When Growth Mixture Models Break:

Identifiability Failures and Misleading Model Evaluation

Presenter: Doria Xiao

Affiliations:

Ph.D., University of California, Berkeley;

Postdoctoral Fellow, Stanford University

Conference Details:

BEAR Seminar

September 9, 2025

BEAR CENTER

Berkeley Evaluation and Assessment Research Center

Outline

What Are GMMs

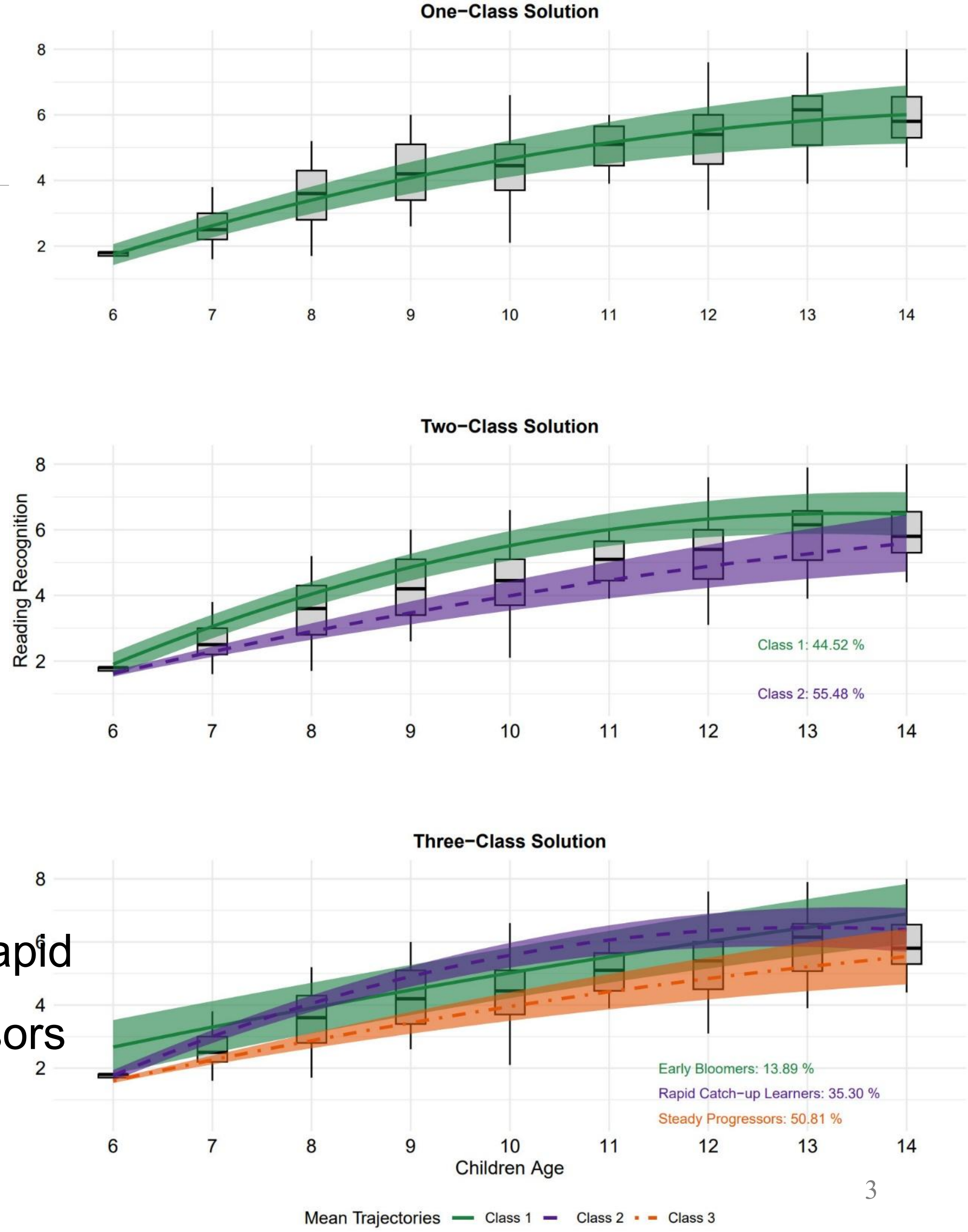
- 1-class: qualitatively homogeneous growth

- 2-class: early vs. late developers

Same data, different interpretations

Which model should we trust?

- 3-class: early bloomers, rapid catch-up, steady progressors



How Each Information Criterion Works

Criterion		Estimator Form Fit term + Penalty term	Penalty/complexity p Optimism due to using data twice
AIC **(Akaike, 1973)**	[1]Stata, Mplus, R/lme4	$-2\log f\big(\, y \mid \hat{\theta}(y) \,\big) + 2p$	Number of parameters p
DIC **(Spiegelhalter et al., 2002)**	[2]Open BUGS, JAGS	$-2\log f\big(\, y \mid \tilde{\theta}(y) \,\big) + 2p_D$	Mean deviance minus plug-in deviance $p_D = \bar{D} - D(\bar{\theta})$
[3]Stan **WAIC** **(Watanabe, 2010)**		$-2\sum_i \log p_{\text{post}}\,(\, y_i \mid y \,) + 2p_{\text{WAIC}}$	Posterior variance of log-likelihood contrib. $p_{\text{WAIC}} = \sum_{i=1}^{N} \text{Var}_{\text{post}}\,[\log f\,(y_i \mid \theta)]$
LOO-CV **(Vehtari et al., 2016)**		$-2\sum_i \log p_{\text{post}}\,(\, y_i \mid y_{-i} \,)$, via PSIS	NA*, leave-one-out reweighting

Plug-in deviance $D(\bar{\theta})$

Posterior predictive

Fit term ≈ log-likelihood or log predictive density

Penalty term ≈ effective number of parameters (adjusts for reuse of data)

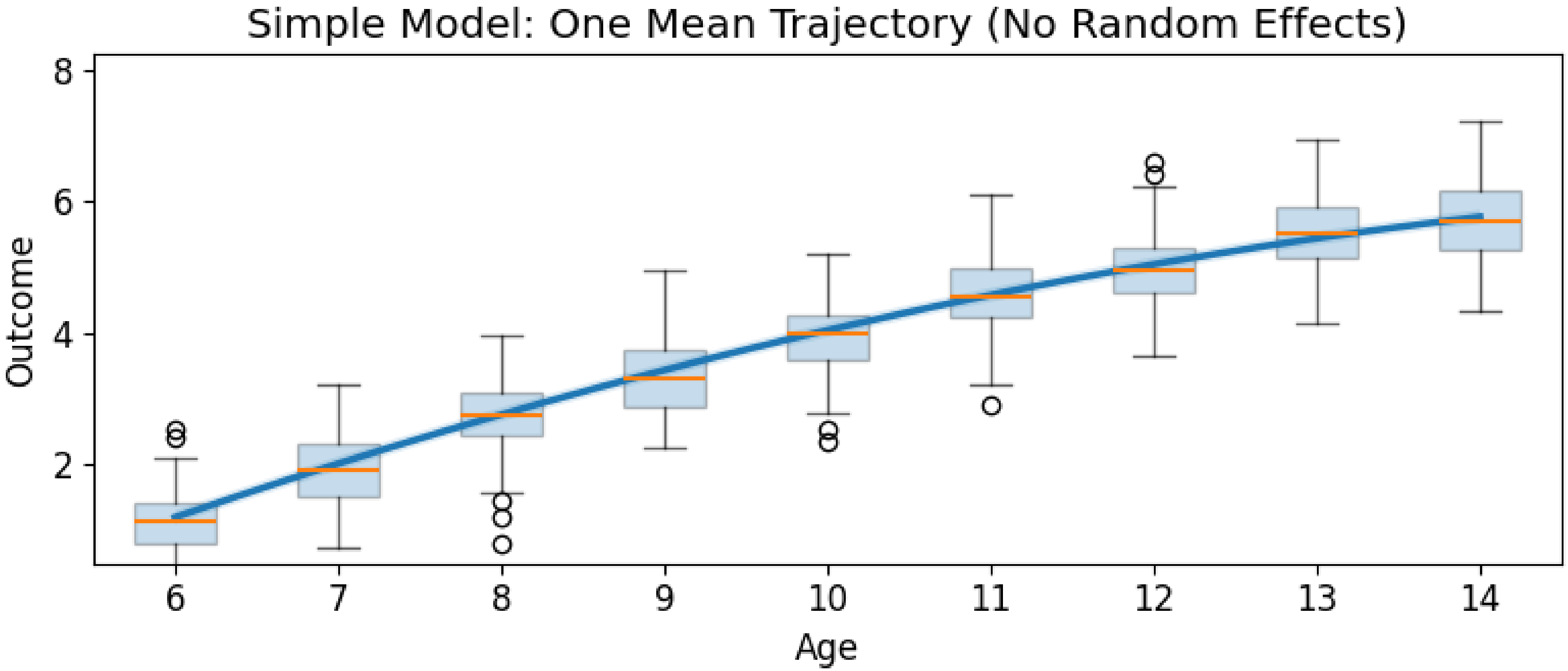
PSIS: Pareto-Smoothed Importance Sampling (Vehtari et al., 2017)

