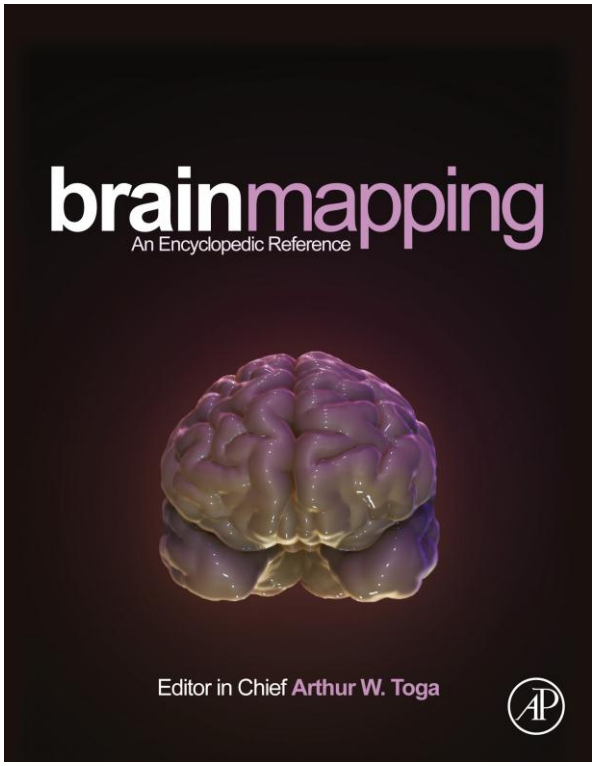

Henson R.N. (2015) Analysis of Variance (ANOVA). In: Arthur W. Toga, editor. *Brain Mapping: An Encyclopedic Reference*, vol. 1, pp. 477-481. Academic Press: Elsevier.

Analysis of Variance (ANOVA)

RN Henson, MRC Cognition and Brain Sciences Unit, Cambridge, UK

Abbreviations

ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
df	Degrees of freedom
GLM	General linear model
MANOVA	Multivariate analysis of variance
OLS	Ordinary least squares
ReML	Restricted maximum likelihood

Introduction

Analysis of variance (ANOVA) is simply an example of the *general linear model* (GLM) that is commonly used for factorial designs. A factorial design is one in which the experimental conditions can be categorized according to one or more *factors*, each with two or more *levels* (Winer, Brown, & Michels, 1991). For example, an experiment might present two types of visual stimuli (e.g., faces and houses), each at three different levels of eccentricity. This would correspond to a 2×3 ANOVA, in which the six conditions correspond to unique combinations of each level of the 'stimulus-type' and 'eccentricity' factors.

In univariate ANOVA, each condition furnishes one measurement (e.g., BOLD response at a given voxel) for each of multiple replications (e.g., subjects). When each level of one or more factors is measured on the same thing, for example, the same subject contributes data to each level, the ANOVA is called a *repeated-measures* ANOVA. Such factors are also called *within-subject* factors, as distinct from *between-subject* factors, for which the levels can be considered independent (ANOVAs that contain both within-subject and between-subject factors are sometimes called *mixed* ANOVAs). A 1×2 repeated-measures ANOVA corresponds to a paired (or dependent samples) *t*-test; 1×2 between-subject ANOVA corresponds to an unpaired (or independent samples) *t*-test. Repeated-measures ANOVAs include additional covariates in the GLM to capture variance across measurements (e.g., between-subject variance), normally reducing the residual error and hence improving statistics for the effects of interest. This is in fact one type of *analysis of covariance*, or ANCOVA, in which the data are adjusted for covariates of no interest (another example covariate might be, e.g., the order in which conditions were measured). Analysis of multiple measurements per condition is also possible (*multivariate ANOVA*, or MANOVA), though this can be formally reduced to a univariate ANOVA with additional factors and proper treatment of the error term (see Kiebel, Glaser, & Friston, 2003), so is not discussed further here. Finally, ANOVA (and the GLM) can be considered special cases of linear mixed-effects (LMEs) models (Chen, Saad, Britton, Pine, & Cox, 2013), though many of the issues to do with error covariance modeling are generalized later in the text.

