MIXED

Overview

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

Notation

The following notation is used throughout this chapter unless otherwise stated:

θ	Overall covariance parameter vector.
θ_G	A vector of covariance parameters associated with random effects.
θ_k	A vector of covariance parameters associated with the k^{th} random effect.
θ_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.
$V(\theta)$	The $n \times n$ covariance matrix for $\mathbf{y}$.
$\dot{V}_s(\theta)$	First derivative of $V(\theta)$ with respect to the s^{th} parameter in θ.
$\ddot{V}_{st}(\theta)$	Second derivative of $V(\theta)$ with respect to the s^{th} and t^{th} parameters in θ.
$R(\theta_R)$	The $n \times n$ covariance matrix for ε.
$\dot{R}_s(\theta_R)$	First derivative of $R(\theta_R)$ with respect to the s^{th} parameter in θ_R
$\ddot{R}_{st}(\theta_R)$	Second derivative of $R(\theta_R)$ with respect to the s^{th} and t^{th} parameters in θ_R.
$G(\theta_G)$	The covariance matrix of random effects.
$\dot{G}_s(\theta_G)$	First derivative of $G(\theta_G)$ with respect to the s^{th} parameter in θ_G.

$\ddot{G}_{st}(\theta_G)$	Second derivative of $G(\theta_G)$ with respect to the s^{th} and t^{th} parameters in θ_G.
$V_k(\theta_k)$	The covariance matrix of the k^{th} random effect for one random subject.
$\dot{V}_{k,s}(\theta_k)$	First derivative of $V_k(\theta_k)$ with respect to the s^{th} parameter in θ_k.
$\ddot{V}_{k,st}(\theta_k)$	Second derivative of $V_k(\theta_k)$ with respect to the s^{th} and t^{th} parameters in θ_k.
$\mathbf{y}$	$n \times 1$ vector of observed values of the dependent variable.
$\mathbf{X}$	$n \times p$ design matrix of fixed effects.
$\mathbf{Z}$	$n \times q$ design matrix of random effects.
$\mathbf{r}$	$n \times 1$ vector of residuals.
β	$p \times 1$ vector of fixed effects parameters.
γ	$q \times 1$ vector of random effects parameters.
ε	$n \times 1$ vector of residual error terms.
$\mathbf{W}_c$	$n \times n$ diagonal matrix of case weights.
$\mathbf{W}_{rw}$	$n \times n$ diagonal matrix of regression weights.

Model

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

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

In this model, we assume that ε is distributed as $N[\mathbf{0}, \mathbf{R}(\theta_R)]$ and γ is independently distributed as $N[\mathbf{0}, G(\theta_G)]$. Therefore $\mathbf{y}$ is distributed as $N[\mathbf{X}\beta, V(\theta)]$, where $V(\theta) = \mathbf{Z}G(\theta_G)\mathbf{Z}^T + R(\theta_R)$. The unknown parameters include the regression parameters in β and covariance parameters in θ. Estimation of these model parameters relies on the use of a Newton-Ralphson or scoring algorithm. When we use either algorithm for finding MLE or REML solutions, we need to compute $V^{-1}(\theta)$ and its derivatives with respect to θ, which are computationally infeasible for large n. Wolfinger et.al.(1994) discussed methods that can avoid direct computation of $V^{-1}(\theta)$. They tackled the problem by using the SWEEP algorithm and exploiting the block diagonal structure of $G(\theta_G)$ and $R(\theta_R)$. 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 $G(\theta_G)$ and this is the actual implementation of the algorithm.

If there are regression weights, the covariance matrix $\mathrm{R}(\theta_R)$ will be replaced by $\mathrm{R}^*(\theta_R) = \mathbf{W}_{rw}^{-1/2}\,\mathrm{R}(\theta_R)\mathbf{W}_{rw}^{-1/2}$. For simpler notations, we will assume that the weights are already included in the matrix $\mathrm{R}(\theta_R)$ 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 $\mathrm{R}(\theta_R)$ with scaled identity structure. In MIXED, only cases with positive case weight and regression weight will be included analysis.

