Working Paper Series, N. 7, July 2012

Department of Statistical Sciences
University of Padua
Italy

Nonparametric Estimation of the Random Effect Variance-Covariance Matrix with Partial Information from the Clusters

Dario Basso
Department of Statistical Sciences
University of Padua
Italy

Abstract: The proposed approach is useful whenever the variability of the response in a linear model can be viewed as the sum of two independent sources of variability, one that is common to all clusters and it is unknown, and another which is assumed to be available and it is cluster-specific. Here by clusters we mean, for instance, second-level units in multi-level models (schools, hospitals etc.), or subjects in repeated measure experiments. The responses here have to be thought as functions of the first-level observations, whose variability is known to depend only on the cluster's specificities. These settings include linear mixed models (LMM) when the estimators of the parameters of interest are obtained conditionally on each cluster. The model may account for additional informations on the clusters, such as covariates, or contrast vectors. An estimator of the common source of variability is obtained from the residual deviance of the (multivariate) model, opportunely re-scaled, through the moment method. An iterative procedure is then suggested (whose initial step depends on the available information), that turns out to be a special case of the EM-algorithm.

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Dario Basso
tel: +39 049 827 4168
dario@unipd.it
http://www.stat.unipd.it/~name

Nonparametric Estimation of the Random Effect Variance-Covariance Matrix with Partial Information from the Clusters

Dario Basso
Department of Statistical Sciences
University of Padua
Italy

Abstract: The proposed approach is useful whenever the variability of the response in a linear model can be viewed as the sum of two independent sources of variability, one that is common to all clusters and it is unknown, and another which is assumed to be available and it is cluster-specific. Here by clusters we mean, for instance, second-level units in multi-level models (schools, hospitals etc.), or subjects in repeated measure experiments. The responses here have to be thought as functions of the first-level observations, whose variability is known to depend only on the cluster's specificities. These settings include linear mixed models (LMM) when the estimators of the parameters of interest are obtained conditionally on each cluster. The model may account for additional informations on the clusters, such as covariates, or contrast vectors. An estimator of the common source of variability is obtained from the residual deviance of the (multivariate) model, opportunely re-scaled, through the moment method. An iterative procedure is then suggested (whose initial step depends on the available information), that turns out to be a special case of the EM-algorithm.

1 Introduction

In many experimental and observational studies, such as multi-level or repeated measure designs, the responses measured on the same cluster are not independent. This is because the one-level units (trials) all share some specific features depending on the cluster. In these situations, linear mixed models are usually assumed to describe the responses, because they account for the correlation between data of the same cluster. In linear mixed models there are two sources of variability, one which is assumed to be common to all clusters and it is due to the random effects, and another which is specific for each cluster. Within this framework, the generation of the response can be viewed as a two-step process: at first, the individual effects are generated from a common probability law, which determine what is to be though as the 'true' individual effects. Later, the final responses are generated conditionally on the random effect realization (and on the individual design matrix) with an additional source of variability due to the individual errors.

Moving on the opposite direction, if the fixed effects are estimated conditionally on the clusters, the conditional variability of their estimators will not depend on the random effects, but only on the variabilty of the individual errors. In a parametric framework, the individual source of variability is usually described by the dispersion parameter, which is assumed to be knwon and equal for all subjects [1]. Similarly, if we assume the dispersion parameters to be known within each cluster (but not necessarely equal), the conditional variability of the estimators will be known as well. This is what is meant by ‘partial information from the clusters’.

The idea behind this work is to describe the marginal distributions of the conditional estimators up to their second moment, through a linear model that eventually takes into account additional covariates of the clusters. The response vector $\mathbf{Y}$ in next section has to be imagined as a data matrix whose rows are independent and heteroscedastic vectors with the estimators of the parameters of interests (e.g. intercept and slope).

2 Variance-Covariance Matrix Estimation

Let $\mathbf{Y}$ be a $n \times p$ data matrix whose rows are independent but not identically distributed random vectors, and whose distributions are continuous and admit finite variance/covariance matrix. Since this assumption is very broad, we will assume that these observations are associated to a $n \times q$ design matrix $\mathbf{Z}$ full of rank, through the linear model

