

# MIXED

---

## Overview

This document summarizes the computational algorithms discussed in Wolfinger, Tobias and Sall (1994).

## Notation

$\boldsymbol{\theta}$	Overall covariance parameter vector
$\boldsymbol{\theta}_G$	A vector of covariance parameters associated with random effects.
$\boldsymbol{\theta}_k$	A vector of covariance parameters associated with the k-th random effect.
$\boldsymbol{\theta}_R$	A vector of covariance parameters associated with the residual term.
$K$	Number of random effects.
$S_r$	Number of repeated subjects.
$S_k$	Number of subjects in k-th random effect.
$\mathbf{V}(\boldsymbol{\theta})$	The $n \times n$ covariance matrix of $\mathbf{y}$ . In the overview document, this matrix is denoted by $\mathbf{V}_y(\boldsymbol{\theta})$ . The subscript is removed here for a cleaner notation.
$\dot{\mathbf{V}}_s(\boldsymbol{\theta})$	First derivative of $\mathbf{V}(\boldsymbol{\theta})$ with respect to the s-th parameter in $\boldsymbol{\theta}$ .
$\ddot{\mathbf{V}}_{st}(\boldsymbol{\theta})$	Second derivative of $\mathbf{V}(\boldsymbol{\theta})$ with respect to s and t-th parameter in $\boldsymbol{\theta}$ .
$\mathbf{R}(\boldsymbol{\theta}_R)$	The $n \times n$ covariance matrix of $\boldsymbol{\epsilon}$ .
$\dot{\mathbf{R}}_s(\boldsymbol{\theta}_R)$	First derivative of $\mathbf{R}(\boldsymbol{\theta}_R)$ with respect to the s-th parameter in $\boldsymbol{\theta}_R$ .
$\ddot{\mathbf{R}}_{st}(\boldsymbol{\theta}_R)$	Second derivative of $\mathbf{R}(\boldsymbol{\theta}_R)$ with respect to s and t-th parameter in $\boldsymbol{\theta}_R$ .

$\mathbf{G}(\boldsymbol{\theta}_G)$	The covariance matrix of random effects.
$\dot{\mathbf{G}}_s(\boldsymbol{\theta}_G)$	First derivative of $\mathbf{G}(\boldsymbol{\theta}_G)$ with respect to s-th parameter in $\boldsymbol{\theta}_G$ .
$\ddot{\mathbf{G}}_{st}(\boldsymbol{\theta}_G)$	Second derivative of $\mathbf{G}(\boldsymbol{\theta}_G)$ with respect to s and t-th parameter in $\boldsymbol{\theta}_G$ .
$\mathbf{V}_k(\boldsymbol{\theta}_k)$	The covariance matrix of the k-th random effect for <i>one random subject</i> .
$\dot{\mathbf{V}}_{k,s}(\boldsymbol{\theta}_k)$	First derivative of $\mathbf{V}_k(\boldsymbol{\theta}_k)$ with respect to s-th parameter in $\boldsymbol{\theta}_k$ .
$\ddot{\mathbf{V}}_{k,st}(\boldsymbol{\theta}_k)$	Second derivative of $\mathbf{V}_k(\boldsymbol{\theta}_k)$ with respect to s and t-th parameter in $\boldsymbol{\theta}_k$ .
$\mathbf{y}$	n by 1 vector of dependent variable.
$\mathbf{X}$	n by p design matrix of fixed effects.
$\mathbf{Z}$	n by q design matrix of random effects.
$\mathbf{r}$	n by 1 vector of residual.
$\boldsymbol{\beta}$	p by 1 vector of fixed effects parameters.
$\boldsymbol{\gamma}$	q by 1 vector of random effects parameters.
$\boldsymbol{\varepsilon}$	n by 1 vector of residual error.
$\mathbf{W}_c$	n by n diagonal matrix of case weights.
$\mathbf{W}_{rw}$	n by n diagonal matrix of regression weights.

## Model

In this document, we assume a mixed effect model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \quad (1)$$

In this model, we assume that  $\boldsymbol{\varepsilon}$  is distributed as  $N[\mathbf{0}, \mathbf{R}(\boldsymbol{\theta}_R)]$  and  $\boldsymbol{\gamma}$  is independently distributed as  $N[\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}_G)]$ . Therefore  $\mathbf{y}$  is distributed as  $N[\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta})]$ , where  $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\theta}_G)\mathbf{Z}^T + \mathbf{R}(\boldsymbol{\theta}_R)$ . The unknown parameters include the regression parameters in  $\boldsymbol{\beta}$  and covariance parameters in

$\boldsymbol{\theta}$ . Estimation of these model parameters relies on the use of a Newton-Raphson or scoring algorithm. When we use either algorithm for finding MLE or REML solutions, we need to compute  $\mathbf{V}^{-1}(\boldsymbol{\theta})$  and its derivatives with respect to  $\boldsymbol{\theta}$ , which are computationally infeasible for large  $n$ . Wolfinger et.al.(1994) discussed methods that can avoid direct computation of  $\mathbf{V}^{-1}(\boldsymbol{\theta})$ . They tackled the problem by using the SWEEP algorithm and exploiting the block diagonal structure of  $\mathbf{G}$  and  $\mathbf{R}$ . In the first half of this document, we will detail the algorithm for mixed model without subject blocking. In second half of the document we will refine the algorithm to exploit the structure of  $\mathbf{G}$  and this is the actual implementation of the algorithm.

If there are regression weights, the covariance matrix  $\mathbf{R}(\boldsymbol{\theta}_R)$  will be replaced by  $\mathbf{R}^*(\boldsymbol{\theta}_R) = \mathbf{W}_{rw}^{-1/2} \mathbf{R}(\boldsymbol{\theta}_R) \mathbf{W}_{rw}^{-1/2}$ . For simpler notations, we will assume that the weights are already included in the matrix  $\mathbf{R}(\boldsymbol{\theta}_R)$  and they will not be displayed in the remainder of this document. When case weights are specified, they will be rounded to nearest integer and each case will be entered into the analysis multiple times depending on the rounded case weight. Since replicating a case will lead to duplicate repeated measures (Note: repeated measures are unique within a repeated subject), non-unity case weights will only be allowed for  $\mathbf{R}(\boldsymbol{\theta}_R)$  with scaled identity structure. In MIXED, only cases with positive case weight and regression weight will be included analysis.

### Fixed Effect Parameterization:

The parameterization of fixed effects is the same as in the GLM procedure.

### Random Effects Parameterization:

If we have  $K$  random effects and there are  $S_k$  random subjects in  $k$ -th random effect, the design matrix  $\mathbf{Z}$  will be partitioned as

$$\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \cdots \quad \mathbf{Z}_K],$$

where  $\mathbf{Z}_k$  is the design matrix of the  $k$ -th random effect. Each  $\mathbf{Z}_k$  can be partitioned further by random subjects as below,

## 4 MIXED

$$\mathbf{Z}_k = [\mathbf{Z}_{k1} \quad \mathbf{Z}_{k2} \quad \cdots \quad \mathbf{Z}_{ks_k}], k=1, \dots, K.$$

Number of columns in the design matrix  $\mathbf{Z}_{kj}$  (j-th random subject of k-th random effect) is equal to number of levels of the k-th random effect variable.

