Working Paper Series, N. 12, December 2011

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

Dario Basso
Department of Statistical Sciences
University of Padua
Italy

Abstract: This work is a new proposal for estimating the variance of the random effects in case the knowledge of the internal variability of the clusters is (or might be) assumed to be known. Here by clusters we mean, for instance, second-level units in multi-level models (schools, hospitals etc.), or subjects in repeated measure experiments. 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. 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 effects 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 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.

Keywords: Iterative Moment Estimation

Department of Statistical Sciences
University of Padua
Italy

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 4151
dario@stat.unipd.it

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

Dario Basso
Department of Statistical Sciences
University of Padua
Italy

Abstract: This work is a new proposal for estimating the variance of the random effects in case the knowledge of the internal variability of the clusters is (or might be) assumed to be known. Here by clusters we mean, for instance, second-level units in multi-level models (schools, hospitals etc.), or subjects in repeated measure experiments. 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. 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 effects 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 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.

Keywords: Iterative Moment Estimation

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. Here by clusters we mean, in general, second-level units like schools, hospitals etc. (or subjects in repeated measure experiments). 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 clusters. 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 thought as the 'true' individual effects. Later, the final responses are generated conditionally on the random effect realization with an additional source of variability due to the individual errors, which may be assumed heteroscedastic among the clusters.

Linea Mixed Models are usually applied to assess some null hypotheses on the fixed effects taking into account that data are collected hierarchically, or sometimes to test for the presence of random effects (e.g. spatial data). In some other situations, we may want to compare the effects of a certain treatment on one or more populations, from which the samples have been drawn. In all these cases, an estimate of both sources of variability is required.

In a parametric framework, the individual source of variability is usually assumed to be the same for all clusters, and the dispersion parameter is assumed to be known [4]. Similarly, it is possible to summaryze the information in each cluster through a conditional estimation of the effects of interest, as well as their variability. In some other situations the data might be aggragated at the second level, and therefore parametric solutions, such as that provided in [1] are not applicable.

In what follows an iterative algorithm to obtain an estimation of the random effect variances is proposed. This method is applicable whenever the variability (or an estimate) of the cluster responses is available. This is what is meant by *partial information from the clusters* here. In Sections 2 and 3 the proposal is introduced, and in Section 4 an example is provided.

2 Variance Estimation

Let $\mathbf{y}$ be a $n \times 1$ vector of independent but not identically distributed random variables, whose distributions are continuous and admit finite second moment. Formally $Y_i \sim F_i(y_i)$, where $F_i(\cdot)$ satisfies $Var(Y_i) < \infty$, $i = 1, \ldots, n$. 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

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\gamma}$ is a $q \times p$ matrix of coefficients, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of independent random errors with zero mean and variance equal to $Var(\epsilon_i) = v_{ii}$. We will denote with $\mathbf{V}$ the variance covariance matrix of $\boldsymbol{\epsilon}$ and write $\boldsymbol{\epsilon} \sim (\mathbf{0}, \mathbf{V})$ to underline that the attention will be focused on the first two moments of the joint distribution.

We will also assume that the error term $\boldsymbol{\epsilon}$ is in fact a sum of two independent random vectors $\mathbf{u}$ and $\boldsymbol{\varepsilon}$, where $\mathbf{u}$ is a source of noise common to all the observations, and $\boldsymbol{\varepsilon}$ is a specific to each observation. Namely, $\mathbf{u} \sim (\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$, and $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots \sigma_n^2)$. The reason of this assumptions will be clearer in the example section. Model (1) can be equivalently viewed in terms of the joint distribution of the response, i.e. $\mathbf{y} \sim (\mathbf{Z}\gamma, \mathbf{V})$, where $\mathbf{V} = diag(\sigma_u^2 + \sigma_i^2)$.

Since the response is heteroscedastic, it is well known [3] that the BLUE estimator of the parameter vector $\boldsymbol{\gamma}$ is given by the weighted least square (WLS) estimator $\hat{\gamma} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, which is obtained by minimizing with respect to $\boldsymbol{\gamma}$ the residual deviance of the general linear model:

$$\mathbf{V}^{-1/2}\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{Z}\boldsymbol{\gamma} + \mathbf{V}^{-1/2}\boldsymbol{\epsilon}. \tag{2}$$

