Maths for Brain Imaging: Lecture 5

W.D. Penny
Wellcome Department of Imaging Neuroscience,
University College, London WC1N 3BG.

November 1, 2006

1 GLMs with multiple covariance components

Given the usual GLM

$$y = X\beta + e \tag{1}$$

where β are the true but unknown parameters, and $Cov(e) = V(\lambda)$ is the error covariance parameterised by unknown parameters λ ie. 'hyperparameters'. These more general models are useful for eg. (a) fMRI analysis allowing for correlated errors and (b) analysis of data from a group of subjects.

We now address two questions

- If we know V how do we estimate β ?
- How do we estimate V ?

The two answers are (i) WLS and (ii) ReML.

2 Weighted Least Squares

If we know V then we can estimate β by maximising the likelihood

$$L = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)^T V^{-1}(y - X\beta)$$

We can derive the normal equations in the usual way, by setting the appropriate derivatives to zero.

$$\frac{dL}{d\beta} = X^T V^{-1} y - X^T V^{-1} X\beta \tag{2}$$

This leads to the solution

$$\hat{\beta}_{ML} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{3}$$

This is often referred to as Weighted Least Squares (WLS), $\hat{\beta}_{ML} = \hat{\beta}_{WLS}$.

For isotropic error covariance $V = \lambda I$, $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$, the Ordinary Least Squares (OLS) solution

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y \tag{4}$$

3 Restricted Maximum Likelihood (ReML)

If we don't know V we can estimate it using ReML. In Maximum Likelihood (ML), λ are estimated by maximising the likelihood

$$p(y|\beta, \lambda) \tag{5}$$

The idea behind ReML is to find λ that maximise the restricted likelihood

$$p(y|\lambda) \tag{6}$$

This does not depend on the parameters β. We can write it as

$$p(y|\lambda) = \int p(y|\beta, \lambda) d\beta \tag{7}$$

We will now use a quadratic identity, derived in the following section, to solve the integral.

3.1 Quadratic Identity

Let $y_t = X\beta$ be the true, but unknown, mean data values. And $\hat{y} = X\hat{\beta}$ be the predictions of the fitted model

Then

$$\begin{aligned}(y - X\beta)^T V^{-1}(y - X\beta) &= (y - y_t)^T V^{-1}(y - y_t) \qquad (8)\\ &= (y - y_t + \hat{y} - \hat{y})^T V^{-1}(y - y_t + \hat{y} - \hat{y})\\ &= (y - \hat{y})^T V^{-1}(y - \hat{y}) + (-y_t + \hat{y})^T V^{-1}(-y_t + \hat{y})\\ &= (y - \hat{y})^T V^{-1}(y - \hat{y}) + (y_t - \hat{y})^T V^{-1}(y_t - \hat{y})\\ &= (y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}) + (X\beta - X\hat{\beta})^T V^{-1}(X\beta - X\hat{\beta})\\ &= (y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}) + (\beta - \hat{\beta})^T (X^T V^{-1} X)(\beta - \hat{\beta})\end{aligned}$$

So, we have shown that

$$(y-X\beta)^T V^{-1}(y-X\beta) = (y-X\hat{\beta})^T V^{-1}(y-X\hat{\beta})+(\beta-\hat{\beta})^T (X^T V^{-1} X)(\beta-\hat{\beta}) \qquad (9)$$

The second term depends on the parameters β but the first does not.

3.2 ReML integral

For an $N \times p$ full-rank design matrix, the likelihood is

$$p(y|\beta,\lambda) = (2\pi)^{-N/2}|V|^{-1/2}\exp\left(\frac{1}{2}(y - X\beta)^T V^{-1}(y - X\beta)\right) \tag{10}$$

Using our quadratic identity we can write

$$\begin{aligned} p(y|\beta,\lambda) &= (2\pi)^{-N/2}|V|^{-1/2} \quad (11) \\ &\times \exp\left(\frac{1}{2}(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta})\right) \\ &\times \exp\left(\frac{1}{2}(\beta - \hat{\beta})^T (X^T V^{-1} X)(\beta - \hat{\beta})\right) \end{aligned}$$