[1]*(StataCorp, 2023; Muthén & Muthén, 1998–2017; Bates et al., 2015)*

[2]*(Surhone et al., 2010; Plummer, 2017)*

[3]*(Stan Dev. Team, 2021)*

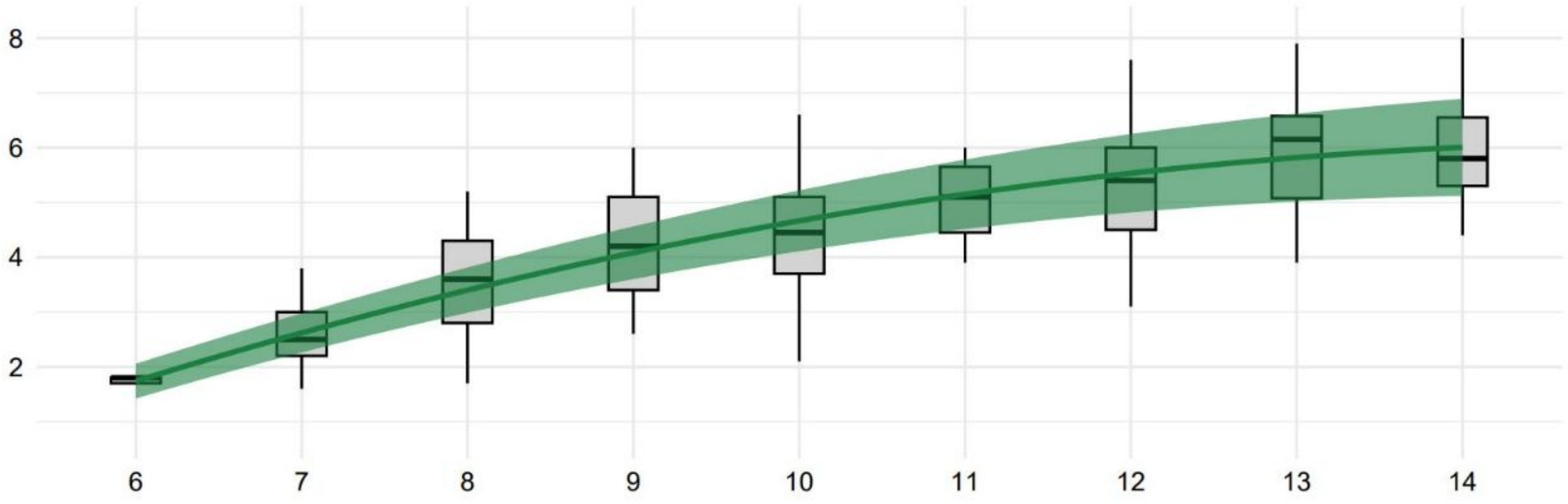
From Simple to Hierarchical to GMM



$$f(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i \mid \boldsymbol{\theta})$$

From Simple to Hierarchical to GMM

Hierarchical Model: One Population Curve with Individual Variation



Conditional Likelihood: $f_c(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{r}) = \prod_{j=1}^{J} f(\boldsymbol{y_j} \mid \boldsymbol{\beta}, \boldsymbol{r_j}) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} f(y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{r_j})$

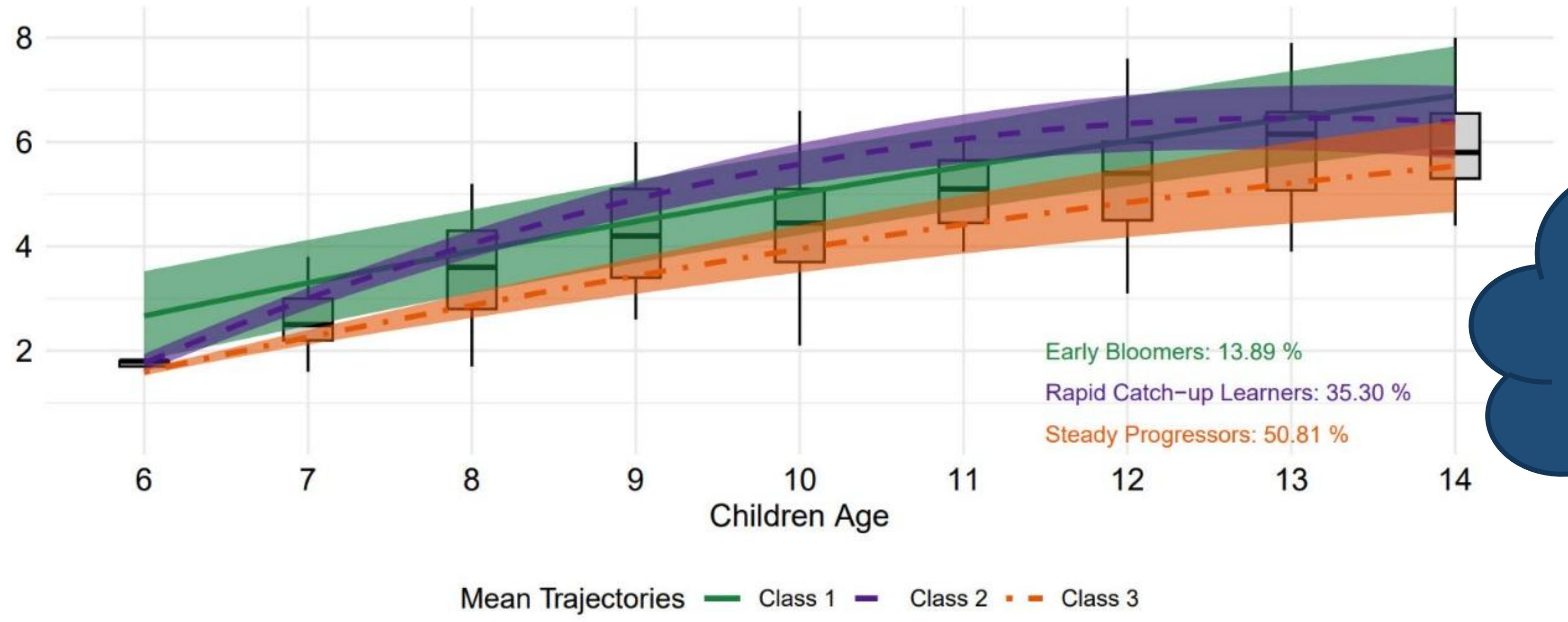
- ➢Conditions on random effects $\boldsymbol{r_j}$ -> Predict at individual level*

Marginal Likelihood: $f_m(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{j=1}^{J} f(\boldsymbol{y_j} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{j=1}^{J} \int f(\boldsymbol{y_j} \mid \boldsymbol{\beta}, \boldsymbol{r_j}) p(\boldsymbol{r_j} \mid \boldsymbol{\Sigma})\, d\boldsymbol{r_j}$

- ➢Integrates over random effects -> Predict at population level*

- Vaida, F., & Blanchard, S. (2005, 06). Conditional Akaike information for mixed-effects models. Biometrika
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P.,&van der Linde, A. (2002). Bayesian measures
of model complexity and fit. Journal of the Royal Statistical Society, Series B,
- Merkle, E., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. Psychometrika

From Simple to Hierarchical to GMM



- What Likelihood Are You Using?
- What Does Your Software Default To?

$$y_{ij} | \boxed{w_j} = k, \boxed{r_{1j}, r_{2j}} \sim N\left(\mu_{ij}^{(k)}, \sigma_e\right);$$

Categorical latent variable (class label) w_j

- Multinomial distribution with probability parameters $\{\lambda^{(1)}, \ldots, \lambda^{(K)}\}$

Continuous latent variables

- $(r_{1j},\ r_{2j})' \,|w_j = k \sim N\,(\mathbf{0}, \mathbf{\Sigma}^{(K)})$ with class-specific covariance matrix $\mathbf{\Sigma}^{(K)}$

Bayesian: The Right Likelihood for GMM for the Right Purpose

Likelihood Type	What Is Integrated or Conditioned?	Prediction Target	Valid for	Common Software
Marginal	Integrates over both latent classes and random effects	Predict outcomes in **new clusters**	Class enumeration	**Stan** (which marginalizes over discrete parameters)
Conditional	Conditions on class memberships and random effects	Predict outcomes **for in-sample clusters**	Model comparison for in-sample clusters	**OpenBUGS**, **JAGS**
Hybrid	Integrates over classes but conditions on random effects	Ambiguous: in-sample clusters, but use prior for class prob.	Theoretically incoherent	Often occurs **by default in Stan** when conditioning on random effects

Bottom line:
Only the **marginal likelihood** aligns with the population-level goal of **class enumeration**.
Conditional usable for some model comparisons, but **hybrid not valid for model comparison**.

Bayesian: Why DIC Breaks in GMMs, How to Fix It

Problem: Traditional DIC

$\mathrm{DIC} = \bar{D} + p_D$, where $p_D = \bar{D} - D(\bar{\theta}) \rightarrow$

$$\mathrm{DIC} = \bar{D} + \bar{D} - D(\bar{\theta}) = \bar{D} + p_D$$

In GMMs:

- Skewed / multimodal posteriors, label switching or degenerate nonidentifiability (Xiao, Rabe-Hesketh, & Skrondal, 2015), leads to poor estimate $\bar{\theta}$
- Plug-in deviance too large
- p_D can be **negative or unstable** (Spiegelhalter et al., 2002; Gelman et al., 2013)

More Stable Alternatives

- DIC_{pV2} *(Gelman , Hwang, and Vehtari, 2014)*:

