literature relates to the choice between the Type II and Type III tests. Ultimately, much of this debate lies within ones feelings about the purpose of hypothesis testing in statistical models. As highlighted by Langsrud (2003) and Fox and Weisberg (2011), the ultimate aim of hypothesis testing should be the desire to answer specific and meaningful questions. Mathematically, this is expressed using linear combinations of the model parameters, however, inference on parameters without a foundation in meaningful questions is arguably counter to the purpose of statistical modeling. The reason that the Type II tests are often argued against is that they are considered solely in relation to parameters from models with interaction terms, rather than in consideration of the questions they ask. Indeed, the Type II tests of parameters from such models necessitate corrections in order to render their values equivalent to a model without interaction terms. This correction depends on the cell frequencies, but ultimately guarantees that the same question is being posed irrespective of the model form. As such, the question itself should therefore be the point of debate for the merit of the test, rather than the specific arithmetic details of how such a question can be posed across different models of the same data. From this perspective, the dependence on cell frequency in Equation (10) should not be considered an indictment about the worth of the hypothesis being tested, particularly as such a question is poorly framed in a model containing interaction terms. This ultimately highlights the limitations of only considering the ANOVA tests as hypotheses about the model parameters, particularly as this is arguably not the most important perspective on hypothesis testing. Such arguments weight heavily in favor of the the Type II tests over the Type III tests. In addition, the Type II tests of main effects are often more powerful than the Type III tests (Lewsey et al., 1997, 2001; Langsrud, 2003), a point of particular note for neuroimaging. As such, the Type II tests deserve serious consideration as the default approach for hypothesis testing in unbalanced ANOVA models. They are principled, sensible, and powerful tests that ask meaningful questions. Unlike the Type III tests, they do not depend on the model constraints when viewed from a model comparison perspective, and similarly, they do not implicitly entertain unrealistic models when testing main effects. When seen from the perspective of the questions posed to the data, it

is difficult to argue against the approach taken by the Type II tests.

## 4. THE UNBALANCED OVERPARAMETERIZED ANOVA MODEL IN NEUROIMAGING SOFTWARE

Now that the different approaches to dealing with unbalanced data in ANOVA models have been discussed, we turn to the practical application of these approaches in neuroimaging software. In this section we present the construction of the Type I–III tests in two of the most popular neuroimaging analysis software packages: SPM ([www.fil.ion.ucl.ac.uk/spm/](http://www.fil.ion.ucl.ac.uk/spm/)) and FSL ([fsl.fmrib.ox.ac.uk/fsl/](http://fsl.fmrib.ox.ac.uk/fsl/)). Though we have limited examples to just these two packages, any analysis software that implements the mass-univariate GLM could be used. In all examples the model is a  $2 \times 2$  between-subject ANOVA with cell counts as given in **Table 10**.

### 4.1. SPM

#### 4.1.1. Using the Flexible Factorial Module

Using an overparameterized design in SPM necessitates the use of the Flexible Factorial module. The SPM design matrix for the overparameterized ANOVA model is shown in **Figure 1**. Here a constant column has been added by specifying a covariate vector of ones. It is notable here that SPM has indicated that none of the parameters from this model will be unique, by providing a gray box per column below the design matrix. This is in keeping with the point made earlier about differing solutions in overparameterized designs, essentially highlighting that the values of the individual  $\beta_{*}.nii$  images cannot be meaningfully interpreted. Once specified, this matrix is available in the SPM.mat file as  $X = \text{SPM.xX.X}$ , for the unfiltered design matrix. This is a convenience, as it is not necessary to specify the design matrix manually for use in a Doolittle factorization. It is, however, necessary to move the intercept column from last to first (e.g.,  $X = [X(:, \text{size}(X, 2)) X(:, 1:\text{size}(X, 2)-1)]$ ).

Using the Doolittle factorization of  $X'X$  provides scaled versions of the Type I weights for **L**. Using the MATLAB function given in the Supplementary Materials, this could be specified very simply as  $W = \text{doolittleWeights}(X)$ . An example of specifying these effects in SPM is given in **Figure 2A**. Note that each of the contrasts were specified on a single line, but have been wrapped within the input box.