What characterizes ANOVA is the focus on a specific set of statistical tests across the conditions (*contrasts*), designed to test the *main effects* of each factor and *interactions* between factors. So in the 2×3 ANOVA example earlier in the text, there would be three such *treatment effects*: (1) the main effect of stimulus type, (2) the main effect of eccentricity, and (3) the interaction between stimulus type and eccentricity. A significant main effect of a factor means that the differences between the levels of that factor are significant (relative to the variability across replications) when averaging over the levels of all other factors. So the main effect of stimulus type would correspond to the difference between faces and houses, regardless of eccentricity. A significant interaction between two factors means that the effect of one factor depends on the levels of the other factor. So an interaction between stimulus type and eccentricity would mean that the difference between faces and houses depends on their eccentricity (or equivalently, that the effect of eccentricity depends on whether the stimulus is a face or house). So, for example, there might be a large difference between faces and houses at low eccentricity but less of a difference (or even a difference in the opposite direction) at high eccentricity (a result that can be followed up by more focused contrasts within each level of a factor, sometimes called *simple effects*). It is arguably difficult to interpret the main effect of a factor if it interacts with other factors (or more generally, to interpret an mth-order interaction if one of the factors is also involved in a significant $(m+1)$-th-order interaction). In such cases, a common strategy is to repeat separate ANOVAs on each level of one of the factors in that interaction, after averaging over the levels of factors not involved in that interaction. More generally, for a K-way ANOVA with K factors, there are K main effects, $K(K-1)/2$ two-way or *second-order* interactions, $K(K-1)(K-2)/6$ three-way or *third-order* interactions, etc., and one *highest-order* K-way interaction (see Section 'Generalization to *K*-Way ANOVAs').

Example 1×4 Between-Subject ANOVA

Consider an ANOVA with one factor A of four levels, each level measured on an independent group of ten subjects. This can be expressed formally as the following GLM:

$$y_{s,a} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \mathbf{x}_3\beta_3 + \mathbf{x}_4\beta_4 + \varepsilon_{s,a}$$

where $y_{s,a}$ refers to the data from the sth subject in the group who received the ath level of factor A, concatenated into a column vector (with $n = 1 \ldots 40$ values in this case); $\mathbf{x}_a$ is a *regressor*, here an indicator variable whose values of 0 or 1 code whether the nth measurement in y comes from the ath level of A; β_a is the parameter for the ath level of A (whose

values are estimated from fitting the model and here correspond to the mean across subjects for that level); and $\varepsilon_{s,a}$ is the residual error for the sth subject and ath level (again derived from fitting the model). Sometimes, a fifth regressor would be added to capture the *grand mean* across all the data, but this is not necessary for the *F*-contrasts considered later in the text. Fitting the model entails estimating the values of the four parameters such that the sum of the squares of the residuals is minimized (the so-called ordinary least squares, or OLS, estimates).

The same equation can be written in matrix format as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim N(0, \mathbf{C}_e) \qquad \mathbf{C}_e = \sigma^2\mathbf{I} \qquad [1]$$

where $\mathbf{X}$ is the *design matrix* in which the four regressors have been combined (shown graphically in Figure 1(a)). The second expression in eqn [1] denotes that the residuals are assumed to be drawn from a zero-mean, multivariate normal (Gaussian) distribution with covariance $\mathbf{C}_e$. In fact, ANOVA normally assumes that the residuals are drawn independently from the same distribution (often termed *independent and identically distributed* (IID), or *white*, residuals), which is what is captured by the third expression in eqn [1], where the error covariance matrix is an N-by-N identity matrix ($\mathbf{I}$) scaled by a single variance term σ^2. One example where this assumption might not hold is when the conditions differ in the variance across replications within each condition (*homogeneity of variance* or *heteroscedasticity*). For example, patients within one group (level) may be more variable than controls in another group (level). Another example arises in repeated-measures ANOVAs, where the conditions may differ in the pairwise covariance between them. Both of these require some form of correction (see Section 'Nonsphericity').

