



A recap on Linear Mixed Models and their hat-matrices

Gianfranco Lovison · Mariangela Sciandra

Abstract This working paper has a twofold goal. On one hand, it provides a recap of Linear Mixed Models (LMMs): far from trying to be exhaustive, this first part of the working paper focusses on the derivation of theoretical results on estimation of LMMs that are scattered in the literature or whose mathematical derivation is sometimes missing or too quickly sketched. On the other hand, it discusses various definitions that are available in the literature for the hat-matrix of Linear Mixed Models, showing their limitations and proving their equivalence.

Keywords Linear Mixed Models · Inference · Hat matrices · Orthogonal Projectors

Riassunto *Questo working paper ha un doppio obiettivo. Da un lato, fornisce un riepilogo sui Modelli Lineari Misti (MLM): lungi dal tentare di essere esaustiva, questa prima parte del working paper si focalizza sulla derivazione di risultati teorici sulla stima dei MLM che sono sparsi in letteratura o la cui giustificazione matematica è a volte mancante o abbozzata troppo frettolosamente. Questa prima parte si articola come segue: dopo aver specificato il Modello Lineare Misto, anche nella sua utile forma "compatta" (stacked), ed aver introdotto i necessari risultati distributivi, vengono derivati formalmente gli stimatori degli effetti fissi e i predittori degli effetti casuali, sia utilizzando l'approccio marginale che quello congiunto. Dapprima, queste derivazioni vengono ottenute sotto l'assunto (poco realistico) che le matrici di varianze/covarianze sia degli*

G. Lovison: Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo,
viale delle Scienze ed. 13, 90128

E-mail: gianfranco.lovison@unipa.it

· M. Sciandra Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo,
viale delle Scienze ed. 13, 90128

E-mail: mariangela.sciandra@unipa.it

errori sia degli effetti casuali siano note; successivamente tale assunto non viene specificato, e si presentano i risultati riguardanti la stima di tali matrici di varianze/covarianze, sia con il metodo ML che con quello REML. Infine, viene discussa la rappresentazione dei MLM mediante “dati aumentati”, dovuta a Hodges (1998), utile per gli sviluppi successivi.

Nella seconda parte, il working paper discute varie definizioni della matrice di proiezione (hat-matrix) che sono disponibili nella letteratura sui Modelli Lineari Misti, evidenziando le loro limitazioni e dimostrandone formalmente, per la prima volta, l'equivalenza. Viene inoltre evidenziato come l'unica matrice di proiezione ortogonale, e dunque simmetrica ed idempotente, sia quella ottenibile dalla rappresentazione di Hodges, un risultato utile per ulteriori sviluppi di ricerca.

Parole chiave Modelli Lineari Misti - Inferenza - Matrici di proiezione - Proiettori ortogonali.

1 The Linear Mixed Models: specification

Let the data have the following structure:

$$y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij} \quad i = 1, \dots, m; \quad j = 1, \dots, n_i; \quad n = \sum_{i=1}^m n_i \quad (1)$$

where: i is the cluster index

j is the individual (within cluster) unit index

y_{ij} is the response variable

\mathbf{x}_{ij} is a vector of p explanatory variables (with fixed parameters)

\mathbf{z}_{ij} is a vector of q explanatory variables (with random parameters)

For the ease of presentation, we assume in this paper $n_i = k \quad \forall i$, so that $n = km$.

We can arrange the data according to the clustered structure as:

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{i,k} \end{bmatrix}, \quad \mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1}^T \\ \vdots \\ \mathbf{x}_{ij}^T \\ \vdots \\ \mathbf{x}_{i,k}^T \end{bmatrix}, \quad \mathbf{Z}_i = \begin{bmatrix} \mathbf{z}_{i1}^T \\ \vdots \\ \mathbf{z}_{ij}^T \\ \vdots \\ \mathbf{z}_{i,k}^T \end{bmatrix} \quad i = 1, \dots, m \quad (2)$$

Suppose the data are generated by a Gaussian Linear Mixed Model (Breslow and Clayton, 1993), specified at cluster level as follows:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i \quad i = 1, \dots, m \quad (3)$$

where: \mathbf{X}_i is a $k \times p$ matrix, $\boldsymbol{\beta}$ is a p -vector of unknown fixed parameters, \mathbf{Z}_i is a $k \times q$ matrix and \mathbf{b}_i is a q -vector of random parameters.

As far as the random components of the model are concerned, we assume that both the random parameters \mathbf{b}_i and the within-cluster errors $\boldsymbol{\epsilon}_i$ are Normally distributed, with variance/covariance matrices which are full-rank unconstrained positive-definite matrices; besides, we assume that the within-cluster errors (conditional) variance/covariance matrices are homogeneous across clusters:

$$\mathbf{b}_i \sim \mathcal{MVN}_q(\mathbf{0}_q, \boldsymbol{\Sigma}_{B_c}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{MVN}_k(\mathbf{0}_k, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c}), \quad \mathbf{b}_i \perp\!\!\!\perp \boldsymbol{\epsilon}_i \quad i = 1, \dots, m$$

where:

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c} = \mathcal{D}[\boldsymbol{\epsilon}_i] \quad \forall i; \quad \boldsymbol{\Sigma}_{B_c} = \mathcal{D}[\mathbf{b}_i]$$

are positive-definite matrices (the index "c" stands here for "at cluster level"). Notice that these assumptions imply that the conditional distribution of the response \mathbf{y}_i , given the random parameters \mathbf{b}_i , is:

$$\mathbf{y}_i | \mathbf{b}_i \sim \mathcal{MVN}_k(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c}) \quad i = 1, \dots, m$$

with:

$$\mathcal{D}[\mathbf{y}_i | \mathbf{b}_i] = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c} \quad \forall i$$

For the subsequent developments, it is convenient to write model (3) in vectorised form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \quad (4)$$

where:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_i \\ \vdots \\ \mathbf{y}_m \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_m \end{bmatrix}, \quad \mathbf{Z} = \bigoplus_{i=1}^m \mathbf{Z}_i = \begin{bmatrix} \mathbf{Z}_1 & \dots & \mathbf{O} & \dots & \mathbf{O} \\ \vdots & & \vdots & & \vdots \\ \mathbf{O} & \dots & \mathbf{Z}_i & \dots & \mathbf{O} \\ \vdots & & \vdots & & \vdots \\ \mathbf{O} & \dots & \mathbf{O} & \dots & \mathbf{Z}_m \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_i \\ \vdots \\ \mathbf{b}_m \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_i \\ \vdots \\ \boldsymbol{\epsilon}_m \end{bmatrix},$$

This representation is called *the stacked form* of the data and the model; here \mathbf{y} and $\boldsymbol{\epsilon}$ are n -dimensional vector, \mathbf{b} is r -dimensional, where $r = m \times q$ is the total number of realisations of the random vector \mathbf{b}_i , \mathbf{X} is $n \times p$, \mathbf{Z} is $n \times r$.

The dispersion matrices for \mathbf{b} and $\boldsymbol{\epsilon}$ (which have dimension $r \times r$ and $n \times n$ respectively), can be written in compact form as:

$$\mathcal{D}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \mathbf{I}_m \otimes \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c}; \quad \mathcal{D}[\mathbf{b}] = \boldsymbol{\Sigma}_B = \mathbf{I}_m \otimes \boldsymbol{\Sigma}_{B_c}$$

whence, the distributional assumptions on $\boldsymbol{\epsilon}$, \mathbf{b} and $\mathbf{y}|\mathbf{b}$ become:

$$\boldsymbol{\epsilon} \sim \mathcal{MVN}_n(\mathbf{0}_n, \boldsymbol{\Sigma}_\epsilon), \quad \mathbf{b} \perp\!\!\!\perp \boldsymbol{\epsilon} \quad (5)$$

$$\mathbf{b} \sim \mathcal{MVN}_r(\mathbf{0}_r, \boldsymbol{\Sigma}_B) \quad (6)$$

$$\mathbf{y}|\mathbf{b} \sim \mathcal{MVN}_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma}_\epsilon) \quad (7)$$

2 Relevant distributions

In what follows, we shall need the following distributions:

- the joint distribution $f(\mathbf{y}, \mathbf{b}; \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_B)$

