Random Coefficient Models[*]

Nicholas T. Longford
SNTL and UPF, Barcelona, Spain

Summary

Random coefficient models are intended for settings with two or more sources of random variation. The widest range of applications is found for them when observational units form natural clusters, such that the units within a cluster are more similar than units in general. Models for independent observations have to be extended to allow for within- and between-cluster variation.

Keywords: *Analysis of variance; borrowing strength; clusters; correlation structure; empirical Bayes analysis; longitudinal analysis; maximum likelihood; ordinary regression.*

[*] N. T. Longford, SNTL and Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain; email: `NTL@sntl.co.uk`

Independence of the observations is a key assumption of many standard statistical methods, such as analysis of variance (ANOVA) and ordinary regression, and some of its extensions. Common examples of data structures that do not fit into such a framework arise in longitudinal analysis, in which observations are made on subjects at subject-specific sequences of time points, and in studies that involve subjects (units) ocurring naturally in clusters, such as individuals within families, schoolchildren within classrooms, employees within companies, and the like. The assumption of independence of the observational units is not tenable when observations within a cluster tend to be more similar than observations in general. Such similarity can be conveniently represented by a positive correlation (dependence).

This article describes an adaptation of the ordinary regression for clustered observations. Such observations require two indices, one for elements within clusters, $i = 1, \ldots, n_j$, and another for clusters, $j = 1, \ldots, m$. Thus, we have $n = n_1 + \cdots + n_m$ elementary units and m clusters. The ordinary regression model

$$y_{ij} \;=\; \mathbf{x}_{ij}\,\boldsymbol{\beta} + \varepsilon_{ij}\,, \qquad (1)$$

with the usual assumptions of normality, independence and equal variance (homoscedasticity) of the deviations ε_{ij}, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, i.i.d., implies that the regressions within the clusters j have a common vector of coefficients $\boldsymbol{\beta}$. This restriction can be relaxed by allowing the regressions to differ in their intercepts. A practical way of defining such a model is by the equation

$$y_{ij} \;=\; \mathbf{x}_{ij}\,\boldsymbol{\beta} + \delta_j + \varepsilon_{ij}\,, \qquad (2)$$

where δ_j, $j = 1, \ldots, m$, are a random sample from a centred normal distribution, $\delta_j \sim \mathcal{N}(0, \sigma_{\mathrm{B}}^2)$, i.i.d., independent from the ε's. This differs from the model for analysis of covariance (ANCOVA) only by the status of the deviations δ_j. In ANCOVA, they are fixed (constant across hypothetical replications), whereas in (2) they are random.

In the model in (2), the within-cluster regressions are parallel — their intercepts are $\beta_0 + \delta_j$, but the coefficients on all the other variables in $\mathbf{x}$ are common to the clusters. A more appealing interpretation of the model is that observations in a cluster are correlated,

$$\mathrm{cor}\left(y_{i_1,j}\,, y_{i_2,j}\right) \;=\; \frac{\sigma_{\mathrm{B}}^2}{\sigma^2 + \sigma_{\mathrm{B}}^2}\,,$$

because they share the same deviation δ_j. Further relaxation of how the within-cluster regressions differ is attained by allowing some (or all) the regression slopes to be specific to the clusters. We select a set of variables in $\mathbf{x}$, denoted by $\mathbf{z}$, and assume that the regressions with respect to these variables differ across the clusters, but are constant with respect to the remaining variables;

$$y_{ij} \;=\; \mathbf{x}_{ij}\,\boldsymbol{\beta} + \mathbf{z}_{ij}\,\boldsymbol{\delta}_j + \varepsilon_{ij}\,, \qquad (3)$$

where $\boldsymbol{\delta}_j$, $j = 1, \ldots, m$, are a random sample from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{B}})$, independent from the ε's. We say that the variables in $\mathbf{z}$ are associated with (cluster-level) variation. The variance of an observation y_{ij}, without conditioning on the cluster j, is

$$\mathrm{var}\,(y_{ij}) \;=\; \sigma^2 + \mathbf{x}_{ij} \boldsymbol{\Sigma}_{\mathrm{B}}\, \mathbf{x}_{ij}^{\top} \,.$$