Significance and *F*-Contrasts

Having fit the model, the main effect of factor A corresponds to the classical statistical test of the null hypothesis that four means of each level are identical, that is, that $\beta_1 = \beta_2 = \beta_3 = \beta_4$. This is tested by constructing an *F-statistic*, which can be expressed in several ways. One way is the mean sum of squares of the treatment effects (β_{1-4} here) divided by the mean sum of squares of the residuals:

$$F(\mathrm{df}_A, \mathrm{df}_\varepsilon) = \frac{SS_A/\mathrm{df}_A}{SS_\varepsilon/\mathrm{df}_\varepsilon}$$

where SS are the sums of squares and df are the *degrees of freedom*. In the present example, with $L = 4$ levels of the factor, $\mathrm{df}_A = L - 1 = 3$ (since there are three ways that four things can differ) and $\mathrm{df}_\varepsilon = N - L = 36$ (i.e., the df in the data minus the df in the model). Given those df, the probability of obtaining that value of F or larger under the null hypothesis, p, can be calculated from the standard F-distribution and declared significant if p is less than a certain value, for example, $p < 0.05$. Note that a significant main effect could result from any pattern of difference across the four means (e.g., there is no requirement of an ordinal relationship across the levels). Note also that F-tests are two-tailed, but there is nothing to prohibit a one-tailed (directional) test of a main effect or interaction if there is only one numerator df in the contrast.

The F-statistic can also be specified by a *contrast matrix*, $\mathbf{c}$, or the so-called F-contrast. For the main effect of A in the present example, $\mathbf{c}$ can be expressed in a number of ways (as long as $\mathrm{rank}(\mathbf{c}) = 3$ to reflect df_A), such as three pairwise differences between the four levels:

$$\mathbf{c} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

The F-statistic can then be expressed in terms of the parameter estimates ($\hat{\boldsymbol{\beta}}$), full design matrix ($\mathbf{X}$), data $\mathbf{y}$, and contrast $\mathbf{c}$ (see Appendix A of Henson & Penny, 2003). Once the use of such F-contrasts is understood, more complicated ANOVAs can be considered, as next.

Example 2 × 2 Within-Subject ANOVA

Consider an ANOVA with two factors A and B, each with two levels, and the resulting four conditions this time measured on

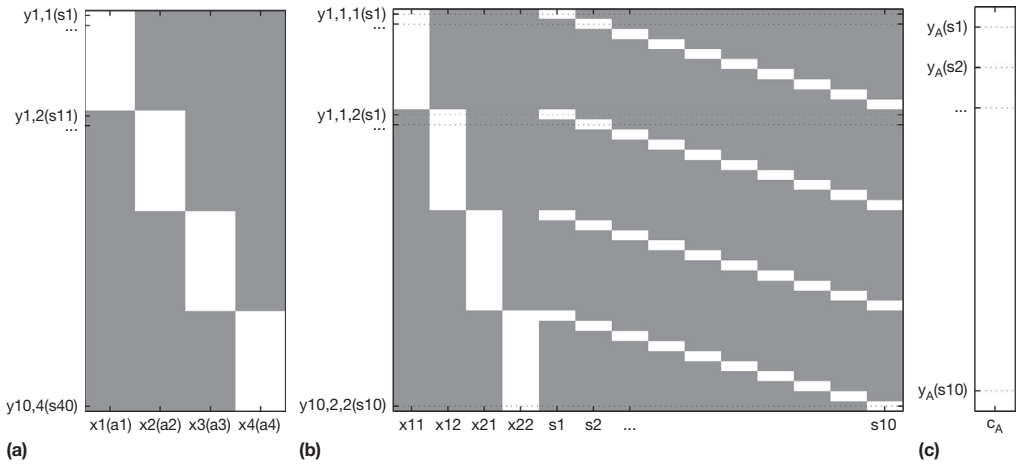

Figure 1 GLM design matrices for example ANOVAs, where white = 1, gray = 0: (a) A 1 × 4 between-subject ANOVA, (b) a 2 × 2 within-subject ANOVA with pooled error, and (c) one of the main effects (or interaction effect) in (b), after premultiplying the data by the contrast for that effect, corresponding to a partitioned error.