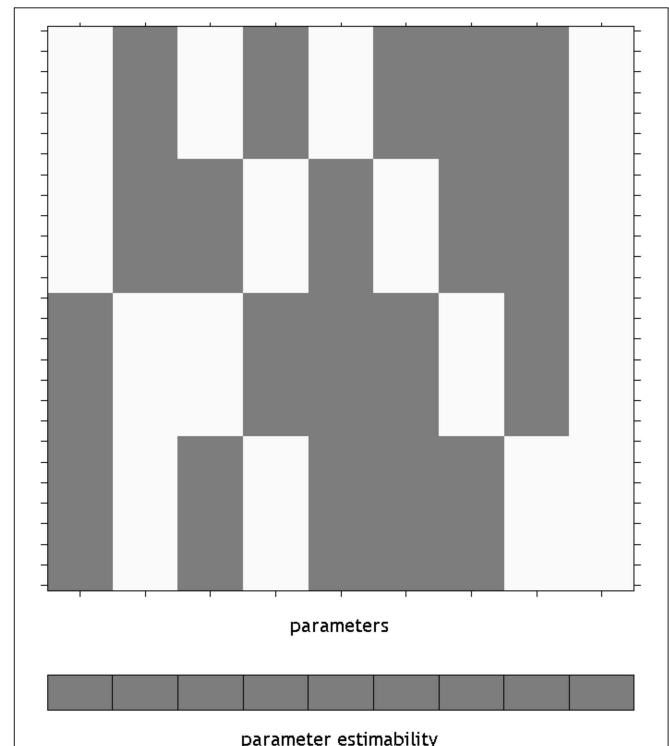
**TABLE 10 | Group numbers for the example unbalanced neuroimaging data.**

		A	A	Total B
		1	2	
B	1	13	14	27
B	2	13	15	28
Total A		26	29	55

Derivation of the Type II tests follows much the same procedure. In this case the Doolittle factorization is performed twice, swapping the ordering of factor A and factor B in **X**. In this instance, the weights provided by the second decomposition are simply re-arranged so that they align with the original ordering of **X**. These weights are shown specified in the SPM contrast manager in **Figure 2B**. As expected, the **L** matrices for the interaction and main effect of B are identical to those in the Type I case.

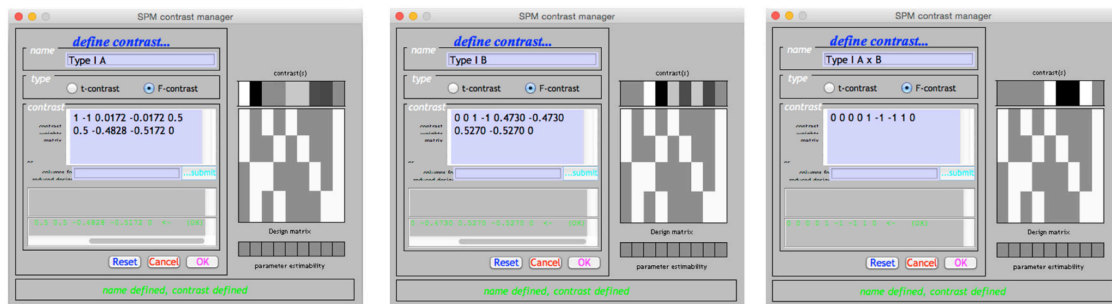
As an additional point, it may be advantageous to make use of the *contrast masking* facility in SPM when exploring Type II main effects. Such an approach allows one to effectively “censor” voxels where a high-order effect is significant. As such, investigations of Type II main effects can be made only in voxels where the higher-order effects are suitably null. If there are multiple higher-order effects, a mask image of all significant higher-order effects could be used.

The Type III tests are those that will seem the most familiar to neuroimaging researchers. Their derivation from a Doolittle factorization is performed after reducing **X** to its unique rows. Again, using the MATLAB function given in the Supplementary Materials, this can be specified on a single line as  $W = \text{doolittleWeights}(\text{unique}(X, \text{'rows'}))$ . Because the Type III tests do not depend on cell frequency, they will

