

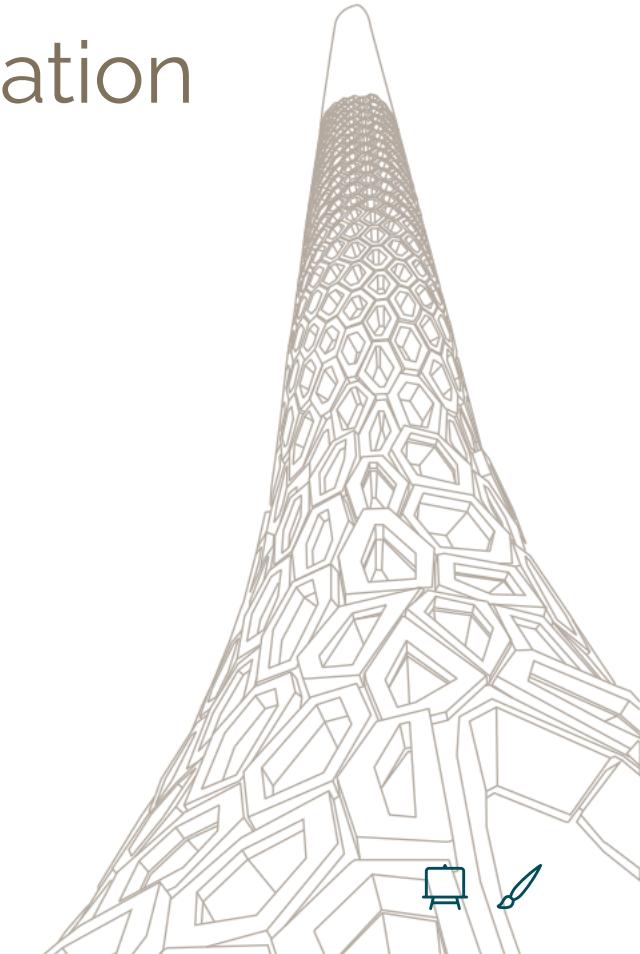
Bias-reduced estimation of structural equation models

Haziq Jamil 

Research Specialist, BAYESCOMP @ CEMSE-KAUST

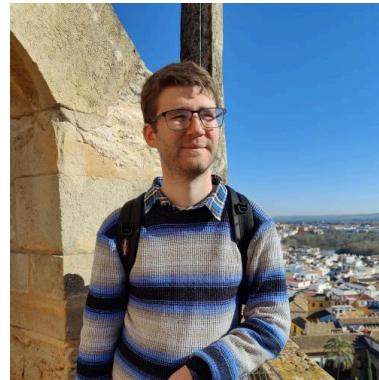
<https://haziqj.ml/sem-bias/>

October 16, 2025





Yves Rosseel
Universiteit Gent | R/{lavaan}



Ollie Kemp
University of Warwick



Ioannis Kosmidis
University of Warwick

Jamil, H., Rosseel, Y., Kemp, O., & Kosmidis, I. (2025). Bias-Reduced Estimation of Structural Equation Models. *Manuscript in Submission.*
arXiv:2509.25419.

- Source: <https://github.com/haziqj/sembias-gradsem>
- Slides: <https://haziqj.ml/sembias-gradsem/slides.pdf>
- R Package: <https://github.com/haziqj/brlavaan>

poll





Context

SEM in a nutshell

Analyse multivariate data $\mathbf{y} = (y_1, \dots, y_p)^\top$ to measure and relate hidden variables $\boldsymbol{\eta} = (\eta_1, \dots, \eta_q)^\top$, $q \ll p$, and uncover complex patterns.

In the social sciences, latent variables are used to represent **constructs**—the *theoretical, unobserved* concepts of interest.



(Psychology)
Personality traits



(Healthcare)
Quality of life



(Political science)
Social trust



(Education)
Competencies

Photo credits: Unsplash @dtravisphd, @impulsq, @ev, @benmullins.





Key issue

"Using SEMs in empirical research is often challenged by small sample sizes."

- Why? Data collection is expensive, time-consuming, or difficult, or all of these!
- Rare populations:
 - **Quetzada, González, and Mecott (2016)**: Identifying factors of adjustment in pediatric burn patients to facilitate appropriate mental health interventions postinjury ($n = 51$).
 - **Figueroa-Jiménez et al. (2021)**: Studying functional connectivity network on individuals with rare genetic disorders ($n = 22$).
 - **Fabbricatore et al. (2023)**: Assessment of psycho-social aspects and performance of elite swimmers ($n = 161$).
 - **Manuela and Sibley (2013)**: Validating self-report measures of identity on a unique cultural group ($n = 143$).
- SEM is desirable, but small $n \Rightarrow$ poor finite-sample performance (esp. bias).



Outline

1. Brief overview of SEMs

- Motivating example
- ML estimation and inference
- Examples of SEMs
 - Two-factor SEM
 - Latent growth models

2. Bias reducing methods

- What is bias?
- A review of bias reduction methods
- Reduced-Bias M -estimation (RBM)
 - Implicit correction
 - Explicit correction

3. Simulation studies and results

poll





Structural equation models



Motivating example

Glycemic control and kidney health

Does poorer glycemic control lead to greater severity of kidney disease?

Observe $p = 6$ variables for each patient:

Indicator	Description	Unit
y_1 HbA1c	3-month avg. blood glucose	%
y_2 FPG	Fasting plasma glucose	mmol/L
y_3 Insulin	Fasting insulin level	μ U/mL
y_4 PCr	Plasma creatinine	μ mol/L
y_5 ACR	Albumin–creatinine ratio	mg/g
y_6 BUN	Blood urea nitrogen	mmol/L

Example adapted from Song and Lee (2012).