Under this partition, the  $\mathbf{G}(\boldsymbol{\theta}_G)$  will be a block diagonal matrix which can be expressed as

$$\mathbf{G}(\boldsymbol{\theta}_G) = \bigoplus_{k=1}^K [\mathbf{I}_{S_k} \otimes \mathbf{V}_k(\boldsymbol{\theta}_k)].$$

It should also be noted that each random effect has its own parameter vector  $\boldsymbol{\theta}_k$ ,  $k=1, \dots, K$ , and there are no functional constraints between elements in these parameter vectors. Thus  $\boldsymbol{\theta}_G = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ .

When there are correlated random effects,  $\mathbf{Z}_{kj}$  will be a combined design matrix of the correlated random effects. Therefore in subsequent sections, each random effect can either be one single random effect or a set of correlated random effects.

### Repeated Subjects:

When the REPEATED subcommand is used,  $\mathbf{R}(\boldsymbol{\theta}_R)$  will be a block diagonal matrix where i-th block is  $\mathbf{R}_i(\boldsymbol{\theta}_R)$ ,  $i=1, \dots, S_R$ . That is,

$$\mathbf{R}(\boldsymbol{\theta}_R) = \bigoplus_{i=1}^{S_R} \mathbf{R}_i(\boldsymbol{\theta}_R)$$

The dimension of  $\mathbf{R}_i(\boldsymbol{\theta}_R)$  will be equal to number of cases in one repeated subject but all  $\mathbf{R}_i(\boldsymbol{\theta}_R)$  share the same parameter vector  $\boldsymbol{\theta}_R$ .

## Likelihood Functions

Recall that the  $-2$  times log likelihood of MLE is

$$\begin{aligned}
-2\ell_{\text{MLE}}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \log |\mathbf{V}(\boldsymbol{\theta})| + \mathbf{r}(\boldsymbol{\theta})^T \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{r}(\boldsymbol{\theta}) \\
&\quad + n \log 2\pi
\end{aligned} \tag{2}$$

and the  $-2$  times log likelihood of REML is

$$\begin{aligned}
-2\ell_{\text{REML}}(\boldsymbol{\theta}) &= \log |\mathbf{V}(\boldsymbol{\theta})| + \mathbf{r}(\boldsymbol{\theta})^T \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{r}(\boldsymbol{\theta}) \\
&\quad + \log |\mathbf{X}' \mathbf{V}^{-1}(\boldsymbol{\theta}) \mathbf{X}| + (n-p) \log 2\pi
\end{aligned} \tag{3}$$

where  $n$  is the number of observations and  $p$  is the rank of fixed effect design matrix. From (2) and (3), we can see that the key component of the likelihood functions are

$$\begin{aligned}
\ell_1(\boldsymbol{\theta}) &= \log |\mathbf{V}(\boldsymbol{\theta})| \\
\ell_2(\boldsymbol{\theta}) &= \mathbf{r}(\boldsymbol{\theta})' \mathbf{V}^{-1}(\boldsymbol{\theta}) \mathbf{r}(\boldsymbol{\theta}) \\
\ell_3(\boldsymbol{\theta}) &= \log |\mathbf{X}' \mathbf{V}^{-1}(\boldsymbol{\theta}) \mathbf{X}|.
\end{aligned} \tag{4}$$

Therefore, in each estimation iteration, we need to compute  $\ell_1(\boldsymbol{\theta})$ ,  $\ell_2(\boldsymbol{\theta})$  and  $\ell_3(\boldsymbol{\theta})$  as well as their 1<sup>st</sup> and 2<sup>nd</sup> derivatives with respect to  $\boldsymbol{\theta}$ .

## Newton & Scoring Algorithms

Covariance parameters in  $\boldsymbol{\theta}$  can be found by maximizing (2) or (3), however, there are no closed form solutions in general. Therefore Newton and scoring algorithms are used to find the solution numerically. The algorithm is outlined as below,

1. Compute starting value and initial log-likelihood (REML or ML).
2. Compute gradient vector  $\mathbf{g}$  and Hessian matrix  $\mathbf{H}$  of the log-likelihood function using last iteration's estimate  $\boldsymbol{\theta}_{i-1}$ . (See later section for computation of  $\mathbf{g}$  and  $\mathbf{H}$ )
3. Compute the new step  $\mathbf{d} = -\mathbf{H}^{-1} \mathbf{g}$ .

4. Let  $\rho = 1$ .
5. Compute estimates of i-th iteration  $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} + \rho \mathbf{d}$ .
6. Check if  $\boldsymbol{\theta}_i$  generates valid covariance matrices and improve the likelihood. If not, reduce  $\rho$  by half and repeat step (5). If this process is repeated for pre-specified no. of times and the stated conditions are still not satisfied, stop.
7. Check convergence of the parameter. If convergence criteria are met, then stop. Otherwise, go back to step (2).

Newton's algorithm performs well if the starting value is close to the solution. In order to improve the algorithm's robustness to bad starting values, the scoring algorithm is used in the first few iterations. This can be done easily by applying different formulae for the Hessian matrix at each iteration. Apart from improved robustness, the scoring algorithm is faster due to the simpler form of the Hessian matrix.

## Convergence Criteria

There are three types of convergence criteria: parameter convergence, log-likelihood convergence and Hessian convergence. For parameter and log-likelihood convergence, they are subdivided into absolute and relative. If we let  $\boldsymbol{\epsilon}$  be some given tolerance level and

$\boldsymbol{\theta}_{s,i}$  be s-th parameter in i-th iteration,

$\ell_i$  be the log-likelihood in log-likelihood in i-th iteration,

$\mathbf{g}_i$  be the gradient vector in i-th iteration,

and  $\mathbf{H}_i$  be the hessian matrix in i-th iteration, then the criteria can be written as follows,

**Absolute parameter convergence:**  $\max_s |\boldsymbol{\theta}_{s,i} - \boldsymbol{\theta}_{s,i-1}| < \boldsymbol{\epsilon}$

**Relative parameter convergence:**  $\max_s \left| \frac{\boldsymbol{\theta}_{s,i} - \boldsymbol{\theta}_{s,i-1}}{\boldsymbol{\theta}_{s,i-1}} \right| < \boldsymbol{\epsilon}$

**Absolute log-likelihood convergence:**  $|\ell_i - \ell_{i-1}| < \boldsymbol{\epsilon}$

**Relative log-likelihood convergence:**  $\left| \frac{\ell_i - \ell_{i-1}}{\ell_{i-1}} \right| < \boldsymbol{\epsilon}$

**Absolute Hessian convergence:**  $\mathbf{g}_i^T \mathbf{H}_i^{-1} \mathbf{g}_i < \varepsilon$

**Relative Hessian convergence:**  $\mathbf{g}_i^T \mathbf{H}_i^{-1} \mathbf{g}_i < \varepsilon |\ell_i|$

## Starting value of Newton's Algorithm

If no prior information is available, we can choose the initial values of  $\mathbf{G}$  and  $\mathbf{R}$  to be the identity. However, it is highly desirable to estimate the scale of the parameter. By ignoring the random effects, and assuming the residual errors are i.i.d. with variance  $\sigma^2$ , we can fit a GLM model and estimate  $\sigma^2$  by the residual sum of squares  $\hat{\sigma}^2$ . Then we choose the starting value of Newton's algorithm to be

$$\mathbf{G}_k = \frac{\hat{\sigma}^2}{K+1} \text{ and } \mathbf{R} = \frac{\hat{\sigma}^2}{K+1}.$$

## Confidence Intervals of Covariance Parameters

The estimate  $\hat{\boldsymbol{\theta}}$  (ML or REML) is asymptotically normally distributed. Its variance covariance matrix can be approximated by  $-2\mathbf{H}^{-1}$ , where  $\mathbf{H}$  is the hessian matrix of the log-likelihood function evaluated at  $\hat{\boldsymbol{\theta}}$ . A simple Wald's type confidence interval for any covariance parameter can be obtained by using the asymptotic normality of the parameter estimates, however it is not very appropriate for variance parameters and correlation parameters that have a range of  $[0, \infty)$  and  $[-1, 1]$  respectively. Therefore these parameters are transformed to parameters that have range  $(-\infty, \infty)$ . Using uniform delta method, see for example (van der Vaart, 1998), these transformed estimates still have asymptotic normal distributions.

