Using MELSM to Identify Consistent and Inconsistent Schools

Marwin Carmo

Background

Mixed-effects location-scale models (MELSM) are an extension of standard mixed-effects models, such that the residual variance is not assumed to be constant but can be modeled, allowing the inclusion of a sub-model to address potential differences in the residual variance. The MELSM estimates simultaneously a model for the means (location) and a model for residual variance (scale). For instance, in an educational research setting, the model defines a multilevel mode for the observed values, y_{ij}, for school j and student i, and a multilevel model for the within-school residual variances, σ_{ij}.

Based on the MESLM ability to estimate residual standard deviations, we can implement the spike-and-slab regularization technique to the scale random effects to identify schools (clusters) whose student-level residual standard deviations are not captured well by scale fixed effects. The idea is to place an indicator variable, $\delta \in [0, 1]$, to the random effects to be subjected to shrinkage where the school-level effect for the scale either retains the random effect or reduces to the fixed effect, according to δ's value. So, for each j school and k random effect, a posterior inclusion probability (PIP) is computed to quantify the probability that a given random effect is included in the model, conditional on the observed data. A PIP greater than 0.75 indicates strong evidence that the school's residual variability deviates significantly from the average due to higher or lower consistency in student performance. This method is termed Spike and slab mixed-effects location-scale model (SS-MELSM) and is implemented the R package `ivd`.

Research question

The motivation for the SS-MELSM is to identify and isolate clusters that display unusual amounts of residual variability. It shares the same general approach of modeling residual variances in a MELSM, such as studying variables that contribute to (in)consistency. The focus, however, is on the *identification* of clustering units that show unusually high or low consistency in academic achievement. It is expected that the clusters attributed PIPs higher than 0.75 have a distinct pattern of within-cluster variability that clearly distinct them from others.

Method

The data for this study comes from The Elementary Education Evaluation System (Saeb), an assessment program conducted by the Brazilian government to evaluate the quality of elementary education across the country. This dataset contains standardized math test scores from 11,386 11th and 12th graders across 160 schools and is part of the `ivd` package.

To keep this illustration simple, I opted for a model that features an intercept-only specification in both the location and scale submodels. The model predicting math achievement for the i-th student within the j-th school is:

$$\begin{aligned} \text{mAch}_{ij} &\sim \mathcal{N}(\mu_j, \sigma_j) \\ \mu_j &= \gamma_0 + u_{0j} \\ \sigma_j &= \exp(\eta_0 + t_{0j}) \end{aligned} \tag{1}$$

$$\mathbf{v} = \begin{bmatrix} u_0 \\ t_0 \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} \tau^2_{u_0} & \tau_{u_0 t_0} \\ \tau_{u_0 t_0} & \tau^2_{t_0} \end{bmatrix}\right)$$

This model defines math achievement (mAch_{ij}) via the fixed intercept parameter γ_0 and the random intercept u_{0j}, capturing the deviation of the j-th school from the fixed effect. Each school's residual standard deviation is modeled as a function of the fixed effect η_0 and the school-specific deviation t_{0j}, both defined on the log scale.

The random effects are assumed to come from the same multivariate Gaussian Normal distribution with zero means and covariance matrix $\mathbf{\Sigma}$. This matrix can be decomposed into $\mathbf{\Sigma} = \boldsymbol{\tau}\mathbf{\Omega}\boldsymbol{\tau}'$, where $\boldsymbol{\tau}$ is a diagonal matrix holding the random-effect standard deviations and $\mathbf{\Omega}$ is the correlation matrix that contains the correlations among all random effects. Next, we can decompose $\mathbf{\Omega}$ via the Cholesky factor $\mathbf{L}$ of $\mathbf{\Omega} = \mathbf{L}'\mathbf{L}$.

With these decomposition, the random effects vector $\mathbf{v}$ by multiplying $\mathbf{L}$ with the standard deviations $\boldsymbol{\tau}$ and scaling it with a standard normally distributed $\mathbf{z}_j$. Then, it can be expanded to include an indicator vector $\boldsymbol{\delta}_j$ of length k (for $1, \ldots, k$ random effects) for each random effect to be subjected to shrinkage:

$$\mathbf{v}_j = \boldsymbol{\tau}\mathbf{L}\boldsymbol{z}_j\boldsymbol{\delta}_j. \tag{2}$$

Each element in $\boldsymbol{\delta}_j$ takes integers $\in \{0, 1\}$ and follows a $\delta_{jk} \sim \text{Bernoulli}(\pi = 0.5)$ distribution. Depending on δ's value, the computations in Equation (2) will either retain the random effect or shrink it to exactly zero. Consequently, the estimated posterior inclusion probability (PIP) quantifies the probability that a given random effect is included in the model, conditional on the observed data. The PIP of any random effect k is determined by the proportion of MCMC samples where $\delta_{jk} = 1$ across the the total number of posterior samples S:

$$Pr(\delta_{jk} = 1|\mathbf{Y}) = \frac{1}{S}\sum_{s=1}^{S}\delta_{jks}. \tag{3}$$