each of ten subjects. One possible GLM for this repeated-measures ANOVA (which uses a single *pooled error*; as explained later) is

$$y_{s,a,b} = \mathbf{x}_{11}\beta_{11} + \mathbf{x}_{12}\beta_{12} + \mathbf{x}_{21}\beta_{21} + \mathbf{x}_{22}\beta_{22} + \mathbf{X}_s\boldsymbol{\beta}_s + \varepsilon_{s,a,b}$$

where $\mathbf{x}_{12}$, for example, indicates whether or not the nth measurement comes from the first level of A and second level of B. The corresponding design matrix is shown in Figure 1(b) (note the order of conditions, in which factor A *rotates* slowest across columns). The matrix $\mathbf{X}_s$, which has one column per subject, captures the mean across conditions for each subject. These covariates of no interest capture a source of variance (between-subject variance) that would otherwise be likely to inflate the residual error (at the price of extra df in the model, i.e., now $df_\varepsilon = N - \text{rank}(X) = 40 - 13 = 27$ for estimating the residuals).

Within this model, we want to test three F-contrasts, where

$$\mathbf{c}_A = [\,1 \quad 1 \quad -1 \quad -1\,] \qquad [2]$$

corresponds to the main effect of A (ignoring an extra ten zeros for the subject effects); the main effect of B is

$$\mathbf{c}_B = [\,1 \quad -1 \quad 1 \quad -1\,]$$

and the interaction is

$$\mathbf{c}_{AB} = [\,1 \quad -1 \quad -1 \quad 1\,]$$

(see Section 'Generalization to *K*-Way ANOVAs').

Nonsphericity

As mentioned in the preceding text, a second consequence of ANOVAs with repeated measures is that the IID assumption in eqn [1] is unlikely to hold, in that the residual for one measurement on one subject is likely to be similar to the residuals for other measurements on that subject, that is, the residuals for repeated measurements are likely to be positively correlated across subjects. This *inhomogeneity of covariance* is another case of *nonsphericity* (in fact, IID is a special case of a spherical $\mathbf{C}_e$; for more precise definition of nonsphericity, see Appendix C of Henson & Penny, 2003). Nonsphericity implies that the *effective* df in the data is less than the number of observations. Standard approximations exist to estimate the degree of nonsphericity and associated loss of df, by estimating a proportion $1/df <= \varepsilon <= 1$ by which the numerator and denominator df of the F-ratio are scaled ($\varepsilon = 1$ corresponding to spherical residuals). Common approximations include the *Greenhouse–Geisser* or *Huynh–Feldt* corrections (Howell, 2002). One problem with these post hoc df corrections however is that they tend to be conservative, since there are rarely sufficient data to estimate ε efficiently (Kiebel et al., 2003).

Pooled and Partitioned Errors

One way of reducing the nonsphericity problem is to *partition* the GLM error term into separate components, with one error term per ANOVA effect. So for the 2×2 ANOVA example earlier in the text, $df_\varepsilon = 27$ for the single pooled error becomes $df_\varepsilon = 9$ for each of the three ANOVA effects. This partitioning can be achieved by premultiplying the data by the F-contrast for each ANOVA effect, for example, for the main effect of A:

$$\mathbf{y}_A = \left(\mathbf{c}_A^{(1)} \otimes \mathbf{I}_n\right)\mathbf{y}$$

where $\otimes$ is the Kronecker product, $\mathbf{I}_n$ is an n-by-n identity matrix for the n subjects per level of A ($n = 10$ here), and $\mathbf{c}_A^{(1)}$ is as defined in eqn [2]. The new data, $\mathbf{y}_A$, can then be fit by the simple design matrix shown in Figure 1(c), with the corresponding F-contrast $\mathbf{c}_A^{(2)} = 1$. The advantage of this procedure is that the error covariance of the new GLM can be estimated as a single scalar, that is, $\mathbf{C}_e = \sigma^2\mathbf{I}_n$, and hence, there are no concerns about nonsphericity, at least for effects like this with one numerator df (i.e., $\text{rank}(\mathbf{c}_A^{(1)}) = 1$). For ANOVA effects with more than one df (e.g., repeated-measures factors with more than two levels), the partitioned error covariance matrices can still be nonspherical (so some form of correction is still necessary), but the degree of nonsphericity is nonetheless normally reduced, owing to the smaller dimensionality of $\mathbf{C}_e$. However, partitioning the error results in less sensitive tests compared with a single pooled error, providing the nonsphericity of that error can be estimated accurately, as discussed next.