Covariance-based approach

Sample correlation matrix looks like this¹:

y_1	y_2	y_3	y_4	y_5	y_6	
y_1	1	0.82	0.8	0.2	0.2	0.2
y_2	0.82	1	0.81	0.19	0.21	0.21
y_3	0.8	0.81	1	0.19	0.2	0.2
y_4	0.2	0.19	0.19	1	0.82	0.82
y_5	0.2	0.21	0.2	0.82	1	0.83
y_6	0.2	0.21	0.2	0.82	0.83	1

- The data suggests clustering of variables
 - y_1, y_2, y_3 measure *glycemic control* (**GlyCon**)
 - y_4, y_5, y_6 measure *kidney health* (**KdnHlt**)
- There is an element of dimension-reduction; much needed for analysing (correlated) multivariate data.
- Easier to hypothesize relationships, e.g.
$$\text{KdnHlt} = \alpha + \beta \text{GlyCon} + \text{error}$$
- SEM is about modelling the covariance structure of the data,
$$\Sigma = \Sigma(\vartheta).$$

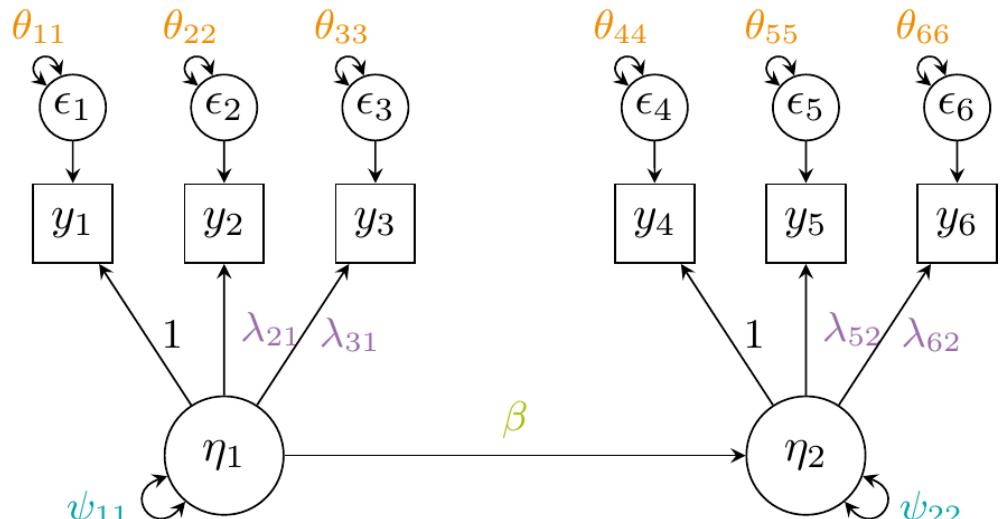
1. Simulated data, from a two-factor SEM ($n = 1000$).



SEM equations

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{pmatrix}$$

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$





ML estimation

- It can be shown that the normal SEM reduces to $\mathbf{y} \sim \mathbf{N}_p(\boldsymbol{\mu}(\vartheta), \boldsymbol{\Sigma}(\vartheta))$, where

$$\begin{aligned}\boldsymbol{\mu}(\vartheta) &= \boldsymbol{\nu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\alpha} \\ \boldsymbol{\Sigma}(\vartheta) &= \underbrace{\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Psi} (\mathbf{I} - \mathbf{B})^{-\top} \boldsymbol{\Lambda}^\top}_{\boldsymbol{\Sigma}^*(\vartheta)} + \boldsymbol{\Theta}\end{aligned}\tag{1}$$



Properties of MLE

- Let $\bar{\vartheta}$ be the true parameter value. Subject to standard regularity conditions (Cox and Hinkley 1979), as $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\vartheta} - \bar{\vartheta}) \xrightarrow{D} N_m \left(\mathbf{0}, [U(\bar{\vartheta})V(\bar{\vartheta})^{-1}U(\bar{\vartheta})]^{-1} \right) \quad (3)$$

where

- $U(\vartheta) = -\mathbb{E} [\nabla \nabla^\top \ell_1(\vartheta)]$ is the *sensitivity matrix*; and
 - $V(\vartheta) = \text{var} [\nabla \ell_1(\vartheta)]$ is the *variability matrix*.
-
- Calculation of SEs are based off estimates of these matrices. The Godambe or "sandwich" matrix gives robust SEs (Satorra and Bentler 1994; Savalei 2014) in cases of model misspecification.
 - If model is correctly specified, $U(\vartheta) = V(\vartheta) = I(\vartheta)$, the Fisher information.



Latent growth curve model (GCM)

- **Longitudinal data:** repeated measurements on individuals i over time, e.g. $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{i10})$, $i = 1, \dots, n$ (Rabe-Hesketh and Skrondal 2008).
- Usually, linear mixed effects models are used, where

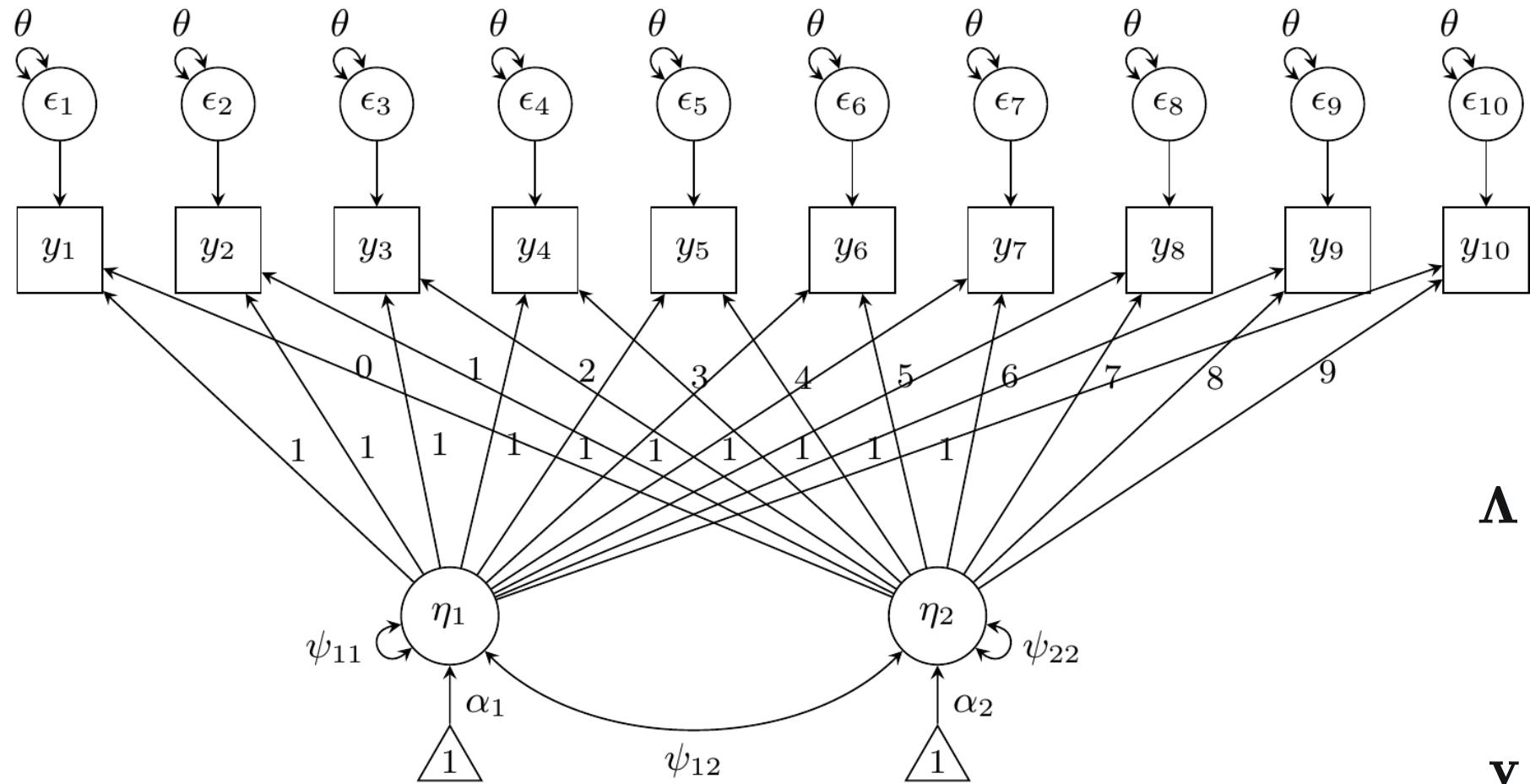
$$y_{it} = \underbrace{(\alpha_1 + \eta_{1i})}_{\text{random int.}} + \underbrace{(\alpha_2 + \eta_{2i})}_{\text{random slope}} \cdot (t - 1) + \epsilon_{it} \quad t = 1, \dots, 10$$