Fixed Effects Parameterization

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

Random Effects Parameterization

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

$$\mathbf{Z} = \left[\mathbf{Z_1} \quad \mathbf{Z_2} \quad \cdots \quad \mathbf{Z}_K\right],$$

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 shown below:

$$\mathbf{Z}_k = \left[\mathbf{Z}_{k\mathbf{1}} \quad \mathbf{Z}_{k\mathbf{2}} \quad \cdots \quad \mathbf{Z}_{kS_k}\right],\ k = 1,\ldots,K.$$

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

Under this partition, the $\mathrm{G}(\theta_G)$ will be a block diagonal matrix which can be expressed as

$$\mathrm{G}(\theta_G) = \oplus_{k=1}^{K}\left[\mathbf{I}_{S_k} \otimes \mathrm{V}_k(\theta_k)\right].$$

It should also be noted that each random effect has its own parameter vector θ_k, $k = 1,\ldots,K$, and there are no functional constraints between elements in these parameter vectors. Thus $\theta_G = (\theta_1,\ldots,\theta_K)$.

Repeated Subjects:

When the REPEATED subcommand is used, $R(\theta_R)$ will be a block diagonal matrix where the i^{th} block is $R_i(\theta_R)$, $i = 1,\ldots,S_R$. That is,

$$R(\theta_R) = \oplus_{i=1}^{S_R} R_i(\theta_R)$$

The dimension of $R_i(\theta_R)$ will be equal to the number of cases in one repeated subject but all $R_i(\theta_R)$ share the same parameter vector θ_R.

Likelihood Functions

Recall that the –2 log-likelihood using maximum likelihood estimation (ML) is

$$-2\ell_{\text{MLE}}(\beta,\theta) = \log|V(\theta)| + r(\theta)^{\text{T}}\, V(\theta)^{-1}\, r(\theta) + n\log 2\pi \qquad \textbf{(2)}$$

and the –2 log-likelihood using restricted maximum likelihood estimation (REML) is

$$-2\ell_{\text{REML}}(\theta) = \log|V(\theta)| + r(\theta)^{\mathbf{T}}\, V(\theta)^{\mathbf{-1}}\, r(\theta) + \log|\mathbf{X'V^{-1}X}| + (n-p)\log 2\pi \qquad \textbf{(3)}$$

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 components of the likelihood functions are

$$\begin{aligned} \ell_1(\theta) &= \log|V(\theta)| \\ \ell_2(\theta) &= r(\theta)^{\text{T}}\, V^{-1}(\theta)\, r(\theta) \\ \ell_3(\theta) &= \log|\mathbf{X}^{\text{T}}\, V^{-1}(\theta)\mathbf{X}|. \end{aligned} \qquad \textbf{(4)}$$

Therefore, in each estimation iteration, we need to compute $\ell_1(\theta)$, $\ell_2(\theta)$ and $\ell_3(\theta)$ as well as their 1st and 2nd derivatives with respect to θ.

Newton & Scoring Algorithms

Covariance parameters in θ 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, as outlined below:

1. Compute the starting parameter values and initial log-likelihood (REML or ML).
2. Compute the gradient vector $\mathbf{g}$ and Hessian matrix $\mathbf{H}$ of the log-likelihood function using the previous iteration's estimate θ_{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 of the i^{th} iteration $\theta_i = \theta_{i-1} + \rho\mathbf{d}$.
6. Check to see if θ_i generates valid covariance matrices and improve the likelihood. If not, reduce ρ by half and repeat step (5). If this process is repeated for a pre-specified number of times and the stated conditions are still not satisfied, stop.
7. Check for convergence of the parameter estimates. 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. Parameter and log-likelihood convergence are further subdivided into absolute and relative. If we let ε be some given tolerance level and