$$\mathrm{DIC}_{pV2} = D(\bar{\theta}) + p_V$$

  - Variance-based penalty,
  - BUT retains plug-in deviance

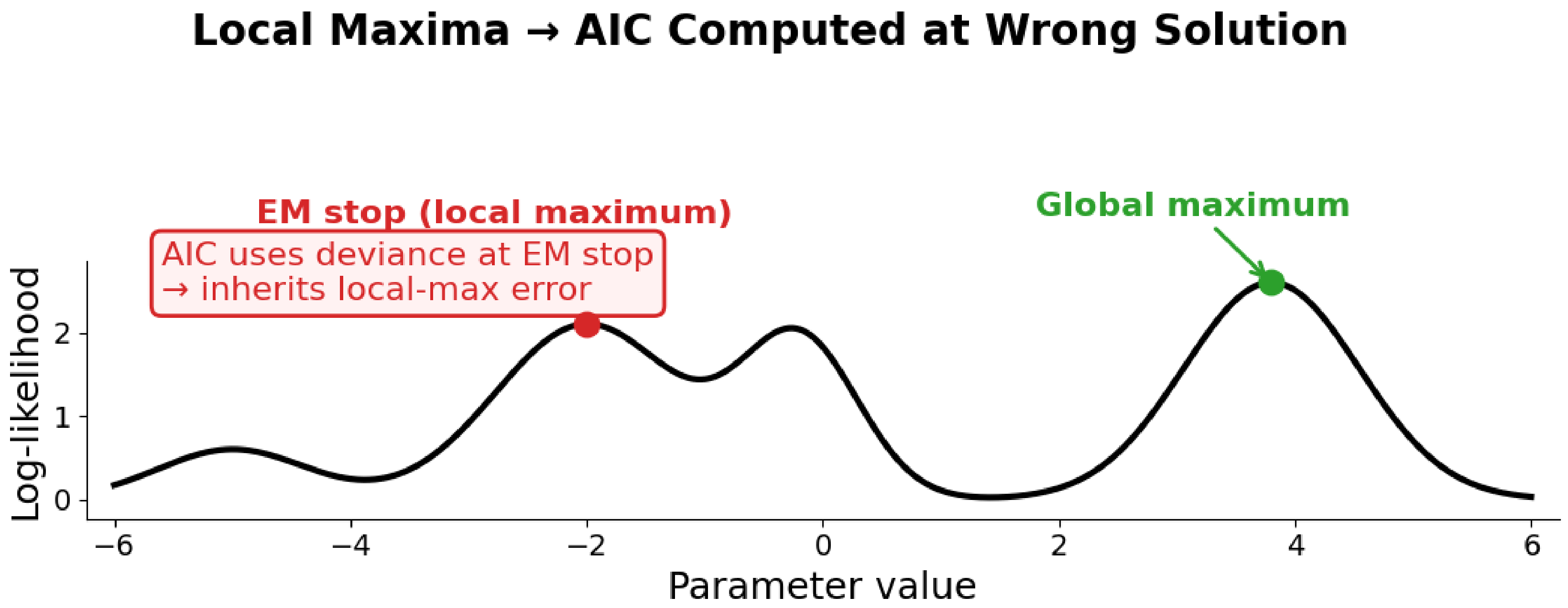
- DIC_{pV} *(we proposed)*:

$$\mathrm{DIC}_{pV} = \bar{D} + p_V$$

  - No plug-in deviance
  - Fully posterior-based

Why AIC Breaks in GMMs, How to Fix It

- Plug-in likelihood depends on EM
- EM can stop at local maxima
- Researchers assume “convergence = global”
- Fix: more iterations, more random starts
- Still no guarantee → trial-and-error



Simulation Conditions & Structure

Purpose

When do model evaluation tools succeed or fail for enumeration?

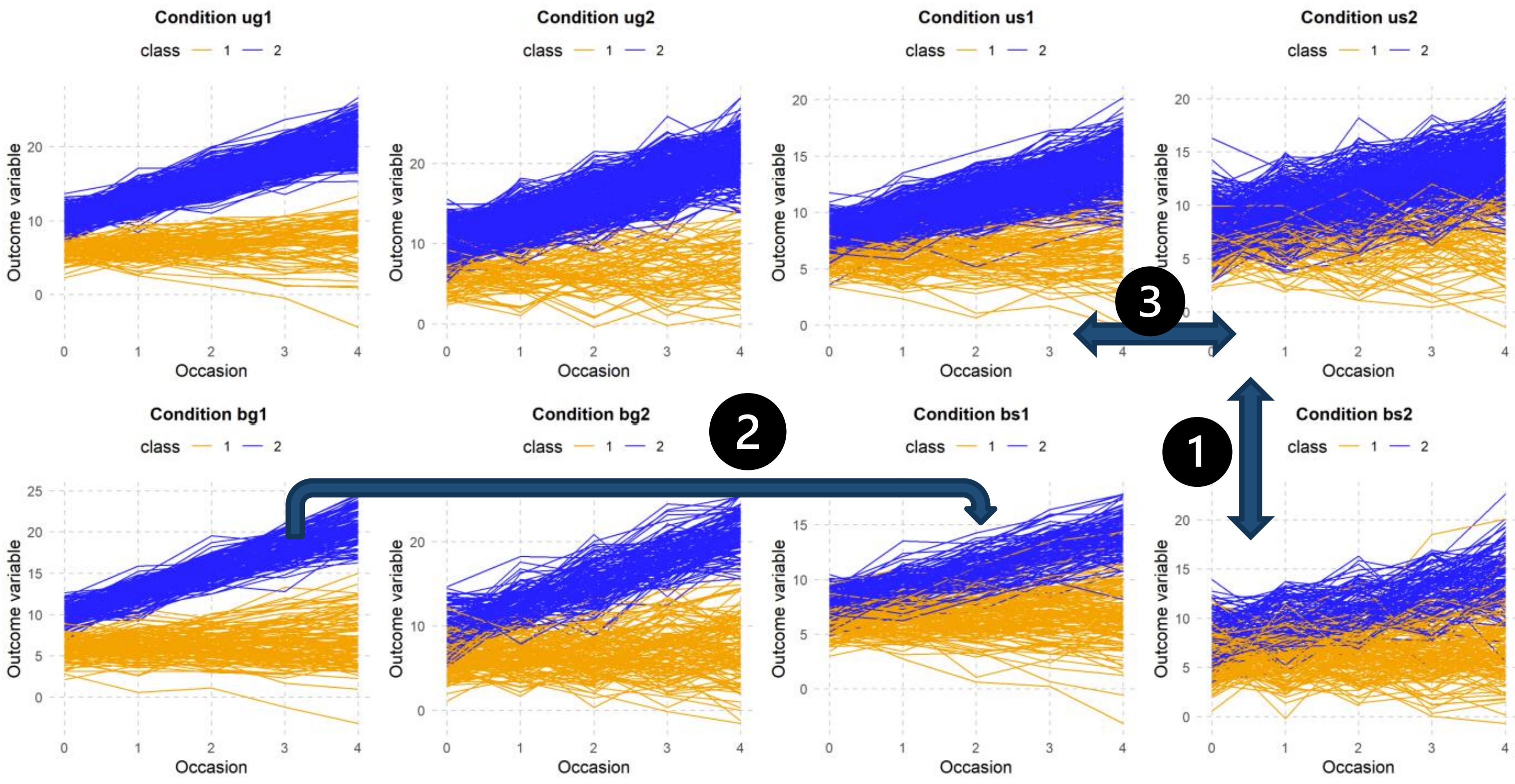


Figure 3.1: Comparison of subject trajectories by class across simulation conditions

Design Factors

Factor	Levels
Class Probabilities and Level-2 Sample Sizes *J*	Balanced ($\lambda = 0.5$) with J = 250 vs. Unbalanced ($\lambda = 0.2$ / 0.8) with J = 400
Class Separation	Strong vs. Weak slope/intercept differences
Residual Variability	Low ($\sigma_e = 1$) vs. High ($\sigma_e = 2$)

Estimation Setup & Fit Criteria

Design Summary
- **Labels**: bg1, ug2, us2, etc. (8 simulation conditions)
- **Replications**: 50 datasets per condition
- **Time Points**: 5 per subject
- **Models Fitted**: 1- to 4-class GMMs
- **Total Fits**: 8 × 50 × 4 = **1,600 per method**

Bayesian Estimation (CmdStan 2.30)
- **MCMC Specs**: 4 chains × 1,000 post-warmup iterations
- **Target**: Marginal likelihood
- **Information Criteria**: DIC, **DIC_pV, DIC_pV2,** WAIC, LOO-CV

Frequentist Estimation (MLE via flexmix)
- **Engine**: R flexmix (Grün & Leisch, 2023)
- **Information Criterion**: AIC

Strengthening EM Estimation

Convergence settings (flexmix, Frequentist)

- **iter.max:** 400 → 5,000

  Prevents premature stopping

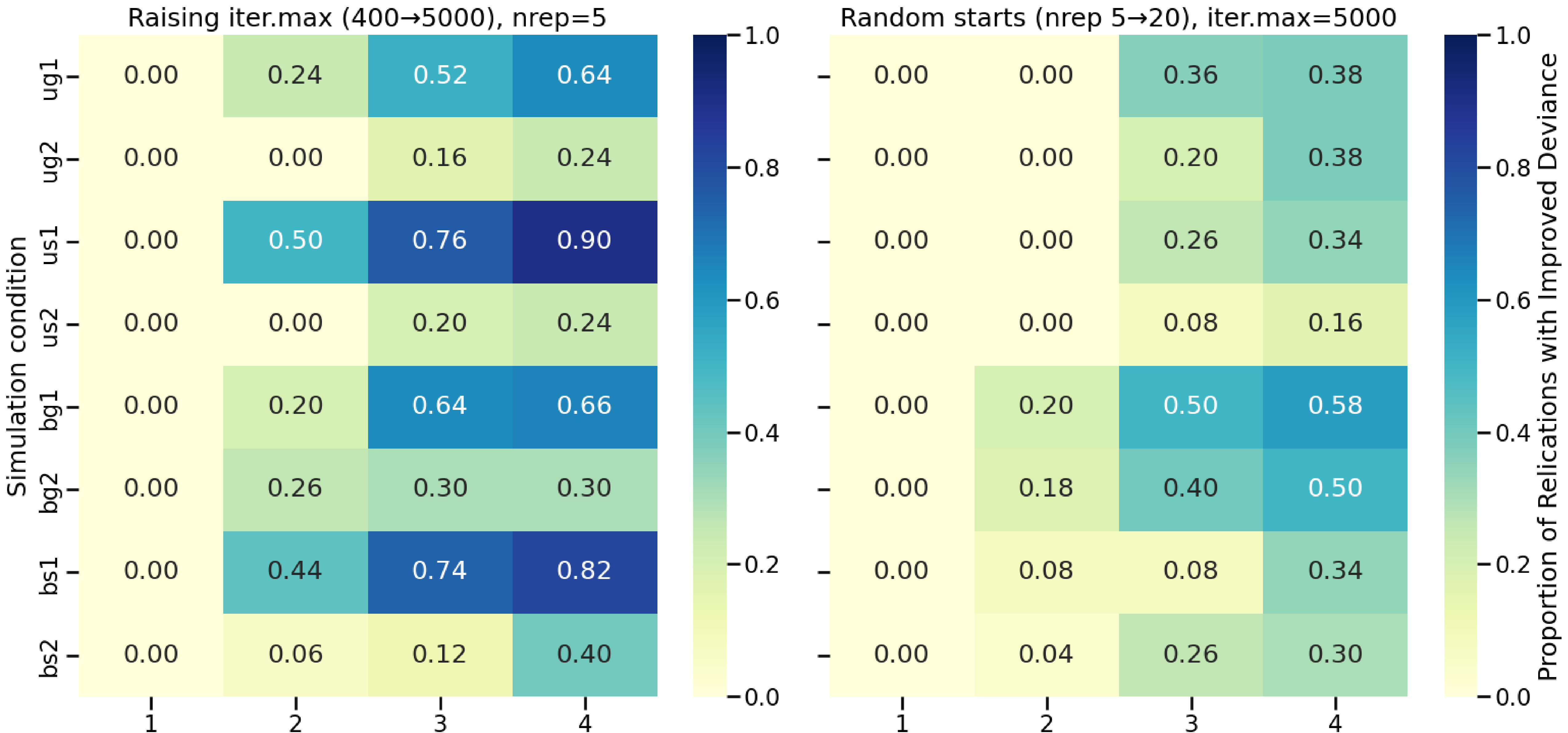
- **nrep:** 5 → 20 (and beyond)

  More chances to escape local maxima

Bottom note:

These fixes improve robustness, but global maximum is not guaranteed.

