Using MELSM to Identify Consistent and Inconsistent Schools

Marwin Carmo

2025-03-20

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. The final goal is to enhance the visualizations currently provided by the `ivd` package.

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 `ivd`.

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:

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

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

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 the eight schools with PIP > 0.75. Panel A shows the posterior distribution of t_{0j}

and in panel B, the posterior of σ_j. The pattern is similar given that σ_j is a function of t_{0j}. Notably, the estimates of schools 46 and 39, respectively the more consistent and inconsistent schools in the sample, "escaped'' the pull of the spike. For other schools, as the PIP of δ_{0j} got closer to 0.75, we can observe a significant posterior mass of t_{0j} on the spike, although a notable portion remains in the slab. These are characteristics of moderate evidence favoring the inclusion of the random effect.

Visualization

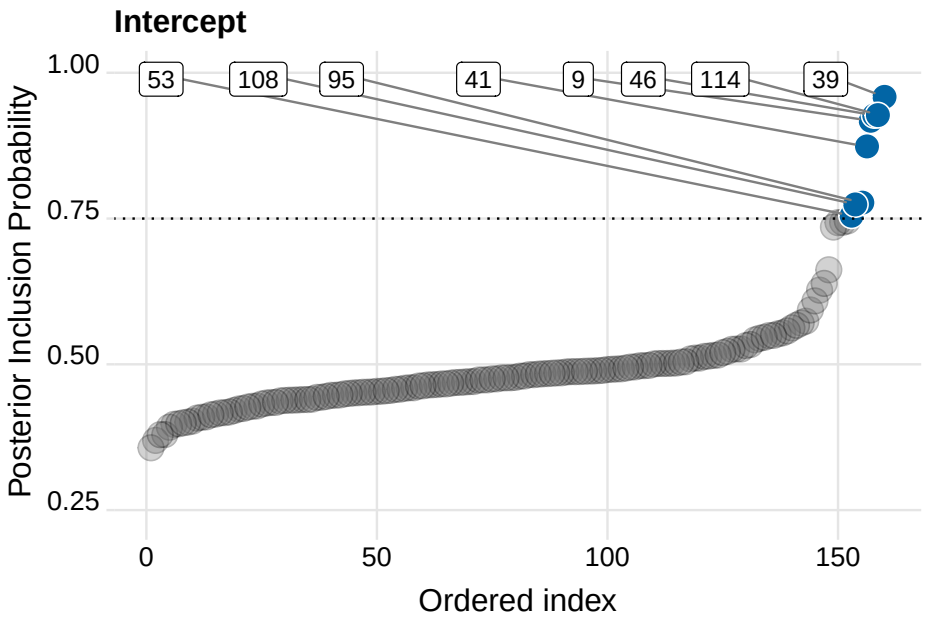

Figure 1: The posterior inclusion probabilities of the scale random effect for the 160 schools in the dataset for Models 1 and 2. A dotted line marks the PIP threshold of 0.75. Schools with PIPs exceeding this threshold are colored and labeled.

Visualization strategies

In this project, I created five ways to visualize the posterior distribution of estimates generated by the `ivd` model. These visualizations were designed to complement the model's summary by focusing on schools with posterior inclusion probabilities (PIPs) exceeding a proposed threshold of "significance." It is important to note that these plots were designed to be displayed individually