If we let $\tilde{\mathbf{y}} = \mathbf{V}^{-1/2}\mathbf{y}$, then $\tilde{\mathbf{y}} \sim (\tilde{\mathbf{Z}}\gamma, \mathbf{I}_n)$, where $\tilde{\mathbf{Z}} = \mathbf{V}^{-1/2}\mathbf{Z}$. The residuals of model (2) can be viewed as the projection of the weighted response $\tilde{\mathbf{y}}$ into the subspace orthogonal to the one spanned by the columns of $\tilde{\mathbf{Z}}$. 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 residuals of model (2) can therefore be written as:

$$\tilde{\mathbf{r}} = (\mathbf{I} - \mathbf{H_V})\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{H_V})\tilde{\epsilon}, \tag{3}$$

where $\tilde{\boldsymbol{\epsilon}}$ is the weighted vector of errors: $\tilde{\boldsymbol{\epsilon}} = \mathbf{V}^{-1/2}\boldsymbol{\epsilon}$. Note that, while the elements of $\tilde{\boldsymbol{\epsilon}}$ are identically distributed up to their second moments, the elements of $\tilde{\mathbf{r}}$ are not. Indeed $Var(\tilde{\mathbf{r}}) = (\mathbf{I} - \mathbf{H_V})Var(\tilde{\boldsymbol{\epsilon}})(\mathbf{I} - \mathbf{H_V}) = (\mathbf{I} - \mathbf{H_V})$, which depends on $\mathbf{V}$. Since $\mathbf{H_V}$ is a projection matrix there exists an $n \times n$ matrix $\mathbf{\Gamma_V}$ that satisfies $\mathbf{\Gamma_V}\mathbf{\Gamma_V}' = (\mathbf{I} - \mathbf{H_V})$. Note that $\mathbf{\Gamma_V}$ is of rank $n - q$ and it is orthogonal, therefore $\mathbf{\Gamma_V}'\mathbf{\Gamma_V} = \mathbf{I}_{n-q}$.

Now consider the vector of re-scaled residuals $\mathbf{r}^* = \mathbf{\Gamma_V}'\tilde{\mathbf{r}}$; its components are i.i.d. up to their second moment, indeed $E[\mathbf{r}^*] = \mathbf{0}$ and

$$Var(\mathbf{r}^*) = \mathbf{\Gamma_V}'Var(\tilde{\mathbf{r}})\mathbf{\Gamma_V} = \mathbf{\Gamma_V}'(\mathbf{\Gamma_V}\mathbf{\Gamma_V}')\mathbf{\Gamma_V} = \mathbf{I}_{n-q}.$$

Note that, in fact, the dimension of the (re-scaled) residuals is only $n - q$, and not n as for the original response $\mathbf{y}$. The re-scaled residuals can be expressed as

$$\mathbf{r}^* = \mathbf{\Gamma_V}'\tilde{\boldsymbol{\epsilon}}, \tag{4}$$

where the equivalence has been obtained by noticing that $\mathbf{\Gamma_V}'(\mathbf{I} - \mathbf{H_V}) = \mathbf{\Gamma_V}'$. A further pre-multiplication for $\mathbf{\Gamma_V}$ would be useless, since it would lead back to model (3). The last transformation vanishes the dependence between the residuals and the matrix $\mathbf{V}$. In fact, this is true only apparently, since the re-scaled residuals of model (4) are obtained from the original response vector $\mathbf{y}$ through a pre-multiplication for the projection matri $\mathbf{H_V}$, that depends on $\mathbf{V}$. Since we are interested on estimating the variability of the common component $\mathbf{u}$, the intuitive idea is to express the uncorrelated residuals back into the original scale; that is, to consider the vector $\mathbf{r} = \mathbf{V}^{1/2}\mathbf{\Gamma_V}'\tilde{\boldsymbol{\epsilon}}$ as a guess for our purpose, because

$$Var(\mathbf{r}) = \mathbf{V}^{1/2}\mathbf{\Gamma_V}'Var(\tilde{\boldsymbol{\epsilon}})\mathbf{\Gamma_V}\mathbf{V}^{1/2} = \mathbf{V}^{1/2}\mathbf{\Gamma_V}'\mathbf{\Gamma_V}\mathbf{V}^{1/2} = \mathbf{V}_{n-q}, \tag{5}$$

where $\mathbf{V}_{n-q} = diag(v_{ii})$ and $v_{ii} = \sigma_u^2 + \sigma_i^2$ if $i \leq n - q$, and zero otherwise.

The quadratic form $\mathbf{r}'\mathbf{r}$ satisfies

$$E[\mathbf{r}'\mathbf{r}] = tr(\mathbf{V}_{n-q}) = \sum_{i=1}^{n-q} v_{ii} = (n-q)\sigma_u^2 + \sum_{i=1}^{n-q} \sigma_i^2. \tag{6}$$

If the elements σ_i^2 are assumed to be known (this is what 'partial information from the clusters' is meant here) then a moment-estimator of σ_u^2 is given by

$$\hat{\sigma}_u^2 = \frac{1}{n-q}\left[\mathbf{r}'\mathbf{r} - \sum_{i=1}^{n-q} \sigma_i^2\right]. \tag{7}$$

Although it may be reasonable, since the estimator in (7) is unbiased and consistent, this does not seem to be a practicable choice, since it only depends on the first $n-q$ elements of $diag(\mathbf{V})$.