The restricted likelihood is then given by

$$\begin{aligned} p(y|\lambda) &= \int p(y|\beta,\lambda) d\beta \quad (12) \\ &= (2\pi)^{-N/2}|V|^{-1/2} \end{aligned}$$

$$
\begin{aligned}
& \times \;\; \exp\left(\frac{1}{2}(y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta})\right) \\
& \times \;\; \int \exp\left(\frac{1}{2}(\beta - \hat{\beta})^T (X^T V^{-1} X)(\beta - \hat{\beta})\right) d\beta \\
& = \;\; (2\pi)^{-N/2} |V|^{-1/2} \\
& \times \;\; \exp\left(\frac{1}{2}(y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta})\right) \\
& \times \;\; (2\pi)^{p/2} |X^T V^{-1} X|^{-1/2}
\end{aligned}
$$

where we've noted that the integral is just the normalising constant for a multivariate Gaussian. Taking logs gives the ReML objective function

$$
\begin{aligned}
L_R(\lambda) &= \log p(y|\lambda) \qquad (13) \\
&= -\frac{N-p}{2} \log 2\pi - \frac{1}{2} \log |V| - \frac{1}{2} \log |X^T V^{-1} X| \\
&- \frac{1}{2}(y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta})
\end{aligned}
$$

This does not depend on the parameters β. It does depend on $\hat{\beta}$, but we can substitute in expressions for this from earlier, so that L_R is just a function of X, y and λ.

3.3 Single variance component

For a single variance component

$$V(\lambda) = \lambda Q \tag{14}$$

an analytic expression for λ can be found

3.3.1 Maximum Likelihood

The likelihood function is

$$L(\lambda) = -\frac{1}{2}\log|V| - \frac{1}{2}r^T V^{-1} r + ... \tag{15}$$

where the residuals are $r = y - X\beta$. The gradient with respect to λ is

$$g = -\frac{1}{2}Tr(V^{-1}Q) + \frac{1}{2}r^T V^{-1} Q V^{-1} r \tag{16}$$

For a single variance component we get the estimate

$$\lambda = \frac{r^T Q^{-1} r}{N} \tag{17}$$

For isotropic errors $Q = I$ we have

$$\lambda = \frac{r^T r}{N} \tag{18}$$

This is biased.

3.3.2 ReML

We write $g_R = \frac{dL_R(\lambda)}{d\lambda}$ as the gradient of the ReML function. This can be shown to be

$$g_R = -\frac{1}{2} Tr(PQ) + \frac{1}{2} y^T P^T Q P y \tag{19}$$

where the projection matrix $P = V^{-1} R_{WLS}$ and

$$R_{WLS} = I - X(X^T V^{-1} X)^{-1} X^T V^{-1} \tag{20}$$

It's the same as the ML gradient but with P instead of V^{-1} - we are working in a subspace of V^{-1} that is orthogonal to the WLS estimates.

Setting $g = 0$ gives

$$\lambda = \frac{r^T Q^{-1} r}{Tr(R)} \tag{21}$$

where

$$R = I - X(X^T Q^{-1} X)^{-1} X^T Q^{-1} \tag{22}$$

is the residual forming matirx and $r = Ry$ are the residuals.

If $Q = I$, ie. isotropic error, $R = I - X(X^T X)^{-1} X^T$ and

$$\lambda = \frac{r^T r}{N - k} \tag{23}$$

which is an *unbiased* estimate of the error variance (unlike the ML estimate).

3.4 Linear constraints

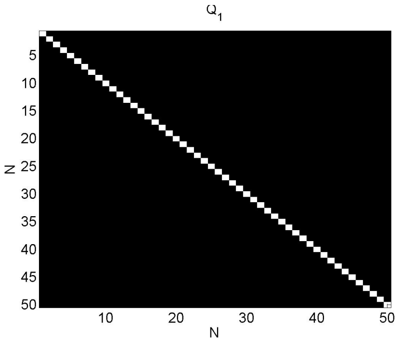
If the error covariance takes the following form