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{E}, \tag{1}$$

where $\boldsymbol{\gamma}$ is a $q \times p$ matrix of coefficients, and $\mathbf{E}$ is a $n \times p$ matrix, whose rows $\mathbf{e}_i'$s are independent random vectors with zero mean and variance $Var(\mathbf{e}_i) = \mathbf{V}_i$. We will write $\mathbf{e}_i \sim (\mathbf{0}, \mathbf{V})$ to underline that the attention will be focused on the first two moments of the multivariate distribution. We will also assume that each row of the error matrix $\mathbf{E}$ is in fact sums of two independent random vectors $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$, where the first component is a random realization from a common distribution, whereas the second is specific for each cluster. Namely, the $\mathbf{u}_i$s are i.i.d. random vectors from a common density function $f_{\mathbf{u}}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{u}})$, and $\boldsymbol{\varepsilon}_i \sim (\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is specific for each cluster. The reason of this assumptions will be clearer in the Section 4. Model (1) can be equivalently viewed in terms of the single row of $\mathbf{Y}$, which are independent random vectors satisfying $\mathbf{y}_i \sim (\mathbf{Z}\gamma, \mathbf{V}_i)$, where $\mathbf{V}_i = \boldsymbol{\Sigma}_{\mathbf{u}} + \boldsymbol{\Sigma}_i$. Note that the error matrix is such that $E[\mathbf{E}'\mathbf{E}] = \sum_{i=1}^n [\boldsymbol{\Sigma}_{\mathbf{u}} + \boldsymbol{\Sigma}_i]$ and $E[\mathbf{E}'\mathbf{E}] = diag\{tr(\boldsymbol{\Sigma}_{\mathbf{u}} + \boldsymbol{\Sigma}_i)\}$, $i = 1 \ldots, n$, where $tr(\cdot)$ denotes the trace operator.

In case the matrices $\mathbf{V}_i$ are assumed to be equal for all the clusters, it is well known that the BLUE estimator of the parameter vector $\boldsymbol{\gamma}$ is given by the weighted least square (WLS) estimator. It is also known that if the error vectors are homoscedastic, the the Ordinary Least Square and the WLS estimators coincide ([3]). In a similar fashion, we consider a $n \times n$ weighting matrix $\mathbf{V} = diag\{tr(\boldsymbol{\Sigma}_{\mathbf{u}} + \boldsymbol{\Sigma}_i)\}$ and the following multiple regression model

$$\mathbf{V}^{-1/2}\mathbf{Y} = \mathbf{V}^{-1/2}\mathbf{Z}\gamma + \mathbf{V}^{-1/2}\mathbf{E}. \tag{2}$$

We denote with $\tilde{\mathbf{E}} = \mathbf{V}^{-1/2}\mathbf{E}$ the current error matrix, whose rows satisfy:

$$E[\tilde{\mathbf{E}}'\tilde{\mathbf{E}}] = \sum_{i=1}^n \frac{\boldsymbol{\Sigma}_{\mathbf{u}} + \boldsymbol{\Sigma}_i}{tr(\boldsymbol{\Sigma}_{\mathbf{u}} + \boldsymbol{\Sigma}_i)} \qquad \text{and} \qquad E[\tilde{\mathbf{E}}\tilde{\mathbf{E}}'] = \mathbf{I}_n.$$

Note that in the univariate case this is exactly equivalent to the general linear model approach. The WLS estimator of $\boldsymbol{\gamma}$ is then given by $\hat{\boldsymbol{\gamma}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$, which is obtained by minimizing with respect to $\boldsymbol{\gamma}$ the residual deviance of model (2). The residuals of model (2) can be viewed as the projection of the weighted response $\tilde{\mathbf{Y}} = \mathbf{V}^{-1/2}\mathbf{Y}$ into the sub-space orthogonal to the one spanned by the columns of $\tilde{\mathbf{Z}} = \mathbf{V}^{-1/2}\mathbf{Z}$. This operation is made possible by the pre-multiplication of the weighted response for the projection matrix $(\mathbf{I}-\mathbf{H_V})$, where $\mathbf{H_V} = \mathbf{V}^{-1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}^{-1/2}$. The matrix $\mathbf{H_V}$ is indeed a projection matrix since it is symmetric and idempotent, and the matrix of residuals of model (2) can therefore be written as:

$$\tilde{\mathbf{R}} = (\mathbf{I} - \mathbf{H_V})\tilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{H_V})\tilde{\mathbf{E}}, \tag{3}$$

Note that the residuals of model (3) are such that $E[\tilde{\mathbf{R}}\tilde{\mathbf{R}}'] = \mathbf{I} - \mathbf{H_V}$ and

$$E[\tilde{\mathbf{R}}'\tilde{\mathbf{R}}] = \sum_{i=1}^{n}\sum_{j=1}^{n}[\mathbf{I} - \mathbf{H_V}]_{ij}E[\tilde{\mathbf{e}}_i\tilde{\mathbf{e}}_j'] = \sum_{i=1}^{n}[\mathbf{I} - \mathbf{H_V}]_{ii}V[\tilde{\mathbf{e}}_i],$$

where $[\mathbf{I} - \mathbf{H_V}]_{ij}$ is the (i,j) element of $\mathbf{I} - \mathbf{H_V}$. Let us denote by ω_i^2 the element in the diagonal of $\mathbf{W}'\mathbf{W}$, where $\mathbf{W} = [\mathbf{I} - \mathbf{H_V}]\mathbf{V}^{-1/2}$. We are now able to express the expectation of the quadratic form as a function of the original errors:

$$E[\tilde{\mathbf{R}}'\tilde{\mathbf{R}}] = \sum_{i=1}^{n}\omega_i^2 V(\mathbf{e}_i) = tr(\mathbf{W}'\mathbf{W})\mathbf{\Sigma_u} + \sum_{i=1}^{n}\omega_i^2\mathbf{\Sigma}_i, \qquad \omega_i^2 = \frac{[\mathbf{I} - \mathbf{H_V}]_{ii}}{tr(\mathbf{\Sigma_u} + \mathbf{\Sigma}_i)}.$$

Finally, through the moment estimation method, we obtain an unbiased estimator of the unknown variance/covariance matrix:

$$\hat{\mathbf{\Sigma}}_{\mathbf{u}} = \frac{1}{tr(\mathbf{W}'\mathbf{W})}\left[\tilde{\mathbf{R}}'\tilde{\mathbf{R}} - \sum_{i=1}^{n}\omega_i^2\mathbf{\Sigma}_i\right]. \tag{4}$$

Note that the same solution can be achieved by considering the eigen decomposition of the projection matrix $\mathbf{I} - \mathbf{H_V} = \mathbf{\Lambda_V}\mathbf{\Lambda}_{\mathbf{V}}'$, in which case the residual matrix can be expressed as $\tilde{\mathbf{R}} = \mathbf{\Lambda}_{\mathbf{V}}'\mathbf{V}^{-1/2}\mathbf{Y}$ and the weight matrix becomes $\mathbf{W} = \mathbf{\Lambda}_{\mathbf{V}}'\mathbf{V}^{-1/2}$. Also note that the estimator is based on a difference, so it can lead to negative definite estimates. In order to avoid this, we will consider the eigen decomposition of $\hat{\mathbf{\Sigma}}_{\mathbf{u}}$ and force the eigenvalues to be nonnegative.

Despite the fact that the weights ω_i^2s in (4) are themselves dependent on the unknown matrix $\mathbf{\Sigma_u}$, this representation will be useful for the iterative estimation algorithm discussed in Section 3, where the ω_i^2s will play the role of weights. Note that here the role usually played by the degrees of freedom in the estimation is replaced by the trace of $\mathbf{W}'\mathbf{W}$, which is generally not an integer.

3 Iterative Estimation

The main assumption in this work is that the individual variance/covariance matrices $\mathbf{\Sigma}_i$'s are known. If so, we can consider as inital guess for the error covariance matrix the matrix $\mathbf{V}_0 = diag\{tr(\mathbf{\Sigma}_i)\}$, in which case the weights will also be known, and obtain a first estimate of $\Sigma_{\mathbf{u}}$. This value can be used to update the guess of $\mathbf{V}$ by adding it to the elements in the diagonal of $\mathbf{V}_0$, obtain an updated vector of weights and repeat these steps iteratively until convergence. Summaryzing, at step s:

- Obtain the vector of modified residuals $\mathbf{R}_s = \mathbf{W}_{s-1}\mathbf{Y}$, where $\mathbf{W}_{s-1}$ has been defined in Section 2 with $\mathbf{V}$ replaced by $\mathbf{V}_{s-1}$, and where $\mathbf{V}_{s-1}$ is the estimate of the error variance/covariance matrix obtained at the previous step (M-step).

- Obtain the weights ω_i^2, i.e. the diagonal elements of $\mathbf{W}'_{s-1}\mathbf{W}_{s-1}$ and the current estimate of σ_u^2 (E-step) as:

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{u}_s} = \frac{1}{\sum_{i=1}^n \omega_i^2}\left[\mathbf{R}'_s\mathbf{R}_s - \sum_{i=1}^n \omega_{i_s}^2 \boldsymbol{\Sigma}_i\right]$$

- Consider the eigen decomposition of $\hat{\boldsymbol{\Sigma}}_{\mathbf{u}_s} = \boldsymbol{\Lambda}_s\boldsymbol{\Psi}_{\boldsymbol{s}}\boldsymbol{\Lambda}'_s$, where $\boldsymbol{\Lambda}_s$ is the eigenvector matrix and $\boldsymbol{\Psi}_{\boldsymbol{s}} = diag\{\psi_{i_s}\}$ is the eigenvalue matrix. Force the current estimate to be nonnegative definite by setting $\hat{\boldsymbol{\Sigma}}_{\mathbf{u}_s} = \boldsymbol{\Lambda}_s\boldsymbol{\Psi}_{\boldsymbol{s}}{}^*\boldsymbol{\Lambda}'_s$, where $\boldsymbol{\Psi}_{\boldsymbol{s}}{}^* = diag\{\psi_{i_s} \cdot I(\psi_{i_s} > 0)\}$ and $I(\cdot)$ is the indicator function. Finally let $\mathbf{V}_s = diag\{tr(\hat{\boldsymbol{\Sigma}}_{\mathbf{u}_s} + \boldsymbol{\Sigma}_i)\}$.