Alternatively, the residual vector $\mathbf{r}$ can be expressed as a function of the errors in the original scale:

$$\mathbf{r}'\mathbf{r} = \boldsymbol{\epsilon}'[(\mathbf{V}^{-1/2}\boldsymbol{\Lambda}_{\mathbf{V}}\mathbf{V}^{1/2})(\mathbf{V}^{1/2}\boldsymbol{\Lambda}'_{\mathbf{V}}\mathbf{V}^{-1/2})]\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'\mathbf{G}'\mathbf{G}\boldsymbol{\epsilon},$$

where $\mathbf{G} = \mathbf{V}^{1/2}\boldsymbol{\Lambda}'_{\mathbf{V}}\mathbf{V}^{-1/2}$. With this in mind, we finally obtain:

$$E[\mathbf{r}'\mathbf{r}] = tr(\mathbf{G}'\mathbf{G}\mathbf{V}) = \sum_{i=1}^{n} g_{ii}v_{ii} = \sigma_u^2 \sum_{i=1}^{n} g_{ii} + \sum_{i=1}^{n} g_{ii}\sigma_i^2, \tag{8}$$

where g_{ii} is the ith diagonal entry of $\mathbf{G}'\mathbf{G}$, which gives raise to the estimator:

$$\hat{\sigma}_u^2 = \frac{1}{tr(\mathbf{G}'\mathbf{G})}\left[\mathbf{r}'\mathbf{r} - \sum_{i=1}^{n} g_{ii}\sigma_i^2\right]. \tag{9}$$

This expression will be useful in the iterative estimation introduced in the next section, because it allows to express the expectation of the quadratic form (8) in terms of the unknown parameter σ_u^2 and the trace elements of the matrix $\mathbf{G}$. An initial guess of the variance/covariance error matrix depending only on the cluster's variability (say $\mathbf{V}_0 = diag\{\sigma_i^2\}$) implies a complete knowledge of $\mathbf{G}$, so an initial estimate of σ_u^2 can be obtained. The matrix $\mathbf{V}$ will be then updated iteratively by adding the current estimate of σ_u^2 the diagonal elements of $\mathbf{V}_0$ until convergence.

3 Iterative Estimation

The results of section 2 hold provided that the individual variances σ_i^2 are known. The relationship between (5) and (8) is given by the equivalence:

$$tr(\mathbf{G}'\mathbf{G}\mathbf{V}) = tr(\mathbf{V}^{-1/2}\boldsymbol{\Lambda}_{\mathbf{V}}\mathbf{V}\boldsymbol{\Lambda}'_{\mathbf{V}}\mathbf{V}^{1/2}) = tr(\boldsymbol{\Lambda}_{\mathbf{V}}\mathbf{V}\boldsymbol{\Lambda}'_{\mathbf{V}}) = tr(\boldsymbol{\Lambda}'_{\mathbf{V}}\boldsymbol{\Lambda}_{\mathbf{V}}\mathbf{V}) = tr(\mathbf{V}_{n-q}),$$

although (7) is a function of the first $n-q$ elements of $diag(\mathbf{V})$, whereas (8) is a weighed sum of all those elements. Note that in (8) the role usually played by the degrees of freedom in the estimation is replaced by the trace of $\mathbf{G}'\mathbf{G}$, which is not an integer.

In order to estimate the unknown parameter σ_u^2, we then propose an iterative algorithm: at the beginning an initial guess for the error variance/covariance matrix is given by $\mathbf{V}_0 = diag(\sigma_i^2)$, $i = 1, \ldots, n$. Then repeat until convergence the following instructions; at step s:

- Estimate the uncorrelated residuals under the original scale as

$$\mathbf{r}_s = \mathbf{V}_{s-1}^{1/2}\boldsymbol{\Lambda}'_{\mathbf{V}_{s-1}}\mathbf{V}_{s-1}^{-1/2}\mathbf{y} = \mathbf{G}_s\mathbf{y},$$

  where $\boldsymbol{\Lambda}_{\mathbf{V}_{s-1}}$ is the matrix of the eigen vectors of $(\mathbf{I} - \mathbf{H}_{\mathbf{V}_{s-1}})$ multiplied by the identiy matrix of order $n-q$, and $\mathbf{H}_{\mathbf{V}_{s-1}}$ is the projection matrix into the residual space defined in section 2 with $\mathbf{V}$ replaced by $\mathbf{V}_{s-1}$ (M-step);

- Obtain the trace of $\mathbf{G}_s'\mathbf{G}_s$ and the current estimate of σ_u^2 (E-step):

$$\hat{\sigma}_{u_s}^2 = \frac{1}{tr(\mathbf{G}_s'\mathbf{G}_s)}\left[\mathbf{r}_s'\mathbf{r}_s - \sum_{i=1}^{n} g_{ii_s}\sigma_i^2\right] \qquad g_{ii_s} = (\mathbf{G}_s'\mathbf{G}_s)_{ii};$$

- Set $\mathbf{V}_s = diag\{I_{u_s} \cdot \hat{\sigma}_{u_s}^2 + \sigma_i^2\}$, where $I_{u_s} = 1$ only if $\hat{\sigma}_{u_s}^2 > 0$.