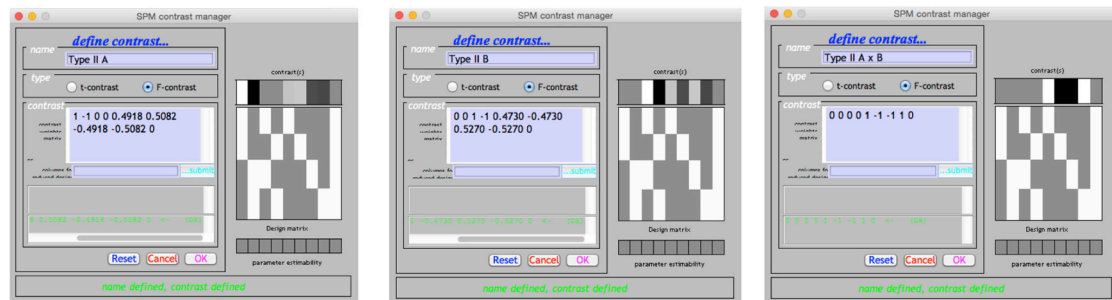


**FIGURE 1 | The design matrix visualization for an overparameterized ANOVA model in SPM.** The use of an overparameterized ANOVA design is possible using the SPM Flexible Factorial module. Here, the blocks below the design matrix are gray, indicating that unique values for the parameters do not exist. The constant is placed on the end of the design as it is not included by default, rather it is added by specifying a covariate of ones.

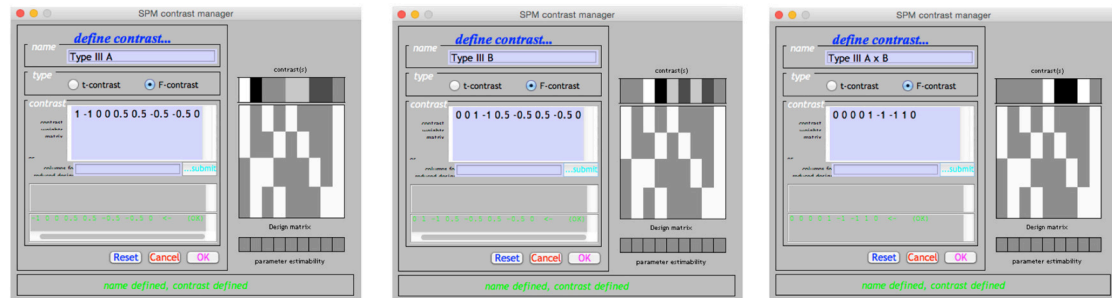
## A Type I



## B Type II



## C Type III



**FIGURE 2 | The Type I–III hypotheses specified in the SPM contrast manager. (A)** Type I ANOVA weights **(B)** Type II ANOVA weights **(C)** Type III ANOVA weights. The green text below the input box for each example indicates that each contrast is an estimable function, as it has passed the estimability test used by SPM.

be the same no matter the number of observations. These tests can therefore be constructed more generically using the number of cells rather than the frequencies within the cells. Generally speaking, this is much easier to do, and likely contributes to why the Type III contrast weights are generally taught for use in neuroimaging software. These tests are shown in the SPM contrast manager in **Figure 2C**.

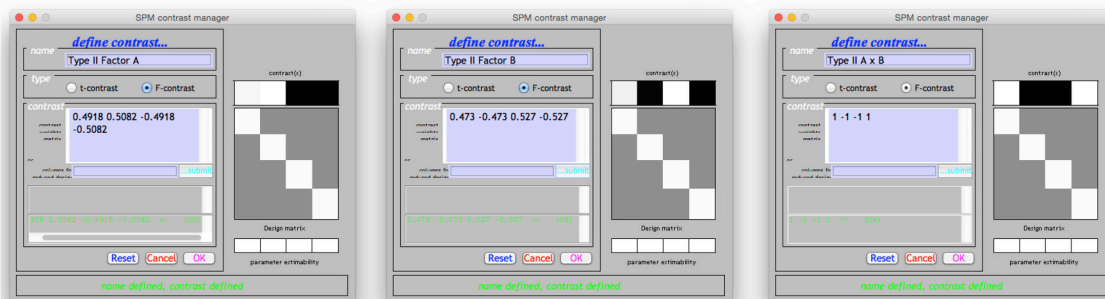
### 4.1.2. Using Other SPM Modules

Although the above example made use of overparameterized designs in the SPM Flexible Factorial module, it is perfectly possible to derive the tests using any of the other models available in SPM. For example, if it is desirable to

instead use the SPM Full Factorial module to specify a cell means design, then the contrasts already derived can easily be adjusted. The method used to convert an overparameterized contrast to a cell means contrast has already been demonstrated in Equations (6 and 7). However, for the current tests, it is more straightforward to simply take the weights associated with the interaction term and apply them to the cell means model. This is demonstrated for the Type II tests in **Figure 3**.