Suppose we are estimation a variance parameter  $\sigma^2$  by  $\hat{\sigma}_n^2$  that is distributed as  $N[\sigma^2, \text{Var}(\hat{\sigma}_n^2)]$  asymptotically. The transformation we used is  $\log(\sigma^2)$  which can correct the skewness of  $\hat{\sigma}_n^2$ , moreover  $\log(\hat{\sigma}_n^2)$  has the range  $(-\infty, \infty)$  which matches that of normal distribution. Using delta method, one

can show that the asymptotic distribution of  $\log(\hat{\sigma}_n^2)$  is  $N[\log(\sigma^2), \sigma^{-4} \text{Var}(\hat{\sigma}_n^2)]$ . Thus, a  $(1-\alpha)100\%$  confidence interval of  $\log(\sigma^2)$  is given by

$$[\log(\hat{\sigma}_n^2) - z_{1-\alpha/2} \sigma_n^{-2} \sqrt{\text{Var}(\hat{\sigma}_n^2)}, \log(\hat{\sigma}_n^2) + z_{1-\alpha/2} \sigma_n^{-2} \sqrt{\text{Var}(\hat{\sigma}_n^2)}]$$

where  $z_{1-\alpha/2}$  is the upper  $(1-\alpha/2)$  percentage point of standard normal distribution. By inverting this confidence interval, a  $(1-\alpha)100\%$  confidence interval for  $\sigma^2$  is given by

$$[\exp(\log(\hat{\sigma}_n^2) - z_{1-\alpha/2} \sigma_n^{-2} \sqrt{\text{Var}(\hat{\sigma}_n^2)}), \exp(\log(\hat{\sigma}_n^2) + z_{1-\alpha/2} \sigma_n^{-2} \sqrt{\text{Var}(\hat{\sigma}_n^2)})]$$

When we need a confidence interval for a correlation parameter  $\rho$ , a possible transformation will be its generalized logit  $\text{arctanh}(\rho) = 0.5 \log[(1+\rho)/(1-\rho)]$ . The resulting confidence interval for  $\rho$  will be

$$[\tanh(\text{arctanh}(\hat{\rho}) - z_{1-\alpha/2} (1-\hat{\rho}^2)^{-1} \sqrt{\text{Var}(\hat{\rho})}), \tanh(\text{arctanh}(\hat{\rho}) + z_{1-\alpha/2} (1-\hat{\rho}^2)^{-1} \sqrt{\text{Var}(\hat{\rho})})]$$

## Fixed and Random Effect Parameters

### Estimation and prediction

After we obtain an estimate of  $\theta$ , best linear unbiased estimator (BLUE) of  $\beta$  and best linear unbiased predictor (BLUP) of  $\gamma$  can be found by solving the mixed model equations, Henderson (1984).

$$\begin{bmatrix} \mathbf{X}^T \hat{\mathbf{R}}^{-1} \mathbf{X} & \mathbf{X}^T \hat{\mathbf{R}}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \hat{\mathbf{R}}^{-1} \mathbf{X} & \mathbf{Z}^T \hat{\mathbf{R}}^{-1} \mathbf{Z} + \hat{\mathbf{G}}^{-1} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{bmatrix} \mathbf{X}^T \hat{\mathbf{R}}^{-1} \mathbf{y} \\ \mathbf{Z}^T \hat{\mathbf{R}}^{-1} \mathbf{y} \end{bmatrix} \quad (5)$$

The solution of (5) can be expressed as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{y} \\ \hat{\boldsymbol{\gamma}} &= \hat{\mathbf{G}} \mathbf{Z}^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) \\ &= \hat{\mathbf{G}} [\mathbf{Z}^T \hat{\mathbf{V}}^{-1} \mathbf{y} - \mathbf{Z}^T \hat{\mathbf{V}}^{-1} \mathbf{X} \hat{\boldsymbol{\beta}}] \end{aligned} \quad (6)$$

The covariance matrix  $\mathbf{C}$  of  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\gamma}}$  is given by

$$\begin{aligned} \mathbf{C} &= \text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \\ &= \begin{bmatrix} \mathbf{X}^T \hat{\mathbf{R}}^{-1} \mathbf{X} & \mathbf{X}^T \hat{\mathbf{R}}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \hat{\mathbf{R}}^{-1} \mathbf{X} & \mathbf{Z}^T \hat{\mathbf{R}}^{-1} \mathbf{Z} + \hat{\mathbf{G}}^{-1} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \hat{\mathbf{C}}_{11} & \hat{\mathbf{C}}_{21}^T \\ \hat{\mathbf{C}}_{21} & \hat{\mathbf{C}}_{22} \end{bmatrix} \end{aligned} \quad (7)$$

where

$$\begin{aligned} \hat{\mathbf{C}}_{11} &= (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \\ \hat{\mathbf{C}}_{21} &= -\hat{\mathbf{G}} \mathbf{Z}^T \hat{\mathbf{V}}^{-1} \mathbf{X} \hat{\mathbf{C}}_{11} \\ \hat{\mathbf{C}}_{22} &= (\mathbf{Z}^T \hat{\mathbf{R}}^{-1} \mathbf{Z} + \hat{\mathbf{G}}^{-1})^{-1} - \hat{\mathbf{C}}_{21} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{Z} \hat{\mathbf{G}} \end{aligned}$$

## Custom Hypotheses

In general, one can construct estimators or predictors for

$$\mathbf{Lb} = [\mathbf{L}_0 \quad \mathbf{L}_1] \begin{bmatrix} \boldsymbol{\beta} \\ \gamma \end{bmatrix} \quad (8)$$

for some hypothesis matrix  $\mathbf{L}$ . Estimators or predictors of  $\mathbf{Lb}$  can easily be constructed by substituting  $\hat{\boldsymbol{\beta}}$  and  $\hat{\gamma}$  into (8) and its variance covariance matrix can be approximated by  $\mathbf{LCL}^T$ . If  $\mathbf{L}_1$  is zero and  $\mathbf{L}_0\boldsymbol{\beta}$  is estimable,  $\mathbf{L}\hat{\boldsymbol{\beta}}$  is called the best linear unbiased estimator of  $\mathbf{L}_0\boldsymbol{\beta}$ . If  $\mathbf{L}_1$  is nonzero and  $\mathbf{L}_0\boldsymbol{\beta}$  is estimable,  $\mathbf{L}\hat{\mathbf{b}}$  is called the best linear unbiased predictor of  $\mathbf{Lb}$ .

