Pairwise stochastic approximation for categorical factor models

IMPS 2024, Prague

Giuseppe Alfonzetti[1] Ruggero Bellio[1] Yunxiao Chen[2] Irini Moustaki[2]

[1]University of Udine [2]London School of Economics

Problem Setting[1]

Model setup:

- **Ordinal** data $\boldsymbol{Y} \in \mathbb{R}^p$, with $Y_i \in \{1, \ldots, m_i\}$, $i = 1, \ldots, p$.

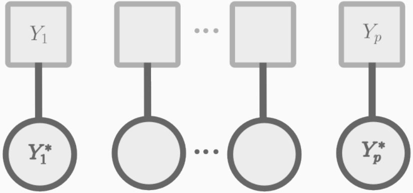
[1] B. Muthén. "A general structural model with dichotomous, ordered categorical and continuous latent variable indicators". In: *Psychometrika* 49.1 (1984), pp. 115–132.

Problem Setting[1]

Model setup:

- **Ordinal** data $\boldsymbol{Y} \in \mathbb{R}^p$, with $Y_i \in \{1, \ldots, m_i\}$, $i = 1, \ldots, p$.
- Discretised **underlying normal response**, i.e. $\boldsymbol{Y}^* \in \mathbb{R}^p$ with

$$Y_i = c \iff \tau_{c-1}^{(i)} \leq Y_i^* < \tau_c^{(i)}, \quad c = 1, \ldots, m_i.$$



[1] B. Muthén. "A general structural model with dichotomous, ordered categorical and continuous latent variable indicators". In: *Psychometrika* 49.1 (1984), pp. 115–132.

Problem Setting[1]

Model setup:

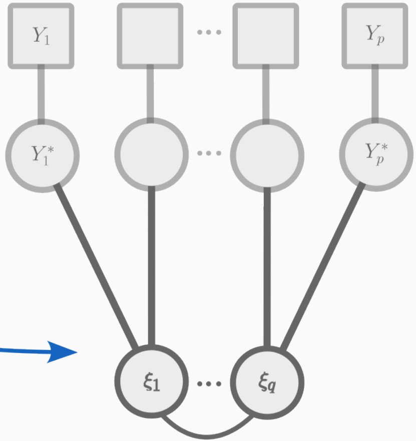
- **Ordinal** data $\boldsymbol{Y} \in \mathbb{R}^p$, with $Y_i \in \{1, \ldots, m_i\}$, $i = 1, \ldots, p$.
- Discretised **underlying normal response**, i.e. $\boldsymbol{Y}^* \in \mathbb{R}^p$ with

$$Y_i = c \iff \tau_{c-1}^{(i)} \le Y_i^* < \tau_c^{(i)}, \quad c = 1, \ldots, m_i.$$

- **Underlying linear factor model:**

$$\boldsymbol{Y}^* = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta},$$

  - $\boldsymbol{\delta} \sim \mathcal{N}_p(0, \boldsymbol{\Sigma}_{\boldsymbol{\delta}})$ and $\boldsymbol{\xi} \sim \mathcal{N}_q(0, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$;
  - $\boldsymbol{\Lambda} = (\lambda_{ij})_{p \times q}$;
  - $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = (\sigma_{rs})_{q \times q}$, $\sigma_{rr} = 1$ for $r = 1, \ldots, q$



[1] B. Muthén. “A general structural model with dichotomous, ordered categorical and continuous latent variable indicators”. In: *Psychometrika* 49.1 (1984), pp. 115–132.

Problem Setting[1]

Model setup:

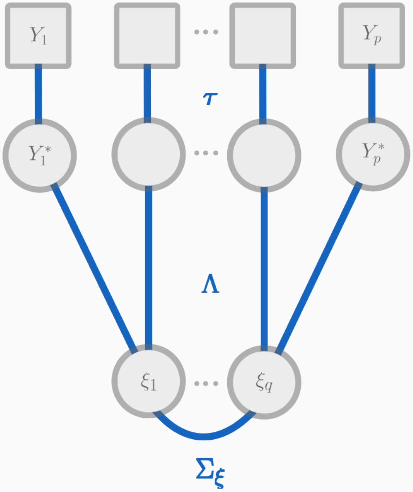
- **Ordinal** data $\boldsymbol{Y} \in \mathbb{R}^p$, with $Y_i \in \{1, \dots, m_i\}$, $i = 1, \dots, p$.
- Discretised **underlying normal response**, i.e. $\boldsymbol{Y}^* \in \mathbb{R}^p$ with

$$Y_i = c \iff \tau_{c-1}^{(i)} \leq Y_i^* < \tau_c^{(i)}, \quad c = 1, \dots, m_i.$$

- **Underlying linear factor model:**

$$\boldsymbol{Y}^* = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta},$$

  - $\boldsymbol{\delta} \sim \mathcal{N}_p(0, \boldsymbol{\Sigma}_{\boldsymbol{\delta}})$ and $\boldsymbol{\xi} \sim \mathcal{N}_q(0, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$;
  - $\boldsymbol{\Lambda} = (\lambda_{ij})_{p \times q}$;
  - $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = (\sigma_{rs})_{q \times q}$, $\sigma_{rr} = 1$ for $r = 1, \dots, q$
- θ collects all the **free parameters** in $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ and $\boldsymbol{\tau}$.