The best way to derive the joint distribution of \mathbf{y} and \mathbf{b} is directly from the specification of the model, which gives us the conditional distribution $f(\mathbf{y}|\mathbf{b})$ in (7) and the marginal distribution $f(\mathbf{b})$ in (6):

$$\begin{aligned} f(\mathbf{y}, \mathbf{b}; \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_B) &= f(\mathbf{y}|\mathbf{b}; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon) f(\mathbf{b}; \boldsymbol{\Sigma}_B) \\ &= |2\pi\boldsymbol{\Sigma}_\epsilon|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})\right\} \\ &\quad \times |2\pi\boldsymbol{\Sigma}_B|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{b})^T \boldsymbol{\Sigma}_B^{-1} (\mathbf{b})\right\} \\ &= (2\pi)^{-\frac{n+r}{2}} |\boldsymbol{\Sigma}_\epsilon|^{-1/2} |\boldsymbol{\Sigma}_B|^{-1/2} \times \\ &\quad \exp\left\{-\frac{1}{2}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{b}^T \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z}\mathbf{b} \right. \\ &\quad \left. + \mathbf{b}^T \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z}\mathbf{b} - \mathbf{b}^T \boldsymbol{\Sigma}_B^{-1} \mathbf{b}]\right\} \end{aligned} \quad (8)$$

Focussing first on the exponential argument in (9), we see that it can be written as a quadratic form:

$$\begin{bmatrix} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T & \mathbf{b}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_\epsilon^{-1} & -\boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \\ -\mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} & \boldsymbol{\Sigma}_B^{-1} + \mathbf{Z} \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z}^T \end{bmatrix} \begin{bmatrix} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \mathbf{b} \end{bmatrix}$$

Using standard results concerning the inverse of partitioned matrices, one gets:

$$\begin{bmatrix} \boldsymbol{\Sigma}_\epsilon^{-1} & -\boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \\ -\mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} & \boldsymbol{\Sigma}_B^{-1} + \mathbf{Z} \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z}^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_\epsilon + \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T & \mathbf{Z} \boldsymbol{\Sigma}_B \\ \boldsymbol{\Sigma}_B \mathbf{Z}^T & \boldsymbol{\Sigma}_B \end{bmatrix}^{-1} = \boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{b}}^{-1} \quad (10)$$

As for the product of determinants in (9), using a known result concerning the determinant of partitioned matrices:

$$|\boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{b}}| = \left| \begin{bmatrix} \boldsymbol{\Sigma}_\epsilon + \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T & \mathbf{Z} \boldsymbol{\Sigma}_B \\ \boldsymbol{\Sigma}_B \mathbf{Z}^T & \boldsymbol{\Sigma}_B \end{bmatrix} \right| = |\boldsymbol{\Sigma}_B| |\boldsymbol{\Sigma}_\epsilon + \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T - \mathbf{Z} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_B^{-1} \boldsymbol{\Sigma}_B \mathbf{Z}^T| = |\boldsymbol{\Sigma}_B| |\boldsymbol{\Sigma}_\epsilon| \quad (11)$$

Substituting (10) and (11) into (9), we find:

$$f(\mathbf{y}, \mathbf{b}; \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_B) = (2\pi)^{-\frac{n+r}{2}} |\boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{b}}|^{-1/2} \exp \left\{ -\frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \ \mathbf{b}^T] \boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{b}}^{-1} \begin{bmatrix} \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \\ \mathbf{b} \end{bmatrix} \right\} \quad (12)$$

i.e.:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{MVN}_{n+r} \left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0}_r \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_\epsilon + \mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T & \mathbf{Z}\boldsymbol{\Sigma}_B \\ \boldsymbol{\Sigma}_B\mathbf{Z}^T & \boldsymbol{\Sigma}_B \end{bmatrix} \right) \quad (13)$$

– the marginal distribution $f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_B)$

Using standard results on the Multivariate Normal distribution, we readily obtain from (13) that the marginal distribution of \mathbf{y} is $\mathbf{y} \sim \mathcal{MVN}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\mathbf{y}})$, with $\boldsymbol{\Sigma}_{\mathbf{y}} = \mathcal{D}[\mathbf{y}] = \mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon$, i.e.:

$$f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_B) = |2\pi \boldsymbol{\Sigma}_{\mathbf{y}}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \quad (14)$$

– the conditional distribution $f(\mathbf{b}|\mathbf{y})$

Again using standard results on the Multivariate Normal distribution (Stein, 1981), and in particular on the conditional distribution of a Normal sub-vector given another Normal sub-vector, we obtain from (13) that the conditional distribution of $\mathbf{b}|\mathbf{y}$ is

$$\mathbf{b}|\mathbf{y} \sim \mathcal{MVN}_r \left(\boldsymbol{\Sigma}_B\mathbf{Z}^T (\mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Sigma}_B - \boldsymbol{\Sigma}_B\mathbf{Z}^T (\mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{Z}\boldsymbol{\Sigma}_B \right) \quad (15)$$

or, applying to the variance/covariance matrix formula (69) in Appendix 1 for the inverse of a Schur complement (Zhang, 2006):

$$\mathbf{b}|\mathbf{y} \sim \mathcal{MVN}_r \left(\boldsymbol{\Sigma}_B\mathbf{Z}^T (\mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), (\boldsymbol{\Sigma}_B^{-1} + \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z})^{-1} \right). \quad (16)$$

For the sake of notational brevity, we shall sometimes use the symbol

$$\boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} = (\boldsymbol{\Sigma}_B^{-1} + \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z})^{-1}.$$

3 Marginal ML estimation of $\boldsymbol{\beta}$ and Empirical Bayesian prediction of \mathbf{b}

As shown in (14), the marginal distribution of \mathbf{y} is $\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\mathbf{y}})$, with $\boldsymbol{\Sigma}_{\mathbf{y}} = \mathcal{D}[\mathbf{y}] = \mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon$. Hence, the marginal likelihood is:

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_B; \mathbf{y}, \mathbf{X}, \mathbf{Z}) = (2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_{\mathbf{y}}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}$$

and the marginal log-likelihood can therefore be written:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_B; \mathbf{y}, \mathbf{X}, \mathbf{Z}) = -\left(\frac{n}{2}\right) \log(2\pi) - \left(\frac{1}{2}\right) \log(|\boldsymbol{\Sigma}_{\mathbf{y}}|) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (17)$$

3.1 With known $\Sigma_{\epsilon}, \Sigma_B$

When $\Sigma_{\epsilon}, \Sigma_B$ are known, $\Sigma_{\mathbf{y}}$ is of course also known, and estimation of β reduces to the solution of a weighted least squares system of equations:

$$\mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \mathbf{X} \beta = \mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \mathbf{y} \quad (18)$$

or:

$$\mathbf{X}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} \mathbf{X} \beta = \mathbf{X}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} \mathbf{y} \quad (19)$$

which yields the (marginal) Maximum Likelihood estimator:

$$\hat{\beta} = [\mathbf{X}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} \mathbf{y} \quad (20)$$

Clearly, the use of the marginal likelihood rules out the possibility of direct prediction of the random parameters \mathbf{b}_i , $i = 1, \dots, m$, since such marginal likelihood is obtained exactly integrating over these random parameters.

The most popular approach for predicting \mathbf{b} is therefore an *empirical Bayesian* one. The necessary ingredient for such an approach is the *posterior distribution* $f(\mathbf{b}|\mathbf{y})$, i.e. the conditional distribution of the random effects realisations given the observations \mathbf{y} . This was shown in equation (16) to be Multivariate Normal, with expected value $E[\mathbf{b}|\mathbf{y}] = \Sigma_B \mathbf{Z}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} (\mathbf{y} - \mathbf{X} \beta)$. Therefore, (point) prediction of \mathbf{b} can be carried out by estimating the posterior mode (or mean, which coincides with the mode owing to Multivariate Normality):

$$\tilde{\mathbf{b}} = E[\widehat{\mathbf{b}}|\mathbf{y}] = \Sigma_B \mathbf{Z}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} (\mathbf{y} - \mathbf{X} \hat{\beta}) \quad (21)$$

Owing to the approach used to derive it, the predictor (21) is often called the *Empirical Bayesian predictor* and denoted in the literature with the acronym EBP (Ando, 2007).