$$\begin{pmatrix} \eta_{1i} \\ \eta_{2i} \end{pmatrix} \sim \mathbf{N}_2 \left(\mathbf{0}, \begin{pmatrix} \psi_{11} & \psi_{12} \\ \cdot & \psi_{22} \end{pmatrix} \right)$$

- α_1 and α_2 are fixed effects (intercept and slope);
- η_{1i} and η_{2i} are correlated random effects (individual deviations);
- $\epsilon_{it} \sim \mathbf{N}(0, \theta)$ are measurement errors.
- Restricted ML is a popular method to estimate such models, with good parameter recovery for variance components (Corbeil and Searle 1976).



Latent GCM as SEM



$$\Lambda = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix}$$

$$\mathbf{y} = \Lambda \boldsymbol{\eta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim N_{10}(\mathbf{0}, \theta \mathbf{I})$$
$$\boldsymbol{\eta} \sim N_2(\boldsymbol{\alpha}, \boldsymbol{\Psi})$$



Bias reduction methods



Poll

Maximum likelihood estimators are known to have which properties?

$$\hat{\vartheta} \rightarrow \vartheta$$



What is bias?

Bias of an estimator

$$\mathcal{B}_{\bar{\vartheta}}(\hat{\vartheta}) = \mathbb{E} [\hat{\vartheta} - \bar{\vartheta}] \quad (4)$$

Consider the stochastic Taylor expansion of $s(\hat{\vartheta}) = \nabla \ell(\vartheta) = 0$ around $\bar{\vartheta}$. For many common estimators including MLE, the bias function is:

$$\mathcal{B}_{\bar{\vartheta}} = \frac{b_1(\bar{\vartheta})}{n} + \frac{b_2(\bar{\vartheta})}{n^2} + \frac{b_3(\bar{\vartheta})}{n^3} + O(n^{-4}). \quad (5)$$

Bias arises because the roots of the score equations are **not exactly centred at $\bar{\vartheta}$** , due to:

- The curvature of the score $s(\vartheta)$ creating asymmetry; and
- The randomness of the score itself.



Illustration

Biased MLE estimator for σ^2

Consider $X_1, \dots, X_n \sim N(0, \sigma^2)$. The MLE for σ^2 is $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$.