$\theta_{s,i}$ be the s^{th} parameter in the i^{th} iteration,

ℓ_i be the log-likelihood in log-likelihood in the i^{th} iteration,

g_i be the gradient vector in the i^{th} iteration,

and H_i be the hessian matrix in the i^{th} iteration,

then the criteria can be written as follows,

Absolute parameter convergence: $\max_s | \theta_{s,i} - \theta_{s,i-1} | < \varepsilon$

Relative parameter convergence: $\max_s \left| \theta_{s,i} - \theta_{s,i-1} \right| < \varepsilon | \theta_{s,i-1} |$

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

Relative log-likelihood convergence: $| \ell_i - \ell_{i-1} | < \varepsilon | \ell_{i-1} |$

Absolute Hessian convergence: $\mathbf{g_i^T H_i^{-1} g_i} < \varepsilon$

Relative Hessian convergence: $\mathbf{g_i^T H_i^{-1} g_i} < \varepsilon | \ell_i |$

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 matrix. However, it is highly desirable to estimate the scale of the variance parameter. By ignoring the random effects, and assuming the residual errors are i.i.d. with variance σ^2, we can fit a GLM model and estimate σ^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 $\hat{\theta}$ (ML or REML) is asymptotically normally distributed. Its variance covariance matrix can be approximated by $-2H^{-1}$, where H is the Hessian matrix of the log-likelihood function evaluated at $\hat{\theta}$. 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 the uniform delta method (van der Vaart, 1998), these transformed estimates still have asymptotic normal distributions.

Suppose we are estimating a variance parameter σ^2 by $\hat{\sigma}_n^2$ that is distributed as $N[\sigma^2, \mathrm{Var}(\hat{\sigma}_n^2)]$ 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 the delta method, one can show that the asymptotic distribution of $\log(\hat{\sigma}_n^2)$ is $N[\log(\sigma^2), \sigma^{-4}\mathrm{Var}(\hat{\sigma}_n^2)]$. Thus, a $(1-\alpha)100\%$ confidence interval of $\log(\sigma^2)$ is given by

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

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 σ^2 is given by

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

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

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

Fixed and Random Effect Parameters

Estimation and prediction

After we obtain an estimate of θ, best linear unbiased estimator (BLUE) of β and best linear unbiased predictor (BLUP) of γ can be found by solving the mixed model equations, Henderson (1984).

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

The solution of (5) can be expressed as

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

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

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

where

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

Custom Hypotheses

In general, one can construct estimators or predictors for

$$\mathbf{Lb} = \begin{bmatrix}\mathbf{L_0} & \mathbf{L_1}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{\gamma}\end{bmatrix} \qquad \textbf{(8)}$$

for some hypothesis matrix $\mathbf{L}$. Estimators or predictors of $\mathbf{Lb}$ can easily be constructed by substituting $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ into (8) and its variance covariance matrix can be approximated by $\mathbf{LCL}^{\mathbf{T}}$. If $\mathbf{L_1}$ is zero and $\mathbf{L_0}\boldsymbol{\beta}$ is estimable, $\mathbf{L}\hat{\mathbf{b}}$ 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

$$\mathbf{F} = \frac{\left(\mathbf{L}\hat{\mathbf{b}} - \mathbf{a}\right)^{\mathrm{T}}(\mathbf{L}\hat{\mathbf{C}}\mathbf{L}^{\mathbf{T}})^{-1}\left(\mathbf{L}\hat{\mathbf{b}} - \mathrm{a}\right)}{\mathrm{q}} \qquad \textbf{(9)}$$

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 $\mathbf{L}\hat{\mathbf{C}}\mathbf{L}^{\mathbf{T}} = \Gamma^{\mathbf{T}}\mathbf{D}\Gamma$ where Γ is an orthogonal matrix of eigenvectors and $\mathbf{D}$ is a diagonal matrix of eigenvalues. If we let $\ell_{\mathbf{m}}$ be the the m^{th} row of $\Gamma\mathbf{L}$, d_{m} be the m^{th} eigenvalues and

$$\nu_{\mathrm{m}} = \frac{2\mathrm{d}_{\mathrm{m}}{}^{2}}{\mathbf{g_m}^{\mathbf{T}}\Sigma(\hat{\theta})^{-1}\mathbf{g_m}}$$

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

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

then the denominator degrees of freedom is given by

$$\nu = \frac{2\,\mathrm{E}}{\mathrm{E} - \mathrm{q}}.$$

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.