4 Joint ML estimation

From (8) the joint likelihood of $\beta, \mathbf{b}, \Sigma_{\epsilon}, \Sigma_B$ is:

$$\begin{aligned} L(\beta, \mathbf{b}, \Sigma_{\epsilon}, \Sigma_B; \mathbf{y}, \mathbf{X}, \mathbf{Z}) &= L(\beta, \Sigma_{\epsilon}|\mathbf{b}; \mathbf{y}, \mathbf{X}, \mathbf{Z}) L(\mathbf{b}, \Sigma_B) = |2\pi \Sigma_{\epsilon}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})^T \Sigma_{\epsilon}^{-1} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})\right\} \\ &\quad \times |2\pi \Sigma_B|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{b})^T \Sigma_B^{-1} (\mathbf{b})\right\} \end{aligned}$$

and the joint log-likelihood can be written:

$$\begin{aligned} \ell(\beta, \mathbf{b}, \Sigma_{\epsilon}, \Sigma_B; \mathbf{y}, \mathbf{X}, \mathbf{Z}) &= \ell(\beta, \Sigma_{\epsilon}|\mathbf{b}; \mathbf{y}, \mathbf{X}, \mathbf{Z}) + \ell(\mathbf{b}, \Sigma_B) \\ &= -\left(\frac{n}{2}\right) \log(2\pi) - \left(\frac{1}{2}\right) \log(|\Sigma_{\epsilon}|) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})^T \Sigma_{\epsilon}^{-1} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}) \\ &\quad - \left(\frac{r}{2}\right) \log(2\pi) - \left(\frac{1}{2}\right) \log(|\Sigma_B|) - \frac{1}{2} \mathbf{b}^T \Sigma_B^{-1} \mathbf{b} \end{aligned}$$

4.1 With known Σ_{ϵ} , Σ_B

If Σ_{ϵ} , Σ_B are known, it is sufficient to differentiate with respect to β and \mathbf{b} to find the score vectors:

$$\begin{aligned}\mathbf{u}(\beta) &= \frac{\partial \ell(\beta, \mathbf{b}; \mathbf{y}, \mathbf{X}, \mathbf{Z}, \Sigma_{\epsilon}, \Sigma_B)}{\partial \beta^T} = \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{y} - \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{X} \beta - \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} \mathbf{b} \\ \mathbf{u}(\mathbf{b}) &= \frac{\partial \ell(\beta, \mathbf{b}; \mathbf{y}, \mathbf{X}, \mathbf{Z}, \Sigma_{\epsilon}, \Sigma_B)}{\partial \mathbf{b}^T} = \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{y} - \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X} \beta - \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} \mathbf{b} - \Sigma_B^{-1} \mathbf{b}\end{aligned}$$

Setting $\mathbf{u}(\beta)$ and $\mathbf{u}(\mathbf{b})$ equal to zero yields the system of Maximum (joint) Likelihood equations:

$$\begin{bmatrix} \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{X} & \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X} & \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{y} \\ \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{y} \end{bmatrix} \quad (22)$$

which are often referred to as the Henderson's *mixed model equations* in the LMM literature (Henderson et al., 1959). Solving (22) yields the ML estimators of β and \mathbf{b} (with known Σ_{ϵ} and Σ_B):

$$\hat{\beta}_J = \{ \mathbf{X}^T [\Sigma_{\epsilon}^{-1} - \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1}] \mathbf{X} \}^{-1} \mathbf{X}^T [\Sigma_{\epsilon}^{-1} - \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1}] \mathbf{y} \quad (23)$$

$$\hat{\mathbf{b}}_J = (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} (\mathbf{y} - \mathbf{X} \hat{\beta}_J) \quad (24)$$

Searle (1992) and Robinson (1991) showed that the predictor (24) obtained from the joint MLE method is the *Best Linear Unbiased Predictor* of \mathbf{b} ; for this reason, (24) is often denoted by the acronym "BLUP" in the Linear Mixed Models literature.

4.2 Marginal MLE and EBP coincide with joint MLE and BLUP in Linear Mixed Models

A very important result, due to Searle et al. (1971), states that the MLE of β and the EBP predictor of \mathbf{b} obtained in the marginal approach coincide in Linear Mixed Models with the MLE and the BLUP obtained in the joint approach.

The identity between the marginal ML estimator $\hat{\beta}_M$ and the joint ML estimator $\hat{\beta}_J$ can be easily proved by applying formula (69) in Appendix 1:

$$\Sigma_{\epsilon}^{-1} - \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} = (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} = \Sigma_{\mathbf{y}} \quad (25)$$

whence

$$\begin{aligned}\hat{\beta}_J &= \{ \mathbf{X}^T [\Sigma_{\epsilon}^{-1} - \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1}] \mathbf{X} \}^{-1} \mathbf{X}^T [\Sigma_{\epsilon}^{-1} - \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1}] \mathbf{y} \\ &= [\mathbf{X}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} \mathbf{y} \\ &= \hat{\beta}_M\end{aligned}$$

In order to prove the identity between the EBP of \mathbf{b} , $\tilde{\mathbf{b}}_M = \Sigma_B \mathbf{Z}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_\epsilon)^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) = \Sigma_B \mathbf{Z}^T \Sigma_{\mathbf{y}}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})$ and the BLUP of \mathbf{b} , $\tilde{\mathbf{b}}_J = (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_\epsilon^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) = \Sigma_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \Sigma_\epsilon^{-1}$ we must show that

$$\Sigma_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \Sigma_\epsilon^{-1} = \Sigma_B \mathbf{Z}^T \Sigma_{\mathbf{y}}^{-1} \quad (26)$$

Post-multiplying $\Sigma_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \Sigma_\epsilon^{-1}$ by $\Sigma_{\mathbf{y}} \Sigma_{\mathbf{y}}^{-1}$, we get:

$$\begin{aligned} (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_\epsilon^{-1} &= (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_\epsilon^{-1} (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_\epsilon) (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_\epsilon)^{-1} \\ &= (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} \Sigma_B \mathbf{Z}^T + \mathbf{Z}^T) (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_\epsilon)^{-1} \\ (\text{inserting } \Sigma_B^{-1} \Sigma_B) &= (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_B^{-1} \Sigma_B \mathbf{Z}^T) (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_\epsilon)^{-1} \\ &= (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} (\mathbf{Z}^T \Sigma_\epsilon^{-1} \mathbf{Z} + \Sigma_B^{-1}) \Sigma_B \mathbf{Z}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_\epsilon)^{-1} \\ &= \Sigma_B \mathbf{Z}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_\epsilon)^{-1} \end{aligned} \quad (27)$$

Thanks to these identities, in Linear Mixed Models it is irrelevant which approach (marginal + EBP vs. joint) is used, and we can drop the subscripts M and J and write simply: $\hat{\boldsymbol{\beta}}$ and $\tilde{\mathbf{b}}$. The result is also crucial for what *it does not imply*: this convenient identity *does not hold* for nonlinear - non-Gaussian mixed models, for example for Generalised Linear Mixed Models (GLMMs). The lack of exchangeability between the joint and the marginal approach is one of the main difficulties in the effort to extend the methods here described for Linear Mixed Models to Generalised Linear Mixed Models.