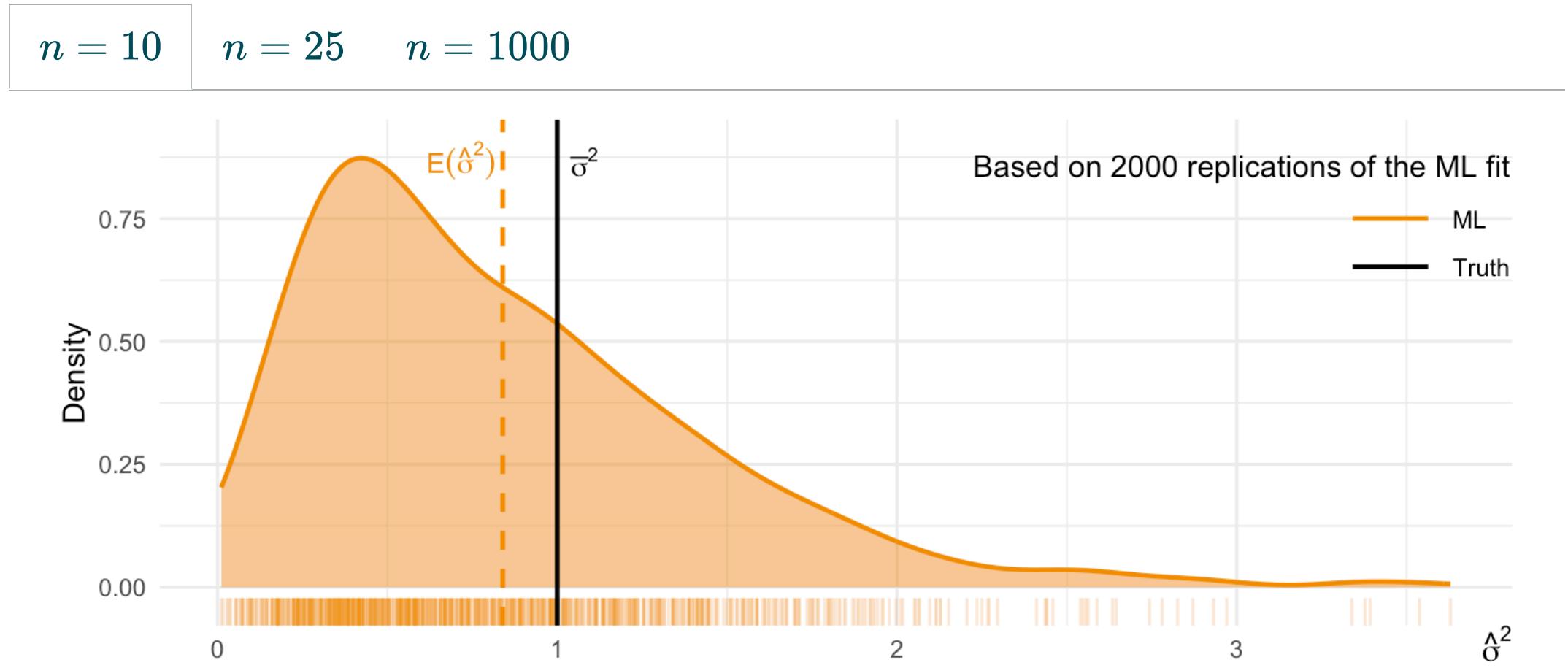
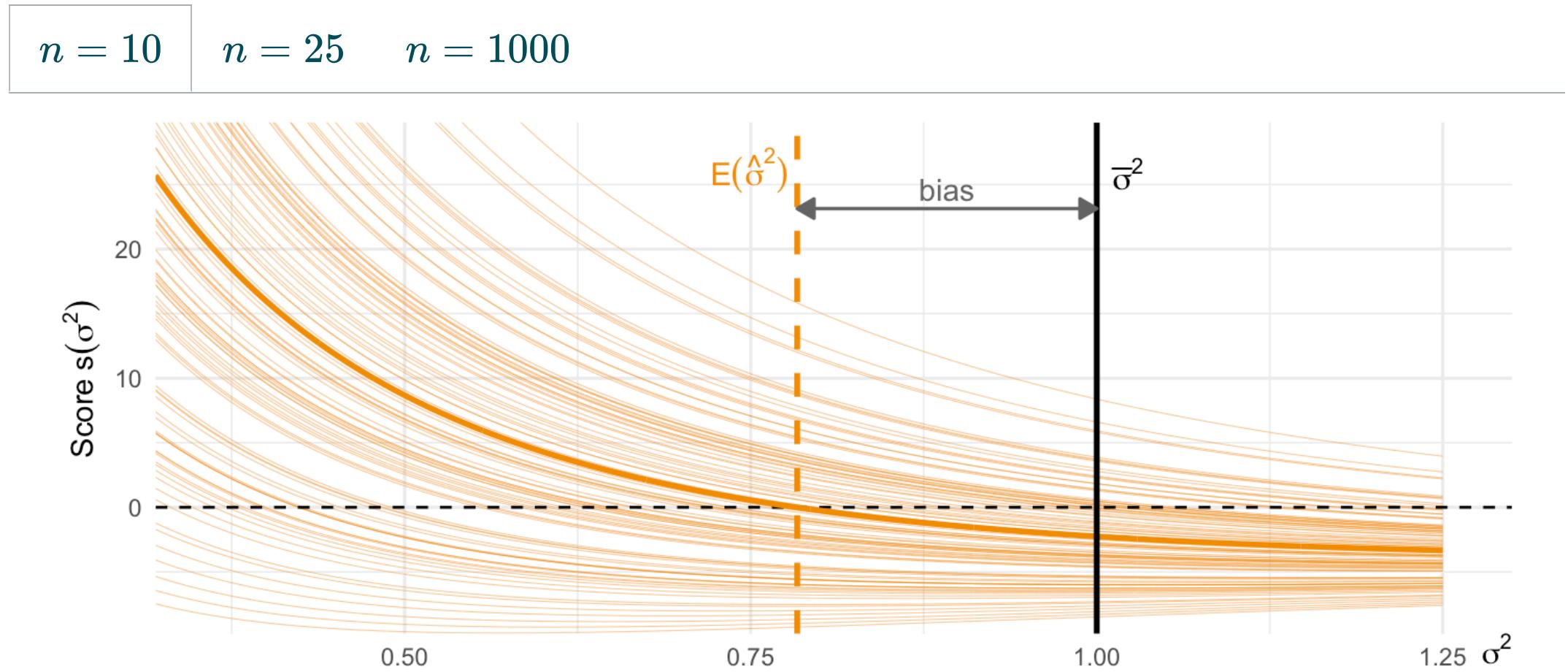




Illustration (cont.)

Score functions are random too

The score is $s(\sigma^2) = \ell'(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n X_i^2$.





If you're interested...

...and love differentiation ❤️🤓

For a comprehensive treatment of bias-reduction methods,

- Start here: Cox and Snell ([1968](#))
- Follow up with: Firth ([1993](#)); Kosmidis and Firth ([2009](#)); Kosmidis ([2014](#))



A review

$$\hat{\vartheta} - \tilde{\vartheta} = \mathcal{B}(\bar{\vartheta}) := \mathbb{E}(\hat{\vartheta} - \bar{\vartheta})$$

estimator ↓
 improved estimator ↓
 bias function ↓
 possibly intractable ↓
 unknown true value ↓