Setting the prior probability of π to 0.5 implies equal prior odds, $Pr(\delta_{jk} = 1)/Pr(\delta_{jk} = 0) = 1$, reflecting no prior information about the presence of a random effect. The Bayes factor for including the kth random effect in the jth school simplifies to:

$$BF_{10_{jk}} = \frac{Pr(\delta_{jk} = 1|\mathbf{Y})}{1 - Pr(\delta_{jk} = 1|\mathbf{Y})}. \tag{4}$$

A posterior inclusion probability $Pr(\delta_j = 1|\mathbf{Y}) \geq 0.75$, corresponding to a $BF_{10_{jk}} \geq 3$, provides evidence for including the random effect over fixed effects alone.

This model was fitted using the `ivd` package using six chains of 3,000 iterations and 12,000 warm-up samples. The results present the posterior estimates of: (i) the PIP for δ_{jt_0}, (ii) the scale random effect SD, τ_{t_0}, (iii) within-school residual SD, σ_j, and (iv) the math scores, μ_j.

Results

The location model yielded an intercept fixed effect of $\gamma_0 = 0.115$ (95% CrI [0.055; 0.174]). This suggests that, on average, schools in this sub-sample performed slightly above the standardized mean score of the larger sample of schools from which it was drawn. In the scale model, the estimated intercept of the residual standard deviation was $\eta_0 = -0.235$ (95% CrI [-0.253; -0.217]) on the log scale. This corresponds to a mean of $\exp(-0.235) = 0.791$ on the standard deviation metric. Figure 1 displays the 160 PIPs for all schools. Notably, eight schools exhibited PIPs greater than 0.75, as indicated by the colored points above the dotted line. These PIPs suggest significant deviations from the average school-level variance.

Figure 2, panel A, shows the distribution of within-school residual standard deviations. Among the eight identified schools, five exhibited lower-than-average standard deviations, indicating *higher consistency* in student performance. Conversely, the remaining three schools showed greater variability, reflected by their positioning farther to the right and above the dotted line, indicating *lower consistency* (larger standard deviations) in math achievement for these schools. This pattern is further elucidated in panel B, which shows a scatterplot of school PIPs to their average math scores. School 46 was not only highly consistent in math achievement, but they also performed better than the average in our sample. We can also observe a mix of more inconsistent and consistent schools within 0.5 standard deviations of the mean. While the average math achievement of these schools is comparable, the within-school variability was substantially different.

Figure 3 shows the posterior distributions of the scale random intercept for four specific schools. Panel A illustrates School 46, a high-achieving and the most consistent school in the sample, that "escaped'' the pull of the spike. In contrast, Panel B shows one of the more inconsistent schools with an average math achievement. This school had a PIP of 0.884 with a small evidence of a spike effect. In contrast, School 38, shown in Panel C, was both average in math achievement and average in the residual standard deviation with a clear spike at zero, indicating that the fixed effect captured this school well.