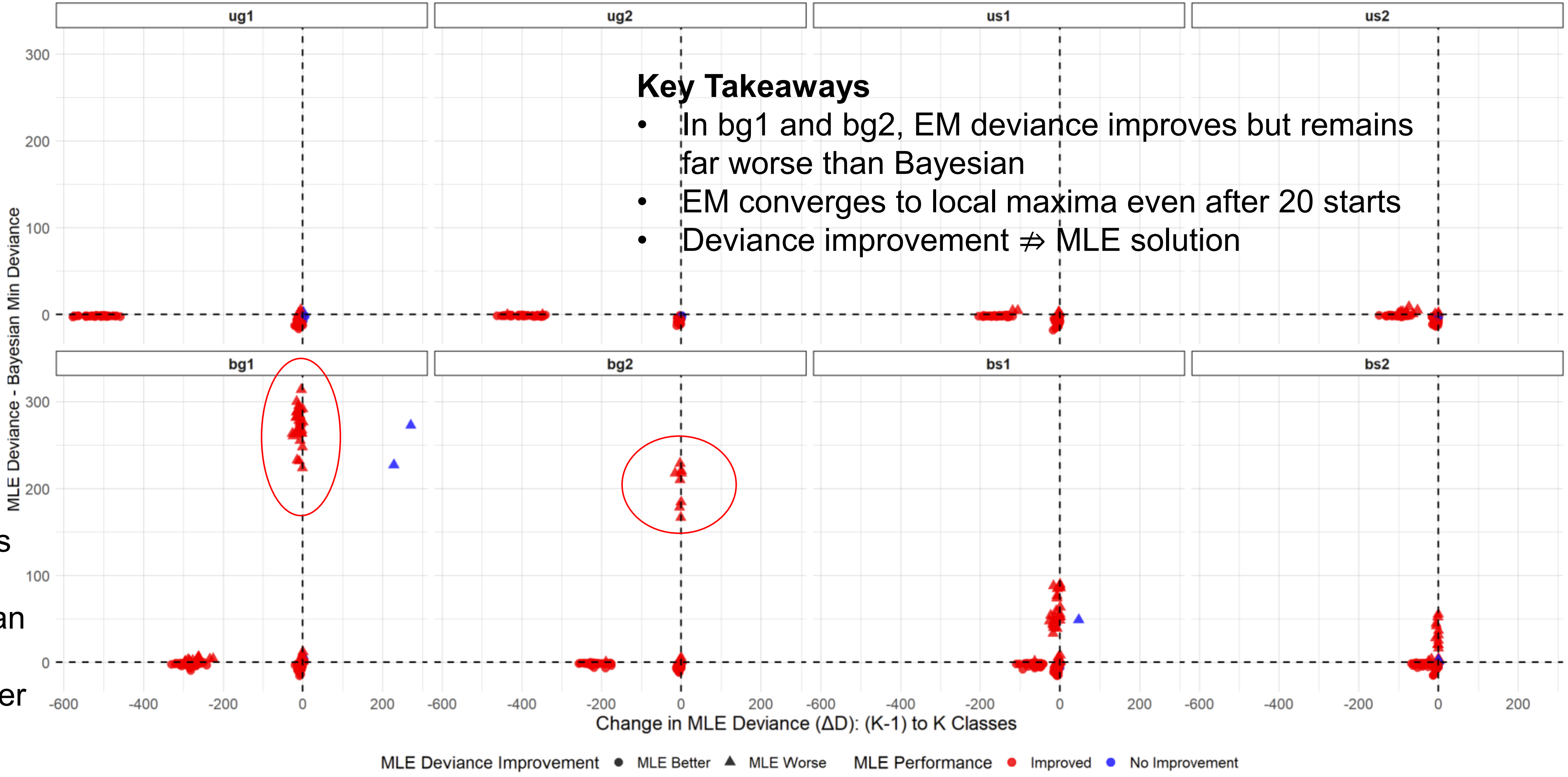
EM Convergence: Iteration Cap and Random Starts



EM Local Maxima, Fit Criteria Fail

Key Takeaways

- In bg1 and bg2, EM deviance improves but remains far worse than Bayesian
- EM converges to local maxima even after 20 starts
- Deviance improvement ⇏ MLE solution

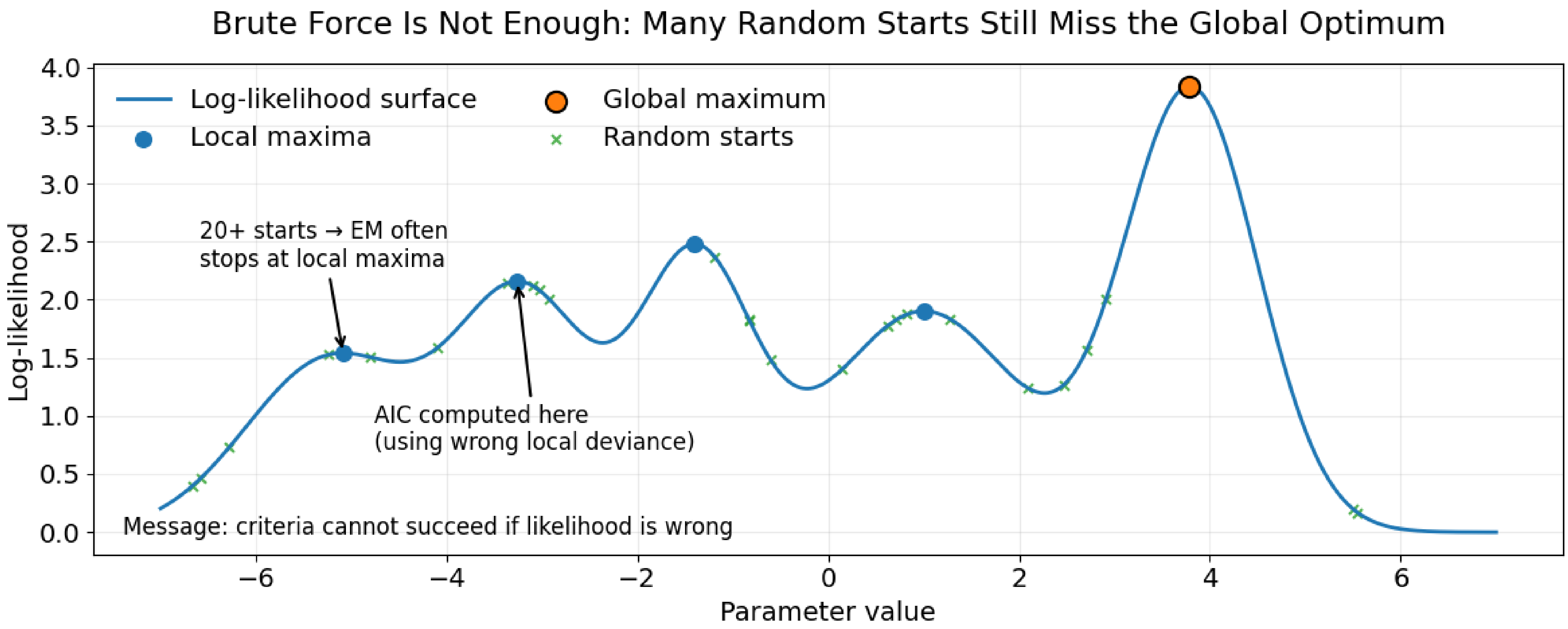


Point: replication
X-axis: the deviance change from add class
- Red: improvement

Y-axis: MLE – Bayesian minimum deviance
- Triangle: MLW worser

Figure 3.7: Comparison of MLE and Bayesian deviance across simulation conditions (based on `nrep = 20`).

Brute Force is Not Enough



False confidence, wrong model.