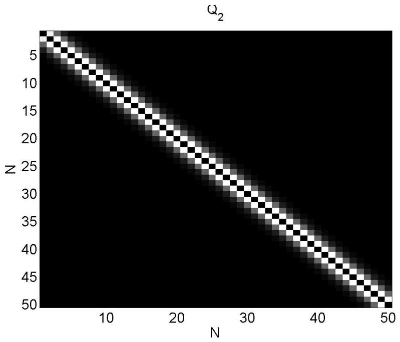
$$V(\lambda) = \sum_k \lambda_k Q_k \tag{24}$$

where Q_k is a known matrix and λ_k is the kth unknown hyperparameter, then λ can be found by maximising the ReML objective function. This could be implemented by any optimisation method eg. one could follow the gradient g where (see appendix)

$$g_k = -\frac{1}{2} Tr(PQ_k) + \frac{1}{2} y^T P^T Q_k P y \tag{25}$$

A better algorithm, based on Fisher scoring, has been derived by Harville [4]. See also Friston et al. [3][1] for applications to brain imaging. This algorithm is implemented in `spm_reml.m`.





3.5 fMRI time series

Correlated fMRI time series can be dealt with by having eg. two covariance components: one for the additive noise, Q_1, and a second for the temporal autocorrelation, Q_2. Q_2 is based on a first-order autoregressive model $e_t = ae_{t-1} + z_t$ with a fixed coefficient $a = 0.2$. Variations in a can be accomodated by specifying a third basis function Q_3 which is a Taylor expansion of Q_2 [1].

4 Hierarchical General Linear Models

Given the hierarchical model (eg. 3 levels)

$$\begin{aligned} y &= X_1\beta_1 + e_1 \\ \beta_1 &= X_2\beta_2 + e_2 \\ \beta_2 &= X_3\beta_3 + e_3 \end{aligned} \tag{26}$$

In the analysis of group data this can enable us, for example, to relate population effects, β_3, to subject effects β_2 to session effects β_1. The vector y contains the data from all trials in all sessions from all subjects.

The error covariances at each level C_3, C_2 and C_1 describe between-subject variance, between-session variance and between-trial variance.

We can substitute from β_2 into the second equation, then β_1 into the first to give a collapsed model

$$y = X\beta_3 + e \tag{27}$$

where

$$\begin{aligned} X &= X_1 X_2 X_3 \quad (28) \\ e &= e_1 + X_1 e_2 + X_1 X_2 e_3 \end{aligned}$$

The error covariance, $Cov(e) = C$ is

$$C = C_1 + X_1 C_2 X_2^T + X_1 X_2 C_3 X_2^T X_1^T \quad (29)$$

The hierarchical structure introduces this particular structure into the error covariance of the collapsed model.

We can then run ReML to estimate the variance components λ (parameters of C). The population effect is then estimated using WLS in the usual way.

4.1 Example

Two-level hierarchical model

$$\begin{aligned} y &= X_1\beta_1 + e_1 \\ \beta_1 &= X_2\beta_2 + e_2 \end{aligned} \tag{30}$$

where β_1 contains subject effects, β_2 the population effect. The errors are between-trial errors e_1 and between-subject errors e_2. In brain imaging such a model is known as a Random-Effects (RFX) analysis, as the subject effects are viewed as random variables (there is a between-subject error). The aim is to make an inference about the population effect.

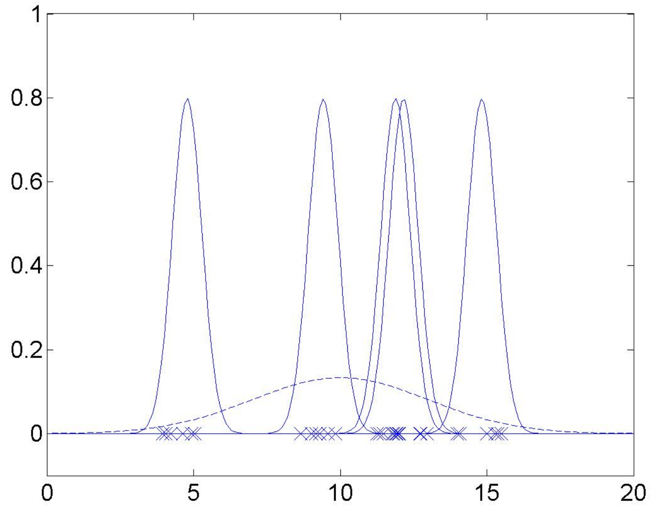
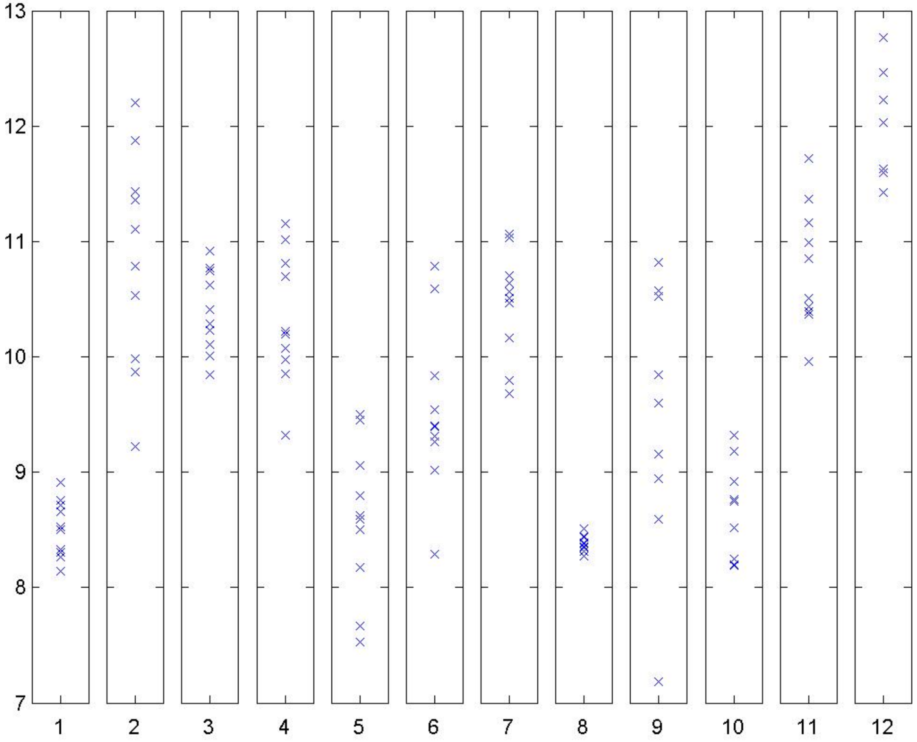
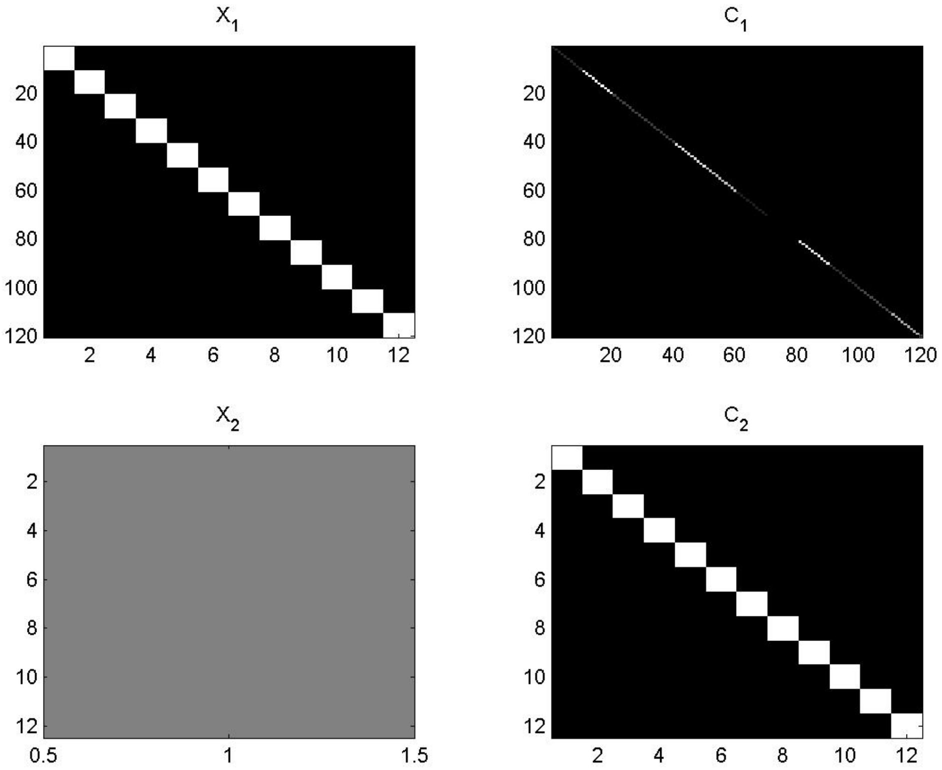
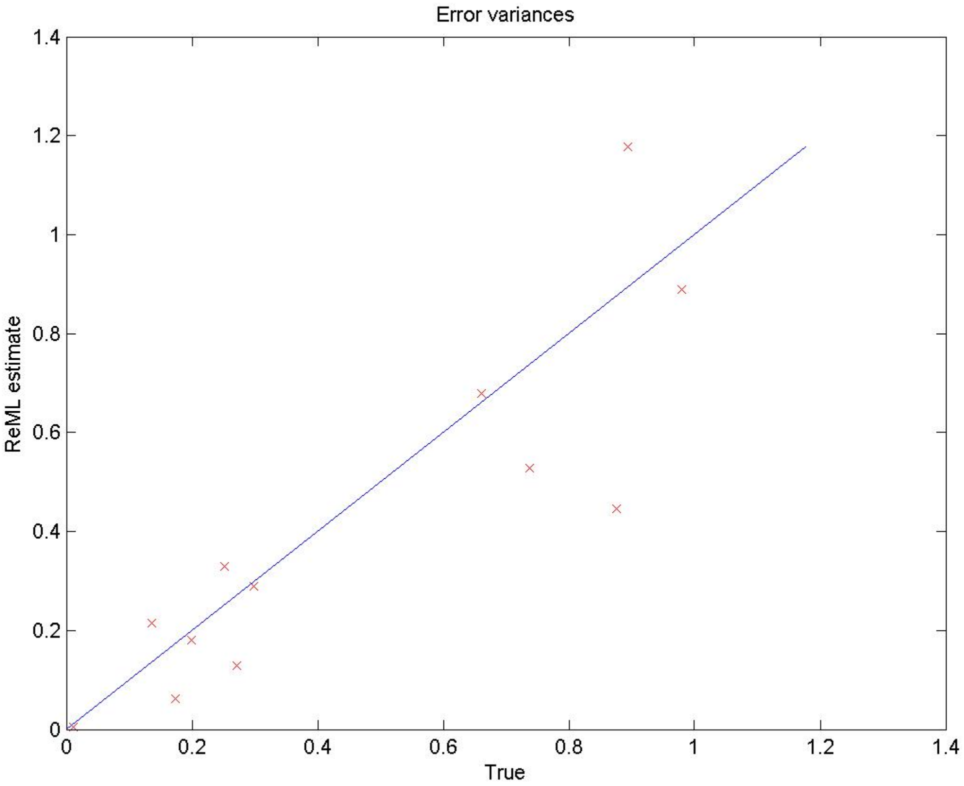


Figure 1: *Synthetic data illustrating the probability model underlying random effects analysis. The dotted line is the Gaussian distribution underlying the second level model with mean* β_2*, the population effect, and variance* σ_b^2*, the between-subject variance. The mean subject effects,* $\beta_1(i)$*, are drawn from this distribution. The solid lines are the Gaussians underlying the first level models with means* $\beta_1(i)$ *and variances* σ_w^2*. In this example the within-subject/between-trial variance is the same for all subjects. The crosses are the observed effects* y_{ij} *which are drawn from the solid Gaussians.*


13
12
11
10
9
8
7
1
2
3
4
5
6
7
8
9
10
11
12


X_1
20
40
60
80
100
120
2
4
6
8
10
12
C_1
20
40
60
80
100
120
20
40
60
80
100
120
X_2
2
4
6
8
10
12
0.5
1
1.5
C_2
2
4
6
8
10
12
2
4
6
8
10
12


Error variances
ReML estimate
True
0
0.2
0.4
0.6
0.8
1
1.2
1.4

4.2 Summary statistic approach

This involves simply taking a Summary Statistic (SS) eg. the mean, from one level and using it as data for the level above. For 'balanced designs', this gives us the correct results on average [5]. This requires the same number of trials per subject and the same between-trial error variance. The two-level hierarchical model is approximated as two separate single level models

$$\begin{aligned} y &= X_1\hat{\beta}_1 + e_1 \\ \hat{\beta}_1 &= X_2\beta_2 + e_2 \end{aligned} \tag{31}$$

The first level effects are estimated for each subject, saved as 'contrast images' and entered as data for a separate 2nd-level model.

5 fMRI data from multiple sessions

This section compares RFX analysis as implemented using SS versus ReML. The dataset comprises 1,200 images that were acquired in 10 sessions of 120 scans each. These data have been described elsewhere [2].

Each session contained a different number of events, so strictly, violates SS assumptoions. The experimental design involved 30-second epochs of single word streams and a passive listening task. The words were concrete, monosyllabic nouns presented at a number of different rates. The word rate was varied pseudo-randomly over epochs within each session. Further details of the paradigm and analysis details are given in [**?**]. The results of the SS and ReML analyses have been thresholded at $p < 0.05$, corrected for the entire search volume.

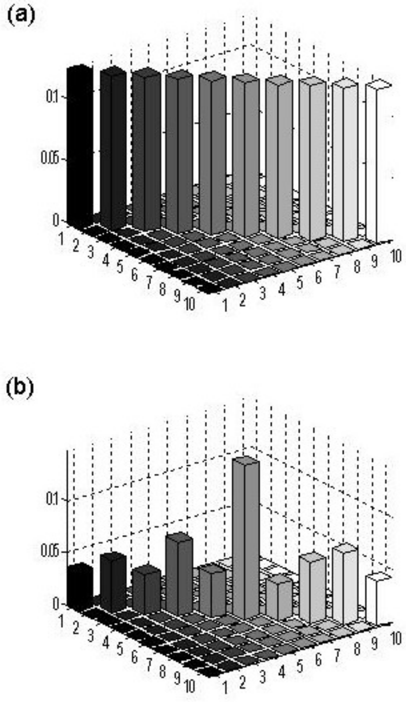


Figure 2: *Within-session variance as (a) assumed by SS and (b) estimated using ReML. This shows that within-session variance can vary by up to a factor of four, although this makes little difference to the final inference.*

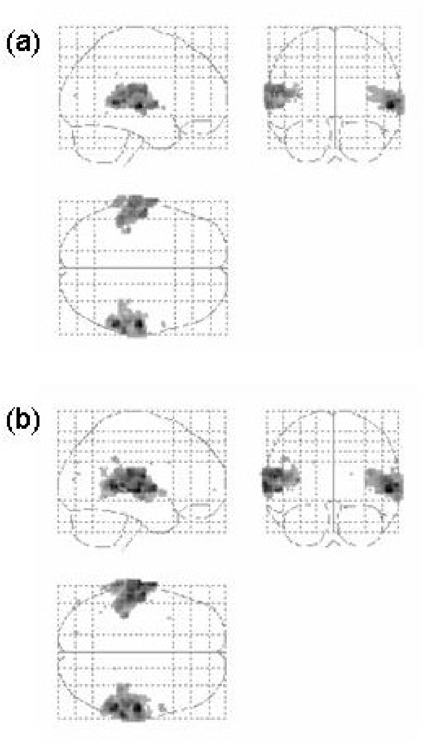


Figure 3: *SPMs showing the effect of words in the population using (a) SS and (b) ReML approaches.*

References

[1] K.J. Friston, D.E. Glaser, R.N.A. Henson, S.J. Kiebel, C. Phillips, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage*, 16:484–512, 2002.

[2] K.J. Friston, O. Josephs, G. Rees, and R. Turner. Non-linear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 39:41–52, 1998.

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

[4] D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[5] W.D. Penny, A.P. Holmes, and K.J. Friston. Random effects analysis. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.