To test the hypothesis  $H_0: \mathbf{Lb} = \mathbf{a}$  for a given vector  $\mathbf{a}$ , we can use the statistic

$$\mathbf{F} = \frac{(\mathbf{L}\hat{\mathbf{b}} - \mathbf{a})^T (\mathbf{L}\hat{\mathbf{C}}\mathbf{L}^T)^{-1} (\mathbf{L}\hat{\mathbf{b}} - \mathbf{a})}{q} \quad (9)$$

where  $q$  is the rank of the matrix  $\mathbf{L}$ . The statistic in (9) has an approximate F distribution. The numerator degrees of freedom is  $q$  and the denominator degree of freedom can be obtained by Satterthwaite (1946) approximation. The method outlined below is similar to Giesbrecht and Burns (1985), McLean and Sanders (1988), and Fai and Cornelius (1996).

### Satterthwaite's Approximation

To find the denominator degrees of freedom of (9), we first perform the spectral decomposition  $\mathbf{L}\hat{\mathbf{C}}\mathbf{L}^T = \boldsymbol{\Gamma}^T \mathbf{D} \boldsymbol{\Gamma}$  where  $\boldsymbol{\Gamma}$  is an orthogonal matrix of eigenvectors and  $\mathbf{D}$  is a diagonal matrix of eigenvalues. If we let  $\ell_m$  be the  $m$ -th row of  $\boldsymbol{\Gamma}\mathbf{L}$ ,  $d_m$  be the  $m$ -th eigenvalues and

$$v_m = \frac{2d_m^2}{\mathbf{g}_m^T \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1} \mathbf{g}_m}$$

where  $\mathbf{g}_m = \left. \frac{\partial \ell_m \mathbf{C} \ell_m^T}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$  and  $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}$  is the covariance matrix of the estimated covariance parameters. If we let

$$E = \sum_{m=1}^q \frac{v_m}{v_m - 2} I(v_m > 2)$$

then the denominator degree of freedom is given by

$$v = \frac{2E}{E - q}.$$

Note that the degrees of freedom can only be computed when  $E > q$ .

## Type I & III Statistics

Type I or III test statistics are special cases of custom hypothesis tests.

### Estimated Marginal Means (EMMEANS)

Estimated marginal means are special cases of custom hypothesis test. The construction of the L matrix for EMMEANS can be found in "Estimated Marginal Means" section of GLM's algorithm document. If Bonferroni or Sidak adjustment is requested for multiple comparisons, they will be computed according to the algorithm detailed in Appendix 10:Post Hoc Tests.

## Saved Values

If predicted values is requested, it will be computed by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\gamma}} \quad (10)$$

using the estimates given in (6).

If fixed predicted values is requested, it will be computed by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \quad (11)$$

If residual is requested, it will be computed by

$$\mathbf{r} = \hat{\mathbf{y}} - \mathbf{y} \quad (12)$$

If standard errors or degrees of freedom are requested for predicted values, a  $\mathbf{L}$  matrix will be constructed for each case and the formula in custom hypothesis section will be used to obtain the requested values.

## Information Criteria

Information criteria are for model comparison, the following criteria are given in smaller is better form. If we let  $\ell$  be the log-likelihood of (REML or ML),  $n$  be total number of cases (or total of case weights if used) and  $d$  is number of model parameters, the formula for various criteria are given as below,

- Akaike information criteria (AIC), Akaike (1974):

$$2\ell - 2d$$

- Finite sample corrected (AICC), Hurvich and Tsai (1989):

$$2\ell - \frac{2d + n}{(n - d - 1)}$$

- Bayesian information criteria (BIC), Schwarz (1978):

$$2\ell - d \log(n)$$

- Consistent AIC (CAIC), Bozdogan (1987):

$$2\ell - d (\log(n) + 1)$$

For REML, the value of  $n$  is chosen to be total number of cases minus number fixed effect parameters and  $d$  is number of covariance parameters. For ML, the value of  $n$  is total number of cases and  $d$  is number of fixed effect parameters plus number of covariance parameters.

## 1st and 2nd Derivatives of $\ell_k(\boldsymbol{\theta})$

In each Newton or scoring iteration we need to compute the 1<sup>st</sup> and 2<sup>nd</sup> derivatives of the components of the log-likelihood  $\ell_k(\boldsymbol{\theta})$ ,  $k=1,2,3$ . Here we let

$\mathbf{g}_k = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_k(\boldsymbol{\theta})$  and  $\mathbf{H}_k = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \ell_k(\boldsymbol{\theta})$ ,  $k = 1,2,3$ , then the 1<sup>st</sup> derivatives with respect to the  $s$ -th parameter in  $\boldsymbol{\theta}$  is given by

$$\begin{aligned} [\mathbf{g}_1]_s &= \text{tr}(\mathbf{V}^{-1} \dot{\mathbf{V}}_s), \\ [\mathbf{g}_2]_s &= -\mathbf{r} \mathbf{V}^{-1} \dot{\mathbf{V}}_s \mathbf{V}^{-1} \mathbf{r}, \\ [\mathbf{g}_3]_s &= -\text{tr}(\tilde{\mathbf{X}}^T \mathbf{V}^{-1} \dot{\mathbf{V}}_s \mathbf{V}^{-1} \tilde{\mathbf{X}}) \end{aligned} \quad (13)$$

and the 2<sup>nd</sup> derivatives with respect to  $s$  and  $t$ -th parameter are given by

$$\begin{aligned} [\mathbf{H}_1]_{st} &= -\text{tr}(\mathbf{V}^{-1} \dot{\mathbf{V}}_s \mathbf{V}^{-1} \dot{\mathbf{V}}_t) + \text{tr}(\mathbf{V}^{-1} \ddot{\mathbf{V}}_{st}), \\ [\mathbf{H}_2]_{st} &= 2\mathbf{r}^T \mathbf{V}^{-1} \dot{\mathbf{V}}_s \mathbf{V}^{-1} \dot{\mathbf{V}}_t \mathbf{V}^{-1} \mathbf{r} \\ &\quad - 2\mathbf{r}^T \mathbf{V}^{-1} \dot{\mathbf{V}}_s \mathbf{V}^{-1} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \dot{\mathbf{V}}_t \mathbf{V}^{-1} \mathbf{r} \\ &\quad - \mathbf{r}^T \mathbf{V}^{-1} \ddot{\mathbf{V}}_{st} \mathbf{V}^{-1} \mathbf{r} \end{aligned} \quad (14)$$

$$\begin{aligned} [\mathbf{H}_3]_{st} &= 2\text{tr}(\tilde{\mathbf{X}}^T \mathbf{V}^{-1} \dot{\mathbf{V}}_s \mathbf{V}^{-1} \dot{\mathbf{V}}_t \mathbf{V}^{-1} \tilde{\mathbf{X}}) \\ &\quad - \text{tr}(\tilde{\mathbf{X}}^T \mathbf{V}^{-1} \dot{\mathbf{V}}_s \mathbf{V}^{-1} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \mathbf{V}^{-1} \dot{\mathbf{V}}_t \mathbf{V}^{-1} \tilde{\mathbf{X}}) \\ &\quad - \text{tr}(\tilde{\mathbf{X}}^T \mathbf{V}^{-1} \dot{\mathbf{V}}_{st} \mathbf{V}^{-1} \tilde{\mathbf{X}}) \end{aligned}$$

where  $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{C}$  for a matrix  $\mathbf{C}$  satisfying  $\mathbf{C}\mathbf{C}^T = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \mathbf{P}$  and  $\mathbf{r} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{y} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ .

## Derivatives w.r.t. Parameters in $\mathbf{G}$

Derivatives with respect to parameters in  $\mathbf{G}$  can be constructed by from the entries of

$$\begin{aligned} \mathbf{W}_1(\mathbf{X}; \mathbf{r}; \mathbf{Z}) &= \begin{bmatrix} \mathbf{W}_1(\mathbf{X}, \mathbf{X}) & \mathbf{W}_1(\mathbf{X}, \mathbf{Z}) & \mathbf{W}_1(\mathbf{X}, \mathbf{r}) \\ \mathbf{W}_1(\mathbf{Z}, \mathbf{X}) & \mathbf{W}_1(\mathbf{Z}, \mathbf{Z}) & \mathbf{W}_1(\mathbf{Z}, \mathbf{r}) \\ \mathbf{W}_1(\mathbf{r}, \mathbf{X}) & \mathbf{W}_1(\mathbf{r}, \mathbf{Z}) & \mathbf{W}_1(\mathbf{r}, \mathbf{r}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Z} & \mathbf{X}^T \mathbf{V}^{-1} \mathbf{r} \\ \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} & \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{r} \\ \mathbf{r}^T \mathbf{V}^{-1} \mathbf{X} & \mathbf{r}^T \mathbf{V}^{-1} \mathbf{Z} & \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r} \end{bmatrix} \end{aligned} \quad (15)$$

The matrix  $\mathbf{W}_1(\mathbf{X}; \mathbf{r}; \mathbf{Z})$  can be computed from  $\mathbf{W}_1(\mathbf{X}; \mathbf{y}; \mathbf{Z})$  given in (27), by using the following relationship,

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\mathbf{b}_0$$

where  $\mathbf{b}_0$  is the current estimate of  $\boldsymbol{\beta}$ .

Using the above formula, we can obtain the following expressions,

$$\begin{aligned} \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r} &= \mathbf{y}^T \mathbf{V}^{-1} \mathbf{y} - \mathbf{y}^T \mathbf{V}^{-1} \mathbf{X} \mathbf{b}_0 \\ &= \mathbf{W}_1(\mathbf{y}, \mathbf{y}) - \mathbf{W}_1(\mathbf{y}, \mathbf{X}) \mathbf{b}_0 \end{aligned} \quad (16)$$

$$\begin{aligned} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{r} &= \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} - \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \mathbf{b}_0 \\ &= \mathbf{W}_1(\mathbf{X}, \mathbf{y}) - \mathbf{W}_1(\mathbf{X}, \mathbf{X}) \mathbf{b}_0 \end{aligned} \quad (17)$$

$$\begin{aligned}
\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{r} &= \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{y} - \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{X} \mathbf{b}_0 \\
&= \mathbf{W}_1(\mathbf{Z}, \mathbf{y}) - \mathbf{W}_1(\mathbf{Z}, \mathbf{X}) \mathbf{b}_0
\end{aligned} \tag{18}$$

In terms of the elements in  $\mathbf{W}_1(\mathbf{X}; \mathbf{r}; \mathbf{Z})$  matrix, we can write down the 1<sup>st</sup> derivatives of  $\ell_1, \ell_2$  and  $\ell_3$  with respect to a parameter  $\theta_s$  of the  $\mathbf{G}$  matrix,

$$\begin{aligned}
[\mathbf{g}_1]_{\mathbf{G},s} &= \text{tr}(\mathbf{W}_1(\mathbf{Z}, \mathbf{Z}) \dot{\mathbf{G}}_s), \\
[\mathbf{g}_2]_{\mathbf{G},s} &= -\mathbf{W}_1(\mathbf{Z}, \mathbf{r})^T \dot{\mathbf{G}}_s \mathbf{W}_1(\mathbf{Z}, \mathbf{r}), \\
[\mathbf{g}_3]_{\mathbf{G},s} &= -\text{tr}(\mathbf{W}_1(\mathbf{X}, \mathbf{Z}) \dot{\mathbf{G}}_s \mathbf{W}_1(\mathbf{Z}, \mathbf{X}) \mathbf{P}),
\end{aligned} \tag{19}$$

For the second derivatives, we first define the following simplification factors

$$\begin{aligned}
\mathbf{H}_{\mathbf{G}1}^{\text{st}} &= -\mathbf{W}_1(\mathbf{Z}, \mathbf{Z}) \dot{\mathbf{G}}_s \mathbf{W}_1(\mathbf{Z}, \mathbf{Z}) \dot{\mathbf{G}}_t + \mathbf{W}_1(\mathbf{Z}, \mathbf{Z}) \ddot{\mathbf{G}}_{st} \\
\mathbf{H}_{\mathbf{G}2}^s &= \mathbf{W}_1(\mathbf{X}, \mathbf{Z}) \dot{\mathbf{G}}_s \mathbf{W}_1(\mathbf{Z}, \mathbf{r}) \\
\mathbf{H}_{\mathbf{G}2}^{\text{st}} &= 2\mathbf{W}_1(\mathbf{r}, \mathbf{Z}) \dot{\mathbf{G}}_s \mathbf{W}_1(\mathbf{Z}, \mathbf{Z}) \dot{\mathbf{G}}_t \mathbf{W}_1(\mathbf{Z}, \mathbf{r}) - \mathbf{W}_1(\mathbf{r}, \mathbf{Z}) \ddot{\mathbf{G}}_{st} \mathbf{W}_1(\mathbf{Z}, \mathbf{r}) \\
\mathbf{H}_{\mathbf{G}3}^s &= \mathbf{W}_1(\mathbf{X}, \mathbf{Z}) \dot{\mathbf{G}}_s \mathbf{W}_1(\mathbf{Z}, \mathbf{X}) \\
\mathbf{H}_{\mathbf{G}3}^{\text{st}} &= 2\mathbf{W}_1(\mathbf{X}, \mathbf{Z}) \dot{\mathbf{G}}_s \mathbf{W}_1(\mathbf{Z}, \mathbf{Z}) \dot{\mathbf{G}}_t \mathbf{W}_1(\mathbf{Z}, \mathbf{X}) - \mathbf{W}_1(\mathbf{X}, \mathbf{Z}) \ddot{\mathbf{G}}_{st} \mathbf{W}_1(\mathbf{Z}, \mathbf{X})
\end{aligned}$$

then second derivatives of  $\ell_1, \ell_2$  and  $\ell_3$  w.r.t.  $\theta_s$  and  $\theta_t$  (in  $\mathbf{G}$ ) are given by

$$\begin{aligned}
[\mathbf{H}_1]_{\mathbf{G},st} &= \text{tr}(\mathbf{H}_{\mathbf{G}1}^{\text{st}}) \\
[\mathbf{H}_2]_{\mathbf{G},st} &= \mathbf{H}_{\mathbf{G}2}^{\text{st}} - 2(\mathbf{H}_{\mathbf{G}2}^s)^T \mathbf{P} \mathbf{H}_{\mathbf{G}2}^t \\
[\mathbf{H}_3]_{\mathbf{G},st} &= \text{tr}[\mathbf{H}_{\mathbf{G}3}^{\text{st}} \mathbf{P}] - \text{tr}[\mathbf{H}_{\mathbf{G}3}^s \mathbf{P} \mathbf{H}_{\mathbf{G}3}^t \mathbf{P}]
\end{aligned} \tag{20}$$

## Derivatives w.r.t. Parameters in R

To compute  $\mathbf{R}$  derivatives, we need to introduce the matrices

$$\mathbf{W}_0^{(1)s} = -\frac{\partial \mathbf{W}_0}{\partial \theta_s}$$

and

$$\mathbf{W}_0^{(2)st} = -\frac{\partial^2 \mathbf{W}_0}{\partial \theta_s \partial \theta_t}$$

where  $\theta_s$  and  $\theta_t$  are the  $s^{\text{th}}$  and  $t^{\text{th}}$  parameters of  $\mathbf{R}$ . Therefore,

$$\begin{aligned} \mathbf{W}_0(\mathbf{A}, \mathbf{B}) &= \mathbf{A}^T \mathbf{R}^{-1} \mathbf{B} \\ \mathbf{W}_0^{(1)s}(\mathbf{A}, \mathbf{B}) &= \mathbf{A}^T \mathbf{R}^{-1} \dot{\mathbf{R}}_s \mathbf{R}^{-1} \mathbf{B} \\ &= -\mathbf{A}^T \left[ \frac{\partial}{\partial \theta_s} \mathbf{R}^{-1}(\boldsymbol{\theta}) \right] \mathbf{B} \\ \mathbf{W}_0^{(2)st}(\mathbf{A}, \mathbf{B}) &= \mathbf{A}^T [\mathbf{R}^{-1} \ddot{\mathbf{R}}_{st} \mathbf{R}^{-1} - \mathbf{R}^{-1} \dot{\mathbf{R}}_s \mathbf{R}^{-1} \dot{\mathbf{R}}_t \mathbf{R}^{-1} - \mathbf{R}^{-1} \dot{\mathbf{R}}_t \mathbf{R}^{-1} \dot{\mathbf{R}}_s \mathbf{R}^{-1}] \mathbf{B} \\ &= -\mathbf{A}^T \left[ \frac{\partial^2}{\partial \theta_s \partial \theta_t} \mathbf{R}^{-1}(\boldsymbol{\theta}) \right] \mathbf{B} \end{aligned}$$

The matrices  $\mathbf{A}$  and  $\mathbf{B}$  can be  $\mathbf{X}$ ,  $\mathbf{Z}$ ,  $\tilde{\mathbf{Z}}$  or  $\mathbf{r}$ , where

$$\tilde{\mathbf{Z}} = \mathbf{Z}\mathbf{M} = \mathbf{Z}(\mathbf{G}^{-1} + \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z})^{-1}, \text{ and}$$

$$\mathbf{r} = [\mathbf{I} - \mathbf{X}(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1}] \mathbf{y} = \mathbf{y} - \mathbf{X}\mathbf{b}_0$$

Remark: The matrix  $(\mathbf{G}^{-1} + \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z})^{-1}$  involved in  $\tilde{\mathbf{Z}}$  can be obtained by pre/post multiply  $(\mathbf{I} + \mathbf{L}^T \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{L})^{-1}$  by  $\mathbf{L}$  and  $\mathbf{L}^T$ .

Using these notations, the 1<sup>st</sup> derivatives of  $\ell_{\mathbf{k}}(\boldsymbol{\theta})$  with respect to a parameter in  $\mathbf{R}$  are as follows,

$$\begin{aligned}
[\mathbf{g}_1]_{\mathbf{R},s} &= \text{tr}(\mathbf{R}^{-1}\dot{\mathbf{R}}_s) - \text{tr}(\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{Z})\mathbf{M}) \\
[\mathbf{g}_2]_{\mathbf{R},s} &= -\mathbf{W}_0^{(1)s}(\mathbf{r}, \mathbf{r}) + 2\mathbf{W}_0(\mathbf{r}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{r}) \\
&\quad - \mathbf{W}_0(\mathbf{r}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{Z})\mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{r}) \\
[\mathbf{g}_3]_{\mathbf{R},s} &= -\text{tr}(\mathbf{H}_{\mathbf{R}3}^s)
\end{aligned} \tag{21}$$

To compute 2<sup>nd</sup> derivatives w.r.t.  $\boldsymbol{\theta}_s$  and  $\boldsymbol{\theta}_t$  (of  $\mathbf{R}$ ), we need to consider the following simplification factors.

$$\begin{aligned}
\mathbf{H}_{\mathbf{R}1}^{\text{st}} &= -\mathbf{R}^{-1}\dot{\mathbf{R}}_s\mathbf{R}^{-1}\dot{\mathbf{R}}_t + \mathbf{R}^{-1}\ddot{\mathbf{R}}_{st} \\
&\quad - \mathbf{W}_0^{(2)\text{st}}(\mathbf{Z}, \mathbf{Z})\mathbf{M} - \mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{Z})\mathbf{M}\mathbf{W}_0^{(1)t}(\mathbf{Z}, \mathbf{Z})\mathbf{M}
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_{\mathbf{R}2}^s &= \mathbf{W}_0^{(1)s}(\mathbf{X}, \mathbf{r}) + \mathbf{W}_0(\mathbf{X}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{Z})\mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{r}) \\
&\quad - \mathbf{W}_0(\mathbf{X}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{r}) - \mathbf{W}_0^{(1)s}(\mathbf{X}, \mathbf{Z})\mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{r})
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_{\mathbf{R}2}^{\text{st}} &= -\mathbf{W}_0^{(2)\text{st}}(\mathbf{r}, \mathbf{r}) \\
&\quad - \mathbf{W}_0(\mathbf{r}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(2)\text{st}}(\mathbf{Z}, \mathbf{Z})\mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{r}) \\
&\quad + 2\mathbf{W}_0(\mathbf{r}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(2)\text{st}}(\mathbf{Z}, \mathbf{r}) \\
&\quad - 2[\mathbf{W}_0(\mathbf{r}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{Z}) - \mathbf{W}_0^{(1)s}(\mathbf{r}, \mathbf{Z})]\mathbf{M} \\
&\quad \times [\mathbf{W}_0^{(1)t}(\mathbf{Z}, \mathbf{Z})\mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{r}) - \mathbf{W}_0^{(1)t}(\mathbf{Z}, \mathbf{r})]
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_{\mathbf{R}3}^s &= \mathbf{W}_0^{(1)s}(\mathbf{X}, \mathbf{X}) - \mathbf{W}_0(\mathbf{X}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{X}) \\
&\quad - \mathbf{W}_0^{(1)s}(\mathbf{X}, \mathbf{Z})\mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{X}) \\
&\quad + \mathbf{W}_0(\mathbf{X}, \tilde{\mathbf{Z}})\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{Z})\mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{X})
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_{R3}^{\text{st}} &= -\mathbf{W}_0^{(2)\text{st}}(\mathbf{X}, \mathbf{X}) \\
&\quad - \mathbf{W}_0(\mathbf{X}, \tilde{\mathbf{Z}}) \mathbf{W}_0^{(2)\text{st}}(\mathbf{Z}, \mathbf{Z}) \mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{X}) \\
&\quad + 2\mathbf{W}_0(\mathbf{X}, \tilde{\mathbf{Z}}) \mathbf{W}_0^{(2)\text{st}}(\mathbf{Z}, \mathbf{X}) \\
&\quad - 2[\mathbf{W}_0(\mathbf{X}, \tilde{\mathbf{Z}}) \mathbf{W}_0^{(1)\text{s}}(\mathbf{Z}, \mathbf{Z}) - \mathbf{W}_0^{(1)\text{s}}(\mathbf{X}, \mathbf{Z})]\mathbf{M} \\
&\quad \times [\mathbf{W}_0^{(1)\text{t}}(\mathbf{Z}, \mathbf{Z}) \mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{X}) - \mathbf{W}_0^{(1)\text{t}}(\mathbf{Z}, \mathbf{X})]
\end{aligned}$$