Saved Values

If predicted values are requested, they are computed by

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

using the estimates given in (6).

If fixed predicted values are requested, they are computed by

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

If residuals are requested, they are computed by

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

Information Criteria

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

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

  $-2\ell + 2\mathrm{d}$

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

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

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

  $-2\ell + \mathrm{d} \times \log(\mathrm{n})$

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

  $-2\ell + \mathrm{d} \times (\log(\mathrm{n}) + 1)$

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

1st and 2nd Derivatives of $\ell_k(\theta)$

In each Newton or scoring iteration we need to compute the 1st and 2nd derivatives of the components of the log-likelihood $\ell_k(\theta)$, k=1,2,3. Here we let $\mathbf{g_k} = \frac{\partial}{\partial\theta}\ell_{\mathbf{k}}(\theta)$ and $\mathbf{H_k} = \frac{\partial^2}{\partial\theta\partial\theta}\ell_{\mathbf{k}}(\theta)$, $\mathrm{k} = 1,2,3$, then the 1st derivatives with respect to the s^{th} parameter in θ are given by

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

and the 2nd derivatives with respect to the s^{th} and t^{th} parameters are given by

$$
[\mathbf{H}_1]_{st} = -\mathbf{tr}(\mathbf{V}^{-1}\dot{\mathbf{V}}_s\mathbf{V}^{-1}\dot{\mathbf{V}}_t) + \mathbf{tr}(\mathbf{V}^{-1}\ddot{\mathbf{V}}_{st}),
$$

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

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

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

Derivatives w.r.t. Parameters in G

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

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

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 β.

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

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

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

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

In terms of the elements in $\mathbf{W}_1(\mathbf{X};\mathbf{r};\mathbf{Z})$ matrix, we can write down the 1st derivatives of ℓ_1, ℓ_2 and ℓ_3 with respect to a parameter θ_s of the **G** matrix,

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

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

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

then second derivatives of ℓ_1, ℓ_2 and ℓ_3 w.r.t. θ_s and θ_t (in **G**) are given by

$$\begin{aligned}[\mathbf{H}_1]_{\mathbf{G},\mathbf{st}} &= \mathbf{tr}(\mathbf{H}^{\mathbf{st}}_{\mathbf{G1}}) \\ [\mathbf{H}_2]_{\mathbf{G},\mathbf{st}} &= \mathbf{H}^{\mathbf{st}}_{\mathbf{G2}} - 2(\mathbf{H}^{\mathbf{s}}_{\mathbf{G2}})^{\mathbf{T}}\mathbf{P}\,\mathbf{H}^{\mathbf{t}}_{\mathbf{G2}} \\ [\mathbf{H}_3]_{\mathbf{G},\mathbf{st}} &= \mathbf{tr}[\mathbf{H}^{\mathbf{st}}_{\mathbf{G3}}\,\mathbf{P}] - \mathbf{tr}[\mathbf{H}^{\mathbf{s}}_{\mathbf{G3}}\,\mathbf{P}\,\mathbf{H}^{\mathbf{t}}_{\mathbf{G3}}\,\mathrm{P}]\end{aligned} \tag{20}$$

Derivatives w.r.t. Parameters in R

To compute **R** derivatives, we need to introduce the matrices

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

and

$$\mathbf{W}_0^{(2)st} = -\frac{\partial^2 \mathbf{W}_0}{\partial \boldsymbol{\theta}_s \partial \boldsymbol{\theta}_t}$$

where θ_s and θ_t are the s[th] and t[th] parameters of R **.** Therefore,

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

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

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

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

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

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

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

To compute 2nd derivatives w.r.t. $\theta_{\mathbf{s}}$ and $\theta_{\mathbf{t}}$ (of $\mathbf{R}$), we need to consider the following simplification factors.