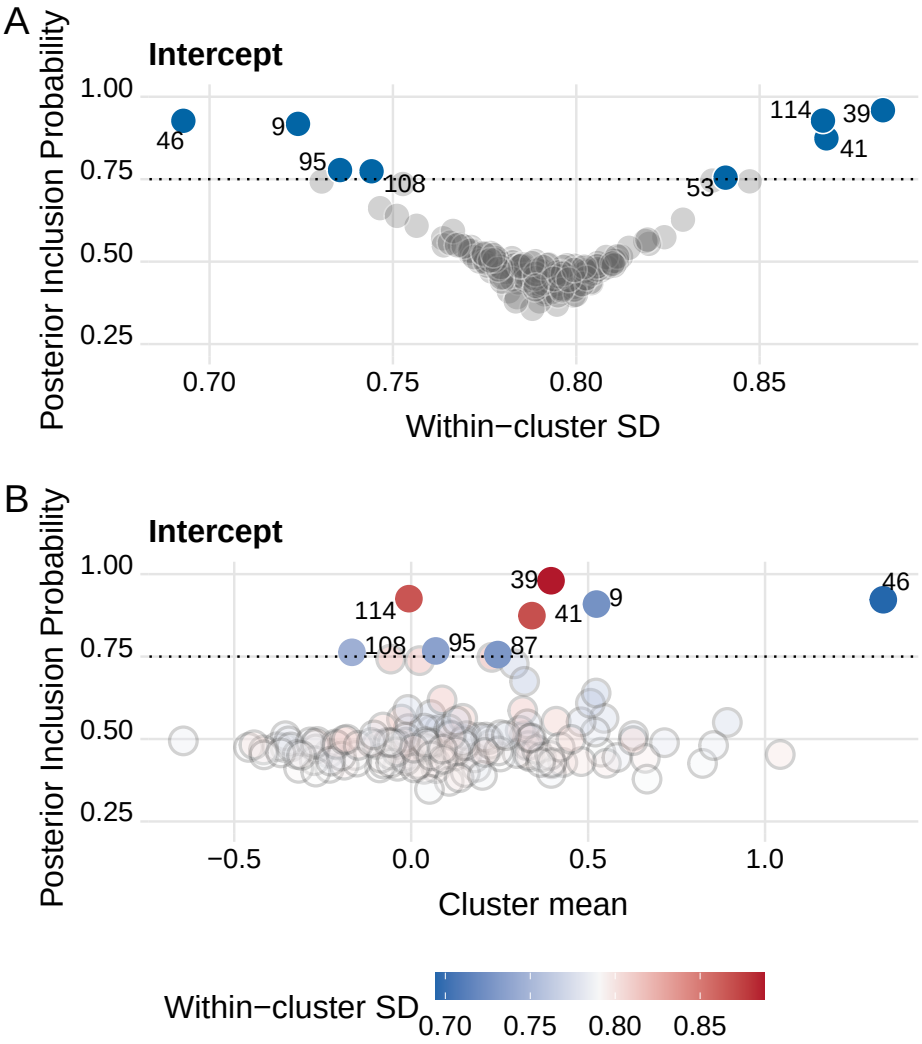

Figure 2: Scatter plots of posterior inclusion probability (PIP) versus within-cluster standard deviation (panel A) and math achievement (panel B). The y-axis represents the PIP for the scale random intercept. In panel A, the x-axis represents the estimated within-cluster SD, and in panel B, the estimated math achievement. In panel A, the plot exhibit a V-shaped pattern, where schools with the lowest and highest SDs are positioned toward the left and right, respectively. In panel B, the color shade represent the school's SD.