We refer to σ^2 and $\mathbf{z}_{ij} \boldsymbol{\Sigma}_{\mathrm{B}}\, \mathbf{z}_{ij}^{\top}$ as the *variance components* (at the elementary and cluster levels, respectively). The principle of invariance with respect to linear transformations of $\mathbf{z}$ implies that the intercept should always be included in $\mathbf{z}$, unless $\mathbf{z}$ is empty, as in the model in (1). The function $V(\mathbf{z}) = \mathbf{z}\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{z}^{\top}$, over the feasible values of $\mathbf{z}$, defines the *pattern of variation*, and it can be described by its behaviour (local minima, points of inflection, and the like). By way of an example, suppose $\mathbf{z}$ contains the intercept and a single variable z. Denote the variances in $\boldsymbol{\Sigma}_{\mathrm{B}}$ by σ_0^2 and σ_{z}^2, and the covariance by $\sigma_{0\mathrm{z}}$. Then

$$V(\mathbf{z}) \;=\; \sigma_0^2 + 2z\sigma_{0\mathrm{z}} + z^2\sigma_{\mathrm{z}}^2 \,, \tag{4}$$

and this quadratic function has a unique minimum at $z^* = -\sigma_{0\mathrm{z}}/\sigma_{\mathrm{z}}^2$, unless $\sigma_{\mathrm{z}}^2 = 0$, in which case we revert to the model in (2) in which $V(\mathbf{z})$ is constant. Figure 1 illustrates four patterns of variation on examples with a single covariate.

The model in (3) is fitted by maximum likelihood (ML) which maximizes the log-likelihood function

$$l\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}_{\mathrm{B}}\right) \;=\; -\frac{1}{2}\sum_{j=1}^{m} \left[\log\left\{\det\left(\mathbf{V}_j\right)\right\} \;+\; \left(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}\right)^{\top} \mathbf{V}_j^{-1} \left(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}\right)\right] \,, \tag{5}$$

where $\mathbf{V}_j$ is the variance matrix of the observations in cluster j, $\mathbf{y}_j$ the vector of the outcomes for the observations in cluster j, and $\mathbf{X}_j$ the corresponding regression design matrix formed by vertical stacking of the rows $\mathbf{x}_{ij}$, $i = 1, \ldots, n_j$. The variation design matrices $\mathbf{Z}_j$, $j = 1, \ldots, m$, are defined similarly; with them, $\mathbf{V}_j = \sigma^2\mathbf{I}_{n_j} + \mathbf{Z}_j\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{Z}_j^{\top}$, where $\mathbf{I}_{n_j}$ is the $n_j \times n_j$ identity matrix. The Fisher scoring algorithm for maximising the log-likelihood in (5) is described in the Appendix; for details and applications, see see Longford (1993), and for an alternative method Goldstein (2000). These and other algorithms are implemented in most standard statistical packages. A key to their effective implementation are closed-form expressions for the inverse and determinant of patterned matrices (Harville, 1997).

Model selection entails two tasks, selecting a set of variables to form $\mathbf{x}$ and selecting its subset to form $\mathbf{z}$. The variables in $\mathbf{x}$ can be defined for elements or clusters; the latter can be defined as being constant within clusters. Inclusion of cluster-level variables in $\mathbf{z}$ does not have an interpretation in terms of varying regression coefficients, so associating them with variation is in most contexts not meaningful. However, the identity in (4) and its generalisations for $\boldsymbol{\Sigma}_{\mathrm{B}}$ with more than two rows and columns indicate that $\mathbf{z}$ can be used for modelling variance heterogeneity. The likelihood ratio test statistic and various information criteria can be used for selecting among alternative models, so long