5 With unknown Σ_ϵ , Σ_B

In practice, Σ_ϵ and Σ_B are very rarely known, and therefore they must be estimated from the data. As well known, there are two approaches in the literature for estimation of Σ_ϵ and Σ_B : Maximum Likelihood (ML) and *Restricted* (or *Residual*) Maximum Likelihood (REML). Actually, in general the so-called *variance components*, i.e. the variances and covariances in Σ_ϵ and Σ_B , depend on a limited number, say s , of parameters, which in this section will be collected in an s -vector and denoted by $\boldsymbol{\phi}$; when needed, we stress this dependence by writing $\Sigma_\epsilon(\boldsymbol{\phi})$ and $\Sigma_B(\boldsymbol{\phi})$. Notice that $\boldsymbol{\phi}$ lies in general in a restricted parametric space, which ensures admissible estimators for $\Sigma_\epsilon(\boldsymbol{\phi})$ and $\Sigma_B(\boldsymbol{\phi})$, i.e.:

$$\boldsymbol{\phi} \in \boldsymbol{\Phi} \quad \text{such that} \quad \Sigma_\epsilon(\boldsymbol{\phi}), \Sigma_B(\boldsymbol{\phi}) \quad \text{are positive definite}$$

5.1 ML estimation of Σ_ϵ and Σ_B

Assuming now $\boldsymbol{\beta}$ known, the marginal log-likelihood (17) can be written as a function of $\boldsymbol{\phi}$, given $\boldsymbol{\beta}$, \mathbf{y} , \mathbf{X} and \mathbf{Z} :

$$\ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z}) = - \binom{n}{2} \log(2\pi) - \left(\frac{1}{2} \right) \log(|\Sigma_{\mathbf{y}}(\boldsymbol{\phi})|) - \frac{1}{2} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta})^T \Sigma_{\mathbf{y}}(\boldsymbol{\phi})^{-1} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}) \quad (28)$$

Since $\ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})$ is a non-linear function of $\boldsymbol{\phi}$, its maximization with respect to $\boldsymbol{\phi}$ requires iterative methods. In particular, following Lindstrom and Bates (1988), we adopt either a Newton-Raphson or a Fisher scoring algorithm, which require explicit expressions for the score vector:

$$\mathbf{u}(\boldsymbol{\phi}) = \frac{\partial \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_s} \end{bmatrix}$$

the observed Fisher information matrix, i.e. the Hessian matrix with negative sign:

$$\mathcal{J}(\boldsymbol{\phi}) = -\frac{\partial^2 \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} = \left\{ -\frac{\partial^2 \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j \partial \phi_k} \right\}$$

and the Fisher information matrix, i.e. the expected value of the observed information matrix:

$$\mathcal{I}(\boldsymbol{\phi}) = \mathbf{E}[\mathcal{J}(\boldsymbol{\phi})] = \mathbf{E} \left[-\frac{\partial^2 \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \right] = \left\{ \mathbf{E} \left[-\frac{\partial^2 \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j \partial \phi_k} \right] \right\}$$

The expressions for the score vector, the Hessian and the Fisher information matrix, which were first derived by Harville (1977), written in matrix form are quite cumbersome. It is more convenient to present them element-wise:

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j} &= -\left(\frac{1}{2}\right) \text{tr} \left[\boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \right] \\ &\quad + \left(\frac{1}{2}\right) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \quad (29)$$

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j \partial \phi_k} &= -\left(\frac{1}{2}\right) \text{tr} \left[\boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial^2 \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k} \right) - \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \\ &\quad + \left(\frac{1}{2}\right) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left[\left(\frac{\partial^2 \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k} \right) - 2 \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \times \\ &\quad \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \quad (30)$$

$$\mathbf{E} \left[-\frac{\partial^2 \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j \partial \phi_k} \right] = \left(\frac{1}{2}\right) \text{tr} \left[\boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \quad (31)$$

Once the derivatives in (29), (30) and (31) have been determined, estimation of $\boldsymbol{\phi}$, and hence estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi})$, can be carried out either by the Newton-Raphson iterative procedure:

$$\hat{\boldsymbol{\phi}}_{(s+1)} = \hat{\boldsymbol{\phi}}_{(s)} - \mathcal{J}(\hat{\boldsymbol{\phi}}_{(s)})^{-1} \mathbf{u}(\hat{\boldsymbol{\phi}}_{(s)}) \quad (32)$$

or by the Fisher scoring iterative procedure:

$$\hat{\boldsymbol{\phi}}_{(s+1)} = \hat{\boldsymbol{\phi}}_{(s)} + \mathcal{I}(\hat{\boldsymbol{\phi}}_{(s)})^{-1} \mathbf{u}(\hat{\boldsymbol{\phi}}_{(s)}) \quad (33)$$

5.2 REML estimation of Σ_{ϵ} and Σ_B

Although consistent, the MLE of variance/covariance parameters are well known to be downward biased in finite samples. This drawback is already known for the MLE $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$ of the variance of a $\mathcal{N}(\mu, \sigma^2)$ distribution, with μ unknown, estimated on an i.i.d. sample of size n . The bias comes from the denominator: dividing the sample deviance by n does not take into account the loss of one degree of freedom caused by the estimation of μ . The obvious remedy, usually taught in basic courses in Statistical Inference, is to correct for this bias, using the corrected estimator $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$, obtained dividing the sample deviance by $(n-1)$, instead of n , which incorporates the right number of degrees of freedom available for estimation of σ^2 after estimating μ , and is therefore unbiased.

It turns out that this simple correction is actually the first, simplest example of application of a more generally strategy: estimation of σ^2 by *REstricted Maximum Likelihood* (often denoted with the acronym *REML*), instead of the standard Maximum Likelihood.

Since the bias for the ML estimators comes from the need to estimate first the unknown parameters in μ , the basic idea in REML estimation is to get rid of the unknown μ parameter when estimating variance/covariance parameters in a Multivariate Normal distribution. This can be achieved by choosing *any* matrix \mathbf{K} such that $\mathbf{E}[\mathbf{K}^T \mathbf{y}] = \mathbf{0}$ and hence $\mathbf{K}^T \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^T \Sigma_{\mathbf{y}} \mathbf{K})$. This choice effectively removes the need of estimating μ , or any parameters modelling μ , before estimating $\Sigma_{\mathbf{y}}$.

The idea of using the REML approach first appeared in the statistical literature in the '50s; Patterson and Thompson (1971) presented a comprehensive treatment of REML theory applied to LMMs. It is important to notice that the estimators obtained for ϕ , and hence for $\Sigma_{\epsilon}(\phi)$ and $\Sigma_B(\phi)$, are invariant to the choice of \mathbf{K} . The most typical choice for \mathbf{K} is the OLS orthogonal projection matrix of the (marginal) residuals:

$$\mathbf{K} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

which clearly satisfies the property $\mathbf{E}[\mathbf{K} \mathbf{y}] = \mathbf{E}[(\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y}] = \mathbf{E}[\mathbf{y} - \mathbf{X} \hat{\beta}_{OLS}] = \mathbf{0}$. For this reason, some authors prefer to interpret the acronym *REML* as *REsidual Maximum Likelihood*, although this seems a rather restrictive interpretation, since this choice of \mathbf{K} matrix is just one of the (actually infinite) equivalent alternatives.

Again, the expressions for the score vector, the Hessian and the Fisher information matrix, which were first derived by Harville (1977), are more conveniently presented element-wise:

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j} &= -\left(\frac{1}{2}\right) \text{tr} \left[\mathbf{M} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \right] \\ &\quad + \left(\frac{1}{2}\right) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \quad (34)$$

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j \partial \phi_k} &= -\left(\frac{1}{2}\right) \text{tr} \left[\mathbf{M} \left(\frac{\partial^2 \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k} \right) - \mathbf{M} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \mathbf{M} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \\ &\quad + \left(\frac{1}{2}\right) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \left[\left(\frac{\partial^2 \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k} \right) - 2 \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \mathbf{M} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \\ &\quad \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \quad (35)$$

$$\mathbf{E} \left[-\frac{\partial^2 \ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}, \mathbf{Z})}{\partial \phi_j \partial \phi_k} \right] = \left(\frac{1}{2}\right) \text{tr} \left[\mathbf{M} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \mathbf{M} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \quad (36)$$

where:

$$\mathbf{M} = \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} - \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \mathbf{X} (\mathbf{x}^T \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}(\boldsymbol{\phi})^{-1}$$

Once the derivatives in (34), (35) and (36) have been determined, estimation of $\boldsymbol{\phi}$, and hence estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi})$, can be carried out either by the Newton-Raphson iterative procedure:

$$\hat{\boldsymbol{\phi}}_{(s+1)} = \hat{\boldsymbol{\phi}}_{(s)} - \mathcal{J}(\hat{\boldsymbol{\phi}}_{(s)})^{-1} \mathbf{u}(\hat{\boldsymbol{\phi}}_{(s)}) \quad (37)$$

or by the Fisher scoring iterative procedure:

$$\hat{\boldsymbol{\phi}}_{(s+1)} = \hat{\boldsymbol{\phi}}_{(s)} + \mathcal{I}(\hat{\boldsymbol{\phi}}_{(s)})^{-1} \mathbf{u}(\hat{\boldsymbol{\phi}}_{(s)}) \quad (38)$$