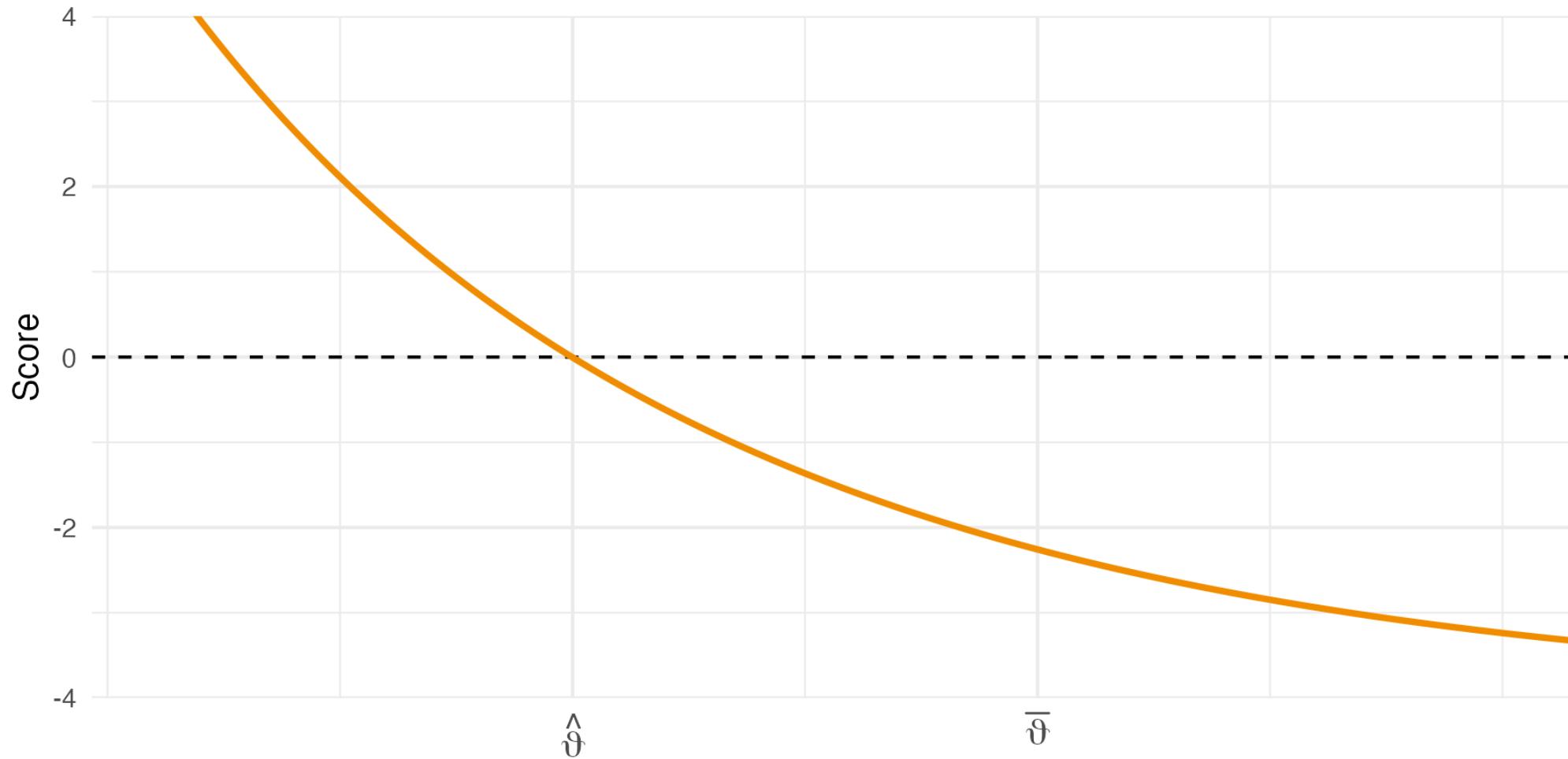
Method	Model	$\mathcal{B}(\bar{\vartheta})$	Type	Requirements		
				$\mathbb{E}(\cdot)$	$\partial \cdot$	$\hat{\vartheta}$
1 Asymptotic bias correction	full	analytical	explicit	✓	✓	✓
2 Adjusted score functions	full	analytical	implicit	✓	✓	✗
3 Bootstrap	partial	simulation	explicit	✗	✗	✓
4 Jackknife	partial	simulation	explicit	✗	✗	✓
5 Indirect inference	full	simulation	implicit	✗	✗	✓
6 Explicit RBM	partial	analytical	explicit	✗	✓	✓
7 Implicit RBM	partial	analytical	implicit	✗	✓	✗

1–Efron (1975), Cordeiro and McCullagh (1991); 2–Firth (1993), Kosmidis and Firth (2009); 3–Efron and Tibshirani (1994), Hall and Martin (1988); 4–Quenouille (1956), Efron (1982); 5–Gourieroux, Monfort, and Renault (1993), MacKinnon and Smith Jr (1998)



Firth's adjusted scores methods

Instead of solving $s(\vartheta) = 0$, solve $s(\vartheta) + \underbrace{B(\vartheta)I(\vartheta)}_{A(\vartheta)} = 0$.





Implicit RBM estimator

Computing $\mathcal{B}(\vartheta)$ and $I(\vartheta)$ can be difficult. Consider

$$s(\vartheta) + A(\vartheta) = 0 \Leftrightarrow \arg \max_{\vartheta} \{\ell(\vartheta) + P(\vartheta)\} \quad (6)$$

where $P(\vartheta)$ is a penalty term constructed such that $A(\vartheta) = \nabla P(\vartheta)$. Kosmidis and Lunardon (2024) show that

$$P(\vartheta) = -\frac{1}{2} \text{tr} \left\{ j(\vartheta)^{-1} e(\vartheta) \right\}$$

where

- $j(\vartheta) = \sum_{i=1}^n \nabla \nabla^\top \ell_i(\vartheta)$ is the observed information;
- $e(\vartheta) = \sum_{i=1}^n \nabla \ell_i(\vartheta) \nabla^\top \ell_i(\vartheta) \nabla \ell_i(\vartheta)$ is the outer-product of scores.

The solution $\tilde{\vartheta}$ to (6) is called the *implicit* RBM estimator (iRBM).

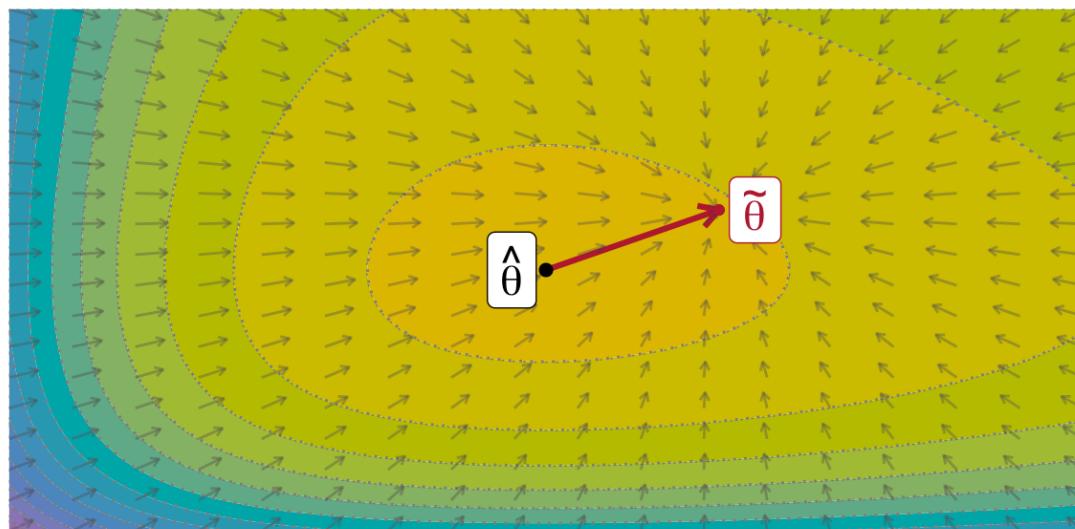


Explicit RBM estimator

Intuitively, by thinking in terms of a Newton-style update, an *explicit* estimator is obtained via

$$\vartheta^* = \hat{\vartheta} + j(\hat{\vartheta})^{-1} A(\hat{\vartheta}).$$

This moves $\hat{\theta}$ in the direction $A(\hat{\theta})$ away from the bias, with step length governed by the curvature $j(\hat{\theta})^{-1}$.



- Operationally, eRBM is simpler and quicker to compute than iRBM—no re-optimisation needed if $\hat{\theta}$ is available.
- However, unlike iRBM, no guarantees that bias correction stays inside the parameter space.