Note that it is possible to obtain negative estimates of σ_u^2 (in this case the estimate will be forced to zero and $\hat{\mathbf{V}} = diag(\sigma_i^2)$), because the estimator is based on a difference of real values. If this does not happen, the converge is ensured by the EM algorithm theory ([2]), of which our 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 likelihood of the model), whereas the expectation step is implicit in the moment estimation of σ_u^2. The algorithm can be stopped after a maximum number of iterations, or as soon as some convergence criterion is satisfied.

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 in n_i trials, $i = 1, \ldots, n$. Assume a linear mixed model with random intercept the $n_i \times 1$ response of each subject:

$$\mathbf{y}_i = \mu_i \cdot \mathbf{1}_{n_i} + \boldsymbol{\varepsilon}_i, \qquad \mu_i = \mu + u_i,$$

where μ is the population mean, and $u_i \sim (0, \sigma_u^2)$ is the random effect related to the ith subject. As far as the error term is concerned, we could either assume heteroscedasticity, i.e. $\boldsymbol{\varepsilon}_i \sim (\mathbf{0}_{n_i}, \sigma_i^2\mathbf{I}_{n_i})$, or homoscedasticity ($\sigma_i^2 = \sigma^2\ \forall\ i$) since, unless the same number of trials is done on each subject, we would obtain the same result: the OLS estimator of μ within each subject is given by the subjects' sample mean $\hat{\mu}_i = \bar{y}_i$, and since the estimation is done conditionally on the subjects we have $\hat{\mu}_i|u_i \sim (\mu + u_i, \sigma_i^2/n_i)$. If we consider the distribution of the OLS estimators among the subjects (i.e. the unconditional distribution of $\hat{\mu}_i$), then $\hat{\mu}_i \sim (\mu, \sigma_u^2 + \sigma_i^2/n_i)$.

Now suppose that an additional variable Z, such as age, is available for the subjects, and that we are interested on seeing whether there is an interaction between age and the response. If we let $\hat{\boldsymbol{\mu}} = [\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n]'$ to be the new vector of responses, and $\mathbf{Z} = [\mathbf{1}_n, \mathbf{z}]$, where $\mathbf{z}$ is the vector with the ages of the subjects, we can consider the analogous of model (1) by replacing $\mathbf{y}$ with $\hat{\boldsymbol{\mu}}$, $\boldsymbol{\gamma} = [\gamma_0, \gamma_1]'$ and $\mathbf{V} = diag\{\sigma_u^2 + \sigma_i^2/n_i\}$. Then testing for no interaction between response and age would be translated into testing the null hypothesis $H_0 : \gamma_1 = 0$. The test can be done, for instance, by comparing the residual deviances of the full and the reduced model. Whatever statistical test is applied, the knowledge of both variance components will be required. In practice, the individual components σ_i^2 should either (assumed to be) known or obtained by some consistent estimators, whereas the common variance component σ_u^2 will be estimated by the iterative algorithm.

The estimation procedure proposed can be easily extended to many other situations, including between-group comparisons (in which case $\mathbf{Z}$ would be a contrast matrix), or when one is interested on possible interactions between treatment and cluster covariates. More generally it can be applied in any case some estimators of the parameters of interest and of their variances is available. Note that this also includes the Generalized Linear Mixed Models (GLMM), provided that these models are fitted within each subject. This is the case when $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, is the model for the response of the ith cluster in the jth trial, where $\mu_{ij} = E[Y_{ij}]$, $g(\cdot)$ is the link function, $\mathbf{x}_{ij}$ is the individual treatment for the jth first level observation, and $\boldsymbol{\beta}_{ij}$ is a $p \times 1$ vector of effects, which can be specified as the sum of a fixed and a random component, i.e. $\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_j + \mathbf{u}_i$ (the fixed part depending on the population and the random part specific for each subject). Note that this estimation method is applied separately to each component of $\boldsymbol{\beta}_{ij}$, and therefore it can be used to estimate the whole variance covariance matrix of the random effects in case these are assumed to be uncorrelated.

References

[1] Bates D., Maechler M., Bolker B. (2011). *Lme4: Linear mixed-effects models using S4 classes.* `http://CRAN.R-project.org/package=lme4`.

[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.

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

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

Department of Statistical Sciences
University of Padua
Italy