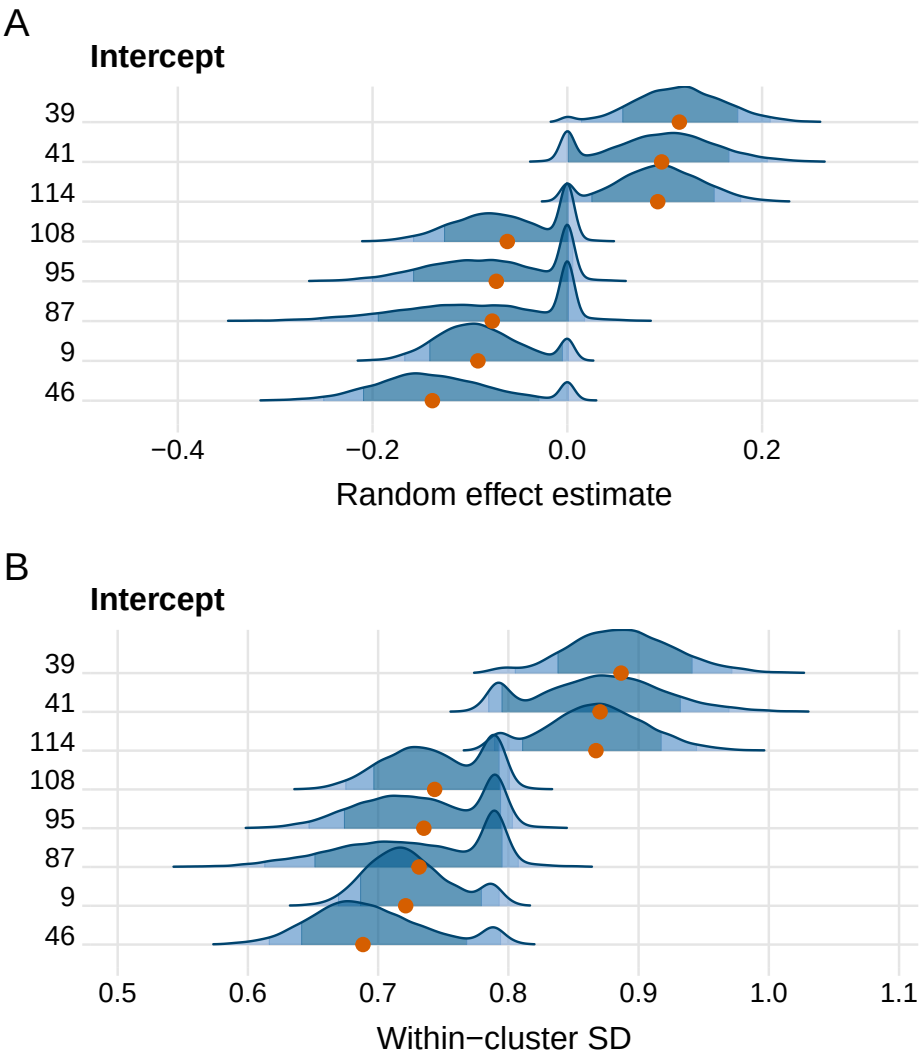

Figure 3: Density plots of the posterior estimates of the scale random effect (panel A), and cluster SD for schools with PIP for the inclusion of the scale random effect larger than 0.75. Schools with lower PIP show more posterior mass of t_0 on the spike (panel A) and, consequently, more posterior mass on the average residual standard deviation. In both panels, color shading represent the limits of posterior probabilities.

as outputs of the `plot()` function in `ivd`. The combination of two plots in a panel, as demonstrated by Figure 2 and Figure 3, is solely for improved organization within this document. Additionally, the titles of the plots reflect only the specific random effect selected by the user

For all plots, color was used as a primary channel to highlight and distinguish schools of interest. In Figure 1 and both panels of Figure 2, schools with high PIPs were colored to emphasize their relevance, while those with "average" variability were shaded in gray to de-emphasize them. This approach aligns with the pre-attentive processing principle (Kazakova 2021; Wilke 2019), where color draws immediate attention to the most important elements of the visualization. Initially, each high-PIP school was assigned a unique color to distinguish them. However, this led to visual clutter due to the large number of high-PIP schools, resulting in too many colors that were difficult to differentiate. To address this, I simplified the design by using a single color for all high-PIP schools and identifying them through direct labeling. This change adheres to the principle of reducing non-data ink (Franconeri et al. 2021), as the colors no longer served a distinct purpose and instead added unnecessary complexity.

In panel B of Figure 2, I employed a sequential color scale to encode the within-cluster standard deviation of schools. This choice was made to better represent the continuous nature of the data compared to an earlier version where size was used. While size can effectively encode quantitative differences, it proved challenging to distinguish between similar sizes and to create a clear legend for interpretation. Additionally, smaller points representing highly consistent schools (e.g., school 46) were nearly invisible. Switching to a sequential color scale improved the clarity of the visualization.

The scatterplots in Figure 2 show the associations between PIPs and τ_{t_0} and μ_j. While no specific trends were expected, this format was chosen to provide a quick overview of how high-PIP schools are distributed across the spectrum of within-cluster standard deviations and cluster means. Similarly, the plot in Figure 1 provides a snapshot of the distribution of estimated PIPs, allowing quick assessment of the concentration and spread of high-PIP schools.

In Figure 3, I used ridgeline plots to visualize the uncertainty of point estimates. Ridgeline plots are particularly effective for comparing multiple distributions, as they allow for easy visual comparison of the posterior distributions of residual standard deviations across high-PIP schools. This choice emphasizes that higher uncertainty (lower PIP) corresponds to more posterior mass concentrated around the spike. Additionally, ridgeline plots scale well if the model had identified a greater number of schools with unusual variability. To enhance interpretability, I applied a sequential color scale to distinguish regions of posterior probabilities. The gradient of blue shades indicates higher probability density near the posterior mean and lower density on the tails, leveraging the perceptual effectiveness of color gradients for representing continuous data. Of note, however, is the lack of a proper label for colors due to difficulties in finding an appropriate solution.

References

Franconeri, Steven L, Lace M Padilla, Priti Shah, Jeffrey M Zacks, and Jessica Hullman. 2021. "The Science of Visual Data Communication: What Works." *Psychological Science in the Public Interest* 22 (3): 110–61. https://doi.org/10.1177/15291006211051956.

Kazakova, Elena V. 2021. "The Psychology Behind Data Visualization Techniques." March 31, 2021. https://towardsdatascience.com/the-psychology-behind-data-visualization-techniques-68ef12865720/?gi=5d2882f99276.

Wilke, Claus O. 2019. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media.