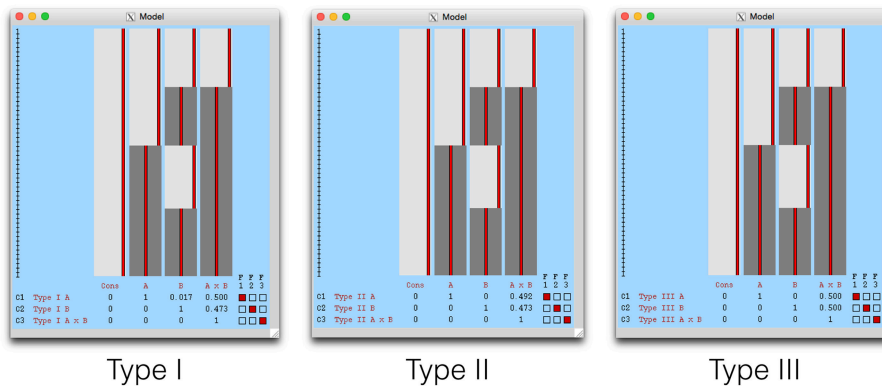
### 4.2. FSL

As mentioned earlier, FSL does not allow overparameterized designs to be used. This means that the overparameterized ANOVA model cannot be specified in FEAT directly. However,

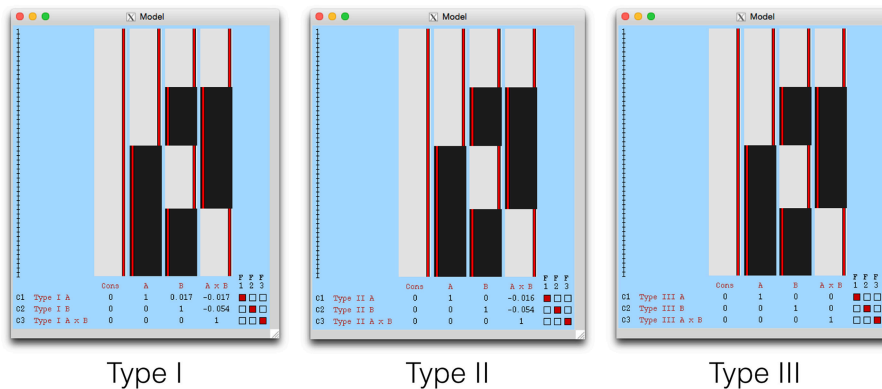


**FIGURE 3 | Specification of the Type II tests for a cell means ANOVA in SPM.** The weights given in the contrast manager were derived from the overparameterized models given earlier. Although a more principled approach can be used to discern these weights, as shown in Equations (6 and 7), in this instance the weights were simply taken from the interaction terms in **Figure 2B**.

**A Treatment coding**



**B Sigma-restricted coding**



**FIGURE 4 | The Type I–III hypotheses specified in FSL FEAT.** The models shown here include **(A)** treatment coding and **(B)** sigma restricted coding. These are both legitimate alternatives to the cell means model, used to render  $X'X$  invertible. The cell means model is not shown as it is identical to SPM, with weights calculated as indicated earlier.

it is possible to start with an overparameterized model, calculate the estimable functions available to test the hypotheses of interest, and then convert these functions into ones useable in a model where  $X'X$  is invertible. Alternatively, both treatment

and sigma-restricted design matrices can be submitted to Doolittle factorization to get the appropriate weights. The design matrix from an FSL FEAT group-level model is stored as a `design.mat` file in the corresponding `*.gfeat` directory.



This is a text file written in the FSL VEST format, and could be read into MATLAB using e.g., the `palm_vestread()` function from the Permutation Analysis of Linear Models (PALM) toolbox (<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/PALM>). The function in the Supplementary Materials could then be used to compute the Doolittle factorization of  $X'X$ , for example `W = doolittleWeights(palm_vestread('./example.gfeat/design.mat'))`. The appropriate weights from `W` can then be entered back into FEAT.

The Type I-III tests are shown in the design visualization from FEAT in **Figure 4**. Given that the Type III tests from a sigma-restricted model are so straightforward, it is perhaps not surprising that this approach is recommended on the FSL GLM wiki ([fsl.fmrib.ox.ac.uk/fsl/fslwiki/GLM](http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/GLM)). As the cell means model is identical to SPM, this has been omitted. In addition, *contrast masking* is also available in FEAT, and as such our comments on the Type II tests in SPM remain relevant for FSL also.