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

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

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

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

$$
\begin{aligned}
\mathbf{H}_{\mathbf{R}3}^{\mathbf{st}} &= -\mathbf{W}_0^{(2)\mathbf{st}}(\mathbf{X},\mathbf{X}) \\
&\quad - \mathbf{W}_0(\mathbf{X},\tilde{\mathbf{Z}})\ \mathbf{W}_0^{(2)\mathbf{st}}(\mathbf{Z},\mathbf{Z})\ \mathbf{W}_0(\tilde{\mathbf{Z}},\mathbf{X}) \\
&\quad + 2\mathbf{W}_0(\mathbf{X},\tilde{\mathbf{Z}})\ \mathbf{W}_0^{(2)\mathbf{st}}(\mathbf{Z},\mathbf{X}) \\
&\quad - 2[\mathbf{W}_0(\mathbf{X},\tilde{\mathbf{Z}})\ \mathbf{W}_0^{(1)\mathbf{s}}(\mathbf{Z},\mathbf{Z}) - \mathbf{W}_0^{(1)\mathbf{s}}(\mathbf{X},\mathbf{Z})]\mathbf{M} \\
&\quad \times [\mathbf{W}_0^{(1)\mathbf{t}}(\mathbf{Z},\mathbf{Z})\ \mathbf{W}_0(\tilde{\mathbf{Z}},\mathbf{X}) - \mathbf{W}_0^{(1)\mathbf{t}}(\mathbf{Z},\mathbf{X})]
\end{aligned}
$$

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

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

G&R cross derivatives

These section gives expressions for the 2nd derivatives of ℓ_1, ℓ_2 and ℓ_3 with respect to a parameter $\theta_{\mathbf{s}}$ in **G** and a parameter $\theta_{\mathbf{t}}$ in **R**. First, we introduce the following simplification terms,

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

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

$$\begin{aligned}\mathbf{H}_{\mathbf{GR}3}^{\mathbf{st}} &= 2[\mathbf{W}_0^{(1)\mathbf{s}}(\mathbf{X},\mathbf{Z}) - \mathbf{W}_0(\mathbf{X},\mathbf{Z})\mathbf{M}\mathbf{W}_0^{(1)\mathbf{s}}(\mathbf{Z},\mathbf{Z})] \\ &\times [\mathbf{M}\mathbf{W}_0(\mathbf{Z},\mathbf{Z}) - \mathbf{I}]\dot{\mathbf{G}}_{\mathbf{t}}[\mathbf{W}_0(\mathbf{Z},\mathbf{Z})\mathbf{M} - \mathbf{I}] \\ &\times \mathbf{W}_0(\mathbf{Z},\mathbf{X})\end{aligned}$$

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

$$[\mathbf{H}_1]_{\mathbf{GR},\mathbf{st}} = \mathbf{tr}(\mathbf{H}_{\mathbf{GR}1}^{\mathbf{st}})$$

$$[\mathbf{H}_2]_{\mathbf{GR},\mathbf{st}} = \mathbf{H}_{\mathbf{GR}2}^{\mathbf{st}} - 2(\mathbf{H}_{\mathbf{G}2}^{\mathbf{s}})^{\mathbf{T}}\mathbf{H}_{\mathbf{R}2}^{\mathbf{t}} \qquad \textbf{23)}$$

$$[\mathbf{H}_3]_{\mathbf{GR},\mathbf{st}} = \mathbf{tr}(\mathbf{H}_{\mathbf{GR}3}^{\mathbf{st}}\,\mathbf{P} - \mathbf{H}_{\mathbf{G}3}^{\mathbf{s}}\,\mathbf{P}\,\mathbf{H}_{\mathbf{R}3}^{\mathbf{t}}\,\mathbf{P})$$