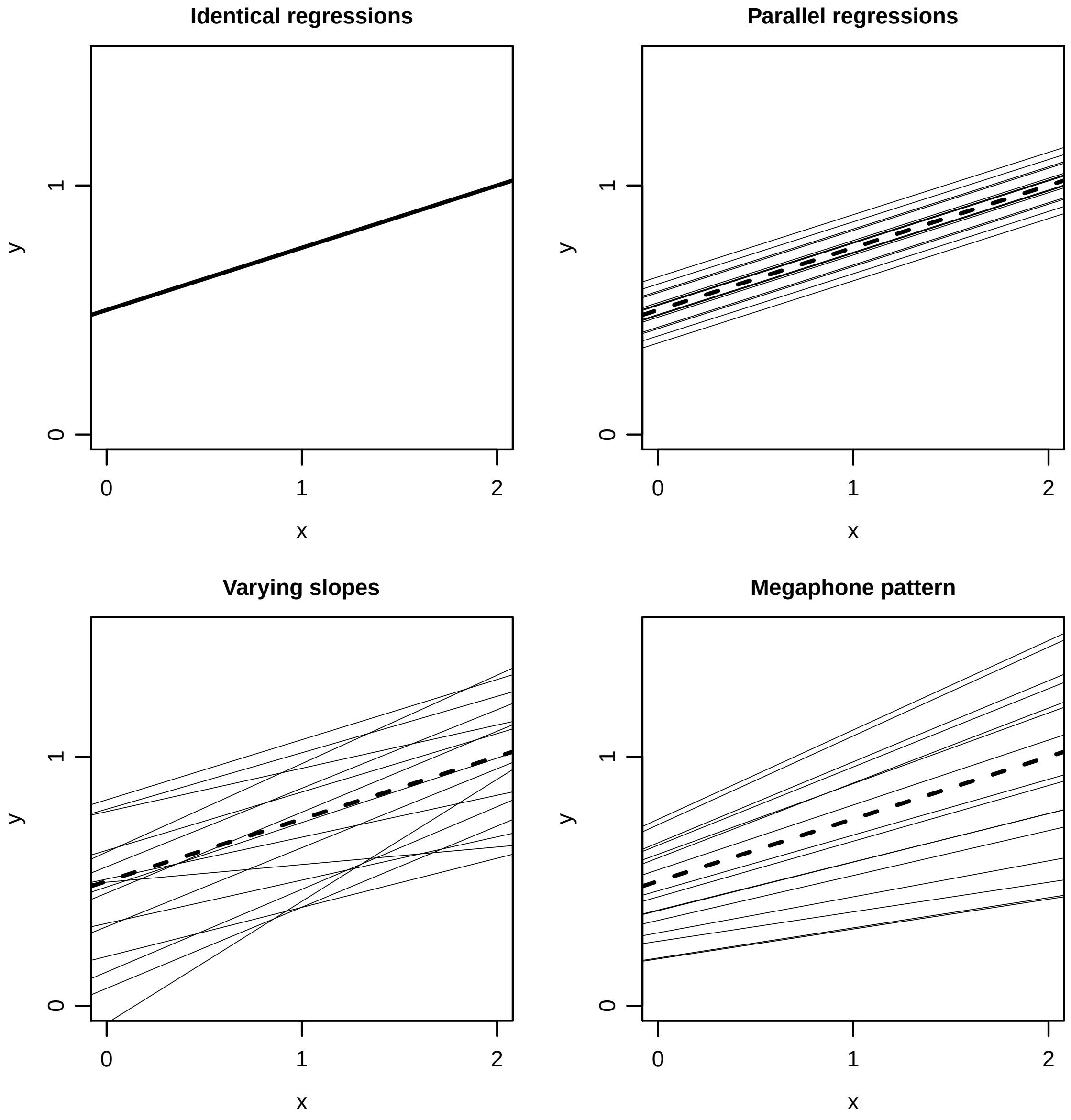

Figure 1: Patterns of variation in random coefficient models with a single covariate. Thick dashes mark the average regression ($\mathbf{x}\boldsymbol{\beta}$) and thin solid lines the within-cluster regressions ($\mathbf{x}\boldsymbol{\beta} + \mathbf{z}\boldsymbol{\delta}_j$). In the megaphone pattern, the matrix $\boldsymbol{\Sigma}_{\mathrm{B}}$ is singular, with $\sigma_{\mathrm{z}}^2 > 0$.

as one is a submodel of the other; that is, the variables in both $\mathbf{x}$ and $\mathbf{z}$ of one model are subsets of (or coincide with) their counterparts in the other model.