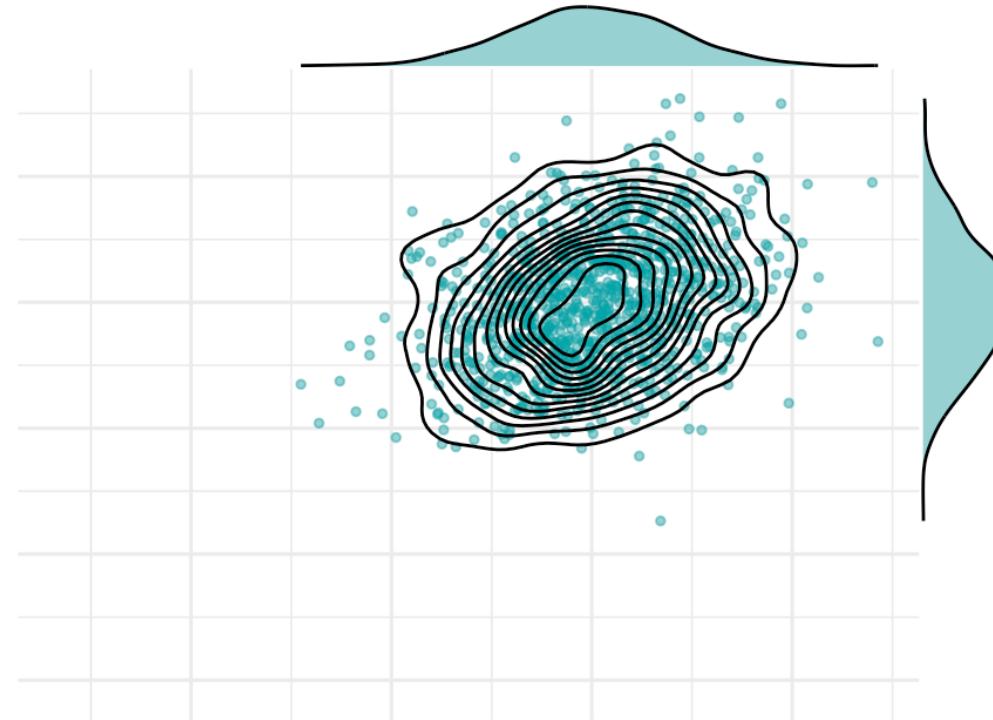
Simulation studies



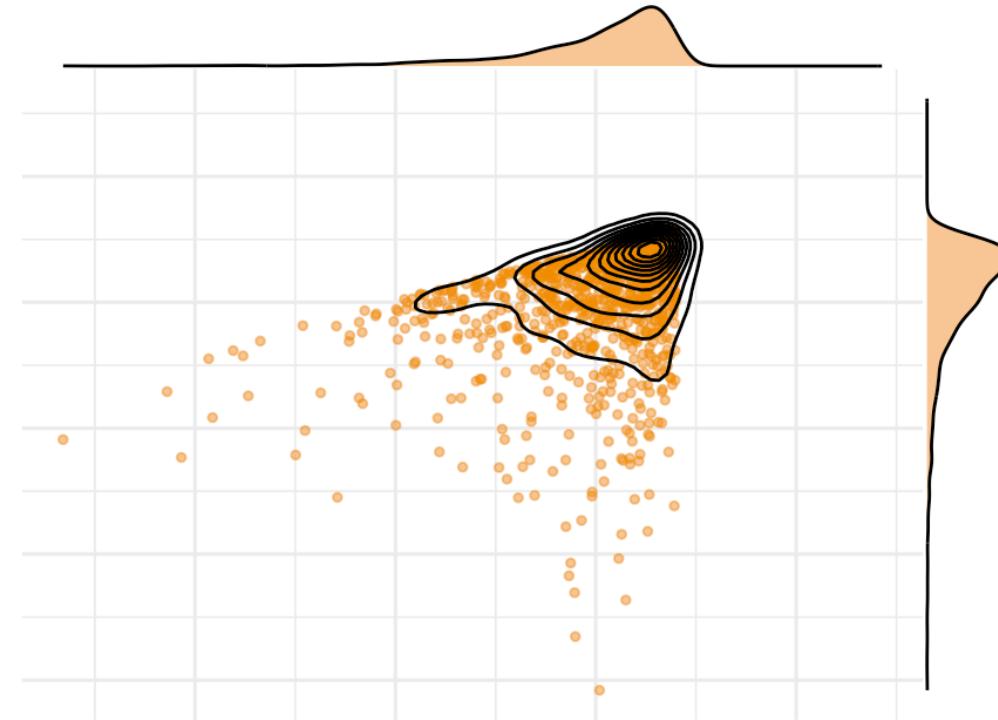
Simulation design

- Sample size: $n \in \{15, 20, 50, 100, 1000\}$
- Item reliability: Low or High ($\text{Rel} = p^{-1} \sum_{j=1}^p \Sigma_{jj}^*/\Sigma_{jj}$)
- Distributional assumption: Normal or Non-normal