Bayesian Estimation: Why DIC Variants Work When DIC Fails

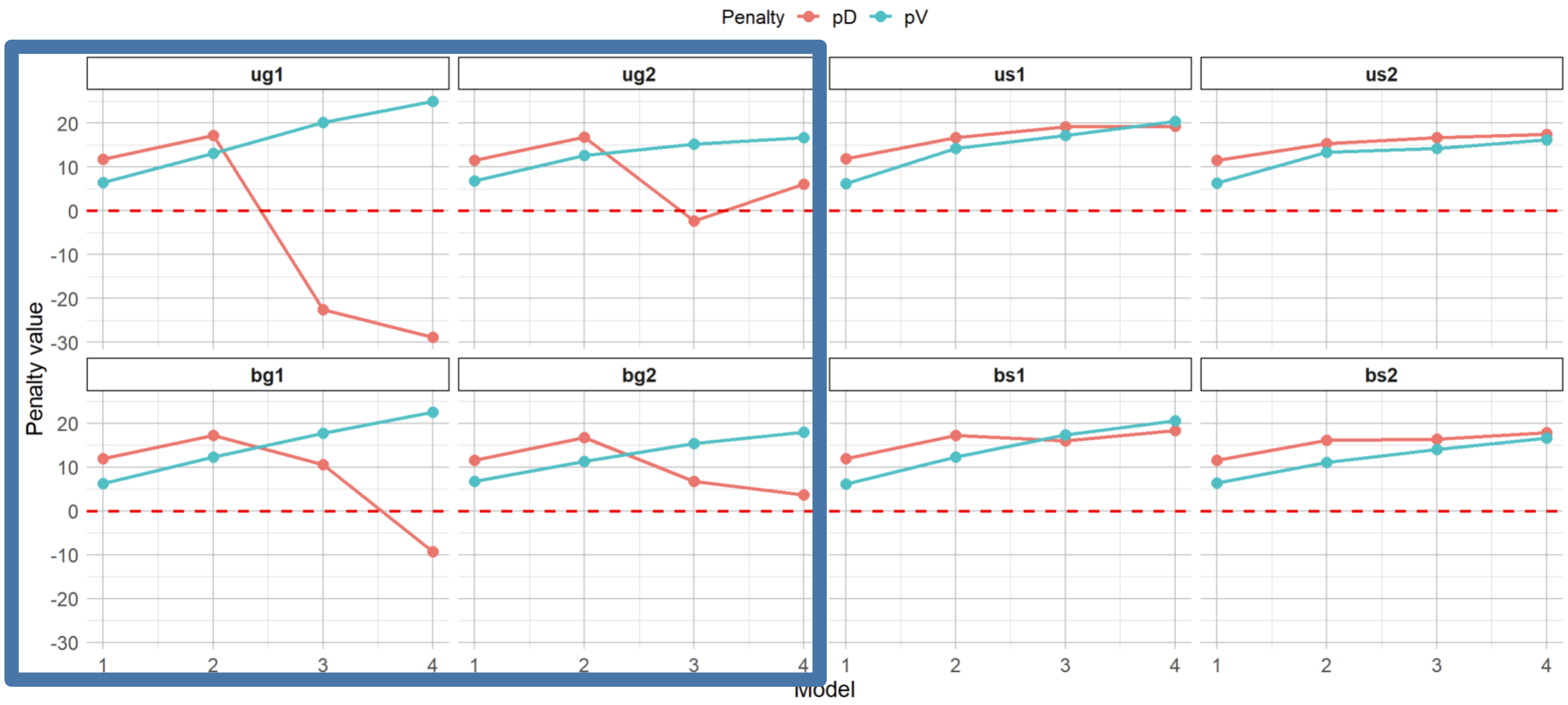
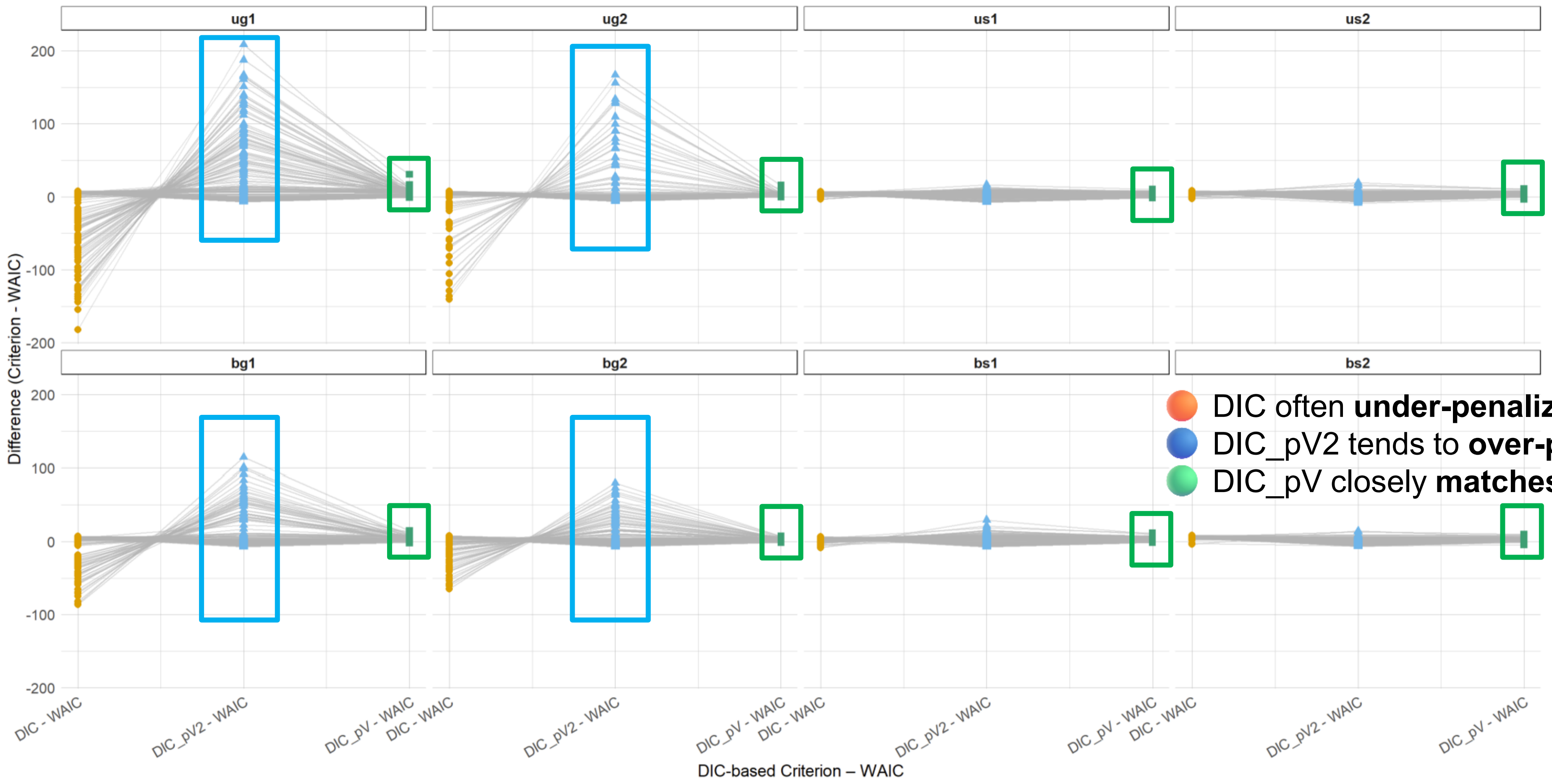


Figure 3.5: Comparison of average DIC penalty terms: Original (p_D) vs. variance-based (p_V) for each simulation Condition.

Traditional DIC penalty (pD) can be negative or unstable
Variance-based penalty (pV) always **positive and stable** across all conditions

Among DIC Variants, DIC_pV Best Aligns with WAIC



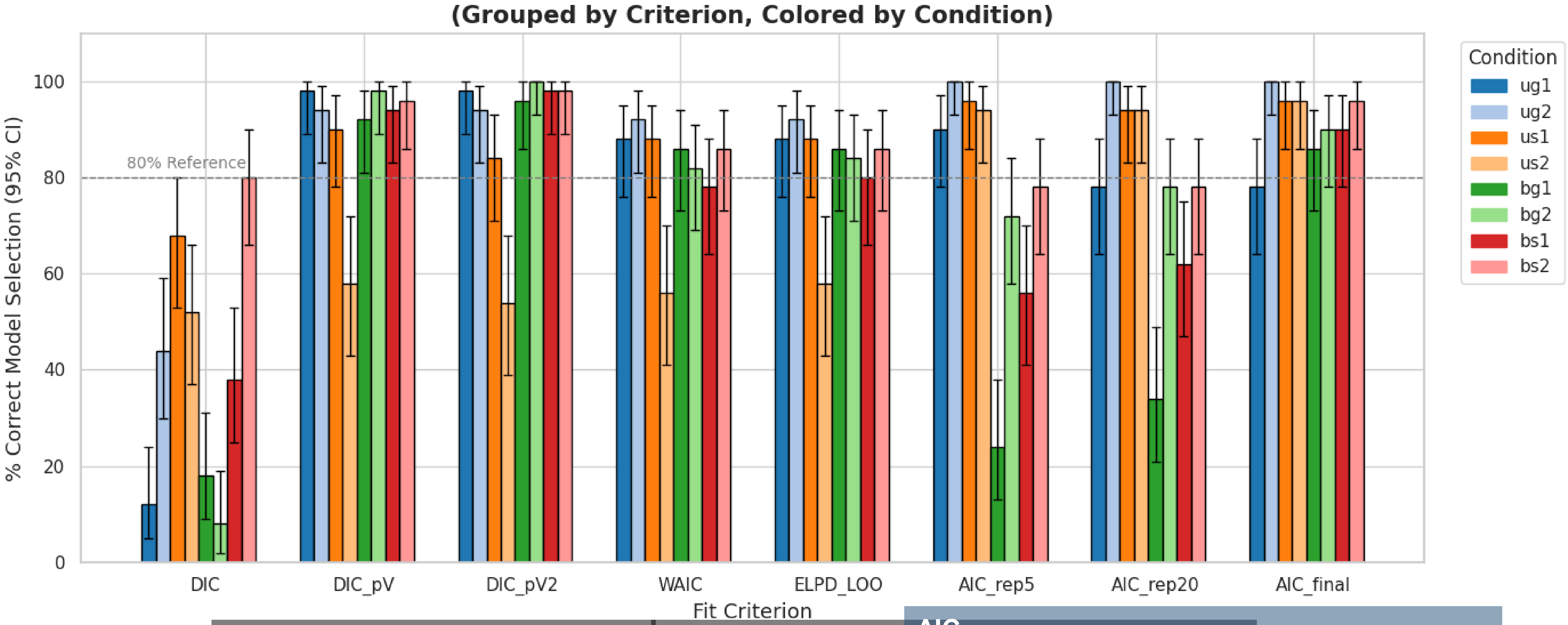
DIC often **under-penalizes** or yields **negative values**

DIC_pV2 tends to **over-penalize**

DIC_pV closely **matches WAIC** with **low variability**

Figure 3.6: Replicate-level comparison of DIC criteria with WAIC for each simulation condition. Each replicate is represented as a path across DIC – WAIC (orange), DIC_{pV2} – WAIC (blue), and DIC_{pV} – WAIC (green).