5.3 Iterative algorithm for estimating $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, $\boldsymbol{\Sigma}_B$ and predicting \mathbf{b}

Summing up, estimation of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, $\boldsymbol{\Sigma}_B$ and prediction of \mathbf{b} proceeds through an iterative process that can be schematised as follows:

- initialise $\boldsymbol{\phi}$, and with the initial value $\boldsymbol{\phi}_{(0)}$ compute $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}_{(0)})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi}_{(0)})$
- considering $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}_{(0)})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi}_{(0)})$ as known, estimate $\boldsymbol{\beta}$ and \mathbf{b} with (23) and (24), to obtain $\hat{\boldsymbol{\beta}}_{(1)}$ and $\tilde{\mathbf{b}}_{(1)}$
- considering $\hat{\boldsymbol{\beta}}_{(1)}$ and $\tilde{\mathbf{b}}_{(1)}$ as known, estimate $\hat{\boldsymbol{\phi}}_{(1)}$ using one of (32), (33), (37) or (38) (depending on whether ML or REML and Newton Raphson or Fisher Scoring is chosen), and compute $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\phi}}_{(1)})$ and $\hat{\boldsymbol{\Sigma}}_B(\hat{\boldsymbol{\phi}}_{(1)})$
-
- iterate the previous steps until convergence

Once the algorithm has reached convergence, we can write the ML (or REML) estimators of ϕ , $\Sigma_{\epsilon}(\phi)$, $\Sigma_B(\phi)$ as $\hat{\phi}$, $\hat{\Sigma}_{\epsilon}(\hat{\phi})$, $\hat{\Sigma}_B(\hat{\phi})$ and the ML (or REML) estimator of β and predictor of \mathbf{b} as:

$$\hat{\beta} = [\mathbf{X}^T(\mathbf{Z}^T \hat{\Sigma}_B \mathbf{Z} + \hat{\Sigma}_{\epsilon})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{Z} \hat{\Sigma}_B \mathbf{Z}^T + \hat{\Sigma}_{\epsilon})^{-1} \mathbf{y} \quad (39)$$

$$\hat{\mathbf{b}} = \hat{\Sigma}_B \mathbf{Z}^T (\mathbf{Z} \hat{\Sigma}_B \mathbf{Z}^T + \hat{\Sigma}_{\epsilon})^{-1} (\mathbf{y} - \mathbf{X} \hat{\beta}) \quad (40)$$

6 Hodges' "augmented data" representation

Hodges (1998) (see also Hodges and Sargent (2001); Vaida and Blanchard (2005)) showed that the joint MLE of β and the BLUP of \mathbf{b} can be obtained as the WLS solution of a unique (general) linear model, through the construction of an "augmented data" response vector. This approach starts from adding to the usual LMM specification (4) the obvious identity: $\mathbf{0} = \mathbf{b} - \mathbf{b}$ and constructing the "augmented response vector:

$$\mathbf{y}_+ = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

Then, model (4) can be compactly written as:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{O} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} + \begin{bmatrix} \epsilon \\ \mathbf{b} \end{bmatrix} \quad (41)$$

or:

$$\mathbf{y}_+ = \mathbf{T} \gamma + \delta$$

with:

$$\delta \sim \mathcal{MVN}_{n+r}(\mathbf{0}_{n+r}, \Sigma_{\delta}), \quad \Sigma_{\delta} = \begin{bmatrix} \Sigma_{\epsilon} & \mathbf{O} \\ \mathbf{O} & \Sigma_B \end{bmatrix}$$

Since Σ_{δ} is a full variance/covariance matrix, the estimator of γ is the solution of the Weighted Least Squares system of equations:

$$\mathbf{T}^T \Sigma_{\delta}^{-1} \mathbf{T} \gamma = \mathbf{T} \Sigma_{\delta}^{-1} \mathbf{y}_+ \quad (42)$$

or, using the extended expressions for \mathbf{T} , \mathbf{y}_+ , Σ_{δ} :

$$\begin{bmatrix} \mathbf{X}^T & \mathbf{O} \\ \mathbf{Z}^T & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \Sigma_{\epsilon}^{-1} & \mathbf{O} \\ \mathbf{O} & \Sigma_B^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{O} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T & \mathbf{O} \\ \mathbf{Z}^T & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \Sigma_{\epsilon}^{-1} & \mathbf{O} \\ \mathbf{O} & \Sigma_B^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

whence:

$$\begin{bmatrix} \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{X} & \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X} & \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{y} \\ \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{y} \end{bmatrix} \quad (43)$$

From (43) we see that Hodges' augmented data weighted least squares system corresponds exactly to Henderson's mixed model equations (22). As a consequence, it yields the same estimators for $\boldsymbol{\beta}$ and \mathbf{b} . To prove it, let us denote:

$$\mathbf{T}^T \boldsymbol{\Sigma}_{\delta}^{-1} \mathbf{T} = \begin{bmatrix} \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{X} & \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{X} & \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{Z} + \boldsymbol{\Sigma}_B^{-1} \end{bmatrix} = \mathbf{U} = \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{11} \end{bmatrix} \quad (44)$$

and

$$\mathbf{U}^{-1} = \begin{bmatrix} \mathbf{U}^{11} & \mathbf{U}^{12} \\ \mathbf{U}^{21} & \mathbf{U}^{11} \end{bmatrix} \quad (45)$$

Since the analytic expression for \mathbf{U}^{-1} is rather cumbersome, it is given in Appendix 2. From (42) we obtain:

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= (\mathbf{T}^T \boldsymbol{\Sigma}_{\delta}^{-1} \mathbf{T})^{-1} \mathbf{T}^T \boldsymbol{\Sigma}_{\delta}^{-1} \mathbf{y}_+ & (46) \\ &= \mathbf{U}^{-1} \mathbf{T}^T \boldsymbol{\Sigma}_{\delta}^{-1} \mathbf{y}_+ = \begin{bmatrix} \mathbf{U}^{11} & \mathbf{U}^{12} \\ \mathbf{U}^{21} & \mathbf{U}^{11} \end{bmatrix} \begin{bmatrix} \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} \\ \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{U}^{11} \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} + \mathbf{U}^{12} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} \\ \mathbf{U}^{21} \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} + \mathbf{U}^{11} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{U}^{11} \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} + \mathbf{U}^{11} \mathbf{U}_{12} \mathbf{U}_{22}^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} \\ -\mathbf{U}_{22}^{-1} \mathbf{U}_{21} \mathbf{U}^{11} \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} + (\mathbf{U}_{22}^{-1} + \mathbf{U}_{22}^{-1} \mathbf{U}_{21} \mathbf{U}^{11} \mathbf{U}^{12^T} \mathbf{U}_{22}^{-1}) \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{U}^{11} \mathbf{X}^T [\boldsymbol{\Sigma}_{\epsilon}^{-1} - \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1}] \mathbf{y} \\ \mathbf{U}_{22}^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} - \mathbf{U}_{22}^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{X} \mathbf{U}^{11} \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} + \mathbf{U}_{22}^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{U}^{12^T} \mathbf{U}_{22}^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{y} \\ \mathbf{U}_{22}^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} [\mathbf{y} - \mathbf{X} (\mathbf{U}^{11} \mathbf{X}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y} + \mathbf{U}^{11} \mathbf{U}_{12} \mathbf{U}_{22}^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{y})] \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{y} \\ (\mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} \mathbf{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_{\epsilon}^{-1} [\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}] \end{bmatrix} \\ &= \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\mathbf{b}} \end{bmatrix} \end{aligned}$$

7 The Hat-matrix of Linear Mixed Models

Unlike what happens for Linear Models, and partly for Generalised Linear Models, the literature on the hat-matrix for Linear Mixed models is rather scarce and scattered. We begin by quoting briefly a few relevant references, and then move to discuss the form of hat-matrix which appears to be the closest generalisation of the Linear Models hat-matrix in terms of mathematical and statistical properties.