Based on these simplification terms, the entries of the Hessian matrices are given by

$$\begin{aligned}
[\mathbf{H}_1]_{R,\text{st}} &= \text{tr}(\mathbf{H}_{R1}^{\text{st}}) \\
[\mathbf{H}_2]_{R,\text{st}} &= \mathbf{H}_{R2}^{\text{st}} - 2(\mathbf{H}_{R2}^{\text{s}})^{\text{T}} \mathbf{P} \mathbf{H}_{R2}^{\text{t}} \\
[\mathbf{H}_3]_{R,\text{st}} &= \text{tr}(\mathbf{H}_{R3}^{\text{st}} \mathbf{P} - \mathbf{H}_{R3}^{\text{s}} \mathbf{P} \mathbf{H}_{R3}^{\text{t}} \mathbf{P})
\end{aligned} \tag{22}$$

## G&R cross derivatives

This section gives expressions for the 2<sup>nd</sup> derivatives of  $\ell_1, \ell_2$  and  $\ell_3$  with respect to a parameter  $\boldsymbol{\theta}_s$  in  $\mathbf{G}$  and a parameter  $\boldsymbol{\theta}_t$  in  $\mathbf{R}$ . First, we introduce the following simplification terms,

$$\begin{aligned}
\mathbf{H}_{GR1}^{\text{st}} &= -\mathbf{W}_0^{(1)\text{s}}(\mathbf{Z}, \mathbf{Z}) \dot{\mathbf{G}}_t \\
&\quad + 2\mathbf{W}_0^{1(\text{s})}(\mathbf{Z}, \mathbf{Z}) \dot{\mathbf{G}}_t \mathbf{W}_0(\mathbf{Z}, \tilde{\mathbf{Z}}) \\
&\quad - \mathbf{W}_0^{(1)\text{s}}(\mathbf{Z}, \mathbf{Z}) \mathbf{W}_0(\tilde{\mathbf{Z}}, \mathbf{Z}) \dot{\mathbf{G}}_t \mathbf{W}_0(\mathbf{Z}, \tilde{\mathbf{Z}})
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_{GR2}^{\text{st}} &= 2[\mathbf{W}_0^{(1)\text{s}}(\mathbf{r}, \mathbf{Z}) - \mathbf{W}_0(\mathbf{r}, \mathbf{Z}) \mathbf{M} \mathbf{W}_0^{(1)\text{s}}(\mathbf{Z}, \mathbf{Z})] \\
&\quad \times [\mathbf{M} \mathbf{W}_0(\mathbf{Z}, \mathbf{Z}) - \mathbf{I}] \dot{\mathbf{G}}_t [\mathbf{W}_0(\mathbf{Z}, \mathbf{Z}) \mathbf{M} - \mathbf{I}] \\
&\quad \times \mathbf{W}_0(\mathbf{Z}, \mathbf{r})
\end{aligned}$$

$$\begin{aligned} \mathbf{H}_{GR3}^{st} &= 2[\mathbf{W}_0^{(1)s}(\mathbf{X}, \mathbf{Z}) - \mathbf{W}_0(\mathbf{X}, \mathbf{Z})\mathbf{M}\mathbf{W}_0^{(1)s}(\mathbf{Z}, \mathbf{Z})] \\ &\quad \times [\mathbf{M}\mathbf{W}_0(\mathbf{Z}, \mathbf{Z}) - \mathbf{I}]\dot{\mathbf{G}}_i[\mathbf{W}_0(\mathbf{Z}, \mathbf{Z})\mathbf{M} - \mathbf{I}] \\ &\quad \times \mathbf{W}_0(\mathbf{Z}, \mathbf{X}) \end{aligned}$$

Based on these simplification terms, the second derivatives are given by

$$[\mathbf{H}_1]_{GR,st} = \text{tr}(\mathbf{H}_{GR1}^{st})$$

$$[\mathbf{H}_2]_{GR,st} = \mathbf{H}_{GR2}^{st} - 2(\mathbf{H}_{G2}^s)^T \mathbf{H}_{R2}^t \quad 23)$$

$$[\mathbf{H}_3]_{GR,st} = \text{tr}(\mathbf{H}_{GR3}^{st} \mathbf{P} - \mathbf{H}_{G3}^s \mathbf{P} \mathbf{H}_{R3}^t \mathbf{P})$$

## Gradient & Hessian of REML

The restricted log likelihood is given by

$$\begin{aligned} -2\ell_{REML}(\boldsymbol{\theta} | \mathbf{y}) &= \log |\mathbf{V}(\boldsymbol{\theta})| + \mathbf{r}(\boldsymbol{\theta})^T \mathbf{V}^{-1}(\boldsymbol{\theta}) \mathbf{r}(\boldsymbol{\theta}) \\ &\quad + \log |\mathbf{X}^T \mathbf{V}^{-1}(\boldsymbol{\theta}) \mathbf{X}| + (n-p) \log 2\pi \end{aligned}$$

where  $p$  is equal to the rank of  $\mathbf{X}$ .

Therefore the  $s^{\text{th}}$  element of the gradient vector is given by

$$[\mathbf{g}]_s = [\mathbf{g}_1]_s + [\mathbf{g}_2]_s + [\mathbf{g}_3]_s$$

and the  $(s,t)$ -th element of the Hessian matrix is given by

$$[\mathbf{H}]_{st} = [\mathbf{H}_1]_{st} + [\mathbf{H}_2]_{st} + [\mathbf{H}_3]_{st}$$

If scoring algorithm is used, the Hessian can be simplified to

$$[\mathbf{H}]_{st} = -[\mathbf{H}_1]_{st} + [\mathbf{H}_3]_{st}.$$

## Gradient & Hessian of MLE

The log likelihood is given by

$$\begin{aligned} -2\ell_{\text{MLE}}(\boldsymbol{\theta} \mid \mathbf{y}) &= \log |\mathbf{V}(\boldsymbol{\theta})| + \mathbf{r}(\boldsymbol{\theta})^T \mathbf{V}^{-1}(\boldsymbol{\theta}) \mathbf{r}(\boldsymbol{\theta}) \\ &\quad + n \log 2\pi \end{aligned}$$

Therefore the  $s^{\text{th}}$  element of the gradient vector is given by

$$[\mathbf{g}]_s = [\mathbf{g}_1]_s + [\mathbf{g}_2]_s$$

and the  $(s,t)^{\text{th}}$  element of the Hessian matrix is given by

$$[\mathbf{H}]_{st} = [\mathbf{H}_1]_{st} + [\mathbf{H}_2]_{st}.$$

If scoring algorithm is used the Hessian can be simplified to

$$[\mathbf{H}]_{st} = -[\mathbf{H}_1]_{st}.$$

It should be noted that the Hessian matrices for the scoring algorithm in both ML and REML are not ‘exact’. In order to speed up calculation, some second derivative terms are dropped. Therefore, they are only used in intermediate step of optimization but not for standard error calculations.