Information Criteria Performance Across Simulation Conditions



DIC often fails in challenging conditions

- Plug-in devia

Variance-based DIC_pV1/pV2

- High accuracy (typically >90%), except in us2

WAIC & LOO-CV

- Robust across all conditions
- Fully Bayesian, no plug-in approximation

AIC

- Highly sensitive to number of starts (nrep)
- Rep5: severe underperformance
- Rep20: improved, but still unstable

Beyond Selection: When DIC Penalty Reveals Failures

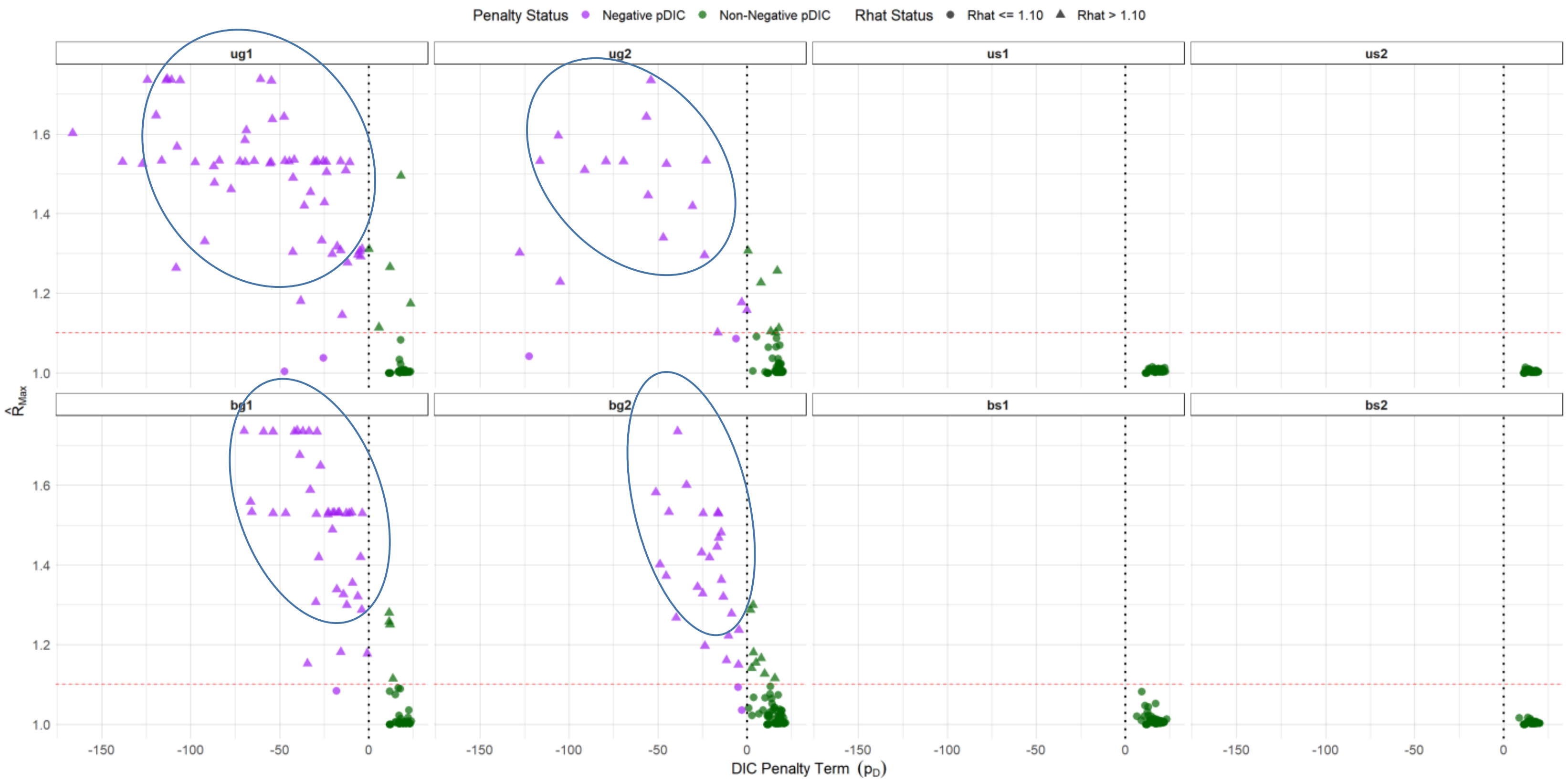


Figure 3.2: Relationship between DIC penalty term (p_D) and convergence diagnostic ($\widehat{R}_{\text{Max}}$) across simulation conditions.

Key Takeaways

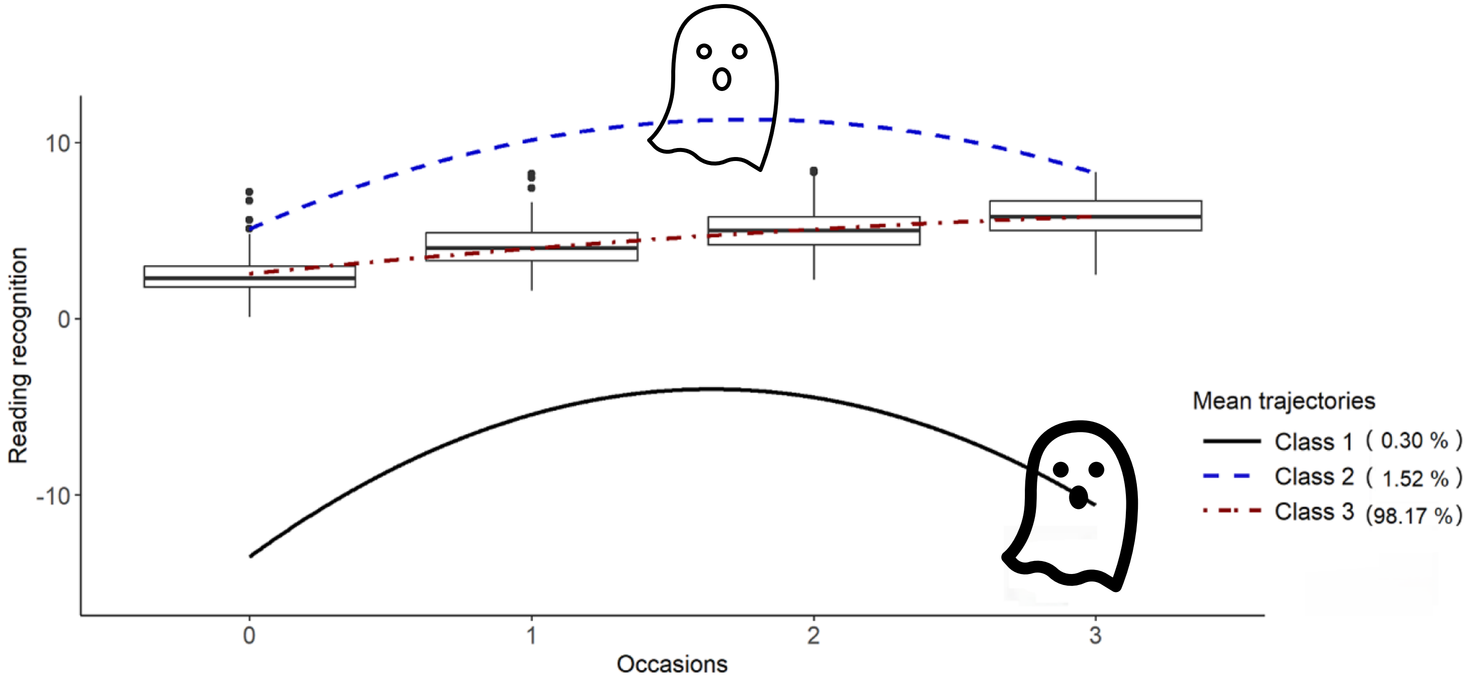
- Negative p_D values **strongly associated with poor convergence** (purple triangles)
- Such failures often stems from **non-identifiability** (Xiao, X., Rabe-Hesketh, S., & Skrondal, A. (2025). Bayesian Identification and Estimation of Growth Mixture Models. *Psychometrika*).

- Panels: simulation conditions
- X-axis: p_D
- Y-axis: the maximum convergence diagnostic $\hat{R}_{\text{Max}}$

- Red dashed line at 1.10: the threshold for acceptable convergence
- Black dotted line at 0: the boundary of p_D
- Point shape: convergence status (circle/triangle)
- Point color: whether p_D is negative/non-negative (purple/green)

Consequences of Minuscule-Class Behavior

- Class weight collapses toward zero ("ghost" class)
- Parameters drift from prior, not learned from data
- Subgroup trajectory becomes incompatible with observed data