7.1 A (short) literature review

Zewotir and Galpin (2007) give an expression for the hat-matrix for the residuals, rather than for the fitted values, of Linear Mixed Models. The model specification they employ is slight

different from ours, since they restrict the variance/covariance matrix of the (conditional) errors to be homoscedastic and the variance/covariance matrix of the random parameters to be also homoscedastic and defined in terms of *relative variances* $\tau_{B_j} = \frac{\sigma_{B_j}^2}{\sigma_\epsilon}$:

$$\mathcal{D}_{ZG}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon};ZG} = \Sigma_\epsilon^2 \mathbf{I}_n \quad \mathcal{D}_{ZG}[\mathbf{b}] = \boldsymbol{\Sigma}_{B;ZG} = \sigma_\epsilon^2 \boldsymbol{\Gamma}_{B;ZG}$$

where $\boldsymbol{\Gamma}_{B;ZG} = \mathbf{I}_m \otimes \text{diag}(\boldsymbol{\tau})$ and $\boldsymbol{\tau} = [\tau_{B_1}, \dots, \tau_{B_j}, \dots, \tau_{B_q}]^T$

With these assumptions, the marginal dispersion matrix of \mathbf{y} is:

$$\mathcal{D}_{ZG}[\mathbf{y}] = \boldsymbol{\Sigma}_{\mathbf{y};ZG} = \sigma_\epsilon (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)$$

Zewotir and Galpin (2007) propose a hat-matrix for the conditional residuals $\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}} - \mathbf{Z} \hat{\mathbf{b}}$:

$$\mathbf{R} = (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)^{-1} - (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)^{-1} \mathbf{X} [\mathbf{X}^T (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)^{-1}$$

such that $\mathbf{r} = \mathbf{R} \mathbf{y}$.

Clearly, the corresponding hat-matrix for the fitted values $\hat{\boldsymbol{\mu}}$ is:

$$\mathbf{H}_{ZG} = \mathbf{I} - (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)^{-1} + (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)^{-1} \mathbf{X} [\mathbf{X}^T (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{I}_n + \mathbf{Z} \boldsymbol{\Gamma}_{B;ZG} \mathbf{Z}^T)^{-1} \quad (47)$$

Demidenko and Stukel (2005) (see also Singer et al. (2004), Nobre and Singer (2011)) move from the idea of *generalised leverage matrix* (Wei et al., 1998), which, for *any* model $\mathbf{y} = \boldsymbol{\mu}(\boldsymbol{\beta}) + \boldsymbol{\epsilon}$, is defined as:

$$\mathbf{H} = \frac{\partial \hat{\boldsymbol{\mu}}}{\partial \mathbf{y}^T} \quad (48)$$

i.e. as the matrix having as generic element $\{h_{ij}\}$ the rate of change of μ_i with respect to the observation y_j . Applying (48) to the Linear Mixed Model (4) we obtain:

$$\begin{aligned} \mathbf{H}_{DS} &= \frac{\partial (\mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{Z} \hat{\mathbf{b}})}{\partial \mathbf{y}^T} = \frac{\partial \{ \mathbf{X} (\mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{y} \}}{\partial \mathbf{y}^T} \\ &\quad + \frac{\partial \{ \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{y} - \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{X} (\mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{y} \}}{\partial \mathbf{y}^T} \\ &= \mathbf{X} [\mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \\ &\quad + \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} - \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{X} [\mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \quad (49) \end{aligned}$$

Demidenko and Stukel (2005) write (49) as the sum of two generalised leverage matrices:

$$\mathbf{H}_{DS} = \mathbf{H}_{DS_1} + \mathbf{H}_{DS_2}$$

where:

$$\mathbf{H}_{DS_1} = \mathbf{X} [\mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \quad (50)$$

is the *generalised marginal leverage matrix* and

$$\begin{aligned} \mathbf{H}_{DS_2} &= \mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T(\mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} - \mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T(\mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1}\mathbf{X}[\mathbf{X}^T(\mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1}\mathbf{X}]^{-1}\mathbf{X}^T(\mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \\ &= \mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T(\mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T + \boldsymbol{\Sigma}_\epsilon)^{-1}[\mathbf{I}_n - \mathbf{H}_{DS_1}] \end{aligned} \quad (51)$$

is the *generalised random-effects leverage matrix*.

In passing, it is worthwhile mentioning that Nobre and Singer (2011), noticing that \mathbf{H}_{DS_2} depends on the generalised marginal leverage matrix \mathbf{H}_{DS_1} , in order to get a leverage matrix for the random effects which is not confounded with that for the fixed effects, propose to use, as an alternative to \mathbf{H}_{DS_2} :

$$\mathbf{H}_{NS_2} = \mathbf{Z}\boldsymbol{\Sigma}_B\mathbf{Z}^T \quad (52)$$

7.2 The Hat-matrix in the Hodges' "augmented-data" representation

The main drawback of the Galpin-Zewotir's and the Demidenko-Stukel's definitions of hat-matrix for Linear Mixed Models is that they are not orthogonal projection matrices. This is readily seen by checking that they are not idempotent:

$$\mathbf{H}_{ZG}\mathbf{H}_{ZG} \neq \mathbf{H}_{ZG}; \quad \mathbf{H}_{DS}\mathbf{H}_{DS} \neq \mathbf{H}_{DS}$$

For the context where these hat-matrices were proposed, that of influence diagnostics, this is not a serious problem, but whenever a hat-matrix is required to operate an orthogonal decomposition of the response vector into a vector of fitted values and a vector of residuals, the requirement of idempotency becomes paramount.

The Hodges's "augmented-data" representation, unlike the Galpin-Zewotir and the Demidenko-Stukel approaches, provides an orthogonal projection hat-matrix. To show this result, we start by defining the Hodges's hat-matrix. From:

$$\hat{\boldsymbol{\mu}}_+ = \mathbf{T}\hat{\boldsymbol{\gamma}} = \mathbf{T}(\mathbf{T}^T\boldsymbol{\Sigma}_\delta^{-1}\mathbf{T})^{-1}\mathbf{T}^T\boldsymbol{\Sigma}_\delta^{-1}\mathbf{y}_+$$

we see that the (unscaled) "augmented-data" hat-matrix is:

$$\mathbf{H}_+ = \mathbf{T}(\mathbf{T}^T\boldsymbol{\Sigma}_\delta^{-1}\mathbf{T})^{-1}\mathbf{T}^T\boldsymbol{\Sigma}_\delta^{-1}$$

\mathbf{H}_+ is not symmetric, but if we consider the scaled version of \mathbf{y}_+ and $\hat{\boldsymbol{\mu}}_+$, $\mathbf{y}_+^* = \boldsymbol{\Sigma}_\delta^{-1/2}\mathbf{y}_+$ and $\hat{\boldsymbol{\mu}}_+^* = \boldsymbol{\Sigma}_\delta^{-1/2}\hat{\boldsymbol{\mu}}_+$, we obtain:

$$\hat{\boldsymbol{\mu}}_+^* = \boldsymbol{\Sigma}_\delta^{-1/2}\mathbf{T}(\mathbf{T}^T\boldsymbol{\Sigma}_\delta^{-1}\mathbf{T})^{-1}\mathbf{T}^T\boldsymbol{\Sigma}_\delta^{-1/2}\mathbf{y}_+^*$$

whence, the scaled "augmented-data" hat-matrix is seen to be:

$$\mathbf{H}_+^* = \boldsymbol{\Sigma}_\delta^{-1/2}\mathbf{T}(\mathbf{T}^T\boldsymbol{\Sigma}_\delta^{-1}\mathbf{T})^{-1}\mathbf{T}^T\boldsymbol{\Sigma}_\delta^{-1/2} \quad (53)$$

The hat-matrix \mathbf{H}_+^* is clearly symmetric and idempotent: as such, it is the matrix which projects orthogonally the scaled "augmented" response vector \mathbf{y}_+^* onto the space of the scaled "augmented" fitted vector $\hat{\boldsymbol{\mu}}_+^*$. In this respect, it can be considered as the closest generalisation of the usual hat-matrix of Linear Models to Linear Mixed Models.