Visualization

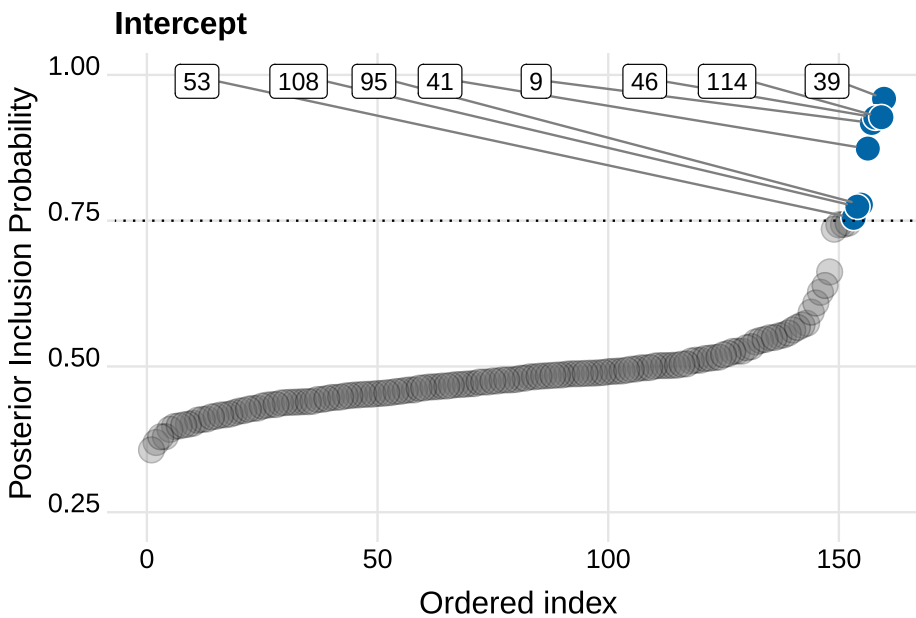


Figure 1: The posterior inclusion probabilities of the scale random effect for the 160 schools in the dataset for Models 1 and 2. A dotted line marks the PIP threshold of 0.75. Schools with PIPs exceeding this threshold are colored.

```
Warning: `stat(quantile)` was deprecated in ggplot2 3.4.0.
i Please use `after_stat(quantile)` instead.
```

```
Picking joint bandwidth of 0.00671
```

```
Picking joint bandwidth of 0.00524
```

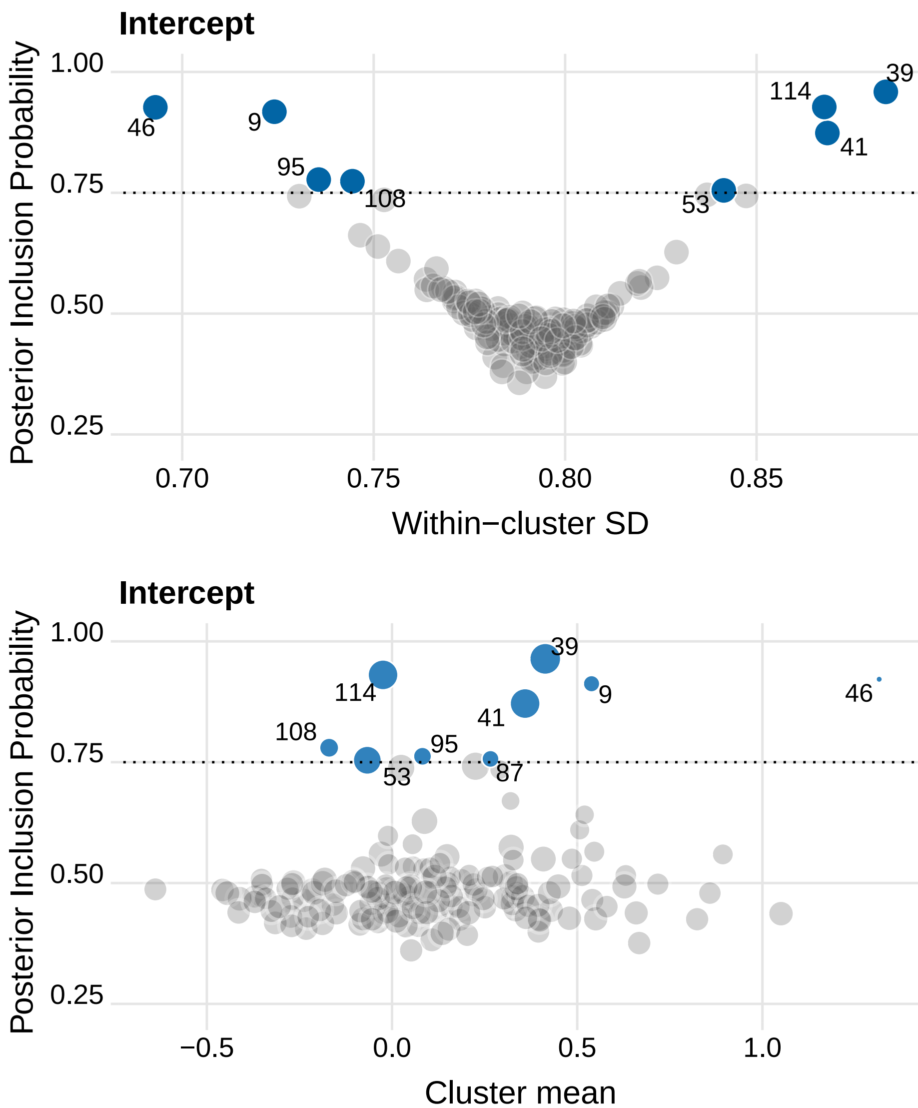
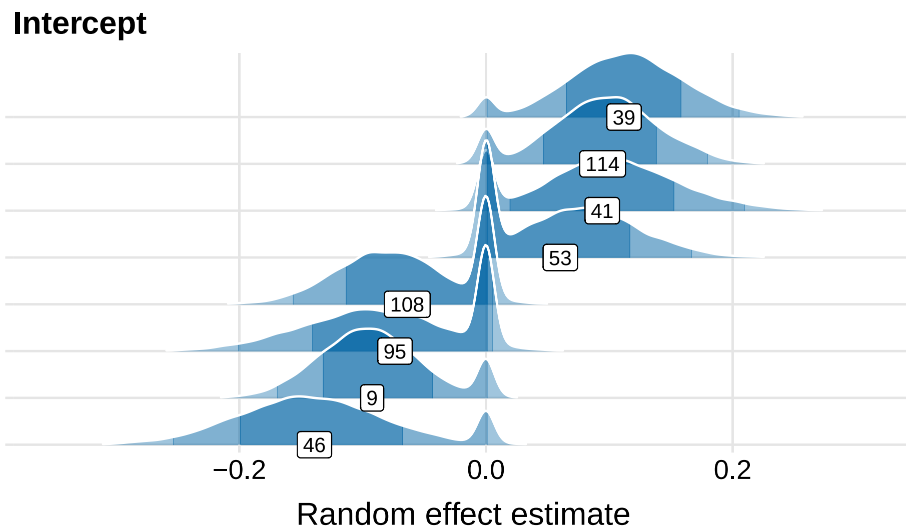