Error Covariance Modeling

Another solution to the nonsphericity problem is to employ a more complex model of $\mathbf{C}_e$:

$$\mathbf{C}_e = \sum\nolimits_i \lambda_i \mathbf{Q}_i$$

where $\mathbf{Q}_i$ are called (*co*)*variance components* and λ_i are their relative weightings, or *hyperparameters*. So for the GLM in Figure 1(b), where there is a single *pooled* error, the structure of the error can be modeled by ten covariance components: four modeling the variance for each condition and six modeling the covariance between each pair of conditions (Figure 2). The hyperparameters ($\boldsymbol{\lambda}$) can be estimated simultaneously with the parameters ($\boldsymbol{\beta}$) using an iterative algorithm, such as ReML (Friston et al., 2002). Once the hyperparameters are estimated, the

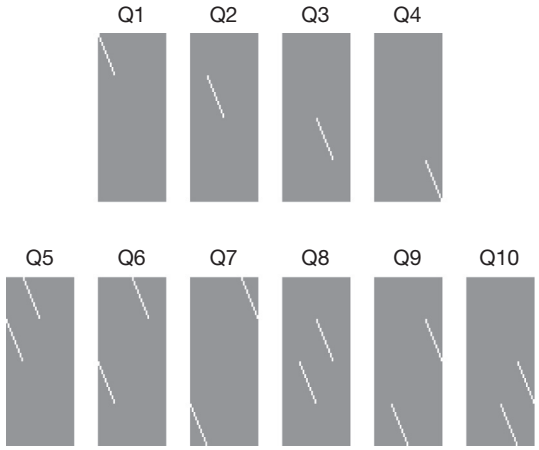

Figure 2 Covariance components for modeling error nonsphericity in a repeated-measures ANOVA with four conditions and ten subjects (data assumed to rotate fastest with subject): Q1–4 model inhomogeneity of variance, while Q5–10 model inhomogeneity of covariance.

estimated error covariance can be constructed, inverted, and multiplied by the data (and model) to *prewhiten* both. This is the most statistically efficient solution, recovering the full df in the data.

However, as also the case for the post hoc df corrections considered in the preceding text, the efficiency with which the hyperparameters can be estimated depends on the precision with which the true error covariance can be estimated from the sample residuals, that is, depends on the amount of data (Kiebel et al., 2003). For neuroimaging data, one approach is to combine data across a large number of voxels, in order to increase the precision of the sample estimate of $\mathbf{C}_e$. These voxels can be selected once as all those showing some evidence of an omnibus experimental effect (Friston et al., 2002), or iteratively in the context of a local neighborhood in a spatially regularized (Bayesian) framework (Woolrich, Jenkinson, Brady, & Smith, 2004). Friston et al. (2002), for example, assumed that the error correlation matrix is identical across those voxels, differing only in a single scaling factor, σ^2, which can be estimated at a voxel-wise level when refitting the model to the prewhitened data, as in eqn [1]. If this assumption holds, then this approach provides maximal sensitivity for the ANOVA effects. (The greater df's also tend to produce smoother maps of residuals, rendering corrections for multiple comparisons across voxels like random field theory less stringent.) Figure 3(a) shows, for example, how this prewhitened, voxel-wide pooled-error approach increases sensitivity to a true effect (blue solid line), relative to partitioning the error (blue dotted line) while maintaining appropriate false-positive control when there is no true effect (overlapping green solid and dotted lines at $p = 0.05$). On the other hand, if one tries to estimate the error correlation voxel-wise rather than voxel-wide, or the true error correlation is not constant across voxels, this approach can produce an increased false-positive rate (red solid and green dotted lines in Figure 3(b) and 3(c)). In sum, this approach to combining data across voxels is more sensitive, but less robust, than partitioning the error or post hoc df corrections.

Generalization to *K*-Way ANOVAs

The examples in the preceding text can be generalized to K-way ANOVAs, with K factors each with L_k levels. Thus, for an L_1-by-L_2-by... L_K ANOVA, there are $\prod_{k=1}^{K} L_k$ conditions, $K!/(m!\,(K-m)!)$ treatment effects of the mth order (where the first-order effects are the main effects), and $2^K - 1$ treatment effects in total. (One should therefore consider correcting the p-values for the number of treatment effects tested, i.e., to allow for the multiple comparison problem in classical statistics.)