Normal



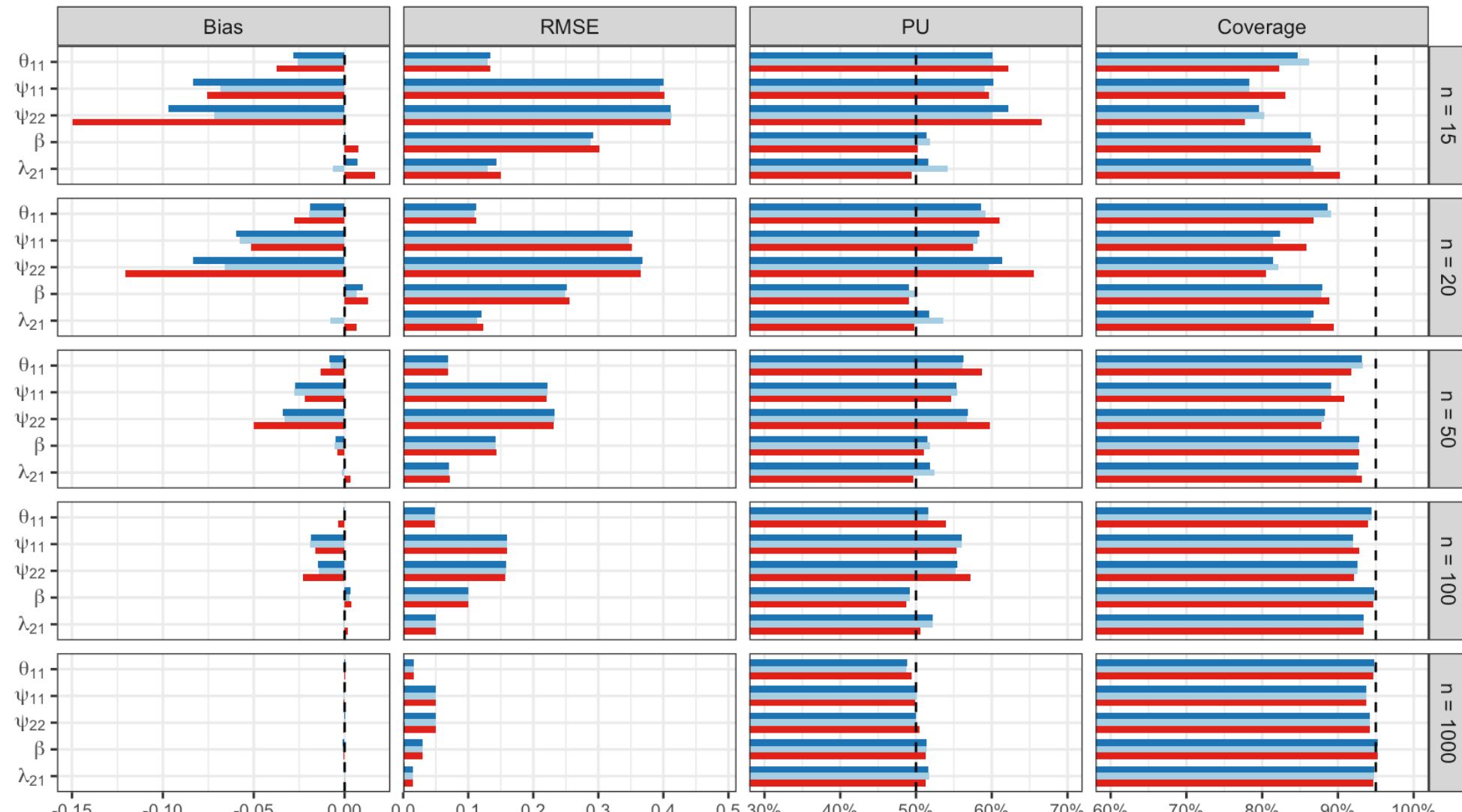
Non-normal





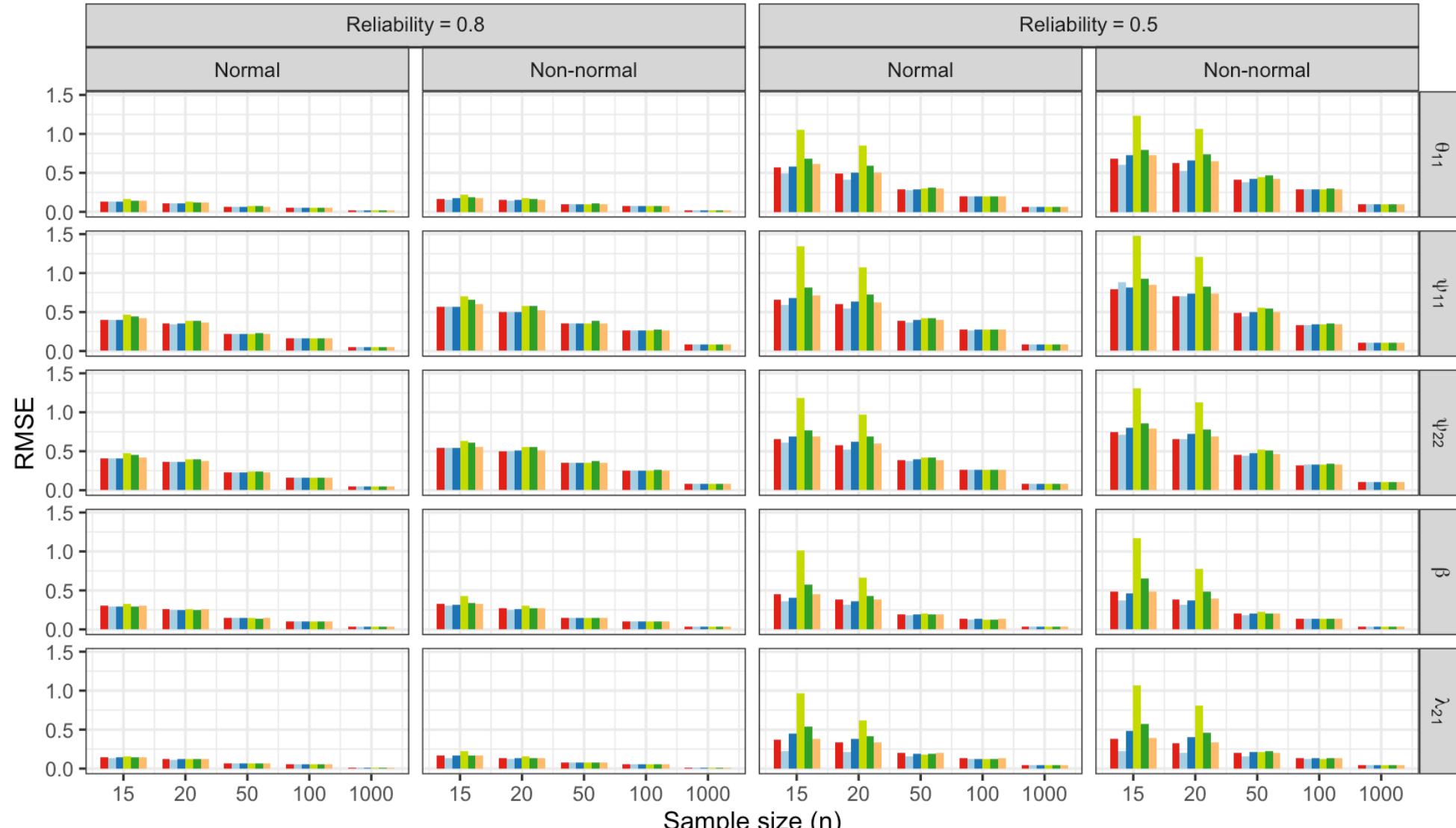
Results: Two-factor SEM

Normal, reliability = 0.8





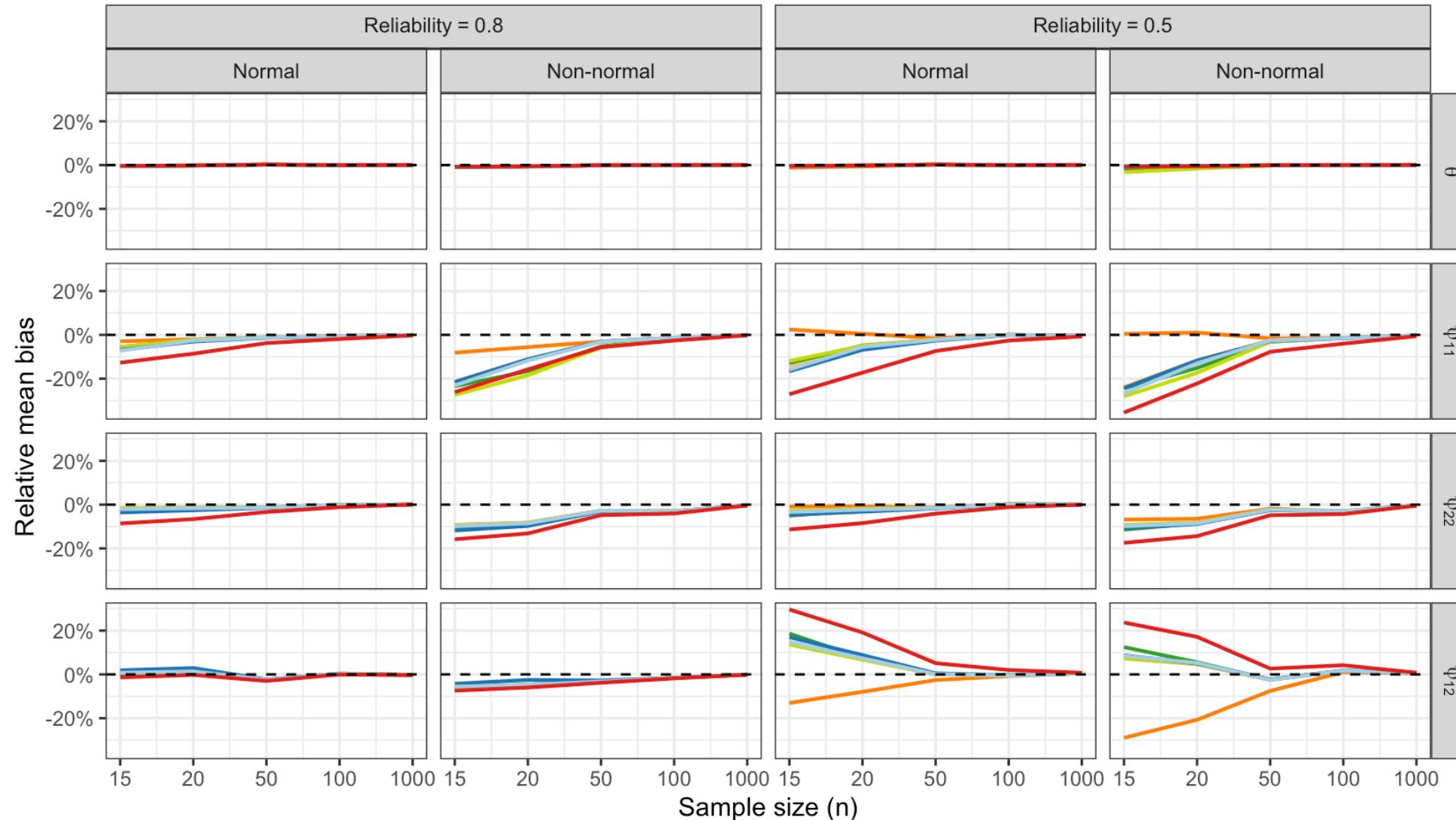
Results: Two-factor SEM (cont.)





Results: Latent GCM

— ML — eRBM — iRBM — Jackknife — Bootstrap — Ozenne et al. — REML





Conclusion



Summary & future work

RBM applied to small sample estimation of SEM show key advantages:

- 🚀 Improved estimator performances (mean & median bias, RMSE, coverage).
- 💻 Computationally efficient (c.f. resampling methods).
- 🤖 Robust to model misspecification.

Future work include

1. Computational improvements for iRBM.
2. Plugin penalties to limit exploration of ill-conditioned regions.
3. Extension to other SEM families, e.g.
 - Path models, mediation models, latent interactions, etc.
 - Alternative to ML estimation e.g. WLS, DWLS, etc.



Software

```
1 # pak:::pak("haziqj/brlavaan")
2 library(brlavaan)
3
4 mod <- "
5   eta1 =~ y1 + y2 + y3
6   eta2 =~ y4 + y5 + y6
7 "
8 fit <- brsem(model = mod, data = dat, estimator.args = list(rbm = "implicit"))
9 summary(fit)
```

brlavaan 0.1.1.9008 ended normally after 88 iterations

Estimator	ML
Bias reduction method	IMPLICIT
Plugin penalty	NONE
Optimization method	NLMINB
Number of model parameters	13
Number of observations	50



شكراً جزيلاً

<https://haziqj.ml/sembias-gradsem>





References

- Corbeil, Robert R., and Shayle R. Searle. 1976. "Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model." *Technometrics* 18 (1): 31–38.
<https://www.tandfonline.com/doi/abs/10.1080/00401706.1976.10489397>.
- Cordeiro, Gauss M., and Peter McCullagh. 1991. "Bias Correction in Generalized Linear Models." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 53 (3): 629–43.
- Cox, D. R., and D. V. Hinkley. 1979. *Theoretical Statistics*. New York: Chapman and Hall/CRC.
<https://doi.org/10.1201/b14832>.
- Cox, D. R., and E. J. Snell. 1968. "A General Definition of Residuals." *Journal of the Royal Statistical Society. Series B (Methodological)* 30 (2): 248–75. <https://www.jstor.org/stable/2984505>.
- Efron, Bradley. 1975. "Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency)." *The Annals of Statistics*, 1189–1242. <https://www.jstor.org/stable/2958246>.
- . 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9781611970319>.
- Efron, Bradley, and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC. <https://doi.org/10.1201/9780429246593>.
- Fabbricatore, Rosa, Maria Iannario, Rosaria Romano, and Domenico Vistocco. 2023. "Component-Based Structural Equation Modeling for the Assessment of Psycho-Social Aspects and Performance of Athletes: Measurement and Evaluation of Swimmers." *AStA Advances in Statistical Analysis* 107 (1–2): 343–67. <https://doi.org/10.1007/s10182-021-00417-5>.
- Figueroa-Jiménez, María Dolores, Cristina Cañete-Massé, María Carbó-Carreté, Daniel Zarabozo-Hurtado,

