SEPARABILITY IN THE FIXED PART OF MULTILEVEL MODELS

G. Van Landeghem and P. Onghena

centrum voor methodologie
van het empirisch pedagogisch onderzoek
departement pedagogische wetenschappen
katholieke universiteit leuven

J. Van Damme

centrum voor secundair en hoger onderwijs
departement pedagogische wetenschappen
katholieke universiteit leuven

This research was supported by the Fonds voor Wetenschappelijk Onderzoek - Vlaanderen.

Correspondence regarding this report can be sent to G. Van Landeghem, K.U.Leuven, Departement Pedagogische Wetenschappen, Centrum voor Methodologie van het Empirisch Pedagogisch Onderzoek, Dekenstraat 2, B-3000 Leuven, België.
E-mail: Georges.VanLandeghem@ped.kuleuven.ac.be

Abstract

Separability in ordinary regression is achieved by partitioning the set of independent variables into mutually orthogonal subsets. The coefficient vector of each subset is separate: its estimate depends only on the response and on the independent variable scores of the subset. This article discusses the feasibility of formulating multilevel models with subsets of separate parameters in the fixed part. Generic sufficient conditions for separability and a series of rules and examples are provided. The search for instances of separability rests on an analysis of the covariance matrix of the multilevel model. Its structure, in terms of its eigenvalue/eigenvector decomposition, explains the role of within-group centered and orthogonalized variables in the discussion of separability and in related topics such as the issue of unbiasedness of fixed part parameters in cases of underspecification.

Key words: multilevel models, orthogonality, centering, separability, unbiasedness.

Introduction

Orthogonality, Centering, Separability, Unbiasedness: an Example

In his discussion of "the crucial importance of centering" in multilevel analysis, Raudenbush (1989a) provides an example of a two-level (pupils within schools) random intercept model, predicting math achievement by means of the socio-economic status (SES, a pupil level variable) and the school mean of the SES (a school level variable). When the pupil level variable is group mean centered (in other words: the SES score of each pupil is adjusted by subtracting the mean SES of his/her school), the point estimate of the coefficient of the SES does not change when the school mean of the SES is removed from the model. This interesting property disappears when the group mean centered SES is replaced by the uncentered variable. Raudenbush (1989a) concludes that using group mean centering "allows consistent estimation of within-group slopes even when the between-group model is misspecified! By using centering, the within-group predictors are orthogonal to all between-group predictors and so cannot be biased by a failure to include the appropriate between-group model."

Indeed, whenever the n elements of a sample (of, say, pupils) are partitioned into m groups (schools, for example), any $n \times 1$ data column P representing school mean centered scores (on some pupil level variable) is orthogonal to any $n \times 1$ data column G containing each basic unit's (pupil's) score on a group (school) level variable: $\mathrm{P}^{\mathsf{T}}\mathrm{G} = 0$ (the superscript T means "transpose"), in words: the inner product of P with G equals zero. Thus Raudenbush's (1989a) example exhibits both a case of orthogonality (between the school mean centered SES and the SES school mean associated with each pupil) and of what we will call "separability": the estimated coefficient of one variable (the school mean centered SES) does not depend on the

scores of some other variable (in this case: the school means variable).

Orthogonality and Separability in Ordinary Regression

The link between orthogonality and separability is familiar and readily verified in the setting of ordinary regression.

Sufficient conditions. When a response and two sets of, respectively, p_1 and p_2 independent variables have been measured for n units, the ordinary regression model postulates the existence of a $p_1 \times 1$ coefficient vector β_1, a $p_2 \times 1$ coefficient vector β_2 and a variance parameter σ^2 which are such that the response data column Y is a realization of a stochastic variable with distribution

$$\mathrm{N}\left(\mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2, \sigma^2\mathrm{I}_{n\times n}\right). \tag{1}$$

The $n \times (p_1 + p_2)$ matrix $[\mathrm{X}_1\mathrm{X}_2]$ which arises by joining the data blocks X_1 and X_2 (which store the measurements on the independent variables) is assumed to be of full rank. The symbol I indicates the unit matrix.

The maximum likelihood estimates of the coefficients can be expressed analytically as

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left([\mathrm{X}_1\mathrm{X}_2]^\mathsf{T}[\mathrm{X}_1\mathrm{X}_2]\right)^{-1}[\mathrm{X}_1\mathrm{X}_2]^\mathsf{T}\mathrm{Y}. \tag{2}$$

With the help of the formula for the inverse of a symmetric two by two block matrix (see, for example, Bryk and Raudenbush, 1992, p. 238) the estimate for β_1 can be written as

$$\begin{aligned}\hat{\beta}_1 = &\left(\left(\mathrm{X}_1^\mathsf{T}\mathrm{X}_1 - \mathrm{X}_1^\mathsf{T}\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\mathrm{X}_1\right)^{-1}\mathrm{X}_1^\mathsf{T}\right. \\ &\left. - \left(\mathrm{X}_1^\mathsf{T}\mathrm{X}_1 - \mathrm{X}_1^\mathsf{T}\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\mathrm{X}_1\right)^{-1}\mathrm{X}_1^\mathsf{T}\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\right)\mathrm{Y}.\end{aligned} \tag{3}$$

Equation (3) demonstrates that when the data satisfy

$$\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\mathrm{X}_1 = 0_{n\times p_1} \tag{4}$$

the formula for $\hat{\beta}_1$ is greatly simplified:

$$\hat{\beta}_1 = (\mathrm{X}_1^\mathsf{T}\mathrm{X}_1)^{-1}\mathrm{X}_1^\mathsf{T}\mathrm{Y}. \tag{5}$$

As X_2 and, consequently, $(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}$ have rank p_2, the conditions (4) are equivalent to

$$\mathrm{X}_2^\mathsf{T}\mathrm{X}_1 = 0_{p_2\times p_1}. \tag{6}$$

When for a given dataset X_1 the dataset X_2 satisfies the p_2p_1 conditions expressed by (6) — a set of conditions which, according to (4), means that in any regression of a column of X_1 on the columns of X_2 all coefficients are zero — the estimate $\hat{\beta}_1$ does not depend on any further details of X_2: according to (5), it can be calculated on the basis of X_1 and Y alone. We describe this situation by saying that the coefficients β_1 of X_1 are *separate* in model (1). The orthogonality conditions $\mathrm{X}_2^\mathsf{T}\mathrm{X}_1 = 0$ are sufficient for the *separability* of the parameter vector β_1 in model (1).

Note that, in this case of ordinary regression, the conditions $\mathrm{X}_2^\mathsf{T}\mathrm{X}_1 = 0$ imply both that β_1 is separate and that β_2 is separate in model (1). Also, as a straightforward extension, note that when the set of explanatory variables in an ordinary regression analysis can be partitioned into m mutually orthogonal subsets ($m \geq 2$), each subset of coefficients is separate.

Advantage of separate coefficients. The orthogonality conditions $\mathrm{X}_2^\mathsf{T}\mathrm{X}_1 = 0$ ensure that the coefficients β_1 are separate within model (1). Moreover, they imply that the coefficients remain separate *and* that their estimation formula (5) remains unchanged when independent variables are removed from the second subset (the columns of X_2). In addition, they ensure the separability of the β_2 coefficients and

hand down the separability property to all models nested within model (1). Thus, in a situation where there are good reasons to assume that model (1) is an over-specification, the search for a better model (among all the models nested within model (1)) is simplified when the orthogonality conditions $X_2^T X_1 = 0$ are satisfied. The specification search among $p_1 + p_2$ variables is then effectively replaced by two searches among p_1 and p_2 variables respectively.

In practice, the columns of $[X_1 X_2]$ are often derived from a larger set of "raw" variables. Obviously, the main consideration in such a data reduction step is to achieve an acceptable operationalization of the relevant concepts. If this leaves enough freedom to introduce sufficient orthogonality relationships (partitioning the set of variables in as many mutually orthogonal subgroups as possible), the complexity of a subsequent specification search will be substantially reduced.

When the number of complete records in a subset X_1 is considerably larger than in the full set of variables $[X_1 X_2]$, and provided that operationalization requirements do not shut the door on the introduction of orthogonality relationships, it may be convenient to keep the coefficients of that subset separate.

A "raw" dataset typically contains a very large number of variables (which will be denoted as "items", because they often are the items from one or several questionnaires). The first task at hand then usually consists of a linear transformation of the data, aimed at replacing the items with a set of variables which operationalize the concepts of interest. The transformation is often chosen by means of factor analyses or similar data reduction techniques. As soon as the coefficients of the transformation have been determined, scores for the (new) variables can be calculated. When some elements in the raw data matrix are missing, the analyst must decide for each score which is based on an incomplete record whether to mark it as missing or to calculate some approximation based on the available data. The decision to reject

or to repair may yield a different outcome for scores on different variables within the same record. In some cases this may result in a subset X_1 of p_1 variables with, say, n_1 of n records usable in a regression, whereas the number n_u of usable records in the complete data matrix $[\mathrm{X}_1\mathrm{X}_2]$ is considerably smaller. In such a situation a researcher may decide to return to the data reduction step and — when this is consistent with the goals of the operationalization — build in the mutual (approximate) orthogonality of the two subsets of variables with respect to the n_u usable records. If the researcher is then prepared to assume that the unknown elements of X_2 are such that they respect (or even bring to perfection) the orthogonality between the two subsets of variables with respect to the previously mentioned n_1 records of $[\mathrm{X}_1\mathrm{X}_2]$, the coefficients β_1 are separate in model (1) applied to those n_1 records. Estimates can be calculated according to (5), using the n_1 records. Thus, although the regression analysis cannot be applied directly to the dataset of n_1 records, due to the missing values in the X_2 block, estimates of β_1 can be calculated as though n_1 usable records were available for the full model.

Separability in the likelihood problem. The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of expression (2) are the solutions of a likelihood problem. This problem consists of finding in $\mathsf{R}^{p_1\times 1} \times \mathsf{R}^{p_2\times 1} \times \mathsf{R}_{>0}$ the value $(\beta_1, \beta_2, \sigma^2)$ that minimizes

$$n \ln \sigma^2 + \frac{1}{\sigma^2}(\mathrm{Y} - \mathrm{X}_1\beta_1 - \mathrm{X}_2\beta_2)^{\mathsf{T}}(\mathrm{Y} - \mathrm{X}_1\beta_1 - \mathrm{X}_2\beta_2). \tag{7}$$

It can be reformulated as a two-step problem. The first step consists of finding in $\mathsf{R}^{p_1\times 1} \times \mathsf{R}^{p_2\times 1}$ the value (β_1, β_2) that minimizes

$$(\mathrm{Y} - \mathrm{X}_1\beta_1 - \mathrm{X}_2\beta_2)^{\mathsf{T}}(\mathrm{Y} - \mathrm{X}_1\beta_1 - \mathrm{X}_2\beta_2). \tag{8}$$

Having obtained the solution $(\hat{\beta}_1, \hat{\beta}_2)$ of the first step, the second step aims at finding σ^2 in $\mathsf{R}_{>0}$, minimizing

$$n \ln \sigma^2 + \frac{1}{\sigma^2}(\mathrm{Y} - \mathrm{X}_1\hat{\beta}_1 - \mathrm{X}_2\hat{\beta}_2)^\mathsf{T}(\mathrm{Y} - \mathrm{X}_1\hat{\beta}_1 - \mathrm{X}_2\hat{\beta}_2). \tag{9}$$

This means that the likelihood problem (7) is special, in the sense that the parameters (β_1, β_2) can be estimated separately; the estimate for the parameter σ^2 is found in a second stage. In general, however, the calculation of an estimate for β_1 (and β_2) involves the complete matrix $[\mathrm{X}_1\mathrm{X}_2]$ of independent variables. This changes when the orthogonality conditions (6) are satisfied. Then, the first stage (8) of the likelihood problem is equivalent to the minimization of $(\mathrm{Y} - \mathrm{X}_1\beta_1)^\mathsf{T}(\mathrm{Y} - \mathrm{X}_1\beta_1)$ in order to find β_1 in $\mathsf{R}^{p_1 \times 1}$ and of $(\mathrm{Y} - \mathrm{X}_2\beta_2)^\mathsf{T}(\mathrm{Y} - \mathrm{X}_2\beta_2)$ in order to find β_2 in $\mathsf{R}^{p_2 \times 1}$. Thus, when $\mathrm{X}_2^\mathsf{T}\mathrm{X}_1 = 0$, in the general problem of finding an estimate for the parameter set $(\beta_1, \beta_2, \sigma^2)$, the problem of finding an estimate for β_1 can be formulated as a separate optimization problem, *involving only* X_1 *and* Y.

In this article, we will use this as the basis of a "technical" definition of separability. Whenever the parameters of a model are estimated by means of an optimization problem and a subset of the parameters can be equivalently estimated by means of an optimization problem involving only a subset of the data, we call this subset of parameters *separate* with respect to the model and the data involved. This "technical" definition implies the previously described "practical" notion of separability, namely that the estimate of the parameters in question can be calculated on the basis of a subset of the data.

Unbiasedness in cases of Underspecification in Ordinary Regression

Orthogonality relationships are also of interest in ordinary regression when it is important to obtain unbiased estimates of particular coefficients in an underspecified model.

Bias and unbiasedness in cases of underspecification. When the response data column Y is a realization of a stochastic variable which is distributed as

$$\mathrm{N}\left(\mathrm{X}_1\beta_{1t} + \mathrm{X}_2\beta_{2t} + \mathrm{X}_3\beta_{3t}, \sigma_t^2\mathrm{I}_{n\times n}\right) \tag{10}$$

(with subscript "t" meaning "true"), but which is assumed to be distributed as

$$\mathrm{N}\left(\mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2, \sigma^2\mathrm{I}_{n\times n}\right) \tag{11}$$

for some unknown $(\beta_1, \beta_2, \sigma^2)$ in $\mathsf{R}^{p_1\times 1} \times \mathsf{R}^{p_2\times 1} \times \mathsf{R}_{>0}$, the bias of the maximum likelihood estimator of (β_1, β_2) is:

$$\left([\mathrm{X}_1\mathrm{X}_2]^\mathsf{T}[\mathrm{X}_1\mathrm{X}_2]\right)^{-1}[\mathrm{X}_1\mathrm{X}_2]^\mathsf{T}\mathrm{X}_3\beta_{3t}. \tag{12}$$

The bias of the estimator of β_1 can be isolated from (12), resulting in:

$$\begin{aligned}&\Big(\left(\mathrm{X}_1^\mathsf{T}\mathrm{X}_1 - \mathrm{X}_1^\mathsf{T}\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\mathrm{X}_1\right)^{-1}\mathrm{X}_1^\mathsf{T}\\ &-\left(\mathrm{X}_1^\mathsf{T}\mathrm{X}_1 - \mathrm{X}_1^\mathsf{T}\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\mathrm{X}_1\right)^{-1}\mathrm{X}_1^\mathsf{T}\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\Big)\,\mathrm{X}_3\beta_{3t}.\end{aligned} \tag{13}$$

This expression resembles (3). Here, however, the aim is not to eliminate the influence of X_2 from (13), but rather to formulate conditions that cause (13) to yield a 0, which means that the estimator of β_1 according to model (11) is unbiased, despite the underspecification of model (11). Expression (13) can be rewritten as:

$$\begin{aligned}&\left(\mathrm{X}_1^\mathsf{T}\mathrm{X}_1 - \mathrm{X}_1^\mathsf{T}\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\mathrm{X}_1\right)^{-1}\mathrm{X}_1^\mathsf{T}\\ &\left(\mathrm{I}_{n\times n} - \mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\right)\mathrm{X}_3\beta_{3t}.\end{aligned} \tag{14}$$

With $[\mathrm{X}_1\mathrm{X}_2]$ of full rank, the matrix

$$\left(\mathrm{X}_1^\mathsf{T}\mathrm{X}_1 - \mathrm{X}_1^\mathsf{T}\mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\mathrm{X}_1\right)^{-1} \tag{15}$$

has rank p_1 (see Searle, Casella and McCulloch, 1991, result (26) on p. 451). Then, unbiasedness of the estimator of β_1 is equivalent to

$$\mathrm{X}_1^\mathsf{T}\left(\mathrm{I}_{n\times n} - \mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\right)\mathrm{X}_3\beta_{3t} = 0_{p_1\times 1} \tag{16}$$

(necessary and sufficient conditions). When the operator $\mathrm{I}_{n\times n} - \mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}$ in the center of the expression is multiplied at its right hand side with a data column, it yields the residual of the regression of this data column onto the columns of X_2. As a consequence, the conditions (16) mean that the residual of the regression of the data column $\mathrm{X}_3\beta_{3t}$ onto X_2 is orthogonal to the columns of X_1. Equivalently, it means that when any column of X_1 is regressed onto X_2, the residual is orthogonal to $\mathrm{X}_3\beta_{3t}$. Note that $\mathrm{X}_3\beta_{3t}$ is the part of the "true" system (10) involving the columns of X_3, which have been omitted in model (11) (underspecification).

Obviously, the conditions $\mathrm{X}_1^\mathsf{T}\left(\mathrm{I}_{n\times n} - \mathrm{X}_2(\mathrm{X}_2^\mathsf{T}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\right)\mathrm{X}_3 = 0_{p_1\times p_3}$ are sufficient for (16) to be true. Those sufficient conditions for unbiasedness have, perhaps, more practical value than (16), as they do not involve β_{3t}, which is unknown in practice.

Unbiasedness and separability. In order to achieve (16), it is sufficient (but not necessary) that $\mathrm{X}_1^\mathsf{T}\mathrm{X}_3 = 0$ and $\mathrm{X}_1^\mathsf{T}\mathrm{X}_2 = 0$. Note that, under those circumstances, the coefficients of X_1 are separate in the (correctly specified) regression of Y on $[\mathrm{X}_1\mathrm{X}_2\mathrm{X}_3]$. Also, the coefficients of X_1 are separate in the (underspecified) model (11). Alternatively, in order to achieve (16), it is also sufficient (but not necessary) that $\mathrm{X}_3^\mathsf{T}\mathrm{X}_1 = 0$ and $\mathrm{X}_3^\mathsf{T}\mathrm{X}_2 = 0$. When those conditions are satisfied, the coefficients of X_3 are separate in the correctly specified regression model, and also in the (underspecified) regression of Y on $[\mathrm{X}_2\mathrm{X}_3]$.

Focus on Separability in Multilevel Models

Datasets are often partitioned according to one or several known classifications, for example: when for each pupil in a sample the present and previous class, the present and previous school, ... have been recorded. Multilevel models have been developed to incorporate such structure into the statistical analyses — see, amongst others, Bryk and Raudenbush (1992), Longford (1993), Kreft and de Leeuw (1998),

Snijders and Bosker (1999) — and often replace the ordinary regression models which have been employed traditionally.

Researchers naturally prefer to retain as much as possible the familiar rules and practices of ordinary regression when performing multilevel analyses. However, some of the properties of ordinary regression that are routinely relied upon may well break down in the more complex multilevel analyses and should not be taken for granted. A case in point are the difficulties that arise when one attempts to extend the concept of explained variance (the R^2 measure in ordinary regression) to multilevel analysis, as was demonstrated by Snijders and Bosker (1994).

Bearing this in mind, we have attempted to clarify the conditions for separability in multilevel models. In ordinary regression, orthogonality turns out to be closely linked to (or, in more precise terms: a sufficient condition for) separability. Raudenbush's (1989a) example is also based on mutually orthogonal variables, but it is a special case of orthogonality, namely between a variable centered within the groups of the higher level and a variable which is constant within each group. If, however, centering, rather than orthogonality in general, is the key to separability in multilevel models, then the question of how to center in a model with more than two levels arises: is a pupil variable centered within classes or rather within schools, and so on. Also, Raudenbush's (1989a) numerical example does not represent a case with random slopes.

Raudenbush (1989a) focussed on the role of centering in multilevel analysis rather than on separability. Subsequently, some further attention has been devoted to the issue of centering by Plewis (1989), Longford (1989), Raudenbush (1989b), Kreft, de Leeuw and Aiken (1995), Schumacker and Bembry (1996), Opdenakker and Van Damme (1997), and Hofmann and Gavin (1998). Neither of those, however, except Hofmann and Gavin (1998), discuss separability as a potential benefit of

centering. Hofmann and Gavin (1998) provide a simulated example of separability in a two-level model, which will be discussed further on.

The issue of centering is referred to by Goldstein (1995) and by Snijders and Bosker (1999), but they do not discuss centering in the light of separability or unbiasedness. Centering is discussed in some detail by Kreft and de Leeuw (1998, section 5.2). They focus on the (non-)equivalence of multilevel models arising from different centering transformations applied to a given dataset, rather than on the other purposes (called "technical purposes" by Kreft and de Leeuw) of centering. Bryk and Raudenbush (1992, pp. 25-28) consider centering and (pp. 117–123 and pp. 204–207) the issue of unbiasedness in cases of underspecification in hierarchical two-level models.

In the previous subsection about unbiasedness in the case of underspecification in ordinary regression it was demonstrated that the issues of separability and of unbiasedness are connected, in that both the conditions for separability and those for unbiasedness can be interpreted in terms of orthogonality between vectors, and because some (orthogonality) conditions yield instances of separability as well as cases of unbiasedness. Nevertheless, the concepts of separability and of unbiasedness in the case of underspecification are not identical. When ordinary regression analysis is extended to multilevel analysis, the issue of unbiasedness in a situation of underspecification (in the fixed part of the model) becomes more complicated, as several possibilities can be considered for the covariance part of the model: the covariance part can be specified correctly or can be misspecified in some way related to the misspecification in the fixed part. Although the problem of separability equally will prove more subtle in the multilevel case, it is easier to delineate. After all, it involves only a particular subset of variables in a particular model, whereas the formulation of a problem of unbiasedness requires indicating a particular "true"

underlying system and a particular postulated model.

In order to keep this contribution within the scope of one article and to give it a clear focus, we will concentrate on the issue of separability (as it was formulated in terms of the likelihood problem) in multilevel models. A rigorous definition of what is denoted as a "multilevel model" in this text, can be found in the appendix.

Generalizing Separability to Multilevel Models

Generalizing the discussion of separability in the ordinary regression model (1) to a model based on

$$\mathrm{N}\left(\mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2, \sigma^2\mathrm{V}\right) \tag{17}$$

where V is a known $n \times n$ positive definite matrix, is straightforward. The least squares problem of minimizing (8) is replaced by the generalized least squares (GLS) problem which aims at minimimizing

$$(\mathrm{Y} - \mathrm{X}_1\beta_1 - \mathrm{X}_2\beta_2)^\mathsf{T}\mathrm{V}^{-1}(\mathrm{Y} - \mathrm{X}_1\beta_1 - \mathrm{X}_2\beta_2). \tag{18}$$

The orthogonality conditions $\mathrm{X}_2^\mathsf{T}\mathrm{X}_1 = 0$ are replaced by

$$\mathrm{X}_2^\mathsf{T}\mathrm{V}^{-1}\mathrm{X}_1 = 0_{p2\times p1}, \tag{19}$$

which is sufficient to transform the optimization of (18) into two separate optimization problems, with solutions

$$\hat{\beta}_1 = (\mathrm{X}_1^\mathsf{T}\mathrm{V}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1^\mathsf{T}\mathrm{V}^{-1}\mathrm{Y} \tag{20}$$

$$\hat{\beta}_2 = (\mathrm{X}_2^\mathsf{T}\mathrm{V}^{-1}\mathrm{X}_2)^{-1}\mathrm{X}_2^\mathsf{T}\mathrm{V}^{-1}\mathrm{Y} \tag{21}$$

meaning that both β_1 and β_2 are separate in the model based on (17).

A further generalization along the same line to a model based on

$$\mathrm{N}\left(\mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2, \Sigma(\theta)\right), \tag{22}$$

where Σ is a mapping of a set T of parameter values into the set of the real $n \times n$ positive definite matrices, runs into difficulties. Although for the full information maximum likelihood (FIML) estimate $(\hat{\beta}_1, \hat{\beta}_2, \hat{\theta})$ it remains true that

$$\mathrm{X}_2^{\mathsf{T}} \Sigma(\hat{\theta})^{-1} \mathrm{X}_1 = 0_{p_2 \times p_1}, \tag{23}$$

is sufficient for

$$\hat{\beta}_1 = (\mathrm{X}_1^{\mathsf{T}} \Sigma(\hat{\theta})^{-1} \mathrm{X}_1)^{-1} \mathrm{X}_1^{\mathsf{T}} \Sigma(\hat{\theta})^{-1} \mathrm{Y} \tag{24}$$

$$\hat{\beta}_2 = (\mathrm{X}_2^{\mathsf{T}} \Sigma(\hat{\theta})^{-1} \mathrm{X}_2)^{-1} \mathrm{X}_2^{\mathsf{T}} \Sigma(\hat{\theta})^{-1} \mathrm{Y}, \tag{25}$$

the conditions (23) lose most of their practical value because they contain $\hat{\theta}$, which is not known a priori (and which implicitly depends on X_1, X_2). The worst of it, however, is that (24) and (25) do not mean anymore that β_1 and β_2 are separate, because in general $\hat{\beta}_1$ and $\hat{\beta}_2$ depend on both X_1 and X_2 via $\hat{\theta}$.

On the other hand, Raudenbush's (1989a) example (see introduction) seems to indicate that it is not altogether impossible to find non-trivial instances of separability in models more general than (17). Indeed, the following property of models based on (22) will prove to be a key to finding more such cases:

Lemma L. If W is a positive definite $n \times n$ matrix and

(a) $\mathrm{X}_2^{\mathsf{T}} \mathrm{W}^{-1} \mathrm{X}_1 = 0_{p_2 \times p_1}$, and

(b) for all $\theta \in \mathsf{T}$:

each column of X_1 is an eigenvector of $\Sigma(\theta)\mathrm{W}^{-1}$

and all those columns are associated with the same eigenvalue,

then

β_1 is separate in the model based on (22)

and $\hat{\beta}_1 = (\mathrm{X}_1^{\mathsf{T}} \mathrm{W}^{-1} \mathrm{X}_1)^{-1} \mathrm{X}_1^{\mathsf{T}} \mathrm{W}^{-1} \mathrm{Y}$.

This Lemma can be verified by transforming the FIML problem for (22) using the

conditions (a) and (b). The likelihood problem then amounts to finding $(\beta_1, \beta_2, \theta)$ in $\mathsf{R}^{p_1 \times 1} \times \mathsf{R}^{p_2 \times 1} \times \mathsf{T}$ minimizing

$$\begin{aligned} &\ln\det\Sigma(\theta) - \lambda(\theta)^{-1}\mathrm{Y}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{Y} + (\mathrm{Y}-\mathrm{X}_2\beta_2)^{\mathsf{T}}\Sigma(\theta)^{-1}(\mathrm{Y}-\mathrm{X}_2\beta_2) \\ &+\lambda(\theta)^{-1}(\mathrm{Y}-\mathrm{X}_1\beta_1)^{\mathsf{T}}\mathrm{W}^{-1}(\mathrm{Y}-\mathrm{X}_1\beta_1), \end{aligned} \tag{26}$$

where $\lambda(\theta)$ is the (positive) eigenvalue of $\Sigma(\theta)\mathrm{W}^{-1}$ associated with X_1. Clearly, for β_1 this comes down to the separate minimization of $(\mathrm{Y}-\mathrm{X}_1\beta_1)^{\mathsf{T}}\mathrm{W}^{-1}(\mathrm{Y}-\mathrm{X}_1\beta_1)$.

When applied to ordinary regression (model (1), where $\Sigma(\theta) = \sigma^2\mathrm{I}$ and $\mathrm{W} = \mathrm{I}$) and GLS regression (model (17), with $\Sigma(\theta) = \sigma^2\mathrm{V}$ and $\mathrm{W} = \mathrm{V}$), condition (a) of Lemma L yields the orthogonality conditions (6) and (19) respectively; condition (b) is obsolete in those cases, as all $n \times 1$ vectors are eigenvectors of $\sigma^2\mathrm{I}$, with eigenvalue σ^2. Note that, while condition (a) is symmetric, condition (b) is not: separability of β_1 is, in contrast with the previous examples, generally not mirrored by separability of β_2.

The next section presents a number of non-trivial instances of separability in multilevel models which originate from Lemma L. They are constructed by searching for eigenspaces of the covariance matrix — see (52) in the appendix —, or subsets thereof, which do not depend on its parameters. Although multilevel models are special cases of model (22), this restriction is not sufficient to make the eigenspaces of the covariance matrix invariant as a matter of course. The parameterized matrix

$$\bigoplus_{i=1}^{n_3} \begin{bmatrix} \sigma_1^2+\sigma_2^2+\sigma_3^2 & \sigma_3^2 & \sigma_3^2 \\ \sigma_3^2 & \sigma_1^2+\sigma_2^2+\sigma_3^2 & \sigma_2^2+\sigma_3^2 \\ \sigma_3^2 & \sigma_2^2+\sigma_3^2 & \sigma_1^2+\sigma_2^2+\sigma_3^2 \end{bmatrix}, \tag{27}$$

for example, is the covariance matrix of a three-level random intercept model (see (53) in the appendix) with in each level-3 group two level-2 groups, one with a single and one with two basic units. It has an eigenspace consisting of the vectors

$\mathrm{K} \otimes \left[\mathrm{f}\left(\frac{\sigma_2^2}{\sigma_3^2}\right) 11\right]^{\mathsf{T}}$, for any $\mathrm{K} \in \mathsf{R}^{n_3 \times 1}$, and with $\mathrm{f}(x) = -\frac{x+1}{2} + \frac{1}{2}\left(x^2 + 2x + 9\right)^{\frac{1}{2}}$. Thus, it makes sense to start a search for eigenspaces that *are* invariant.

Examples of Separability in Multilevel Models

Completely Balanced Completely Hierarchical Random Intercept Model

As a first example, consider a special case of the random intercept model (see (53) in the appendix). The case is defined by two requirements. Firstly, the model is assumed to be completely hierarchical. This means that the second classification is nested within the third, the third is nested within the fourth, and so on — the basic level is always nested within the second classification. (It is in such a completely hierarchical model that all the classifications can be appropriately designated as "levels". Nevertheless, we apply the customary but somewhat inappropriate term "multilevel" model to the most general model defined in the appendix.) Secondly, the model is required to be completely balanced, meaning that each classification is balanced, that is: all groups within the classification have equal size.

The covariance matrix of this rigidly structured instance of model (53) can (after a permutation of the basic units) be expressed as:

$$\bigoplus_{i_{\mathrm{g}}=1}^{n_{\mathrm{g}}} \left(\sigma_g^2 \mathrm{J}_{q_{\mathrm{g}} \times q_{\mathrm{g}}} + \bigoplus_{i_{\mathrm{g}-1}=1}^{s_{\mathrm{g}-1}} \left(\sigma_{g-1}^2 \mathrm{J}_{q_{\mathrm{g}-1} \times q_{\mathrm{g}-1}} + \ldots + \bigoplus_{i_2=1}^{s_2} \left(\sigma_2^2 \mathrm{J}_{q_2 \times q_2} + \sigma_1^2 \mathrm{I}_{q_2 \times q_2} \right) \right) \right) \quad (28)$$

with J indicating a matrix of 1's, s_j denoting the number of level-j groups within a level-(j+1) group and q_j indicating the size of a level-j group. Its eigenvalues and eigenvectors can be derived analytically. They are listed in Table 1. There are g different eigenvalues. The eigenspace of the l-th eigenvalue consists of all level-l vectors (that is: vectors that are constant within each of the n_l level-l groups, but which are not constant within all level-$l+1$ groups) that are centered at level $l+1$. The latter means that the sum of their elements within each level-$l+1$ group

Table 1.

Completely Balanced Completely Hierarchical Random Intercept Model

Eigenvalue	Multiplicity	Eigenspace: the $n \times 1$ vectors ...
σ_1^2	$n - n_2$	... centered at level 2
$\sigma_1^2 + q_2\sigma_2^2$	$n_2 - n_3$	... of level 2 centered at level 3
$\vdots$	$\vdots$	$\vdots$
$\sum_{h=1}^{l} q_h\sigma_h^2$	$n_l - n_{l+1}$	... of level l centered at level $l+1$
$\vdots$	$\vdots$	$\vdots$
$\sum_{h=1}^{g-1} q_h\sigma_h^2$	$n_{g-1} - n_g$	... of level $g-1$ centered at level g
$\sum_{h=1}^{g} q_h\sigma_h^2$	n_g	... of level g

equals zero. The multiplicity number in the second column of Table 1 can be easily derived: the set of all vectors constant within level-l groups has n_l degrees of freedom; the centering requirement imposes n_{l+1} restrictions, excluding, amongst others, the non-zero vectors which are constant within all level-$l+1$ groups. Note that the multiplicity numbers add up to n, which indicates that Table 1 lists the complete eigenvalue/eigenvector decomposition of the $n \times n$ covariance matrix.

The most interesting feature of Table 1 is that, although the eigenvalues are dependent on the parameters, the eigenspaces are not. This provides opportunities to apply Lemma L and to derive a set of rules concerning separability in a completely balanced completely hierarchical random intercept model.

First, consider the case of a subset of the explanatory variables of the model, all centered at level 2. Then every variable (data column) in the subset is an eigenvector of the covariance matrix (28), associated with the eigenvalue σ_1^2 (Table 1). Also, the subset is orthogonal — in the sense of: inner product zero — to every explanatory variable of level l, $l \geq 2$, due to the centering at level 2. Thus, if the subset is

orthogonal to the remaining level-1 explanatory variables in the model (if present), the conditions of Lemma L are satisfied. Consequently, the following rule for basic level variables applies (and is denoted with the "B" of "basic level"):

> *Rule B.* In a completely balanced completely hierarchical random intercept model, the coefficients of a subset of (basic level) explanatory variables centered at level 2 and orthogonal to the remaining level-1 explanatory variables, are separate.

Secondly, Table 1 points at a similar application of Lemma L to variables of intermediate levels:

> *Rule I.* For an intermediate level l in a completely balanced completely hierarchical random intercept model, the coefficients of a subset of explanatory level-l variables centered at level $l+1$ and orthogonal to the remaining variables of level l and of lower levels, are separate.

Note that, for a lower level variable to be orthogonal to a set of level-l variables, it is sufficient (but not necessary) that the lower level variable is centered at level l or at a lower level. A basic level variable, for example, is orthogonal to all variables of level 3 (and of higher levels) when it is centered at level 2 or at level 3. Finally, a similar rule can be derived for top level variables:

> *Rule T.* In a completely balanced completely hierarchical random intercept model, the coefficients of a subset of top level explanatory variables orthogonal to the remaining explanatory variables, are separate.

The intercept term $\mathrm{J}_{n\times 1}$ can be regarded as a top level variable. A data column is orthogonal to $\mathrm{J}_{n\times 1}$ when the sum of its elements equals zero, that is: when it is "grand mean centered". Thus, when all the (measured) explanatory variables in

a completely balanced completely hierarchical random intercept model are grand mean centered, the intercept is separate.

It is instructive to consider the special case of a completely balanced completely hierarchical random intercept model where all the explanatory variables of the basic and intermediate levels are centered at the next (higher) level. In such a case, the FIML estimates of the fixed effects at any given level can be obtained simply as the least squares approximation of the response by means of the variables at that level. Within the set of explanatory variables of each level, further separation can be achieved by introducing orthogonality, as in ordinary regression.

In order to illustrate a part of their discussion of centering decisions in organizational research, Hofmann and Gavin (1998) created a balanced two-level dataset of 150 basic units classified in 15 groups, containing a response Y, four level-1 independent variables $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$ centered at level 2, and one level-2 variable G. The dataset was analysed by means of models based on

$$\mathrm{N}\left(\mathrm{J}_{150\times 1}\alpha + \mathrm{X}\beta + \mathrm{G}\gamma, \sigma_1^2 \mathrm{I}_{150\times 150} + \sigma_2^2 \bigoplus_{i=1}^{15} \mathrm{J}_{10\times 10}\right) \tag{29}$$

with, successively, X = A, X = [AB], X = [ABC], X = [ABCD]. Hofmann and Gavin (1998) found that the estimate for γ remained constant within this sequence of analyses. From the point of view developed in the present text, this is explained by Rule T, which implies that the coefficients of the subset $\{\mathrm{J}, \mathrm{G}\}$ of independent variables are separate in the models based on (29). The FIML estimates for those coefficients are, in all those models:

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\gamma} \end{bmatrix} = \left([\mathrm{JG}]^{\mathsf{T}}[\mathrm{JG}]\right)^{-1}[\mathrm{JG}]^{\mathsf{T}}\mathrm{Y}. \tag{30}$$

When (29) with X = [ABCD] is defined as the correctly specified model, the other three models can be regarded as underspecified in the fixed part (while being cor-

rectly specified in the random part). Thus, the illustration of Hofmann and Gavin (1998) also represents an example of unbiasedness in a case of underspecification: the estimator (see (30)) for α and γ in those underspecified models is not only unbiased, it is even identical to the estimator based on the correctly specified model, due to the separability of the coefficients α and γ.

Partially Balanced Completely Hierarchical Random Intercept Model

The prerequisite of balancedness which was introduced in the previous subsection may well be at odds with the requirement of optimal utilization of a given dataset. When this prerequisite is dropped in the case of (29), that is: when the analysis is based on:

$$\mathrm{N}\left(\mathrm{J}\alpha + \mathrm{X}\beta + \mathrm{G}\gamma, \sigma_1^2\mathrm{I}_{n\times n} + \sigma_2^2 \bigoplus_{i=1}^{n_2} \mathrm{J}_{q_{2\mathrm{i}}\times q_{2\mathrm{i}}}\right), \tag{31}$$

with q_{2i} denoting the size of the i-th level-2 group (and with X and G as before), the FIML estimates of the intercept, γ, σ_1^2 and σ_2^2 are connected by

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\gamma} \end{bmatrix} = \left([\mathrm{JG}]^{\mathsf{T}}\mathrm{D}[\mathrm{JG}]\right)^{-1}[\mathrm{JG}]^{\mathsf{T}}\mathrm{DY} \tag{32}$$

with $\mathrm{D} = \bigoplus_{i=1}^{n_2}\left(\hat{\sigma_1^2} + q_{2i}\hat{\sigma_2^2}\right)^{-1}\mathrm{I}_{q_{2\mathrm{i}}\times q_{2\mathrm{i}}}$. When q_{2i} is not dependent on i (balancedness), (32) is reduced to (30), implying separability for $\{\alpha, \gamma\}$. In general, however, $\hat{\alpha}$ and $\hat{\gamma}$ are dependent on the columns of X via $\hat{\sigma_1^2}$ and $\hat{\sigma_2^2}$. Thus, removing the requirement of balancedness causes at least a partial breakdown of the set of rules established in the previous section.

In the (general) random intercept model, based on the covariance matrix $\sum_{l=1}^{g}\sigma_l^2\mathrm{W}^{(l)}(\mathrm{W}^{(l)})^{\mathsf{T}}$ (see appendix), the matrix $\mathrm{W}^{(l)}(\mathrm{W}^{(l)})^{\mathsf{T}}$, which defines the l-th classification, can be expressed as $\mathrm{P}^{(l)}\left(\bigoplus_{i=1}^{n_{\mathrm{l}}}\mathrm{J}_{q_{\mathrm{li}}\times q_{\mathrm{li}}}\right)\mathrm{P}^{(l)}$ by means of a suitable $n\times n$ permutation matrix $\mathrm{P}^{(l)}$. It is sufficiently tractable for an analytical derivation

of its eigenvalue/eigenvector decomposition, which is presented in Table 2, assuming that the l-th classification contains $k^{(l)}$ different group sizes $r_1, r_2, \ldots, r_{k^{(l)}}$, with multiplicity $m_1, m_2, \ldots, m_{k^{(l)}}$, respectively. Whereas one can conceive that in a par-

Table 2.

Eigenvalue/Eigenvector Decomposition of $W^{(l)}(W^{(l)})^T$

Eigenvalue	Multiplicity	Eigenspace: the $n \times 1$ vectors ...	
0	$n - n_l$	... centered within the groups	E_0^l
r_1	m_1	... constant within each group of size r_1	$E_{r_1}^l$
		and zero in the other $n_l - m_1$ groups	
$\vdots$	$\vdots$	$\vdots$	
r_i	m_i	... constant within each group of size r_i	$E_{r_i}^l$
		and zero in the other $n_l - m_i$ groups	
$\vdots$	$\vdots$	$\vdots$	
$r_{k^{(l)}}$	$m_{k^{(l)}}$	... constant within each group of size $r_{k^{(l)}}$	$E_{r_{k^{(l)}}}^l$
		and zero in the other $n_l - m_{k^{(l)}}$ groups	

The groups referred to are the groups of the l-th classification, which contains $k^{(l)}$ different group sizes.

ticular research context the within-group centering of a data column according to the l-th classification (making the transformed column a member of E_0^l in Table 2) may be compatible with the substantive requirements of an operationalization, the usefulness of an eigenspace $E_{r_i}^l$, with possibly only a few groups of size r_i, seems doubtful. When the l-th classification is balanced, however, the eigenspaces $E_{r_1}^l, \ldots, E_{r_{k^{(l)}}}^l$ merge into a single eigenspace $E_{q_l}^l$ of all $n \times 1$ vectors which are constant within each group. Any measurement that consists of obtaining one value for each of the n_l groups of the l-th classification yields a $n \times 1$ data column belonging

to $\mathsf{E}^{l}_{q_l}$.

Table 3 represents the interesting eigenspaces in the case of a completely hierarchical g-level random intercept model with each of the f lower levels ($f < g$) balanced (hence the term "partially balanced"). The nestedness of the classifica-

Table 3.

Partially Balanced Completely Hierarchical Random Intercept Model: Eigenspaces

1	2	3	...	$f-1$	f	$f+1$	...	g
	$\mathsf{E}^2_0 \subset$	$\mathsf{E}^3_0 \subset$	$\ldots \subset$	$\mathsf{E}^{f-1}_0 \subset$	$\mathsf{E}^f_0 \subset$	$\mathsf{E}^{f+1}_0 \subset$	$\ldots \subset$	E^g_0
$\mathsf{E}^1_1 \supset$	$\mathsf{E}^2_{q_2} \supset$	$\mathsf{E}^3_{q_3} \supset$	$\ldots \supset$	$\mathsf{E}^{f-1}_{q_{f-1}} \supset$	$\mathsf{E}^f_{q_f}$			

The levels 1,2,...,f are each balanced. Only the "interesting" eigenspaces have been indicated. Note that $\mathsf{E}^1_1 = \mathsf{R}^{n\times 1}$.

tions is a prerequisite for the inclusion relationships indicated in Table 3. These relationships can be exploited by observing that the covariance matrix of the random intercept model consists of a linear combination of the matrices $\mathrm{W}^{(1)}(\mathrm{W}^{(1)})^{\mathsf{T}}$, $\mathrm{W}^{(2)}(\mathrm{W}^{(2)})^{\mathsf{T}}$, ..., $\mathrm{W}^{(g)}(\mathrm{W}^{(g)})^{\mathsf{T}}$. When g eigenspaces $\mathsf{E}_1, \mathsf{E}_2, \ldots, \mathsf{E}_g$ associated with matrices $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_g$ and eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_g$, respectively, have a cross-section $\mathsf{E}_1 \cap \mathsf{E}_2 \cap \ldots \cap \mathsf{E}_g$ containing a non-zero vector, then any linear combination $\sum_{i=1}^{g} a_i \mathrm{A}_i$ has $\sum_{i=1}^{g} a_i \lambda_i$ as an eigenvalue, with its eigenspace encompassing this cross-section. Consider then, for any $j \in \{1, 2, \ldots, f\}$, the set $\mathsf{E}^j_{q_j} \cap \mathsf{E}^{j+1}_0$, which consists of all $n \times 1$ columns representing level-j data centered at level $j+1$. According to Table 3, this set is included in an eigenspace of $\mathrm{W}^{(l)}(\mathrm{W}^{(l)})^{\mathsf{T}}$, for every $l \in \{1, 2, \ldots, g\}$. Consequently, $\mathsf{E}^j_{q_j} \cap \mathsf{E}^{j+1}_0$ is included in an eigenspace of the covariance matrix of the partially balanced completely hierarchical random intercept model. The set $\mathsf{E}^j_{q_j} \cap \mathsf{E}^{j+1}_0$ is not dependent on the parameters of the covariance matrix. Thus, Lemma L can be applied, which implies that

Rule B is extended to the partially balanced completely hierarchical random intercept model. Rule I holds for the balanced intermediate levels ($l \leq f$).

Referring again to the model based on (31) as an example, it is clear that Rule B guarantees that β is separate. (Remember that the columns of X are centered at level 2.) It is also instructive to consider the problem of underspecification when a response generated by the "true" system

$$\mathrm{N}\left(\mathrm{J}\alpha_t + [\mathrm{ABCD}]\beta_t + \mathrm{G}\gamma_t, \sigma^2_{1t}\mathrm{I}_{n\times n} + \sigma^2_{2t}\bigoplus_{i=1}^{n_2}\mathrm{J}_{q_{2\mathrm{i}}\times q_{2\mathrm{i}}}\right) \tag{33}$$

is modeled as (31) with X consisting of one, two or three columns of [ABCD]. The bias of the estimator of α and γ is

$$\left([\mathrm{JG}]^{\mathsf{T}}\hat{\Sigma}^{-1}[\mathrm{JG}]\right)^{-1}[\mathrm{JG}]^{\mathsf{T}}\hat{\Sigma}^{-1}[\mathrm{ABCD}]\beta_t \tag{34}$$

with $\hat{\Sigma} = \hat{\sigma^2_1}\mathrm{I}_{n\times n} + \hat{\sigma^2_2}\bigoplus_{i=1}^{n_2}\mathrm{J}_{q_{2\mathrm{i}}\times q_{2\mathrm{i}}}$, where $\hat{\sigma^2_1}$ and $\hat{\sigma^2_2}$ are estimates according to the underspecified model. As $[\mathrm{ABCD}]\beta_t$ is an eigenvector of $\hat{\Sigma}$ (whatever the values of $\hat{\sigma^2_1}$ and $\hat{\sigma^2_2}$ are!) and is orthogonal to [JG], this bias is equal to zero. This example seems to indicate that the line of reasoning of the present text, which is founded on the eigenvalue/eigenvector decomposition of the covariance matrix, can be fruitfully applied to a discussion of unbiasedness in cases of underspecification. Note also that although α and γ are separate in neither the correctly specified nor the underspecified model, their estimator derived from the underspecified model is unbiased.

Basic Level Variables in a Random Intercept Model

As the random intercept model is, by definition (see appendix), balanced at the basic level, Rule B extends to any completely hierarchical random intercept model.

Rule B can be extended further. Consider the case of a random intercept model, with the single restriction that, apart from the basic level, an additional classification (call it the second) is nested within all the remaining classifications (3rd, 4th, ...,

g-th). A dataset of pupils nested within classes, each of which is, per subject, taught by one subject teacher, serves as an example. The class level is nested within each subject teacher classification, but the teacher classifications are not nested amongst themselves. In such a case E_0^2 (see Table 2) is included in $\mathsf{E}_0^3, \mathsf{E}_0^4, \ldots, \mathsf{E}_0^g$. Using the properties of eigenvalues and eigenvectors and Lemma L as in the previous subsection, it follows that Rule B applies to such a model: basic level variables centered at level 2 and orthogonal to the remaining basic level variables, are separate.

Similarly, Rule B can be extended to any random intercept model, provided that it makes sense to center variables within the groups of all the classifications (excepting the basic level). In a dataset of pupils nested within schools and neighbourhoods, separability of pupil level variables may be achieved by centering them within each school/neighbourhood combination.

Multilevel Model with Random Intercept Structure at the Basic Level

Variables involved in the random part. According to the definition of a multilevel model which is adopted in this text (see appendix), any of the available independent variables may be involved in the random part of the model. Restrictions on the set of potential parameter values constitute an essential part of the definition of subtypes of the multilevel model. Such restrictions determine which of the independent variables are effectively involved in the random part. In the extreme case of a random intercept model, for each classification only a single parameter of the random part is allowed to assume non-zero values, causing only the constant vector J, which implements the intercept, to be involved in the random part and excluding the other (proper) independent variables.

Orthogonalization within groups. Random intercept models have been discussed from the point of view of separability in the previous subsections. It was demon-

strated that, under certain conditions, the coefficients of centered variables are separate. A $n \times 1$ data column B is *centered within the groups of the l-th classification* when the sum of its components within each group equals zero. This also means that $(\mathrm{W}^{(l)})^{\mathsf{T}}\mathrm{B} = 0_{n_l \times 1}$, that is: the vector B is *orthogonal* to the vector $\mathrm{J}_{n\times 1}$ *within each group* of the l-th classification. In more general multilevel models the $n \times n_l$ matrix $\mathrm{W}^{(l)}$ of indicator variables is replaced by (or rather: incorporated in) the matrices ${}^{(l)}\mathrm{Z}^{(h_i)}, i = 1, 2, \ldots, r_l$ (assuming that the columns of the matrix X of independent variables indicated by $h_1, h_2, \ldots, h_{r_l}$ are effectively involved in the l-th term of (52)). The matrix ${}^{(l)}\mathrm{Z}^{(h)}$ incorporates both the group membership data for the l-th classification and the data of the h-th independent variable, via ${}^{(l)}\mathrm{Z}^{(h)}_{ij} = \mathrm{W}^{(l)}_{ij}\mathrm{X}_{ih}$. When

$$\left[{}^{(l)}\mathrm{Z}^{(h_1)}{}^{(l)}\mathrm{Z}^{(h_2)} \ldots {}^{(l)}\mathrm{Z}^{(h_{r_l})}\right]^{\mathsf{T}} \mathrm{B} = 0_{n_l r_l \times 1}, \tag{35}$$

the vector B is *orthogonal* to each of the independent variables involved in the l-th term of (52), *within each group* of the l-th classification. This is described by stating that B is *orthogonalized within the groups of the l-th classification with regard to the matrix* $[\mathrm{X}_{\cdot h_1}\mathrm{X}_{\cdot h_2} \ldots \mathrm{X}_{\cdot h_{r_l}}]$. The orthogonalization of the data column B amounts to replacing it with

$$\left(\mathrm{I}_{n\times n} - \left[{}^{(l)}\mathrm{Z}^{(h_1)}{}^{(l)}\mathrm{Z}^{(h_2)} \ldots {}^{(l)}\mathrm{Z}^{(h_{r_l})}\right]\left[{}^{(l)}\mathrm{Z}^{(h_1)}{}^{(l)}\mathrm{Z}^{(h_2)} \ldots {}^{(l)}\mathrm{Z}^{(h_{r_l})}\right]^{+}\right)\mathrm{B}, \tag{36}$$

where "+" indicates the Moore-Penrose inverse (Searle, Casella and McCulloch, 1991, pp. 447–448), just like centering B means: replacing it with

$$\left(\mathrm{I}_{n\times n} - \mathrm{W}^{(l)}(\mathrm{W}^{(l)})^{+}\right)\mathrm{B}. \tag{37}$$

The i-th component of the $n \times 1$ vector

$$\mathrm{W}^{(l)}(\mathrm{W}^{(l)})^{+}\mathrm{B} = \mathrm{W}^{(l)}\left((\mathrm{W}^{(l)})^{\mathsf{T}}\mathrm{W}^{(l)}\right)^{-1}(\mathrm{W}^{(l)})^{\mathsf{T}}\mathrm{B} \tag{38}$$

is the mean of B in the group of the l-th classification which contains the i-th basic unit.

Note that when a variable is within-group centered or, a fortiori, when a variable is within-group orthogonalized, a part of its information content is removed. This may cause a shift in its interpretation (in the sense that after the transformation another concept is associated with it). Whether this shift is admissible or desirable depends on the research context. The transformation may also yield a "variable" (in the extreme case: a column of zeros) which is uninterpretable.

Sufficient conditions for separability. This concept of within-group orthogonalization enables the formulation of sufficient conditions for separability in multilevel models with non-trivial independent variables (that is: other than the constant vector J) involved in the random part ("O" refers to "orthogonalization"):

> *Rule O.* Consider a subset of the explanatory variables in a multilevel model with a random intercept structure at the basic level. If the variables of the subset are orthogonal to the remaining explanatory variables in the fixed part and, for each classification in the model (except the basic level), the variables of the subset are orthogonalized within the groups of the classification at hand, with regard to the variables effectively involved in the corresponding term of the covariance matrix, then the coefficients of this subset are separate.

In the covariance matrix of a multilevel model with a random intercept structure at the basic level:

$$\sigma_1^2 \mathrm{I}_{n\times n} + \sum_{l=2}^{g} \left(\sum_{h=1}^{p} \sum_{h'=1}^{p} \Omega_{hh'}^{(l)} {}^{(l)}\mathrm{Z}^{(h)} \left({}^{(l)}\mathrm{Z}^{(h')} \right)^{\mathsf{T}} \right) \tag{39}$$

each matrix ${}^{(l)}\mathrm{Z}^{(h')}$ which corresponds to a potentially non-zero $\Omega^{(l)}_{hh'}$ is orthogonal to the subset of variables mentioned in the Rule (those are the within-group orthogonalization conditions). Thus, the variables of the subset are eigenvectors of (39) with eigenvalue σ_1^2. Lemma L then proves the rule.

An example. Consider an example based on the two-level model discussed by Bryk and Raudenbush (1992) (see also the appendix). The n basic units are classified in n_2 groups, as described by the $n \times n_2$ matrix W of indicator variables. The $n \times p$ matrix of independent variables X is structured: it is constructed on the basis of measurements for p_1 basic level variables — stored in a $n \times (p_1 + 1)$ data matrix $\mathrm{P}^{(d)}$, including $\mathrm{J}_{n\times 1}$ — and for p_2 group level variables — stored in a $n_2 \times (p_2 + 1)$ data matrix $\mathrm{G}^{(d)}$, including $\mathrm{J}_{n_2\times 1}$. For the purposes of this illustration, it is assumed that the matrix $\mathrm{P}^{(d)}$ is partitioned into two blocks (two subsets of variables), $\mathrm{P}^{(d)} = [\mathrm{P}_a^{(d)}\mathrm{P}_b^{(d)}]$, containing p_a and p_b variables, respectively ($p_a + p_b = p_1 + 1$). In the notation used by Bryk and Raudenbush (1992, pp. 225-226), X can be expressed as: $\mathrm{X} = [\mathrm{P}_a\mathrm{G}_a\ \mathrm{P}_b\mathrm{G}_b]$ where $\mathrm{P}_a = [\mathrm{Z}^{(1)}\mathrm{Z}^{(2)}\ldots\mathrm{Z}^{(p_\mathrm{a})}]$, $\mathrm{P}_b = [\mathrm{Z}^{(p_\mathrm{a}+1)}\mathrm{Z}^{(p_\mathrm{a}+2)}\ldots\mathrm{Z}^{(p_\mathrm{a}+p_\mathrm{b})}]$ with $\mathrm{Z}^{(h)}_{ij} = \mathrm{W}_{ij}\mathrm{P}^{(d)}_{ih}$, and $\mathrm{G}_a = [\bigoplus_{i=1}^{p_\mathrm{a}}\mathrm{G}^{(d)}_{.1}\ldots\bigoplus_{i=1}^{p_\mathrm{a}}\mathrm{G}^{(d)}_{.p_2+1}]$, $\mathrm{G}_b = [\bigoplus_{i=1}^{p_\mathrm{b}}\mathrm{G}^{(d)}_{.1}\ldots\bigoplus_{i=1}^{p_\mathrm{b}}\mathrm{G}^{(d)}_{.p_2+1}]$. The $n\times(p_1+1)(p_2+2)$ matrix X contains the constant column $\mathrm{J}_{n\times 1}$, the $p_1 + p_2$ "main" variables and the p_1p_2 cross-level interaction variables. The "full" covariance matrix adopted by Bryk and Raudenbush (1992) can be expressed as:

$$\sigma^2\mathrm{I}_{n\times n} + \sum_{h=1}^{p_\mathrm{a}+p_\mathrm{b}}\sum_{h'=1}^{p_\mathrm{a}+p_\mathrm{b}}\Omega_{hh'}\mathrm{Z}^{(h)}(\mathrm{Z}^{(h')})^{\mathsf{T}} \tag{40}$$

with $\sigma^2 \in \mathsf{R}_{>0}$ and $\Omega \in \mathsf{R}^{(p_\mathrm{a}+p_\mathrm{b})\times(p_\mathrm{a}+p_\mathrm{b})}_{\text{pos. semidef.}}$. It involves (only) the intercept vector and the main basic level variables in the second term. Consider now the restricted model

based on

$$\mathrm{N}\left(\mathrm{P}_a\mathrm{G}_a\beta_a + \mathrm{P}_b\mathrm{G}_b\beta_b, \sigma^2\mathrm{I}_{n\times n} + \sum_{h=1}^{p_\mathrm{a}}\sum_{h'=1}^{p_\mathrm{a}}\Omega_{hh'}\mathrm{Z}^{(h)}(\mathrm{Z}^{(h')})^\mathsf{T}\right) \tag{41}$$

where only the columns of $\mathrm{P}_a^{(d)}$ are involved in the group level term of the covariance matrix. According to Rule O, the conditions

$$[\mathrm{Z}^{(1)}\mathrm{Z}^{(2)}\ldots\mathrm{Z}^{(p_\mathrm{a})}]^\mathsf{T}\mathrm{P}_b\mathrm{G}_b = 0_{n_2 p_\mathrm{a}\times(p_2+1)p_\mathrm{b}}, \tag{42}$$

which mean that $\mathrm{P}_b\mathrm{G}_b$ is orthogonalized within the groups of the second level with regard to $\mathrm{P}_a^{(d)}$, and

$$(\mathrm{P}_b\mathrm{G}_b)^\mathsf{T}\mathrm{P}_a\mathrm{G}_a = 0_{(p_2+1)p_\mathrm{b}\times(p_2+1)p_\mathrm{a}}, \tag{43}$$

which express that $\mathrm{P}_b\mathrm{G}_b$ is orthogonal to the remaining independent variables in the fixed part, are sufficient for β_b to be separate in (41).

One can verify that the joint conditions (42) and (43) are equivalent to

$$\mathrm{P}_b^\mathsf{T}\mathrm{P}_a = 0_{p_\mathrm{b}\times p_\mathrm{a}}. \tag{44}$$

According to Bryk and Raudenbush (1992, p. 226), (44) is also sufficient for the unbiasedness of the estimator of β_b according to the model

$$\mathrm{N}\left(\mathrm{P}_b\mathrm{G}_b\beta_b, \sigma^2\mathrm{I}_{n\times n} + \sum_{h=p_\mathrm{a}+1}^{p_\mathrm{a}+p_\mathrm{b}}\sum_{h'=p_\mathrm{a}+1}^{p_\mathrm{a}+p_\mathrm{b}}\Omega_{hh'}\mathrm{Z}^{(h)}(\mathrm{Z}^{(h')})^\mathsf{T}\right), \tag{45}$$

which is underspecified in both the fixed and the random part when the correct specification is

$$\mathrm{N}\left(\mathrm{P}_a\mathrm{G}_a\beta_a + \mathrm{P}_b\mathrm{G}_b\beta_b, \sigma^2\mathrm{I}_{n\times n} + \sum_{h=1}^{p_\mathrm{a}+p_\mathrm{b}}\sum_{h'=1}^{p_\mathrm{a}+p_\mathrm{b}}\Omega_{hh'}\mathrm{Z}^{(h)}(\mathrm{Z}^{(h')})^\mathsf{T}\right). \tag{46}$$

Concluding Remarks and Summary

Scope of the results presented above

Residual maximum likelihood estimation. The multilevel model defined in the appendix implies that the distribution of the response, when it has been projected

onto and represented within the $n - p$ dimensional orthogonal complement of the p columns of the independent variables matrix X, depends on the parameters of the random part of (52) only. Thus, those parameters can be estimated by applying the maximum likelihood principle to this "residual" response and its distribution, resulting in the "residual maximum likelihood" (REML) estimates of the parameters of the random part (see Searle, Casella and McCulloch, 1991, chapter 6). The parameters of the fixed part may — as suggested by Kreft and de Leeuw (1998, p. 132), and by Searle, Casella and McCulloch (1991, p. 254) — be estimated in a second step, by means of a GLS procedure based on (52), after the substitution of the random part parameters with their REML estimates.

When, in the case of a model based on $\mathrm{N}(\mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2, \Sigma(\theta))$ (22), the conditions of Lemma L are satisfied and (β_1, β_2) is estimated by means of GLS based on $\mathrm{N}(\mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2, \Sigma(\hat{\theta}))$, the estimate for β_1 is equal to $(\mathrm{X}_1^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{Y}$, which implies that β_1 is separate. The property is not dependent on the definition of the estimate $\hat{\theta}$. Thus, when separability has been proven on the basis of Lemma L, this result is not influenced by the choice between FIML or REML (+ GLS). Moreover, when a subset of coefficients is separate, their estimates according to the two competing principles coincide.

Sufficient conditions. This text does not provide a complete set of rules (in the sense of: sufficient *and* necessary conditions) for separability in multilevel models. Instead, generic sufficient conditions were formulated and a list of examples was provided.

As a final example, consider the analysis of a quasi-experiment involving two treatments administered to subjects at several sites, as described by Raffe (1991). In the data matrix X = [J P G T], T is a basic level variable indicating the treatment received. Its coefficient, which quantifies the differential effect of the two treatments

on the response, is the focus of the analysis. A set of basic level variables (columns of P) and site level variables (columns of G) qualify the meaning of the treatment effect and alleviate — together with the multilevel structure of the model — the influence of population differences between the sites.

In a two-level random intercept model, it is sufficient (Rule B) to center the columns of [P T] within the sites in order to make their coefficients separate, shielding their estimates from unknown or badly measured site characteristics. The within-site centering of the indicator variable T (see also Bryk and Raudenbush, 1992, p. 28) replaces the 1's (0's) of its usual implementation with (minus) the fraction of subjects receiving treatment 0 (1) in the site.

Raffe (1991), however, takes into account the possibility of differences in the treatment effect between sites, and introduces a site level random term in the slope of T, thus involving T in the random part. With the previous examples in mind, it is no surprise that the coefficient of T is not separate in this model (despite a grand mean centering of P and within-site centering of T), as is shown by Table 4 in Raffe's (1991) text. Also, Rule O cannot be applied to make the coefficient of T separate: within-site orthogonalization of T with regard to T leaves only a zero vector. Nevertheless, it can be shown that a within site centered treatment variable T is an eigenvector of the covariance matrix of this model when all the sites have the same size and the same number of subjects receiving treatment 1 (0), and the site level random parts of the intercept and the slope of T are not correlated ($\Omega^{(2)}$ diagonal, see appendix). It is then sufficient that $\mathrm{P}^{\mathsf{T}}\mathrm{T} = 0$ (by centering P within each site/treatment combination, for example) to make the coefficient of T separate. Note that this instance of separability is founded on Lemma L, but is not included in Rule O.

Centered Variables

A $n \times 1$ data column centered within the groups of a classification characterised by the $n \times n_l$ matrix $\mathrm{W}^{(l)}$ (consisting of the indicator variables defining the n_l groups) is orthogonal to $\mathrm{W}^{(l)}$. Consequently, it is an eigenvector of $\mathrm{W}^{(l)}(\mathrm{W}^{(l)})^{\mathsf{T}}$, which is a building block of the covariance matrix of a random intercept model involving this classification. This is the foundation of the special role of centered variables in the present discussion about separability and in related accounts of unbiasedness (see the present text and Bryk and Raudenbush, 1992, pp. 204–207). This structural relationship between centered variables and the random intercept model probably also explains the relatively large share of attention that has been devoted in the multilevel analysis literature to issues of centering, as compared to the other aspects of the transformation of raw data into operationalizations of concepts. Those preliminary steps are usually outside the scope of books and texts concerned with multilevel analysis.

Centering may be applied in order to "shield" variables measured at a balanced level from (badly measured or incomplete) data for variables measured at other levels in a random intercept model (see Rules B and I). With that potential use of centering in mind, it is interesting to note that the centering transformation itself requires only group membership data. It does not require any information about the variables which the analyst wants to avoid, except for the level associated with them.

If the efforts of, amongst others, Bryk and Raudenbush (1992), Kreft and de Leeuw (1995) and Hofmann and Gavin (1998) have been successful, analysts will think carefully about whether or not, and how, to center their variables when preparing a multilevel analysis. The range of centering (or, for that matter, orthogonalization) options broadens quickly when more classifications become involved. Consider,

for example, the case of a dataset of n pupils within n_2 classes within n_3 schools, with p_1 pupil variables ($n \times p_1$ data matrix P) and p_2 class characteristics ($n_2 \times p_2$ data matrix C). For each pupil variable, the analyst faces at least two options, namely to center within classes or rather within schools. Also, if the analyst wants to center the class variables at the school level, it is possible to apply the centering transformation to the $n_2 \times p_2$ data matrix C or rather to the "expanded" $n \times p_2$ class characteristics matrix $\mathrm{W}^{(2)}\mathrm{C}$ (where $\mathrm{W}^{(2)}$ defines the class level). In the former case, all classes within a school have equal weight, in the latter classes are weighted according to their number of pupils. If the prevalent argument for centering the pupil variables happens to be the advantage of separability, Rule B dictates that they should be centered within classes rather than (merely) within schools. Unfortunately, Rule I is not able to distinguish between centering C or centering $\mathrm{W}^{(2)}\mathrm{C}$, because it is only applicable when all class sizes in the dataset are equal, in which case the two centering options are equivalent.

Centering decisions in multilevel analysis may have repercussions on the (success of) the *operationalization* of the concepts of interest, may constitute a choice between different model *specifications*, may help to obtain the advantages of *separability and unbiasedness*, and may, allegedly, even have a significant influence on the *numerical* properties of the iterative algorithms employed to estimate the models. It is important to note that separability (and unbiasedness), despite its "natural" link with centering (via the eigenvalue/eigenvector structure of the covariance matrix of the random intercept model) is only one item in this list, which was stated in what is, in our opinion, the order of decreasing priority.

Summary

We formulated sufficient conditions for separability of components of the parameter vector β in a model based on $\mathrm{Y} \sim \mathrm{N}(\mathrm{X}\beta, \Sigma(\theta))$ and the maximum likelihood principle. The application of those generic sufficient conditions to the multilevel model required a search for invariant eigenspaces of the multilevel model covariance matrix. It resulted in a list of examples, each stating rules (sufficient conditions) for separability in particular subtypes of the multilevel model. The list ranges from the rigidly structured completely balanced completely hierarchical random intercept model to the fairly general multilevel model with a random intercept structure at the basic level.

Two conclusions stand out. First, the special role (with regard to separability, but also, for example, in the issue of unbiasedness in cases of underspecification) of within-group orthogonalized and centered variables in multilevel models, can be traced back to the eigenvalue/eigenvector decomposition of the covariance matrix. Second, in unbalanced multilevel models, the possibility to build in separability seems to be limited to coefficients of basic level variables, under stringent orthogonality requirements.

In a broader perspective, those results illustrate the typical breakdown of symmetries and balances when ordinary regression is replaced with multilevel modeling, implying the need to reformulate the rules directing data analyses. This article is an attempt to contribute to this effort.

Appendix

The range of models indicated by the term "multilevel model" is not unequivocal in the literature. Instead, the definition of the scope of the term and the notation are usually geared to the purpose of interest. Therefore, this appendix provides a

rigorous definition of the multilevel model referred to in this article.

Ingredients of the multilevel model. It is assumed that measurements for p "independent" variables and for a "response variable" have been obtained for n (basic) units of observation. The data on the independent variables are stored in the $n \times p$ matrix X, which is assumed to have the maximal rank, p. The symbol Y indicates the $n \times 1$ response data column.

It is also assumed that the set of n basic units is partitioned in g different ways. The first partition is the "basic level", which trivially allocates each basic unit to its exclusive group (a classification consisting of $n_1 = n$ groups). The $g - 1$ remaining classifications ("levels") partition the basic units into $n_2, n_3, \ldots, n_g$ groups respectively. At this stage no nesting relationships between those $g - 1$ non-trivial classifications are postulated. The group membership status of each basic unit according to the $g - 1$ non-trivial classifications must be explicitly available if a multilevel analysis is to be applied. Such membership data can be represented in several ways. Here, it is assumed that group membership is stored by means of scores on indicator variables (0/1), one variable for each group of each classification. For the l-th classification ($l = 1, 2, \ldots, g$) the j-th column of the $n \times n_l$ matrix $\mathrm{W}^{(l)}$ represents the indicator variable for the j-th group: every basic unit belonging to the group scores a 1, the other units score a 0.

Multilevel models are parametric. In the definition of multilevel models which is used here, the parameters are: a $p \times 1$ column β of "fixed effects" and g $p \times p$ matrices $\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(g)}$ (one for each classification) for the "random part" of the model. In order to define a particular multilevel model the set of potential values of the parameter set $(\beta, \Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(g)})$ must be specified. This set of potential values is indicated by the symbol D. The discussion of multilevel models in this article is

limited to cases where D is a subset of

$$\mathsf{R}^{p\times 1} \times \mathsf{H} \times \underbrace{\mathsf{R}^{p\times p}_{\text{pos. semidef.}} \times \ldots \times \mathsf{R}^{p\times p}_{\text{pos. semidef.}}}_{g-1 \text{ factors}} \tag{47}$$

where $\mathsf{R}^{p\times 1}$ is the set of all real $p \times 1$ vectors, $\mathsf{R}^{p\times p}_{\text{pos. semidef.}}$ is the set of the positive semidefinite real $p \times p$ matrices, and H denotes the subset of $\mathsf{R}^{p\times p}_{\text{pos. semidef.}}$ consisting of the matrices Ω_1 that satisfy the condition

$$\mathrm{X}_{i.}\Omega_1(\mathrm{X}_{i.})^{\mathsf{T}} > 0 \tag{48}$$

for every row (record) $\mathrm{X}_{i.}$ in the data matrix X.

Formulation with latent variables. When the ingredients p, n, X, Y, g, n_1, n_2, $\ldots$, n_g, $\mathrm{W}^{(1)}$, $\mathrm{W}^{(2)}$, $\ldots$, $\mathrm{W}^{(g)}$ and D have been specified, the multilevel model can be formulated. The model postulates the existence of a parameter value $(\beta, \Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(g)})$ from D and the existence, for each value $l \in \{1, 2, \ldots, g\}$, of a series of p stochastic vectors, each of dimension $n_l \times 1$: ${}^{(l)}U^{(1)}, {}^{(l)}U^{(2)}, \ldots, {}^{(l)}U^{(p)}$. Each of the stochastic vectors is normally distributed with mean zero. The variance/covariance relationships between the gp stochastic vectors are determined by:

$$\mathrm{cov}({}^{(l')}U^{(h')}, {}^{(l'')}U^{(h'')}) = \begin{cases} 0_{n_{l'}\times n_{l''}} & \text{when } l' \neq l'' \\ \Omega^{(l')}_{h'h''}\mathrm{I}_{n_{l'}\times n_{l'}} & \text{when } l' = l'' \end{cases} \tag{49}$$

The model then states that the response vector Y is a realization of the $n \times 1$ stochastic variable

$$\mathrm{X}\beta + \sum_{l=1}^{g}\sum_{h=1}^{p} {}^{(l)}\mathrm{Z}^{(h)(l)}U^{(h)}. \tag{50}$$

In this expression, the matrices ${}^{(l)}\mathrm{Z}^{(h)}$ implement the group structure and the involvement of the independent variables in the random part of the model:

$${}^{(l)}\mathrm{Z}^{(h)}_{ij} = \mathrm{W}^{(l)}_{ij}\mathrm{X}_{ih} \tag{51}$$

which means that the j-th column of the $n \times n_l$ matrix ${}^{(l)}\mathrm{Z}^{(h)}$ is obtained by copying the h-th column of X and replacing the values of units that do not belong to the j-th group of the l-th classification by zero.

Formulation without latent variables. This article focusses on the estimation of the parameters β, $\Omega^{(1)}$, $\Omega^{(2)}$, ..., $\Omega^{(g)}$ only, and not on the calculation of scores for the latent variables ${}^{(l)}U^{(h)}$. For such a purpose, the multilevel model can be formulated without any explicit involvement of the latent variables. It then postulates the existence of a parameter value $(\beta, \Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(g)})$ from D which is such that the response vector Y is a realization of a $n \times 1$ stochastic variable which is distributed as

$$\mathrm{N}\left(\mathrm{X}\beta, \sum_{l=1}^{g} \sum_{h=1}^{p} \sum_{h'=1}^{p} \Omega^{(l)}_{hh'} {}^{(l)}\mathrm{Z}^{(h)} ({}^{(l)}\mathrm{Z}^{(h')})^{\mathsf{T}}\right). \tag{52}$$

The latter formulation of the multilevel model, without the latent variables, is implied by the previous latent variables formulation.

Note that the requirement that D is a subset of (47) guarantees that the covariance matrix in (52) is positive definite for any admissible value of the parameters.

Subtypes derived from the full model. The definition of multilevel models which is adopted here is fairly general, as no a priori limit is imposed on the number of different classifications, nor is there any a priori assumption about nesting relationships (except for the nestedness of the basic level within all other "levels"). Moreover, there is no a priori difference between the treatment of variables that can vary between any pair of basic units ("basic level" variables or "level-1" variables) and, say, variables that are constant within each group of the l-th classification ("level-l" variables) or indeed the constant vector $\mathrm{J}_{n\times 1}$ (all elements equal to 1) that implements the intercept. This is in line with the general formulation of multilevel

models proposed by Goldstein (1995, p. 20), who notes that "any of the explanatory variables may be measured at any of the levels".

In a meaningful discussion of separability, however, it proves necessary to define and consider special types of multilevel models. Such subtypes are defined by means of assumptions about the classifications (nesting relationships between classifications, balancedness within classifications) and by specifying the set D of potential parameter values. A choice of the set D that is particularly helpful in the discussion of separability is the one that reduces the general multilevel model to a so-called random intercept model. Assuming that the r-th column of X is the constant $n \times 1$ vector of 1's (which implements the intercept in the model), the reduction to a random intercept model is achieved by requiring that all elements of each parameter matrix $\Omega^{(l)}$ are zero, excepting the (r,r)-th element, which will be denoted as σ_l^2. The resulting model then postulates the existence of parameter values $\beta \in \mathsf{R}^{p\times 1}$, $\sigma_1^2 \in \mathsf{R}_{>0}$ and $\sigma_2^2, \sigma_3^2, \ldots, \sigma_g^2 \in \mathsf{R}_{\geq 0}$ such that Y is a realization of a stochastic variable distributed as

$$\mathrm{N}\left(\mathrm{X}\beta, \sum_{l=1}^{g} \sigma_l^2 \mathrm{W}^{(l)} (\mathrm{W}^{(l)})^{\mathsf{T}}\right). \tag{53}$$

Between the two extremes of a "full" model which is defined by choosing D equal to the set (47) and a random intercept model with only g scalar parameters in the random part, lies a broad range of models that are specified by choosing combinations of parameter components which are to be kept fixed at given values. The hierarchical two- and three-level models (a basic level nested within a second level, which is nested within a third, for example: pupils within classes within schools) discussed by Bryk and Raudenbush (1992) make an important example of such intermediate models.

In the most general three-level model defined by Bryk and Raudenbush (1992), some (say: p_1) independent variables have been measured at the basic level (meaning

that they are not constant within the groups of the second level), some (p_2) at level 2 (they are constant within level-2 groups, but not within all groups of the third level), and some (p_3) at level 3 (they are constant within level-3 groups, but not constant overall). In its fullest form, the datamatrix X contains not only the constant column $J_{n\times 1}$ and the $p_1 + p_2 + p_3$ "main" variables, but also the $p_1p_2 + p_2p_3 + p_3p_1$ second order cross-level interaction variables and the $p_1p_2p_3$ third order cross-level interaction variables. The $p \times p$ (with $p = 1+p_1+p_2+p_3+p_1p_2+p_2p_3+p_3p_1+p_1p_2p_3$) parameter matrix $\Omega^{(1)}$, however, is restricted as in the random intercept model, with only one non-zero element (at the row and column associated with the constant vector $J_{n\times 1}$). Secondly, the elements of the $p \times p$ matrix $\Omega^{(2)}$ are allowed to be non-zero in the block corresponding to the intercept vector $J_{n\times 1}$ and the p_1 main level-1 variables only. Thirdly, the elements of the $p \times p$ parameter matrix $\Omega^{(3)}$ can be different from zero in the block associated with the intercept, the p_1 main level-1 variables, the p_2 main level-2 variables and the p_1p_2 interaction variables between the basic and the second level, but not elsewhere. None of the main level-3 variables nor any interaction variable involving a level-3 variable is represented in the random part.

References

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park: Sage Publications.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Arnold.

Hofmann, D.A., & Gavin, M.B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24* (5), 623–641.

Kreft, I.G.G., & de Leeuw, J. (1998). *Introducing multilevel modeling.* London: Sage Publications.

Kreft, I.G.G., de Leeuw, J., & Aiken, L.S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30* (1), 1–21.

Longford, N.T. (1989). To center or not to center. *Multilevel Modelling Newsletter, 1* (3), 7.

Longford, N.T. (1993). *Random coefficient models.* Oxford: Clarendon Press.

Opdenakker, M.-C., & Van Damme, J. (1997). Centreren in multilevel analyse: Implicaties van twee centreringsmethoden voor het bestuderen van schooleffectiviteit [Centering in multilevel analysis: Implications of two centering methods for the study of school effectiveness]. *Tijdschrift voor Onderwijsresearch, 22* (4), 264–290.

Plewis, I. (1989). Comment on "centering" predictors. *Multilevel Modelling Newsletter, 1* (3), 6.

Raffe, D. (1991). Assessing the impact of a decentralized initiative: The British technical and vocational initiative. In S. W. Raudenbush & J. D. Willms, (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 149–166). Orlando: Academic Press.

Raudenbush, S.W. (1989a). "Centering" predictors in multilevel analysis: Choices and consequences. *Multilevel Modelling Newsletter, 1* (2), 10–12.

Raudenbush, S.W. (1989b). A response to Longford and Plewis. *Multilevel Modelling Newsletter, 1* (3), 8-10.

Schumacker, R.E., & Bembry, K. (1996). Empirical characteristics of centering methods for level-1 predictor variables in HLM. *Multiple Linear Regression Viewpoints, 23* (1), 1–8.

Searle, S.R., Casella, G., & McCulloch, C.E. (1991). *Variance components.* New York: Wiley.

Snijders, T.A.B., & Bosker, R.J. (1994). Modeled variance in two-level models. *Sociological Methods and Research, 22* (3), 342–363.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage Publications.