Random coefficients can be applied to a range of models much wider than ordinary regression. In principle, we can conceive any *basis model*, characterized by a vector of parameters, which applies to every cluster. A subset of these parameters is constant across the clusters and the remainder varies according to a model for cluster-level variation. The latter model need not be a multivariate normal distribution, although suitable alternatives to it are difficult to identify. The basis model itself can be complex, such as a random coefficient model itself. This gives rise to three- or, generally, *multilevel models*, in which elements are clustered within two-level units, these units in three-level units, and so on.

Generalized linear mixed models (GLMM) have generalized linear models (McCullagh and Nelder, 1989) as their basis; see Pinheiro and Bates (2000). For cluster j we posit the model

$$g\left\{\mathrm{E}(\mathbf{y}_j \,|\, \boldsymbol{\delta}_j)\right\} \;=\; \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\delta}_j \,,$$

with a (monotone) link function g, which transforms the expected outcomes to a linear scale; the outcomes (elements of $\mathbf{y}_j$) are conditionally independent given $\boldsymbol{\delta}_j$ and have a specified distribution, such as binary or gamma. Some advantage, in both interpretation and computing, is gained by using *canonical* link functions, for which the conditional likelihood has a compact set of sufficient statistics. For example, the logistic link, $g(p) = \log(p) - \log(1-p)$, is the canonical link for binary outcomes, and the logarithm is the canonical link for Poisson outcomes. Models with the (standard) normality assumptions correspond to the link $g(y) = y$.

Without conditioning, the likelihood for non-normally distributed $\mathbf{y}_j$ has an intractable form. An established approach to fitting GLMM maximises a (tractable) normal-like approximation to the log-likelihood. This algorithm can be described as an iteratively reweighted version of the Fisher scoring (or another) algorithm for fitting the (normal) linear mixed model. With the advent of modern computing, using numerical quadrature and other methods for numerical integration has because feasible, especially for fitting models with univariate deviations δ_j. An alternative framework for GLMM and a computational algorithm for fitting them are constructed by Lee and Nelder (2001).

Random coefficient models are well suited for analysing surveys in which clusters arise naturally as a consequence of the organisation (design) of the survey and the way the studied population is structured. They can be applied also in settings in which multiple observations are made on subjects, as in longitudinal studies (Molenberghs and Verbeke, 2000). In some settings it is contentious as to whether the clusters should be regarded as fixed or random. For example, small-area estimation (Rao, 2003) is concerned with inferences about districts or another partition of a country when some (or all) districts are represented in the analysed national survey by small subsamples. In one perspective, district-level quantities, such as their means of a variable, should be regarded as fixed because they are the inferential targets, fixed across hypothetical replications. When they are assumed to be random the (random coefficient) models are often more parsimonious than their fixed-effects (ANCOVA) counterparts, because the number of parameters involved does not depend on the number of clusters.

Borrowing strength (Robbins, 1955, Efron and Morris, 1972) is a general principle for efficient inference about each cluster (district) by exploiting the similarity of the clusters. It is the foundation of the empirical Bayes analysis, in which the between-cluster variance matrix plays a role similar to the Bayes prior for the within-cluster regression coefficients. The qualifier 'empirical' refers to using a data-based estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{B}}$ in place of the unknown $\mathbf{\Sigma}_{\mathrm{B}}$.

References

Efron, B., and Morris, C. N. (1972). Limiting the risk of Bayes and empirical Bayes estimators — Part II: empirical Bayes case. *Journal of the American Statistical Association* **67**, 1286–1289.

Goldstein, H. (2000). *Multilevel Statistial Models.* 2nd Edition. London: Edward Arnold.