Gradient & Hessian of REML

The restricted log likelihood is given by

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

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

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

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

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

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

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

$$[\mathbf{H}]_{\mathbf{st}} = -[\mathbf{H}_{\mathbf{1}}]_{\mathbf{st}} + [\mathbf{H}_{\mathbf{3}}]_{\mathrm{st}} .$$

Gradient & Hessian of MLE

The log likelihood is given by

$$-2\ell_{\mathbf{MLE}}(\theta \mid \mathbf{y}) = \log \mid \mathbf{V}(\theta) \mid + \mathbf{r}(\theta)^{\mathbf{T}}\mathbf{V}^{-1}(\theta)\mathbf{r}(\theta) + n \log 2\pi$$

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

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

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

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

If scoring algorithm is used the Hessian can be simplified to

$$[\mathbf{H}]_{\mathbf{st}} = -[\mathbf{H}_{\mathbf{1}}]_{\mathbf{st}} .$$

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: $\mathbf{W_0(X;y;Z)}$, $\mathbf{W_1(X;y;Z)}$, $\mathbf{W_0^A(X;y;Z)}$, $\mathbf{W_1^A(X;y;Z)}$, $\mathbf{W_{b0}(X;y)}$, and $\mathbf{W_{b1}(X;y)}$. The SWEEP operator (see for example Goodnight (1979)) is used in constructing these matrices. Basically, the SWEEP operator performs the following transformation

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

The steps needed to construct these matrices are outlined below,

STEP 1:

Construct

$$\mathbf{W_0(X;y;Z)} = \begin{bmatrix} \mathbf{X^TR^{-1}X} & \mathbf{X^TR^{-1}y} & \mathbf{X^TR^{-1}Z} \\ \mathbf{y^TR^{-1}X} & \mathbf{y^TR^{-1}y} & \mathbf{y^TR^{-1}Z} \\ \mathbf{Z^TR^{-1}X} & \mathbf{Z^TR^{-1}Z} & \mathbf{y^TR^{-1}y} \end{bmatrix} \qquad \textbf{(24)}$$

STEP 2:

Construct $\mathbf{W_0^A(X;y;Z)}$ which is an augmented version of $\mathbf{W_0(X;y;Z)}$. It is given by the following expression.

$$\mathbf{W_0^A(X;y;Z)} = \begin{bmatrix} \mathbf{I + L^TZ^TR^{-1}ZL} & \mathbf{L^TW_0(Z,\cdot)} \\ \mathbf{W_0(\cdot,Z)L} & \mathbf{W_0} \end{bmatrix} \qquad \textbf{(25)}$$

where **L** is the lower-triangular Cholesky root of **G**, i.e. $\mathbf{G=LL^T}$ and $\mathbf{W_0(Z,\,.\,)}$ is the rows of $\mathbf{W_0}$ corresponding to **Z.**

STEP 3:

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

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

where

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

and

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

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

STEP 4:

Consider the following submatrix $\mathbf{W}_{\mathbf{b}0}(\mathbf{X};\mathbf{y})$ of $\mathbf{W}_1(\mathbf{X};\mathbf{y};\mathbf{Z})$,

$$\mathbf{W}_{\mathbf{b}0}(\mathbf{X};\mathbf{y}) = \begin{bmatrix} \mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X} & \mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{y}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X} & \mathbf{y}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{y} \end{bmatrix}. \qquad \textbf{(28)}$$

Sweeping $\mathbf{W_{b0}(X;y)}$ by pivoting on diagonal elements of $\mathbf{X^TV^{-1}X}$ will give us

$$\mathbf{W_{b1}(X;y)} = \begin{bmatrix} (\mathbf{X^TV^{-1}X}\ )^- & \mathbf{b_0} \\ \mathbf{b_0^T} & \ell_2(\theta) \end{bmatrix} \qquad \textbf{(29)}$$

where $\mathbf{b_0}$ is an estimate of β_0 in current iteration. After this step, we will obtain $\ell_2(\theta)$ and $\ell_3(\theta) = |\ \mathbf{X^TV^{-1}X}\ |$.

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.

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.