The F-contrasts for each treatment effect can be built from two types of *component* contrast matrix $\mathbf{m}_k$ and $\mathbf{d}_k$ for the kth factor:

$$\mathbf{m}_k = \mathbf{1}_{L_k} \quad \mathbf{d}_k = \mathrm{orth}(\mathrm{diff}(\mathbf{I}_{L_k}))^T$$

where $\mathbf{1}_{L_k}$ is a row vector of L_k ones, $\mathbf{I}_{L_k}$ is an L_k-by-L_k identity matrix, $\mathbf{P}^T$ is the transpose of matrix $\mathbf{P}$, $\mathrm{diff}(\mathbf{P})$ is a matrix of column differences of a matrix $\mathbf{P}$, and $\mathrm{orth}(\mathbf{P})$ is the orthonormal basis of $\mathbf{P}$. The component $\mathbf{m}_k$ can be thought of as the *common effect* of the kth factor and the component $\mathbf{d}_k$ can be thought of as the *differential effect* for the kth factor. The F-contrast for the mth-order interaction between the first f

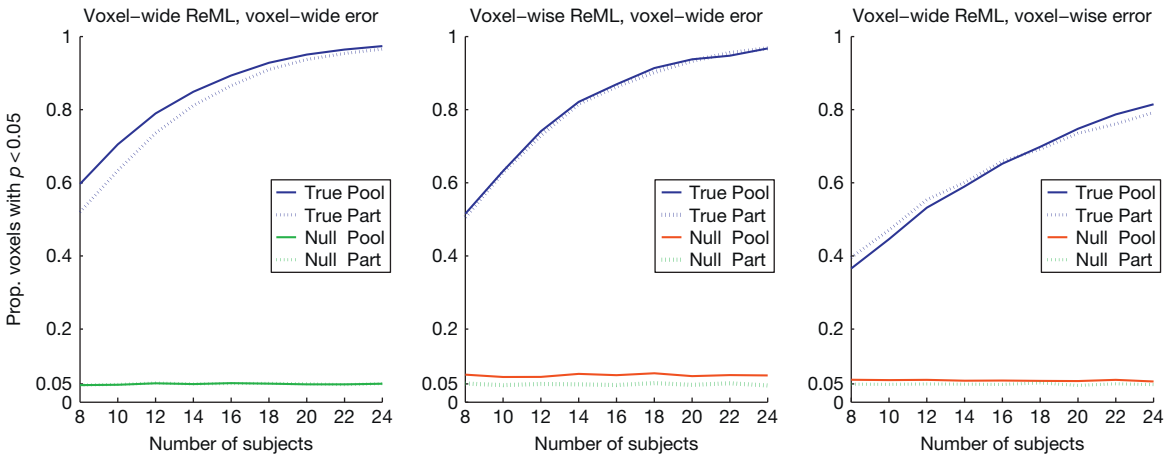

Figure 3 Sensitivity and bias for various treatments of the error in a 2 × 2 repeated-measures ANOVA, in which there is a *true* main effect of A (blue lines), but no main effect of B (red or green lines). The proportion of 10 000 voxels whose p-values exceed $p < 0.05$ are plotted against the number of subjects. In (a), the true error correlation is constant across voxels. The solid lines arise when averaging residual covariances across all voxels and estimating the nonsphericity of the single pooled-error term (*Pool*) using ReML and the ten covariance components depicted in Figure 2 (see text for details); the dotted lines reflect the same effects estimated using a partitioned error (*Part*). Note the pooled error is more sensitive to the main effect of A while maintaining the same control of false-positives (at expected chance proportion of 0.05) for the main effect of B. In (b), the error nonsphericity is estimated for each voxel separately, and the inefficiency of this estimation no longer results in a gain in sensitivity for the pooled relative to partitioned error, and there is now an increased false-positive rate (red line). In (c), the true error correlation varies across voxels but is still estimated by averaging residuals across voxels. This also results in a loss of sensitivity and (modest) increase in false-positive rates for pooled relative to partitioned error. The code for these simulations is available at http://www.mrc-cbu.cam.ac.uk/wp-content/uploads/2013/05/check_pooled_error.m.