Harville, D. (1997). *Matrix Algebra from a Statistician's Perspective.* New York: Springer-Verlag.

Lee, Y., and Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika* **88**, 987–1004.

Longford, N. T. (1993). *Random Coefficient Models.* Oxford: Oxford University Press.

Magnus, J. R., and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* New York: Wiley.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd Edition. London: Chapman and Hall.

Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-Effects Models in* S *and* Splus. New York: Springer-Verlag.

Rao, J. N. K. (2003). *Small Area Estimation.* New York: Wiley and Sons.

Robbins, H. (1955). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* **1**, 157–164. Berkeley, CA: University of California Press.

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

Appendix. Fisher scoring algorithm

This Appendix describes a method for fitting a random coefficient model by maximum likelihood. We prefer to use the scaled variance matrices $\mathbf{W}_j = \sigma^{-2}\mathbf{V}_j$ and $\mathbf{\Omega} = \sigma^{-2}\mathbf{\Sigma}_{\rm B}$, so that $\mathbf{W}_j = \mathbf{I}_{n_d} + \mathbf{Z}_j\mathbf{\Omega}\mathbf{Z}_j^\top$ does not depend on σ^2. The log-likelihood in (5) is equal to

$$l\left(\boldsymbol{\beta}, \sigma^2, \mathbf{\Omega}\right) \;=\; -\frac{1}{2}\sum_{j=1}^{m}\left[n\log\left(\sigma^2\right) + \log\left\{\det\left(\mathbf{W}_j\right)\right\} \;+\; \frac{1}{\sigma^2}\,\mathbf{e}_j^\top\mathbf{W}_j^{-1}\mathbf{e}_j\right]\,. \tag{6}$$

where $\mathbf{e}_j = \mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}$ is the vector of residuals for cluster j. We have the following closed-form expressions for the inverse and determinant of $\mathbf{W}_j$:

$$\begin{aligned} \mathbf{W}_j^{-1} &= \mathbf{I}_{n_d} - \mathbf{Z}_j\mathbf{\Omega}\mathbf{G}_j^{-1}\mathbf{Z}_j^\top \\ \det\left(\mathbf{W}_j\right) &= \sigma^{2n_d}\det\left(\mathbf{G}_j\right)\,, \end{aligned} \tag{7}$$

where $\mathbf{G}_j = \mathbf{I}_r + \mathbf{Z}_j^\top\mathbf{Z}_j\mathbf{\Omega}$. All the matrices $\mathbf{G}_j$ have the same dimension, $r \times r$, where r is the number of variables in $\mathbf{Z}$.

Assuming that the log-likelihood l has a single maximum, it can be found as the root of the score vector. By matrix differentiation we obtain

$$\frac{\partial l}{\partial\boldsymbol{\beta}} \;=\; \frac{1}{\sigma^2}\sum_{j=1}^{m}\mathbf{X}_j^\top\mathbf{W}_j^{-1}\mathbf{e}_j\,,$$

and so the maximum likelihood estimator of $\boldsymbol{\beta}$ is the generalised least-squares estimator

$$\hat{\boldsymbol{\beta}} \;=\; \left(\sum_{j=1}^{m}\mathbf{X}_j^\top\hat{\mathbf{W}}_j^{-1}\mathbf{X}_j\right)^{-1}\sum_{j=1}^{m}\mathbf{X}_j^\top\hat{\mathbf{W}}_j^{-1}\mathbf{y}_j\,. \tag{8}$$

The matrices $\hat{\mathbf{W}}_j$ are the estimated versions of $\mathbf{W}_j$, with estimator $\hat{\mathbf{\Omega}}$ substituted for $\mathbf{\Omega}$. Estimation of $\mathbf{\Omega}$ is described below.

The elementary-level (residual) variance σ^2 is estimated by the root of its score,

$$\frac{\partial l}{\partial\sigma^2} \;=\; -\frac{1}{2}\left(\frac{n}{\sigma^2} - \frac{1}{\sigma^4}\sum_{j=1}^{m}\mathbf{e}_j^\top\mathbf{W}_j^{-1}\mathbf{e}_j\right)\,,$$

which is

$$\hat{\sigma}^2 \;=\; \frac{1}{n}\sum_{j=1}^{m}\mathbf{e}_j^\top\hat{\mathbf{W}}_j^{-1}\mathbf{e}_j\,.$$

The elements of $\mathbf{\Omega}$ are estimated by the Fisher scoring algorithm. Let these elements, without any redundancy, comprise the vector $\boldsymbol{\omega}$. In most applications, $\boldsymbol{\omega}$ comprises the $\frac{1}{2}r\times(r+1)$ unique elements, r variances and $\frac{1}{2}r \times (r-1)$ covariances. The Fisher scoring algorithm proceeds by iterations that update the estimate of ω, based on the score vector $\mathbf{s}$ and the expected information matrix $\mathbf{H}$ evaluated at the current solution. The vector $\mathbf{s}$ and matrix $\mathbf{H}$ are derived by matrix differentiation. See Magnus and Neudecker (1988) and Harville (1997) for background.

Let $\mathbf{i}_h$ be the $(r \times 1)$ indicator vector for element h. For example, when $r = 5$, $\mathbf{i}_2 = (0,1,0,0,0)^\top$. The element of $\boldsymbol{\omega}$ that corresponds to the (scaled) covariance (h, h') in $\boldsymbol{\Omega}$ can be expressed as

$$\omega \,=\, \mathbf{i}_h^\top \boldsymbol{\Omega} \mathbf{i}_{h'} \,,$$

and $\partial\boldsymbol{\Omega}/\partial\omega = \mathbf{i}_h \, \mathbf{i}_{h'}^\top + \mathbf{i}_{h'} \, \mathbf{i}_h^\top$. If we replace each (scaled) variance in $\boldsymbol{\omega}$ by its half, then this expression holds also for these half-variance parameters. With this parametrisation, noting that $\partial\mathbf{W}_j/\partial\omega = \mathbf{Z}_j \, \partial\boldsymbol{\Omega}/\partial\omega \, \mathbf{Z}_j^\top$, we have

$$\begin{aligned}
\frac{\partial l}{\partial \omega} &= -\frac{1}{2}\sum_{j=1}^{m}\left\{ \operatorname{tr}\left(\mathbf{W}_j^{-1}\frac{\partial \mathbf{W}_j}{\partial\omega}\right) - \frac{1}{\sigma^2}\,\mathbf{e}_j^\top \mathbf{W}_j^{-1}\frac{\partial \mathbf{W}_j}{\partial\omega}\mathbf{W}_j^{-1}\mathbf{e}_j \right\} \\
&= \sum_{j=1}^{m}\left(-\mathbf{i}_h^\top bUU_j \, \mathbf{i}_{h'} + \frac{1}{\sigma^2}\, \mathbf{u}_j^\top \mathbf{i}_h \, \mathbf{u}_j^\top \mathbf{i}_{h'} \right) \\
&= \sum_{j=1}^{m}\left(-U_{j,hh'} + \frac{1}{\sigma^2}\, u_{j,h} \, u_{j,h'} \right) \,,
\end{aligned}$$

where $bUU_j \,=\, \mathbf{Z}_j^\top \mathbf{W}_j^{-1} \mathbf{Z}_j$, $\mathbf{u}_j \,=\, \mathbf{Z}_j^\top \mathbf{W}_j^{-1} \mathbf{e}_j$, $U_{j,hh'}$ is the (h, h')-element of bUU_j and $u_{j,h}$ the h-element of $\mathbf{u}_j$. Multiplying by a vector $\mathbf{i}_h$ amounts to extracting an element. Thus, evaluation of the score $\partial l/\partial\omega$ requires calculation of quadratic forms $\mathbf{q}^\top \mathbf{W}_j^\top \mathbf{Z}_j$. From (7) we have, for an arbitrary $n_d \times 1$ vector $\mathbf{q}$, the identity

$$\mathbf{Z}_j^\top \mathbf{W}_j^{-1} \mathbf{q} \,=\, \mathbf{G}_j^{-1} \mathbf{Z}_j^\top \mathbf{q} \,,$$

so we do not have to form the matrices $\mathbf{W}_j$ and do not have to invert any matrices of large size. Further differentiation yields the expression

$$\begin{aligned}
\frac{\partial^2 l}{\partial\omega_1 \, \partial\omega_2} &= \sum_{j=1}^{m}\left[\mathbf{i}_{h_1}^\top bUU_j \, \frac{\partial\boldsymbol{\Omega}}{\partial\omega_2} \, bUU_j \mathbf{i}_{h_1'} \,-\, \frac{2}{\sigma^2}\left\{ \mathbf{u}_j^\top \mathbf{i}_{h_1} \, \mathbf{u}_j^\top \frac{\partial\boldsymbol{\Omega}}{\partial\omega_2} \, bUU_j \mathbf{i}_{h_1'} + \mathbf{u}_j^\top \mathbf{i}_{h_1'} \, \mathbf{u}_j^\top \frac{\partial\boldsymbol{\Omega}}{\partial\omega_2} \, bUU_j \mathbf{i}_{h_1} \right\}\right] \\
&= \sum_{j=1}^{m} \Big\{ U_{j,h_1 h_2} \, U_{j,h_1' h_2'} + U_{j,h_1' h_2} \, U_{j,h_1 h_2'} \\
&\qquad - \frac{2}{\sigma^2}\left(u_{j,h_1} \, u_{j,h_2} \, U_{j,h_1' h_2'} + u_{j,h_1} \, u_{j,h_2'} \, U_{j,h_1' h_2} + u_{j,h_1'} \, u_{j,h_2} \, U_{j,h_1 h_2'} + u_{j,h_1'} \, u_{j,h_2'} \, U_{j,h_1 h_2} \right) \Big\}
\end{aligned}$$

for ω_1 and ω_2 associated with the respective elements $(h_1 \,, h_1')$ and $(h_2 \,, h_2')$ of $\boldsymbol{\Omega}$. The Hessian matrix, its negative expectation, comprises elements

$$\begin{aligned}
H\left(\omega_1 \,, \omega_2\right) &= \sum_{j=1}^{m}\left(\mathbf{i}_{h_1}^\top bUU_j \mathbf{i}_{h_2} \, \mathbf{i}_{h_1'}^\top bUU_j \mathbf{i}_{h_2'} \,+\, \mathbf{i}_{h_1'}^\top bUU_j \mathbf{i}_{h_2} \, \mathbf{i}_{h_1}^\top bUU_j \mathbf{i}_{h_2'} \right) \\
&= \sum_{j=1}^{m}\left(U_{j,h_1 h_2} \, U_{j,h_1' h_2'} + U_{j,h_1' h_2} \, U_{j,h_1 h_2'} \right) \,,
\end{aligned}$$

which are cluster-level totals of the cross-products of various elements of bUU_j. In each iteration t, the estimate of $\boldsymbol{\omega}$ is updated as

$$\hat{\boldsymbol{\omega}}^{(t)} \,=\, \hat{\boldsymbol{\omega}}^{(t-1)} + r\hat{\mathbf{H}}_t^{-1}\hat{\mathbf{s}}_t \,,$$

where $r = 1$, unless the updated matrix $\mathbf{\Omega}^{(t)}$ is not positive definite. One way to avoid this is to keep halving r until the updated matrix $\mathbf{\Omega}^{(t)}$ is positive definite. Having to do so in many (or all) iterations is usually a sign of having included too many variables in $\mathbf{Z}$, and the model should be revised accordingly. An alternative approach estimates the Cholesky (or another) decomposition of $\mathbf{\Omega}$. The iterations are terminated when the norm of the updating vector $\mathbf{H}^{-1}\mathbf{s}$ is sufficiently small.

Based on an article from Lovric, Miodrag (2011), *International Encyclopedia of Statistical Science.* Heidelberg: Springer Science + Business Media, LLC.