It is useful to derive the extended form of \mathbf{H}_+^* . Using (45), we can write:

$$\mathbf{H}_+^* = \boldsymbol{\Sigma}_\delta^{-1/2} \mathbf{T} \mathbf{U}^{-1} \mathbf{T}^T \boldsymbol{\Sigma}_\delta^{-1/2} \quad (54)$$

$$= \begin{bmatrix} \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{X} & \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{Z} \\ \mathbf{O} & -\boldsymbol{\Sigma}_B^{-1/2} \end{bmatrix} \begin{bmatrix} \mathbf{U}^{11} & \mathbf{U}^{12} \\ \mathbf{U}^{21} & \mathbf{U}^{11} \end{bmatrix} \begin{bmatrix} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1/2} & \mathbf{O} \\ \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} & -\boldsymbol{\Sigma}_B^{-1} \end{bmatrix} \quad (55)$$

$$= \begin{bmatrix} \mathbf{H}_{+11}^* & \mathbf{H}_{+12}^* \\ \mathbf{H}_{+21}^* & \mathbf{H}_{+22}^* \end{bmatrix} \quad (56)$$

where:

$$\mathbf{H}_{+11}^* = \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{X} \mathbf{U}^{11} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1/2} + \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{X} \mathbf{U}^{12} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1/2} + \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{Z} \mathbf{U}^{21} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1/2} + \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{Z} \mathbf{U}^{22} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1/2}$$

$$\mathbf{H}_{+12}^* = -\boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{X} \mathbf{U}^{12} \boldsymbol{\Sigma}_B^{-1/2} - \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{Z} \mathbf{U}^{22} \boldsymbol{\Sigma}_B^{-1/2}$$

$$\mathbf{H}_{+21}^* = -\boldsymbol{\Sigma}_B^{-1/2} \mathbf{U}^{21} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1/2} - \boldsymbol{\Sigma}_B^{-1/2} \mathbf{U}^{22} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1/2}$$

$$\mathbf{H}_{+22}^* = \boldsymbol{\Sigma}_B^{-1/2} \mathbf{U}^{22} \boldsymbol{\Sigma}_B^{-1/2}$$

Also, it is easy to check that the scaled "augmented" fitted vector $\hat{\boldsymbol{\mu}}_+^*$ turns out to be:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_+^* &= \mathbf{H}_+^* \mathbf{y}_+^* = \begin{bmatrix} \mathbf{H}_{+11}^* & \mathbf{H}_{+12}^* \\ \mathbf{H}_{+21}^* & \mathbf{H}_{+22}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{H}_{+11}^* \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{y} \\ \mathbf{H}_{+21}^* \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{X} \hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_\epsilon^{-1/2} \mathbf{Z} \tilde{\mathbf{b}} \\ -\boldsymbol{\Sigma}_B^{-1/2} \tilde{\mathbf{b}} \end{bmatrix} \end{aligned} \quad (57)$$

i.e. the scaled "augmented" fitted vector $\hat{\boldsymbol{\mu}}_+^*$ has two components: the conditional (within clusters) scaled fitted values $\widehat{\boldsymbol{\mu}}|\tilde{\mathbf{b}}^* = \boldsymbol{\Sigma}_\epsilon^{-1/2} (\mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{Z} \tilde{\mathbf{b}})$ and the scaled predicted random parameters (with negative sign) $-\boldsymbol{\Sigma}_B^{-1/2} \tilde{\mathbf{b}}$.

7.3 Equivalence between \mathbf{H}_{ZG} , \mathbf{H}_{DS} and \mathbf{H}_{+11}

It is natural to wonder whether, although apparently quite different, \mathbf{H}_{ZG} , \mathbf{H}_{DS} and the Hodges' hat-matrix are equivalent. It is possible to prove that it is actually the case, as long as the different assumptions made for the Zewotir-Galpin hat-matrix are accounted for and only the upper-left block of the "unscaled" version of the Hodges's hat-matrix, denoted by \mathbf{H}_{+11} , is used in the comparison.

Consider the the (unscaled) "augmented-data" Hodges' hat-matrix:

$$\mathbf{H}_+ = \mathbf{T}(\mathbf{T}^T \boldsymbol{\Sigma}_\delta^{-1} \mathbf{T})^{-1} \mathbf{T}^T \boldsymbol{\Sigma}_\delta^{-1}$$

By a derivation parallel to the one used in (54), (55), (56) for the "scaled" version, we get:

$$\mathbf{H}_+ = \mathbf{T} \mathbf{U}^{-1} \mathbf{T}^T \boldsymbol{\Sigma}_\delta^{-1} \quad (58)$$

$$= \begin{bmatrix} \mathbf{X} \mathbf{U}^{11} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} + \mathbf{X} \mathbf{U}^{12} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} + \mathbf{Z} \mathbf{U}^{21} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} + \mathbf{Z} \mathbf{U}^{22} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} & -\mathbf{X} \mathbf{U}^{12} \boldsymbol{\Sigma}_B^{-1} - \mathbf{Z} \mathbf{U}^{22} \boldsymbol{\Sigma}_B^{-1} \\ -\mathbf{U}^{12} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} - \mathbf{U}^{22} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} & \mathbf{U}^{22} \boldsymbol{\Sigma}_B^{-1/2} \end{bmatrix} \quad (59)$$

$$= \begin{bmatrix} \mathbf{H}_{+11} & \mathbf{H}_{+12} \\ \mathbf{H}_{+21} & \mathbf{H}_{+22} \end{bmatrix} \quad (60)$$

As observed by Vaida and Blanchard (2005), the upper left block \mathbf{H}_{+11} is the hat-matrix of the (conditional) within-cluster fitted values $\widehat{\boldsymbol{\mu}}|\mathbf{b} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\mathbf{b}}$, but it is not an orthogonal projection matrix, since it is neither symmetric nor idempotent. We now set out to prove that it is this block of the "unscaled" Hodges' hat-matrix \mathbf{H}_+ which is equivalent to Zewotir-Galpin \mathbf{H}_{ZG} and to Demidenko-Stukel \mathbf{H}_{DS} .

Result 1

Under the restrictive assumptions made by Zewotir and Galpin (2007):

$$\mathbf{H}_{ZG} = \mathbf{H}_{+11} \quad (61)$$

Proof Using (73), (75), (77) and (80) in Appendix 2, we can write \mathbf{H}_{+11} as follows:

$$\begin{aligned} \mathbf{H}_{+11} = & \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} - \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} - \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \\ & + \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} + \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \end{aligned} \quad (62)$$

But:

$$\begin{aligned} & \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} - \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} \\ & - \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} + \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} = \\ & \quad \left[\mathbf{I}_n - \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \right] \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \left[\mathbf{I}_n - \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \right] \boldsymbol{\Sigma}_\epsilon^{-1} \end{aligned}$$

and

$$\mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} = \mathbf{I}_n - (\mathbf{I}_n - \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1})$$

whence (62) can be written:

$$\mathbf{H}_{+11} = \mathbf{I}_n - (\mathbf{I}_n - \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1}) + \left[\mathbf{I}_n - \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \right] \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{X}^T \left[\mathbf{I}_n - \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \right] \boldsymbol{\Sigma}_\epsilon^{-1} \quad (63)$$

It is immediate to recognise that $\left[\mathbf{I}_n - \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \right] = \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_y^{-1}$. This follows from (25):

$$\boldsymbol{\Sigma}_y^{-1} = \boldsymbol{\Sigma}_\epsilon^{-1} - \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} (\mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1} = \boldsymbol{\Sigma}_\epsilon^{-1} - \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{b|y} \mathbf{Z}^T \boldsymbol{\Sigma}_\epsilon^{-1}$$

whence:

$$\Sigma_{\epsilon} \Sigma_{\mathbf{y}}^{-1} = \Sigma_{\epsilon} [\Sigma_{\epsilon}^{-1} - \Sigma_{\epsilon}^{-1} \mathbf{Z} \Sigma_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \Sigma_{\epsilon}^{-1}] = \mathbf{I}_n - \mathbf{Z} \Sigma_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \quad (64)$$

$$\Sigma_{\mathbf{y}}^{-1} \Sigma_{\epsilon} = \mathbf{I}_n - \Sigma_{\epsilon}^{-1} \mathbf{Z} \Sigma_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \quad (\text{by symmetry}) \quad (65)$$

and (63) can be written as:

$$\mathbf{H}_{+11} = \mathbf{I}_n - \Sigma_{\epsilon} \Sigma_{\mathbf{y}}^{-1} + \Sigma_{\epsilon} \Sigma_{\mathbf{y}}^{-1} \mathbf{X} (\mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \quad (66)$$