## Cross Product matrices

During the estimation we need to construct several cross product matrices in each iteration, namely:  $\mathbf{W}_0(\mathbf{X};\mathbf{y};\mathbf{Z})$ ,  $\mathbf{W}_1(\mathbf{X};\mathbf{y};\mathbf{Z})$ ,  $\mathbf{W}_0^A(\mathbf{X};\mathbf{y};\mathbf{Z})$ ,  $\mathbf{W}_1^A(\mathbf{X};\mathbf{y};\mathbf{Z})$ ,  $\mathbf{W}_{b0}(\mathbf{X};\mathbf{y})$ , and  $\mathbf{W}_{b1}(\mathbf{X};\mathbf{y})$ . The sweeping operator (see for example Goodnight (1979)) is used in constructing these matrices. Basically, the sweeping operator performs the following transformation

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{bmatrix} \Rightarrow \begin{bmatrix} \mathbf{A}^- & \mathbf{A}^- \mathbf{B} \\ -\mathbf{B}' \mathbf{A}^- & \mathbf{C} - \mathbf{B}' \mathbf{A}^- \mathbf{B} \end{bmatrix}.$$

The steps needed to construct these matrices are outlined below,

### STEP 1:

Construct

$$\mathbf{W}_0(\mathbf{X};\mathbf{y};\mathbf{Z}) = \begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{y} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \\ \mathbf{y}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{y}^T \mathbf{R}^{-1} \mathbf{y} & \mathbf{y}^T \mathbf{R}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{y} & \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \end{bmatrix} \quad (24)$$

### STEP 2:

Construct  $\mathbf{W}_0^A(\mathbf{X};\mathbf{y};\mathbf{Z})$  which is an augmented version of  $\mathbf{W}_0(\mathbf{X};\mathbf{y};\mathbf{Z})$ . It is given by the following expression.

$$\mathbf{W}_0^A(\mathbf{X};\mathbf{y};\mathbf{Z}) = \begin{bmatrix} \mathbf{I} + \mathbf{L}^T \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{L} & \mathbf{L}^T \mathbf{W}_0(\mathbf{Z}, \cdot) \\ \mathbf{W}_0(\cdot, \mathbf{Z}) \mathbf{L} & \mathbf{W}_0 \end{bmatrix} \quad (25)$$

where  $\mathbf{L}$  is the lower-triangular Cholesky root of  $\mathbf{G}$ , i.e.  $\mathbf{G} = \mathbf{L} \mathbf{L}^T$  and  $\mathbf{W}_0(\mathbf{Z}, \cdot)$  is the rows of  $\mathbf{W}_0$  corresponding to  $\mathbf{Z}$ .

**STEP 3:**

Sweeping  $\mathbf{W}_0^A(\mathbf{X}; \mathbf{y}; \mathbf{Z})$  by pivoting on diagonal elements in upper-left partition will give us the matrix  $\mathbf{W}_1^A(\mathbf{X}; \mathbf{y}; \mathbf{Z})$ , which is shown below.

$$\mathbf{W}_1^A(\mathbf{X}; \mathbf{y}; \mathbf{Z}) = \begin{bmatrix} \mathbf{W}_1^A(\mathbf{1}, \mathbf{1}) & \mathbf{W}_1^A(\mathbf{1}, \mathbf{1})\mathbf{L}^T\mathbf{W}(\mathbf{Z}, \cdot) \\ -\mathbf{W}_0(\cdot, \mathbf{Z})\mathbf{L}\mathbf{W}_1^A(\mathbf{1}, \mathbf{1}) & \mathbf{W}_1(\mathbf{X}; \mathbf{y}; \mathbf{Z}) \end{bmatrix} \quad (26)$$

where

$$\mathbf{W}_1(\mathbf{X}; \mathbf{y}; \mathbf{Z}) = \begin{bmatrix} \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{V}^{-1}\mathbf{Z} & \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{V}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z} & \mathbf{Z}^T\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{y}^T\mathbf{V}^{-1}\mathbf{X} & \mathbf{y}^T\mathbf{V}^{-1}\mathbf{Z} & \mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} \end{bmatrix} \quad (27)$$

and

$$\mathbf{W}_1^A(\mathbf{1}, \mathbf{1}) = (\mathbf{I} + \mathbf{L}^T\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\mathbf{L})^{-1}.$$

During the sweeping, if we accumulate the log of the  $i$ -th diagonal element just before  $i$ -th sweep, we will obtain  $\log |\mathbf{I} + \mathbf{L}^T\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\mathbf{L}| = \log |\mathbf{V}| - \log |\mathbf{R}|$  as a by-product. Thus, adding to this quantity by  $\log |\mathbf{R}|$  will give us  $\ell_1(\boldsymbol{\theta})$ .

**STEP 4:**

Consider the following submatrix  $\mathbf{W}_{b_0}(\mathbf{X}; \mathbf{y})$  of  $\mathbf{W}_1(\mathbf{X}; \mathbf{y}; \mathbf{Z})$ ,

$$\mathbf{W}_{b_0}(\mathbf{X}; \mathbf{y}) = \begin{bmatrix} \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{y}^T\mathbf{V}^{-1}\mathbf{X} & \mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} \end{bmatrix}. \quad (28)$$

Sweeping  $\mathbf{W}_{b_0}(\mathbf{X}; \mathbf{y})$  by pivoting on diagonal elements of  $\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}$  will give us

$$\mathbf{W}_{b_1}(\mathbf{X}; \mathbf{y}) = \begin{bmatrix} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} & \mathbf{b}_0 \\ \mathbf{b}_0^T & \ell_2(\boldsymbol{\theta}) \end{bmatrix} \quad (29)$$

where  $\mathbf{b}_0$  is an estimate of  $\boldsymbol{\beta}_0$  in current iteration. After this step, we will obtain  $\ell_2(\boldsymbol{\theta})$  and  $\ell_3(\boldsymbol{\theta}) = |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|$ .

## Reference

- Akaike, H. (1974), A New Look at the Statistical Model Identification, *IEEE Transaction on Automatic Control*, AC -19, 716 -723.
- Bozdogan, H. (1987), *Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions*, Psychometrika, 52, 345 - 370.
- Fai, A.H.T. and Cornelius, P.L. (1996), Approximate F-tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-plot Experiments, *Journal of Statistical Computation and Simulation*, 54, 363 -378.
- Giesbrecht, F.G. and Burns, J.C. (1985), Two-Stage Analysis Based on a Mixed Model: Large-sample Asymptotic Theory and Small-Sample Simulation Results, *Biometrics*, 41, 477 -486.
- Goodnight, J. H. (1979). A Tutorial on the SWEEP Operator. *The American Statistician*, 33(3), 149-158.
- Henderson, C.R. (1984), *Applications of Linear Models in Animal Breeding*, University of Guelph.
- Hurvich, C. M., and Tsai, C-L. (1989). *Regression and Time Series Model Selection in Small Samples*, Biometrika 76, 297-307.
- McLean, R.A. and Sanders, W.L. (1988), Approximating Degrees of Freedom for Standard Errors in Mixed Linear Models, *Proceedings of the Statistical Computing Section, American Statistical Association*, New Orleans, 50 -59.

**24** MIXED

Schwarz, G. (1978), *Estimating the Dimension of a Model*, *Annals of Statistics*, 6, 461 -464.

Wolfinger, R., Tobias, R., Sall J., (1994), Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM J. Sci. Comput.*, Vol.15, No. 6, pp. 1294-1310.