If the algorithm does not stop because all the eigenvalues of $\hat{\boldsymbol{\Sigma}}_{\mathbf{u}}$ are negative (in which case $\hat{\boldsymbol{\Sigma}}_{\mathbf{u}} = \mathbf{0}$), the converge is ensured by the EM algorithm theory [2], of which this algorithm is a special case. Indeed the maximization step is represented by the WLS estimation of the residuals, which are obtained by minimizing the residual deviance (hence maximizing the related likelihood), whereas the expectation step is implicit in the moment estimation. The algorithm can be stopped after a maximum number of iterations, or as soon as some convergence criterion is satisfied; for instance, given a precision value δ, when $|tr(\boldsymbol{\Psi}_{\boldsymbol{s}}{}^*) - tr(\boldsymbol{\Psi}_{\boldsymbol{s+1}}{}^*)| < \delta$.

4 Example

In this section a simple example where such estimation method is applied will be explained. Consider a repeated measure experiment where a response variable Y is measured on n subjects observed on n_i trials, $i = 1, \ldots, n$, that depend on the levels/values of a certain treatment/covariate $\mathbf{x}_i$. Additionally, suppose that another covariate is available for each cluster, say the age, $\mathbf{z} = [z_1, \ldots, z_n]'$. Assume a linear mixed model with random intercept and slope for the $n_i \times 1$ response of each subject holds:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i$$

where $\mathbf{X}_i = [\mathbf{1}_{n_i}, \mathbf{x}_i]$ is the $n_i \times 2$ design matrix for the ith subject, $\boldsymbol{\beta}$ is a 2×1 vector of fixed effects, $\mathbf{u}_i$ is a 2×1 vector of random effects from a common probability law, i.e. $\mathbf{u}_i \sim (\mathbf{0}, \boldsymbol{\Sigma}_u)$, and $\boldsymbol{\varepsilon}_i$ is a $n_i \times 1$ vector of i.i.d. errors, so $\boldsymbol{\varepsilon}_i \sim (\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$. Now consider the OLS estimator of $\boldsymbol{\beta}_i$ obtained on the ith cluster: $\hat{\boldsymbol{\beta}}_i = (\mathbf{X}'_i\mathbf{X}_i)^{-1}\mathbf{X}'_i\mathbf{y}_i$.

Conditionally on the realization of the random effect $\mathbf{u}_i$, we have the usual result $\hat{\boldsymbol{\beta}}_i|\mathbf{u}_i \sim (\mathbf{X}_i[\boldsymbol{\beta} + \mathbf{u}_i], \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = \sigma_i^2(\mathbf{X}'_i\mathbf{X}_i)^{-1}$ only depends on the ith cluster features. If the dispersion parameter σ_i^2 is assumd to be known, so is the conditional covariance matrix $\boldsymbol{\Sigma}_i$. As far as the unconditional distribution of $\hat{\boldsymbol{\beta}}_i$ is concerned, we have $\hat{\boldsymbol{\beta}}_i \sim (\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_i)$.

Now suppose that we are interested on testing wheter there is a significant interaction between age and treatment, we can do this by considering a further linear model:

$$\mathbf{B} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{E}, \qquad \mathbf{B}_{n\times 2} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1' \\ \vdots \\ \hat{\boldsymbol{\beta}}_n' \end{bmatrix}, \quad \mathbf{Z}_{n\times 2} = [\mathbf{1}_n, \mathbf{z}], \quad \boldsymbol{\Gamma}_{2\times 2} = \left[\frac{\boldsymbol{\Gamma}_0'}{\boldsymbol{\Gamma}_1'}\right]$$

and $\mathbf{E}$ is a $n \times 2$ data matrix whose rows are independent random vectors satisfying $\mathbf{e}_i \sim (\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{u} + \boldsymbol{\Sigma}_i)$. The hypothesis of no interaction between age and treatment can be thus translated into the null $\boldsymbol{\Gamma}_1 = \mathbf{0}$.

References

[1] McCulloch C. E., Searle S. R. (2001). *Generalized, linear, and mixed models*. Wiley-New York.

[2] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm”. *Journal of the Royal Statistical Society, Series B*, **39(1)**, 1-38.

[3] Mardia K. V., Kent J.T., Bibby J.M. (1980). *Multivariate Analysis*. Elsevier Science.

Working Paper Series
Department of Statistical Sciences, University of Padua

UNIVERSITA DEGLI STUDI DI PADOVA DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy