UCLA Statistics

Multilevel Analysis, with Extensions

Jan de Leeuw

May 26, 2010

- We start by reviewing the research on *multilevel analysis* that has been done in psychometrics and educational statistics, roughly since 1985.
- The canonical reference (at least I hope so) is De Leeuw and Meijer (eds), *Handbook of Multilevel Analysis*, Springer, 2008.
- Although I have been asked to specifically review the social science applications, I will also try to establish some links with environmental statistics and space-time analysis.
- This is actually easier than expected, because the multilevel model is a special case of the *mixed linear model* or the *hierarchical linear model*.
- Here, and throughout the Handbook, we use the Van Dantzig (or Dutch) Convention: random variables are underlined.

Random Coefficient Models

Basics

Let's start simple. Suppose we have m groups, n_j observations in group j, and p predictors in each of the groups.

$$\begin{aligned}\underline{y}_j &= X_j\underline{b}_j + \underline{\epsilon}_j,\\ \underline{b_j} &= \beta + \underline{\delta}_j.\end{aligned}$$

Thus

$$\underline{y}_j = X_j\beta + X_j\underline{\delta_j} + \underline{\epsilon}_j,$$

and

$$\begin{aligned}\mathbf{E}(\underline{y}_j) &= X_j\beta,\\ \mathbf{V}(\underline{y}_j) &= X_j\Omega_jX_j' + \Sigma_j.\end{aligned}$$

Random Coefficient Model

Stacked

$$\begin{bmatrix}\underline{y}_1\\ \vdots\\ \underline{y}_m\end{bmatrix} = \begin{bmatrix}X_1\\ \vdots\\ X_m\end{bmatrix}\begin{bmatrix}\beta\end{bmatrix} + \begin{bmatrix}X_1 & 0 & \cdots & 0\\ 0 & X_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & X_m\end{bmatrix}\begin{bmatrix}\underline{\delta}_1\\ \vdots\\ \underline{\delta}_m\end{bmatrix} + \begin{bmatrix}\underline{\epsilon}_1\\ \vdots\\ \underline{\epsilon}_m\end{bmatrix}$$

An RC model is a random intercept (RI) model if only the intercept has a random component. Thus the first column of all X_j has all ones (is equal to u_j), and only the first of the elements of $\underline{\delta}_j$ has non-zero variance.

$$\begin{bmatrix}\underline{y}_1\\ \vdots\\ \underline{y}_m\end{bmatrix} = \begin{bmatrix}X_1\\ \vdots\\ X_m\end{bmatrix}\begin{bmatrix}\beta\end{bmatrix} + \begin{bmatrix}\underline{\delta}_1 u_1\\ \vdots\\ \underline{\delta}_m u_m\end{bmatrix} + \begin{bmatrix}\underline{\epsilon}_1\\ \vdots\\ \underline{\epsilon}_m\end{bmatrix}$$

Random Coefficient Models

Identification

It is common, and in fact necessary, to make some additional assumptions.

- $\beta_j = \beta$
- $\Omega_j = \Omega$, often diagonal.
- $\Sigma_j = \Sigma$, often scalar.

In addition we often assume that the disturbances $\underline{\epsilon}_j$ and $\underline{\delta}_j$ are jointly multivariate normal and mutually uncorrelated.

This is needed for likelihood inference, and it is an excuse to stay away from higher order moments.

Slopes-as-Outcomes Models

Basics

In educational statistics multilevel analysis was introduced as a way to formalize *contextual analysis* by embedding it in the hierarchical linear model. The leading example is n_j students in m schools. Suppose we have p student-level predictors, and we have q school-level predictors. We assume

$$\underline{y}_{ij} = \sum_{s=1}^{p} x_{ijs}\underline{b}_{js} + \underline{\epsilon}_{ij},$$

$$\underline{b}_{js} = \sum_{r=1}^{q} z_{jr}\beta_{sr} + \underline{\delta}_{js}.$$

This elementwise formulation of the model shows how to extend SAO to more than two levels (students in classrooms in schools in districts).

Slopes-as-Outcomes Models

Matrix Form

Define $\overline{Z}_j$ as the $p \times pq$ matrix $\overline{Z}_j = z_j' \oplus \cdots \oplus z_j'$, where all z_j have length q. Also do some obvious stacking. Then we can write

$$\begin{aligned} \underline{y}_j &= X_j \underline{b}_j + \underline{\epsilon}_j, \\ \underline{b}_j &= \overline{Z}_j \beta + \underline{\delta}_j. \end{aligned}$$

Using the columns x_s^j and the vectors z_j it follows that the fixed part of the model has the block-rank-one structure given by *cross-level interactions*.

$$\begin{bmatrix} \mathbf{E}(\underline{y}_1) \\ \vdots \\ \mathbf{E}(\underline{y}_m) \end{bmatrix} = \begin{bmatrix} X_1 B z_1 \\ \vdots \\ X_m B z_m \end{bmatrix} = \begin{bmatrix} x_1^1 z_1' & x_2^1 z_1' & \cdots & x_p^1 z_1' \\ x_1^2 z_2' & x_2^2 z_2' & \cdots & x_p^2 z_2' \\ \vdots & \vdots & \ddots & \vdots \\ x_1^m z_m' & x_2^m z_m' & \cdots & x_p^m z_m' \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Slopes-as-Outcomes Models

Essence

This is the core of SAO models: we fit cross-level interactions to a ragged array of data Y in which the rows are uncorrelated and the columns have covariance structure $\mathbf{V}(\underline{y}_j) = X_j\Omega_j X_j' + \Sigma_j$.

And then one searches over the submodels that simultaneously set various cross-level interaction effects and various covariances in Ω_j and Σ_j equal to zero. The default in most cases is to use $\Omega_j = \mathbf{diag}(\Omega)$ and $\Sigma_j = \sigma^2 I$.

The next step is to become “state-of-the-art” and get applied researchers to buy our expensive software package.

Slopes-as-Outcomes Models

Truth ?

It would be foolhardy, maybe even insane, to pretend that models such as SAO actually describe what goes on in classrooms. We merely have a descriptive device that gives one formalization of the dependence between students in the same classroom, as well as the cross-level interactions from contextual analysis.

Of course if we have reliable prior information we should incorporate it into the device to reduce both bias and variance, to be able to talk to our colleagues, and to get published. But in this case the prior information is that some students are in the same class and that some variables may be related to school outcomes.

SAO provides a framework, maybe even a “language”, to imbed this (rather minimal) prior information. There are many functionally similar frameworks around in statistics.

Slopes-as-Outcomes Models

Generalizations

Some fairly straightforward generalizations follow, once we realize that the SAO model is just a mixed linear model.

- We can generalize to nonlinear mixed models.
- We can generalize to generalized mixed linear models.

Both generalizations are straightforward, although computationally far from trivial.

Somewhat more intricate are

- Non-nested (crossed) designs.
- Multivariate outcomes.
- Correlated observations.
- Covariance structures.

Slopes-as-Outcomes Models

Non-nested Designs

Non-nested designs occur in educational statistics, for example, if we keep track of both the student's primary and secondary schools. Clearly those classifications are not hierarchical, and the multilevel model becomes more complicated.

Suppose, for example, we measure PM-10 in a number of years and at a number of observation points located in some cells of a rectangular grid. A simple random intercept model would be

$$\underline{y}_i = \sum_{s=1}^{p} x_{is}\beta_s + \underline{\delta}_{t(i)} + \underline{\nu}_{l(i)} + \underline{\epsilon}_i,$$

where $t(i)$ is the year and $l(i)$ is the location of observation i. This can be extended to SAO models if we have regressors to describe time and space.

Growth Curve Models

Basics

Before we start making matters even more complicated, let us first treat a more simple special case. Suppose all X_j are the same. An example could be m objects measured at n time points. X could contain a basis of polynomials or Fourier coefficients to code time points, while Z would have characteristics of individuals.

The data are a realization of an $n \times m$ matrix-valued random variable $\underline{Y}$ (often assumed matrix-variate normal). Then

$$\begin{aligned}\mathbf{E}(\mathbf{vec}(\underline{Y})) &= (X \otimes Z)\beta,\\ \mathbf{V}(\mathbf{vec}(\underline{Y})) &= I \otimes (X\Omega X' + \Sigma).\end{aligned}$$

This is a generalization of the classical Pothoff-Roy model, which has $\Omega = 0$. There can be missing data, of course.

Growth Curve Models

Loss Function

The negative log-likelihood loss function for an even more general growth curve model is

$$\mathcal{D} = m \log \mathbf{det}(\Sigma) + n \log \mathbf{det}(\Omega) + \\ + \mathbf{tr}\ \Sigma^{-1}(Y - XBZ')\Omega^{-1}(Y - XBZ')',$$

which shows that the dispersions of groups (objects) and individuals (time points) are *separable*.

In growth curve analysis we usually model the expectations and keep the covariances simple. But we can also work the other way around and use elaborate dispersion and simple expectation models. That is in the tradition of structural equation and multivariate time series modeling. They are combined in the R package `leopold`, with Wei-Tan Tsai.

Separable Models

Elaborate Means

The mean structure $\mathbf{E}(\underline{Y}) = XBZ'$ in the growth curve model has both X and Z known. If X and/or Z are (partially) unknown, we can incorporate

- principal component analysis (perhaps non-negative),
- reduced rank regression analysis,
- correspondence analysis,
- canonical correspondence analysis,
- fixed score factor analysis.

Although these models *seem* very different, they are all basically matrix approximation methods minimizing the same loss function. And they naturally fit into the same block relaxation or alternating least squares algorithm.

Separable Models

Elaborate Dispersions

- In linear models, such as growth curve models, we typically have simple dispersion structures such as $\Sigma = \sigma^2 I$. But it also makes sense to try, for example, AR(1) or more general Toeplitz forms for Σ.
- Thus we can have, say, AR(1) form for Σ, and some factor analytic or spatial covariance structure for Ω.
- There is an obvious trade off allocating parameters to the means and allocating parameters to the dispersions.
- If there are too many parameters in both modes, maximum likelihood runs into incidental parameter problems, such as Neyman-Scott bias or Anderson-Rubin degeneracy. The boundary, where the log likelihood becomes unbounded, is interesting.

Separable Models

Higher Order

The growth curve model with

$$\begin{aligned} \mathbf{E}(\underline{y}) &= (X \otimes Z)\beta, \\ \mathbf{V}(\underline{y}) &= \Sigma \otimes \Omega, \end{aligned}$$

easily generalizes to $K-$dimensional arrays as

$$\begin{aligned} \mathbf{E}(\underline{y}) &= (X_1 \otimes X_2 \otimes \cdots \otimes X_K)\beta, \\ \mathbf{V}(\underline{y}) &= \Sigma_1 \otimes \Sigma_2 \otimes \cdots \otimes \Sigma_K, \end{aligned}$$

where we can have covariance structures for each of the Σ_k, and elaborate mean structures for X_k and β as well. This can be used to cover various forms of array decomposition, such as ICA. The BR/ALS algorithm, in the next version of `leopold`, remains basically the same.

Separable Models

Truth ?

The modes of the multi-array can be defined in various ways. One mode can be cross-sectional, if multiple variables are measured on the same individuals, or at the same time an location. One mode can be spatial, another temporal. Or, alternatively, one mode can be longitude and another latitude.

As in the students-in-schools context, most processes observed in nature are not separable. But neither are regressions linear, distributions normal, and variables conditionally independent. Models are false, but means and variances are still useful summaries. Same for regression, GLM, and PCA/ICA.

Truth is elusive. The question is if the matrix approximations help to make description simpler and/or prediction better, and how much bias is traded off against how much variance. Ultimately, the client/scientist should decide, preferably in a reproducible way.