Now, recalling that under the restrictive assumptions of Zewotir and Galpin (2007) paper:

$$\Sigma_{\mathbf{y};ZG} = \sigma_{\epsilon} (\mathbf{I}_n + \mathbf{Z} \Gamma_{B;ZG} \mathbf{Z}^T); \quad \Sigma_{\epsilon;ZG} = \sigma_{\epsilon}^2 \mathbf{I}_n; \quad \Sigma_{\epsilon;ZG} \Sigma_{\mathbf{y};ZG}^{-1} = (\mathbf{I}_n + \mathbf{Z} \Gamma_{B;ZG} \mathbf{Z}^T)^{-1}$$

formula (66) yields:

$$\begin{aligned} \mathbf{H}_{+11} &= \mathbf{I}_n - (\mathbf{I}_n + \mathbf{Z} \Gamma_{B;ZG} \mathbf{Z}^T)^{-1} \\ &\quad + (\mathbf{I}_n + \mathbf{Z} \Gamma_{B;ZG} \mathbf{Z}^T)^{-1} \mathbf{X} \left[\mathbf{X}^T \frac{1}{\sigma_{\epsilon}^2} (\mathbf{I}_n + \mathbf{Z} \Gamma_{B;ZG} \mathbf{Z}^T)^{-1} \mathbf{X} \right]^{-1} \mathbf{X}^T \frac{1}{\sigma_{\epsilon}^2} (\mathbf{I}_n + \mathbf{Z} \Gamma_{B;ZG} \mathbf{Z}^T)^{-1} \\ &= \mathbf{H}_{ZG} \end{aligned}$$

Q.E.D.

Result 2

$$\mathbf{H}_{DS} = \mathbf{H}_{+11} \quad (67)$$

Proof We can re-write (63) as:

$$\begin{aligned} \mathbf{H}_{+11} &= \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \left[\mathbf{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \right] \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \\ &\quad + \mathbf{Z} \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \mathbf{Z} \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \mathbf{X} (\mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \left[\mathbf{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \right] \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \end{aligned}$$

From (65), we can substitute $\left[\mathbf{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \mathbf{Z} \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \right] \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}$, whence:

$$\begin{aligned} \mathbf{H}_{+11} &= \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{Z} \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \left[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \right] \\ &= \mathbf{H}_{DS_1} + \mathbf{Z} \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} (\mathbf{I}_n - \mathbf{H}_{DS_1}) \end{aligned}$$

Since from (26) we know that $\mathbf{Z} \boldsymbol{\Sigma}_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} = \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}$, it follows that:

$$\mathbf{H}_{+11} = \mathbf{H}_{DS_1} + \mathbf{Z} \boldsymbol{\Sigma}_B \mathbf{Z}^T \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{I}_n - \mathbf{H}_{DS_1}) = \mathbf{H}_{DS}$$

Q.E.D.

A Appendix 1: A useful matrix algebra result

In this Appendix we recall, without proof, a useful result concerning the inverse of a Schur complement $(\mathbf{A} - \mathbf{C} \mathbf{B} \mathbf{C}^T)^{-1}$ and of the associated form $(\mathbf{A} + \mathbf{C} \mathbf{B} \mathbf{C}^T)^{-1}$. The Schur complement, and its inverse, are very important in the derivations of theoretical results, in particular in Linear and Linear Mixed Models. The reader can find a formal proof in Henderson and Searle (1981) (equation 17).

Let \mathbf{A} and \mathbf{B} be nonsingular symmetric matrices of order n and m respectively, and let \mathbf{C} be an $n \times m$ rectangular matrix. Then:

$$(\mathbf{A} - \mathbf{C} \mathbf{B} \mathbf{C}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{C} (\mathbf{B}^{-1} - \mathbf{C}^T \mathbf{A}^{-1} \mathbf{C})^{-1} \mathbf{C}^T \mathbf{A}^{-1} \quad (68)$$

$$(\mathbf{A} + \mathbf{C} \mathbf{B} \mathbf{C}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{C} (\mathbf{B}^{-1} + \mathbf{C}^T \mathbf{A}^{-1} \mathbf{C})^{-1} \mathbf{C}^T \mathbf{A}^{-1} \quad (69)$$

B Appendix 2: The inverse of matrix U in Hodges' representation

Among various possible representations (see e.g. Seber (2008)) of the inverse of a partitioned matrix, we choose the following:

$$U^{11} = (U_{11} - U_{12}U_{22}^{-1}U_{21})^{-1} \\ = [\mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{X} - \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X}]^{-1} \quad (70)$$

$$= \{\mathbf{X}^T [\Sigma_{\epsilon}^{-1} - \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1}] \mathbf{X}\}^{-1} \quad (71)$$

$$= [\mathbf{X}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{Z} \Sigma_B \mathbf{Z}^T + \Sigma_{\epsilon})^{-1} \quad (72)$$

$$= (\mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \quad (73)$$

$$U^{12} = -U^{11}U_{12}U_{22}^{-1} \\ = -[\mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{X} - \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X}]^{-1} \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \quad (74)$$

$$= -(\mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} \Sigma_{\mathbf{b}|\mathbf{y}} \quad (75)$$

$$U^{21} = U^{12T} = U_{22}^{-1}U_{21}U^{11} = \\ = -(\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X} [\mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{X} - \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X}]^{-1} \quad (76)$$

$$= -\Sigma_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X} (\mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \quad (77)$$

$$U^{22} = U_{22}^{-1} + U_{22}^{-1}U_{21}U^{11}U^{12T}U_{22}^{-1} \\ = (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \\ + (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X} [\mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{X} - \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X}]^{-1} \quad (78)$$

$$\mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} (\mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} + \Sigma_B^{-1})^{-1} \quad (79)$$

$$= \Sigma_{\mathbf{b}|\mathbf{y}} + \Sigma_{\mathbf{b}|\mathbf{y}} \mathbf{Z}^T \Sigma_{\epsilon}^{-1} \mathbf{X} (\mathbf{X}^T \Sigma_{\mathbf{y}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma_{\epsilon}^{-1} \mathbf{Z} \Sigma_{\mathbf{b}|\mathbf{y}} \quad (80)$$

References

- Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, 443–458.
- Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* 88(421), 9–25.
- Demidenko, E. and T. A. Stukel (2005). Influence analysis for linear mixed-effects models. *Statistics in medicine* 24(6), 893–909.
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72(358), 320–338.
- Henderson, C. R., O. Kempthorne, S. R. Searle, and C. Von Krosigk (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15(2), 192–218.

- Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices. *Siam Review* 23(1), 53–60.
- Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(3), 497–536.
- Hodges, J. S. and D. J. Sargent (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 367–379.
- Lindstrom, M. J. and D. M. Bates (1988). Newtonraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* 83(404), 1014–1022.
- Nobre, J. S. and J. M. Singer (2011). Leverage analysis for linear mixed models. *Journal of Applied Statistics* 38(5), 1063–1072.
- Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 545–554.
- Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, 15–32.
- Searle, J. R. (1992). *The rediscovery of the mind*. MIT press.
- Searle, J. R., J. L. Austin, P. Strawson, H. Grice, N. Chomsky, J. J. Katz, N. Goodman, and H. Putnam (1971). *The philosophy of language*, Volume 39. Oxford University Press London.
- Seber, G. A. (2008). *A matrix handbook for statisticians*, Volume 15. John Wiley & Sons.
- Singer, J. M., J. S. Nobre, and H. C. Sef (2004). Regression models for pretest/posttest data in blocks. *Statistical Modelling* 4(4), 324–338.
- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1135–1151.
- Vaida, F. and S. Blanchard (2005). Conditional akaike information for mixed-effects models. *Biometrika* 92(2), 351–370.
- Wei, B.-C., Y.-Q. Hu, and W.-K. Fung (1998). Generalized leverage and its applications. *Scandinavian Journal of statistics* 25(1), 25–37.
- Zewotir, T. and J. S. Galpin (2007). A unified approach on residuals, leverages and outliers in the linear mixed model. *Test* 16(1), 58–75.
- Zhang, F. (2006). *The Schur complement and its applications*, Volume 4. Springer Science & Business Media.