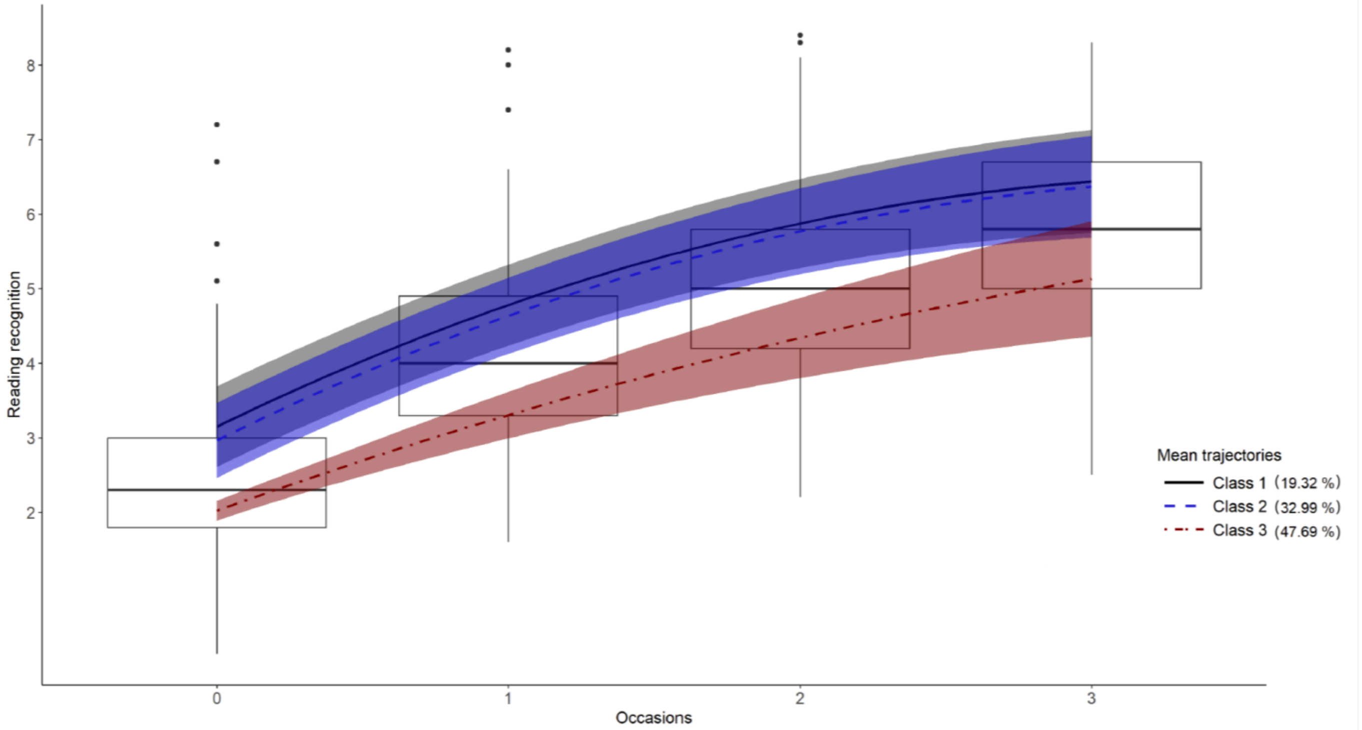


Two classes fit nonsensical curves (tiny class weights).
Highlight "phantom" class trajectories with dotted lines.

Demo: *https://DoriaXiao.github.io/BayesianIdentification/Demo.html*

Consequences of Twinlike-Class Behavior

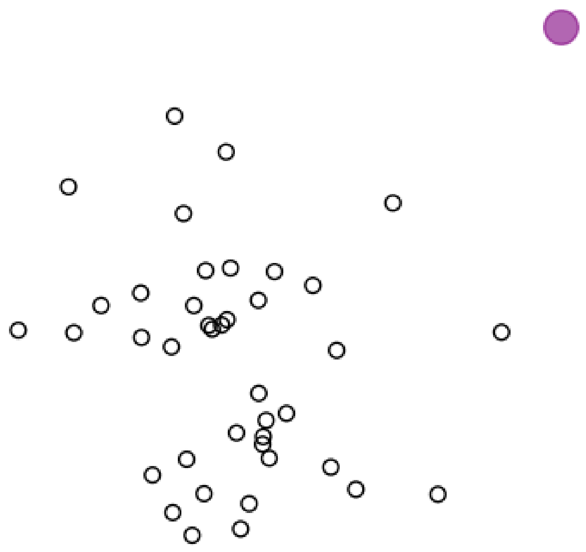
- One subgroup split into two nearly identical classes
- Creates illusion of distinct learner types
- Leads to overestimating heterogeneity



Same subgroup split in two → inflates # of learner types.

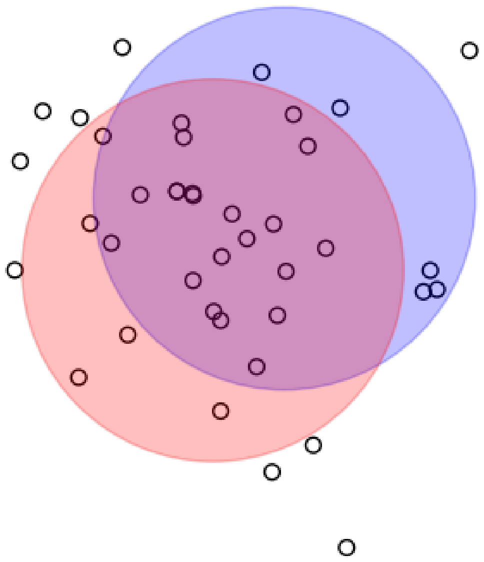
What Nonidentifiability Looks Like in GMMs

Minuscule-class behavior



A phantom tiny class with weight → 0
parameters borrowed from the prior

Twinlike-class behavior



One real group → model invents two
overlapping, indistinguishable classes

- Standard convergence ($\hat{R}$) may not detect these
- **Distinguishability Index (DI):** near zero flags collapsed or indistinguishable classes

Negative p_D as a Diagnostic

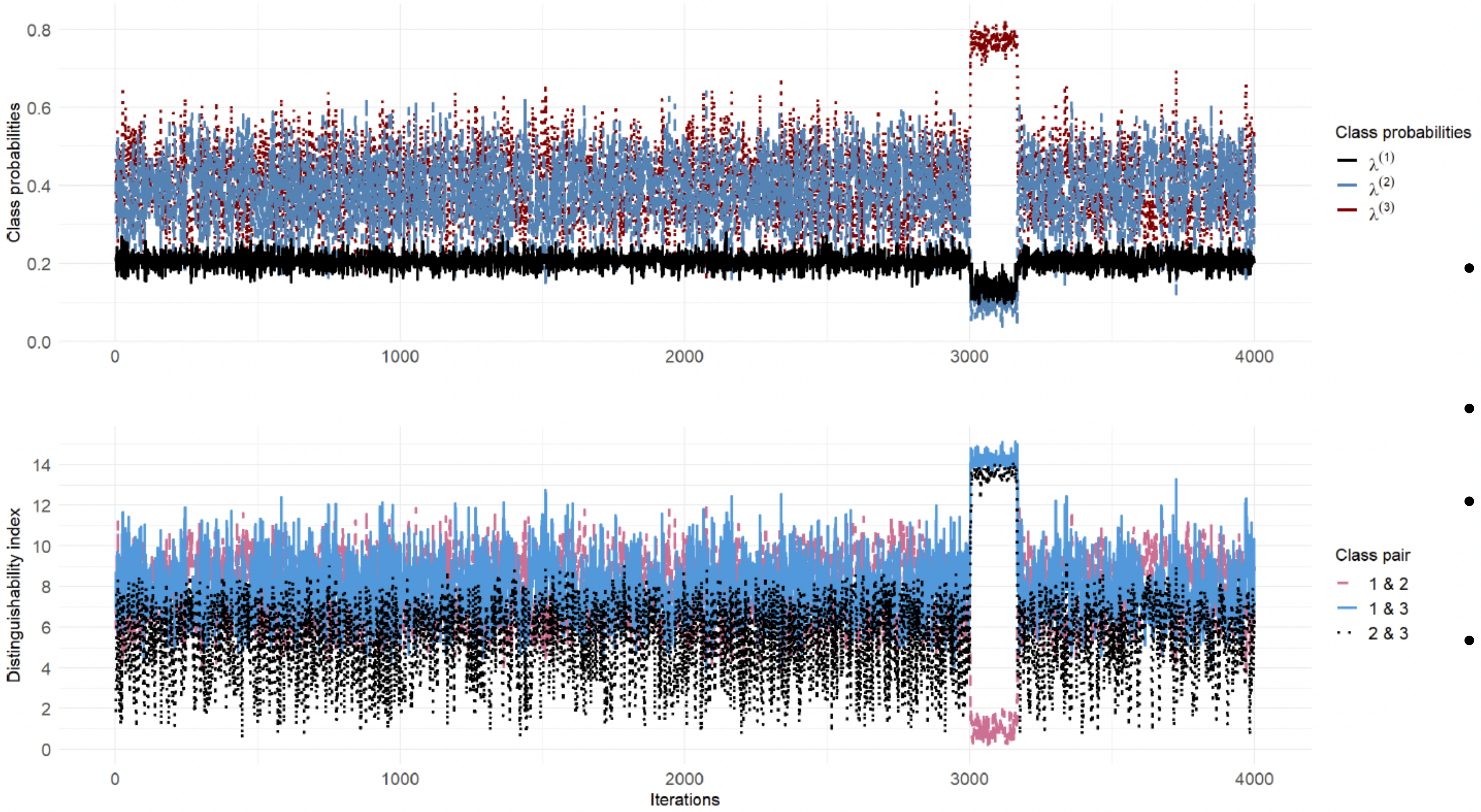


Figure 3.3: Traceplots of class probabilities and distinguishability index (DI) for Condition ug2, replicate 21, 3-class model.

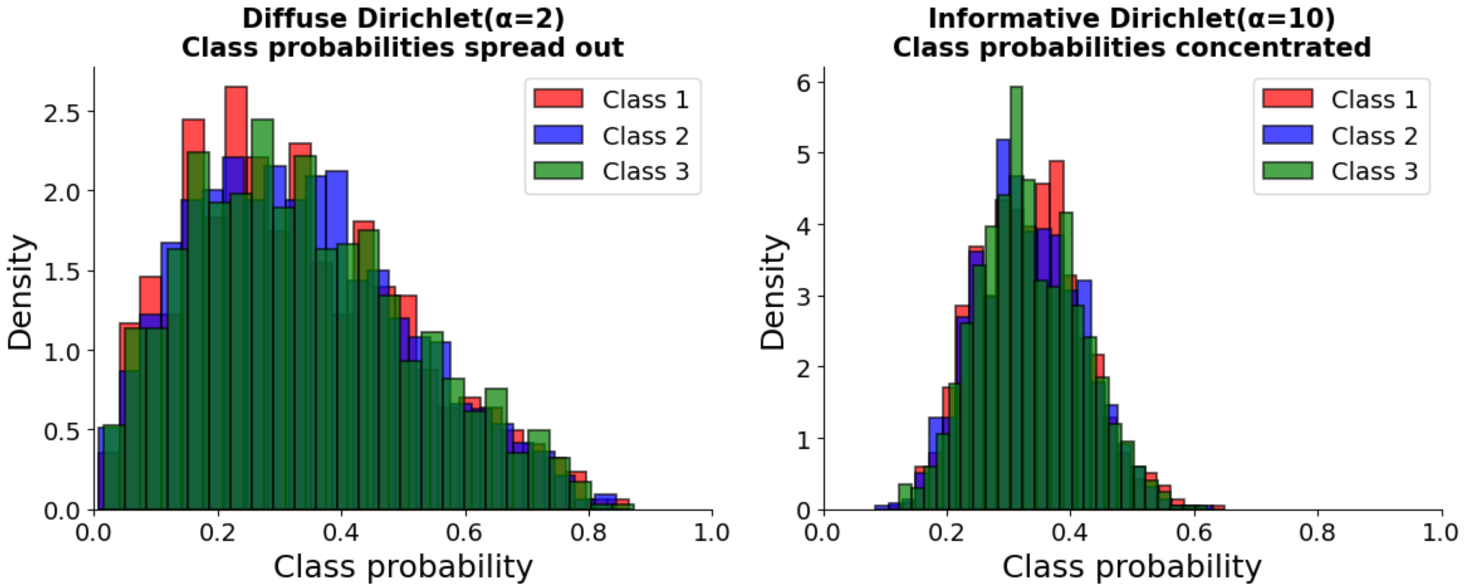
- Example: 3-class model with acceptable $\hat{R}_{\text{Max}}$ = 1.04
- But p_D strongly negative −122
- Chain switches between degenerate solutions, undermining stability
- Distinguishability Index shows classes collapsing