Figure 2: Scatter plots of posterior inclusion probability (PIP) versus within-cluster standard deviation (SD). The y-axis represents the PIP for the scale random intercept, and the x-axis represents the within-cluster SD. The plots exhibit a V-shaped pattern, where schools with the lowest and highest SDs are positioned toward the left and right, respectively, while those with average SDs are clustered near the center.



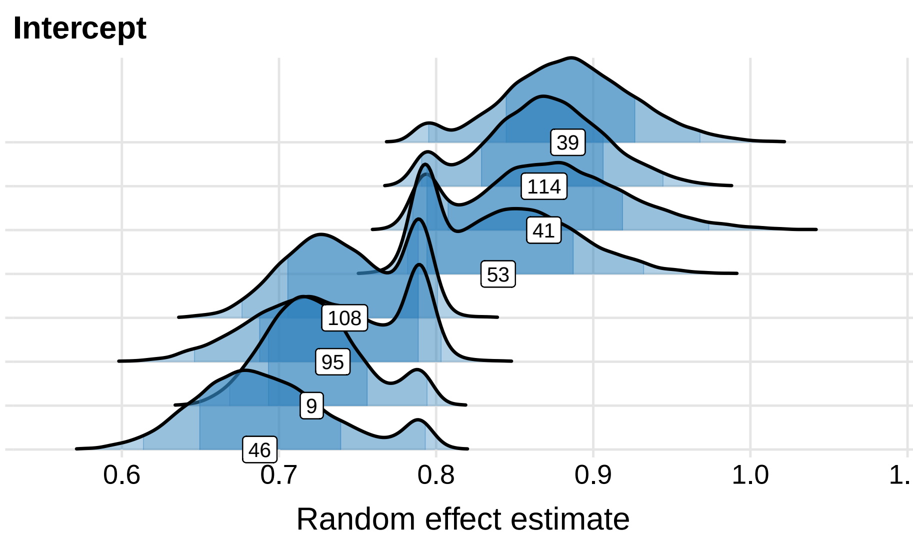


Figure 3: Scatter plots of posterior inclusion probability (PIP) versus within-cluster standard deviation (SD). The y-axis represents the PIP for the scale random intercept, and the x-axis represents the within-cluster SD. The plots exhibit a V-shaped pattern, where schools with the lowest and highest SDs are positioned toward the left and right, respectively, while those with average SDs are clustered near the center.

Visualization strategies

In this project I created five ways to visualize the posterior distribution of estimates produced by the `ivd` model. They were designed to complement the information provided by the model's summary by focusing on the schools with PIPs exceeding the proposed threshold of "significance."

For all plots, color was used as a tool to highlight particular schools. In Figure 1 and both panels of Figure 2, those with unnusual levels of variability were colored to emphasize to the user that these are the relevant units, opposed to the gray shade given to schools with "average" variability that do not provide relevant information to the purposes of the analysis. In previous versions of these plots, a unique color was assigned to each high-PIP school to distinguish them. However, following Wilke's () recommendation, now all high-PIP schools have the same color and are identified by direct labeling. Using different colored points could lead to too many different colors to distinct. Adding colors and direct labels then added irrelevant information given that colors now were not needed to distinguish the points.

The associations among PIPs and τ_{t_0} and μ_j are represented by scatterplots, named `funnel` and `outcome` plots, respectively. Although no trends are expected to be seen, this option was chosen to give a quick snapshot of how schools with high PIP are located across the within-cluster SD and the cluster mean spectrum. Similarly, the plot in Figure 1 glances on the distribution of estimated PIPs.

- Funnel plot
  - color
  - shape, linetype
  - direct labeling
  - line for reference
  - overlapping points: alpha, nudge
  - larger labels so text is visible
- Outcome plot
  - proportions? (dot size)
  -
- Sugarloaf
  - point estimates
  - uncertainty
  - direct labeling
  - color