## 5. SUMMARY

In this paper we have given an overview of the use of estimable functions in the GLM, paying particular attention to the different forms of hypothesis tests available for unbalanced ANOVA models. Though this issue has a long history in statistics, it has seemingly not been considered in the neuroimaging literature. Despite the fact that Type III tests have been settled on exclusively in popular software packages, there may be merit in the Type I and II tests for certain designs and approaches. Indeed, the Type II tests in particular provide greater power for investigating main effects, and arguably provide a more sensible hypothesis testing

framework via the principle of marginality. Understanding the difference between these tests, and their derivation, allows for greater flexibility in the use of the GLM in neuroimaging, particularly in unbalanced designs. We have also touched on the use of Type I tests in regressions models, suggesting that these forms of hypotheses could be useful in ordered designs where the test of each coefficient is only adjusted for those that precede it. Using the overparameterized ANOVA model as a base, we have shown how all these tests can be derived from a generic framework that can be adjusted to suit any form of coding used. Such an approach allows immediate application irrespective of the software package used, but also provides a key focus on hypothesis testing as the single most important aspect of using the GLM with neuroimaging data.

## AUTHOR CONTRIBUTIONS

MM wrote the article, conceived the idea for the article, and produced the code used to create the figures and supplementary material for the article.

## FUNDING

This work was supported by a MRC Centenary Early Career Award (MR/J500410/1).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fnins.2016.00270>

## REFERENCES

- Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, eds E. Parzen, K. Tanabe, and G. Kitagawa (New York, NY: Springer), 199–213. doi: 10.1007/978-1-4612-1694-0\_15
- Christensen, R. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer Texts in Statistics, 4th Edn. New York, NY: Springer. doi: 10.1007/978-1-4419-9816-3
- Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*, 2nd Edn. Los Angeles, CA: Sage.
- Fox, J., and Weisberg, S. (2011). *An R Companion to Applied Regression*, 2nd Edn. Thousand Oaks, CA: SAGE Publications.
- Gaylor, D., Lucas, H., and Anderson, R. (1970). Calculation of expected mean squares by the abbreviated doolittle and square root methods. *Biometrics* 26, 641–655. doi: 10.2307/2528712
- Goodnight, J. H. (1979). A tutorial on the sweep operator. *Am. Stat.* 33, 149–158.
- Kempthorne, O. (1952). *The Design and Analysis of Experiments*. New York, NY: Wiley.
- Langsrud, Ø. (2003). Anova for unbalanced data: use type II instead of type III sums of squares. *Stat. Comput.* 13, 163–167. doi: 10.1023/A:1023260610025
- Lewsey, J. D., Gardiner, W. P., and Gettinby, G. (1997). A study of simple unbalanced factorial designs that use type II and type III sums of squares. *Commun. Stat. Simulat. Comput.* 26, 1315–1328. doi: 10.1080/03610919708813442
- Lewsey, J. D., Gardiner, W. P., and Gettinby, G. (2001). A study of type II and type III power for testing hypotheses from unbalanced factorial designs. *Commun. Stat. Simulat. Comput.* 30, 597–609. doi: 10.1081/SAC-100105081
- McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd Edn. Hoboken, NJ: Wiley.
- Mumford, J. A., Poline, J.-B., and Poldrack, R. A. (2015). Orthogonalization of regressors in fMRI models. *PLoS ONE* 10:e0126255. doi: 10.1371/journal.pone.0126255
- Nelder, J. (1977). A reformulation of linear models. *J. R. Stat. Soc. Ser. A* 140, 48–77. doi: 10.2307/2344517
- Nelder, J. A. (1994). The statistics of linear models: back to basics. *Stat. Comput.* 4, 221–234. doi: 10.1007/BF00156745
- Nelder, J. A., and Lane, P. W. (1995). The computer analysis of factorial experiments: in memoriam—frank yates. *Am. Stat.* 49, 382–385.
- Searle, S. R. (1987). *Linear Models for Unbalanced Data*. New York, NY: Wiley.
- Venables, W. (1998). "Exegeses on linear models," in *S-Plus User's Conferences* (Washington, DC).
- Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *J. Am. Stat. Assoc.* 29, 51–66. doi: 10.1080/01621459.1934.10502686

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 McFarquhar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.