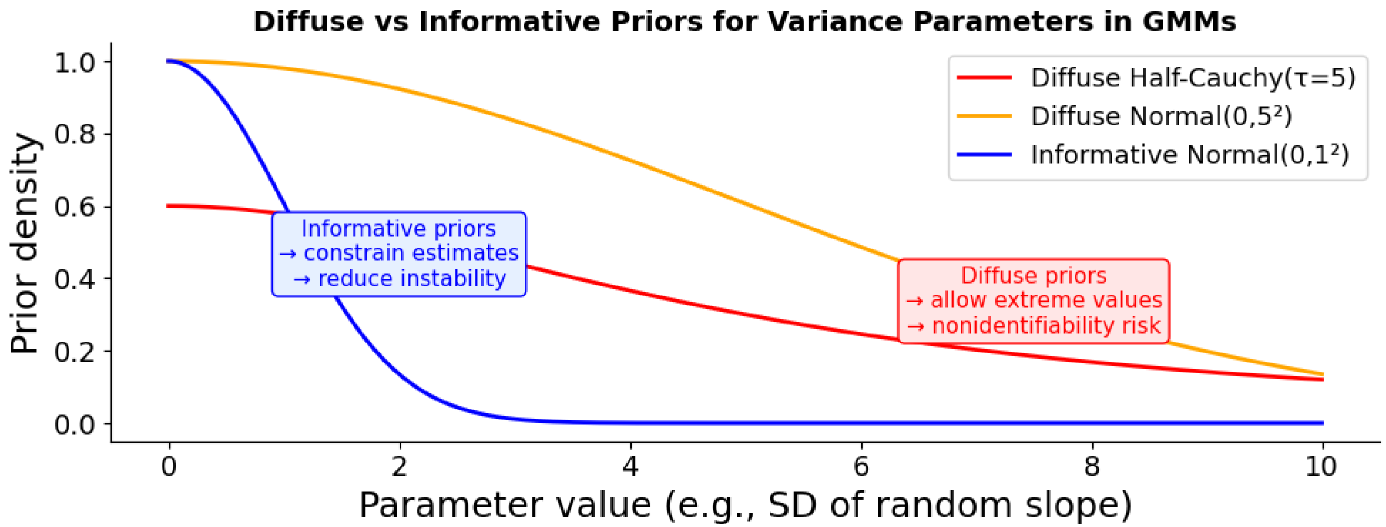
From Weak Priors to Nonidentifiability

- Priors too weak → model can't separate classes.

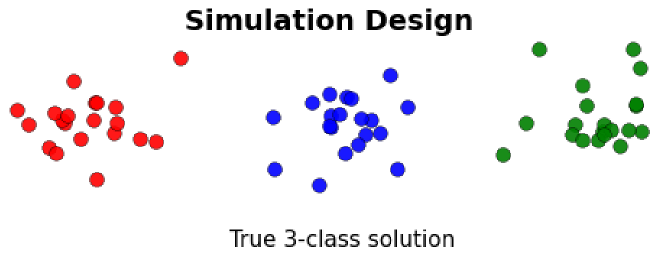
Example defaults

- Class probs: Dirichlet(α = 2)
- SDs: half-Cauchy(τ) or half-Normal(τ)



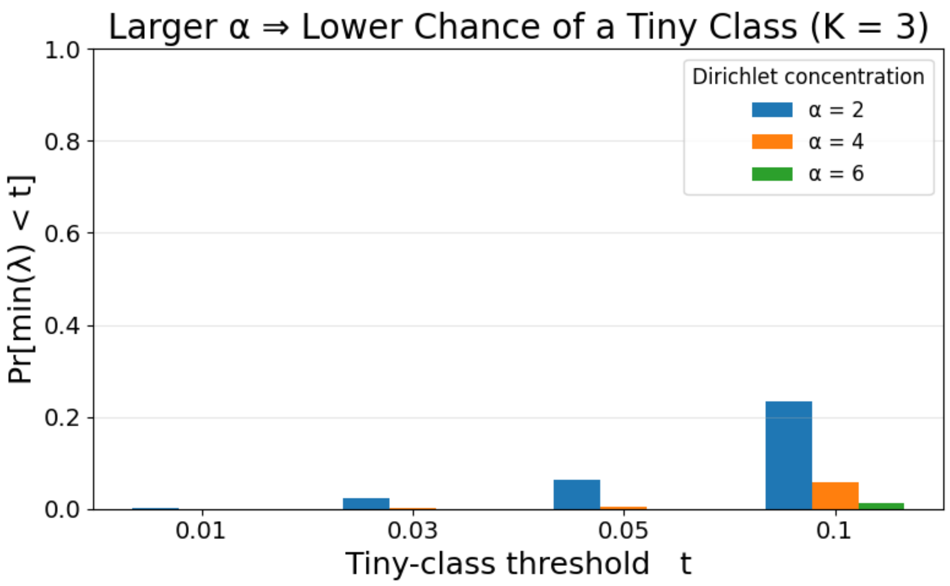


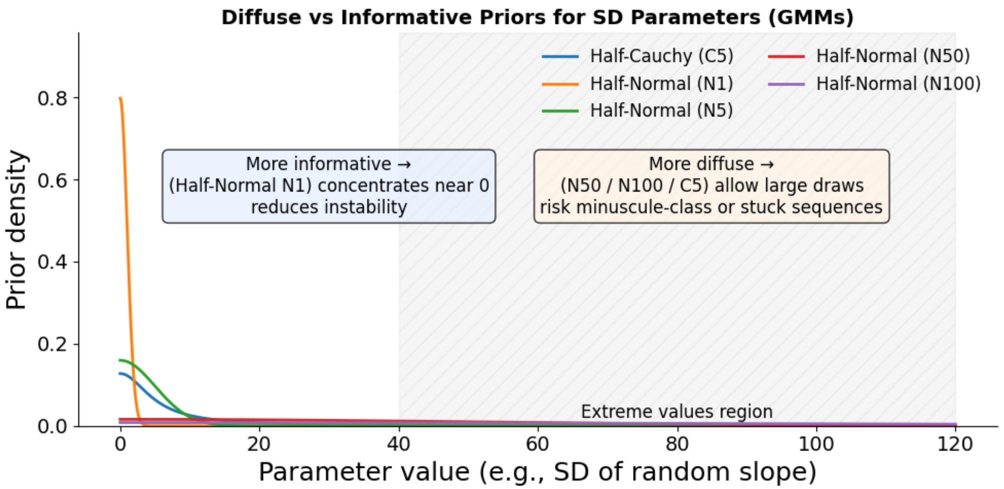
Simulation Highlight: Testing Priors for Stability



200 chains × 1,000 iterations

Simulated responses for a real data using a well-behaved 3-class solution.





Findings:

- Vague priors → more stuck chains and minuscule-class behavior
- Informative priors → fewer failures, more stable estimation

Practical Prior Choices

Prefer half-Normal for SDs over half-Cauchy

Choose α to reflect plausible class balance

Use weakly informative fixed-effect priors with realistic scales

Tune priors in response to diagnostics

Minimal Workflow You Can Use Tomorrow

Fit with reasonable priors → Check: $\hat{R}$, negative p_D, DI → If unstable, tighten priors slightly and refit → Report sensitivity to prior settings

Model-Selection Checklist

- Use marginal-likelihood-based criteria: WAIC, LOO-CV, or DIC_pV
- Avoid plug-in DIC
- For MLE: check/ensure convergence to global maximum carefully.
- Negative DIC penalties = diagnostic red flag

Seatbelts on, chairs filled, thermostat set, stethoscope ready.

https://doriaxiao.github.io/

SCAN ME

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016.

Grün, B., & Leisch, F. (2023). *flexmis: Flexible mixture modeling*. https:// CRAN.R-project.org/package=flexmix (R package version 2.3-19)

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*(1), 302–317. https://doi.org/10.1111/j.1751-9004.2007.00054.x

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*(3), 802–829. https://doi.org/10.1007/s11336-019-09679-0

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus user's guide. In *www.statmodel.com* (8th ed.). Muthén & Muthén.

Plummer, M. (2017). *JAGS version 4.3.0 User Manual*. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/Manuals

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stan Development Team. (2021). *CmdStan User's Guide: Version 2.30*. https://mc-stan.org/docs/cmdstan-guide/index.html

StataCorp. (2023). *Stata statistical software: Release 18*. College Station: StataCorp LLC.

Surhone, L. M., Tennoe, M. T., & Henssonow, S. F. (2010). *OpenBUGS*. Betascript Publishing.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*(2), 351–370. https://doi.org/10.1093/biomet/92.2.351

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27, 1413–1432.

Wardenaar, K. J. (2020). Latent class growth analysis and growth mixture modeling using R: A tutorial for two R-packages and a comparison with mplus. PsyArXiv. https://doi.org/10.31234/osf.io/m58wx

Xiao, X., Rabe-Hesketh, S., & Skrondal, A. (2025). Bayesian Identification and Estimation of Growth Mixture Models. *Psychometrika*, 1–34. doi:10.1017/psy.2025.11