factors (assuming that the first factor rotates slowest in the data and design matrix) is then given by

$$\mathbf{c} = \mathbf{d}_1 \otimes \mathbf{d}_2 \otimes \ldots \mathbf{d}_f \otimes \mathbf{m}_{K-f+1} \otimes \mathbf{m}_{K-f+2} \otimes \ldots \mathbf{m}_K$$

So for the 2×2 ANOVA considered previously in the text,

$$\mathbf{m}_k = [1 \quad 1] \quad \mathbf{d}_k = \left[-1/\sqrt{2} \quad 1/\sqrt{2}\right] \quad \equiv [1 \quad -1]$$

(the latter equivalence shown for simplicity, since the sign and overall scaling of an *F*-contrast do not matter). We can then construct the previous *F*-contrasts for the 2×2 example, with the main effect of factor A:

$$\mathbf{c}_A = \mathbf{d}_1 \otimes \mathbf{m}_2 = [1 \quad -1] \otimes [1 \quad 1] = [1 \quad 1 \quad -1 \quad -1]$$

the main effect of factor B:

$$\mathbf{c}_B = \mathbf{m}_1 \otimes \mathbf{d}_2 = [1 \quad 1] \otimes [1 \quad -1] = [1 \quad -1 \quad 1 \quad -1]$$

and the interaction:

$$\begin{aligned}\mathbf{c}_{AB} = \mathbf{d}_1 \otimes \mathbf{d}_2 &= [1 \quad -1] \otimes [1 \quad -1] = [1 \quad -1 \quad -1 \quad 1] \\ &\equiv [1 \quad -1 \quad 0 \quad 0] - [0 \quad 0 \quad -1 \quad 1]\end{aligned}$$

(where the final equivalence indicates how such an interaction can be thought of as a difference of differences, or difference of two simple effects). This procedure can be generalized to any ANOVA, and the resulting contrasts can be used to partition the error (for repeated measures) and/or construct an *F*-statistic and corresponding *p*-value.

Acknowledgments

This work was supported by the UK Medical Research Council (MC_US_A060_5PR10).

See also: **INTRODUCTION TO METHODS AND MODELING**: Contrasts and Inferences; The General Linear Model; Topological Inference.

References

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., & Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *NeuroImage*, *73*, 176–190.

Friston, K. J., Glaser, D. E., Henson, R. N., Kiebel, S., Phillips, C., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: Applications. *Neuroimage*, *16*, 484–512.

Henson, R. N., & Penny, W. (2003). *ANOVAs and SPM*. Technical Report, Wellcome Department of Imaging Neuroscience.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Duxbury Press.

Kiebel, S. J., Glaser, D. E., & Friston, K. J. (2003). A heuristic for the degrees of freedom of statistics based on multiple hyperparameters. *NeuroImage*, *20*, 466–478.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*. McGraw-Hill.

Woolrich, M. W., Jenkinson, M., Brady, J. M., & Smith, S. M. (2004). Fully Bayesian spatio-temporal modeling of FMRI data. *IEEE Transactions on Medical Imaging*, *23*, 213–231.

Relevant Websites

http://afni.nimh.nih.gov/sscc/gangc/ANOVA.html – ANOVA in AFNI software, and extension to LME models: http://afni.nimh.nih.gov/sscc/gangc/lme.html.

https://en.wikipedia.org/wiki/Analysis_of_variance – wikipedia, classical perspective.

http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/GLM – ANOVA in FSL.

http://www.mrc-cbu.cam.ac.uk/wp-content/uploads/2013/05/check_pooled_error.m – Matlab code used to calculate efficiency in examples here.

http://www.mrc-cbu.cam.ac.uk/personal/rik.henson/personal/HensonPenny_ANOVA_03.pdf – GLM perspective and implementation in SPM.

http://nmr.mgh.harvard.edu/harvardagingbrain/People/AaronSchultz/GLM_Flex.html – software toolbox for partitioned error models in SPM.

http://surfer.nmr.mgh.harvard.edu/fswiki/LinearMixedEffectsModels – LME in Freesurfer.