[1] B. Muthén. "A general structural model with dichotomous, ordered categorical and continuous latent variable indicators". In: *Psychometrika* 49.1 (1984), pp. 115–132.

Likelihood inference

Consider the **joint p-variate probability** of observing the pattern $\mathbf{c} = (c_1, ..., c_p)$

$$\pi_{\mathbf{c}}(\boldsymbol{\theta}) = P(\boldsymbol{Y} = \boldsymbol{c}; \boldsymbol{\theta}) = \int_{\tau^{(1)}_{c_1-1}}^{\tau^{(1)}_{c_1}} \cdots \int_{\tau^{(p)}_{c_p-1}}^{\tau^{(p)}_{c_p}} \phi_p(\boldsymbol{y}^*; \boldsymbol{\Lambda}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}^\top + \boldsymbol{\Sigma}_\delta) d\boldsymbol{y}^*,$$

where $\phi_p(x; \Sigma)$ is the density of a p-dimensional normal distribution at x, with mean zero and variance Σ.

Likelihood inference

Consider the **joint p-variate probability** of observing the pattern $\mathbf{c} = (c_1, ..., c_p)$

$$\pi_{\mathbf{c}}(\boldsymbol{\theta}) = P(\boldsymbol{Y} = \boldsymbol{c}; \boldsymbol{\theta}) = \int_{\tau^{(1)}_{c_1-1}}^{\tau^{(1)}_{c_1}} \cdots \int_{\tau^{(p)}_{c_p-1}}^{\tau^{(p)}_{c_p}} \phi_p(\boldsymbol{y}^*; \boldsymbol{\Lambda}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}^\top + \boldsymbol{\Sigma}_\delta) d\boldsymbol{y}^*,$$

where $\phi_p(x; \Sigma)$ is the density of a p-dimensional normal distribution at x, with mean zero and variance Σ.

It requires evaluating a **p-dimensional integral!**

Pairwise Likelihood inference[2] (i)

Pairwise Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_{PML} = \arg\max_{\boldsymbol{\theta}} \; p\ell(\boldsymbol{\theta}), \quad \text{where} \quad p\ell(\boldsymbol{\theta}) = \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} n_{c_i c_j}^{(ij)} \log\left\{\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})\right\},$$

with $n_{c_i c_j}^{(ij)} = \sum_{l=1}^{n} \mathbf{1}_{\{y_{li}=c_i, y_{lj}=c_j\}}$ and $\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}) = P\left(Y_i = c_i, Y_j = c_j; \boldsymbol{\theta}\right)$.

[2] M. Katsikatsou et al. "Pairwise likelihood estimation for factor analysis models with ordinal data". In: *Computational Statistics and Data Analysis* 56 (2012), pp. 4243–4258.

Pairwise Likelihood inference[2] (i)

Pairwise Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_{PML} = \arg\max_{\boldsymbol{\theta}} \; p\ell(\boldsymbol{\theta}), \quad \text{where} \quad p\ell(\boldsymbol{\theta}) = \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} n_{c_i c_j}^{(ij)} \log \left\{ \pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}) \right\},$$

with $n_{c_i c_j}^{(ij)} = \sum_{l=1}^{n} \mathbf{1}_{\{y_{li}=c_i, y_{lj}=c_j\}}$ and $\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}) = P\left(Y_i = c_i, Y_j = c_j; \boldsymbol{\theta}\right)$.

✔ It only involves **bidimensional integrals**

[2] M. Katsikatsou et al. "Pairwise likelihood estimation for factor analysis models with ordinal data". In: *Computational Statistics and Data Analysis* 56 (2012), pp. 4243–4258.

Pairwise Likelihood inference[2] (i)

Pairwise Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_{PML} = \arg\max_{\boldsymbol{\theta}} \; p\ell(\boldsymbol{\theta}), \quad \text{where} \quad p\ell(\boldsymbol{\theta}) = \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} n_{c_i c_j}^{(ij)} \log\left\{\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})\right\},$$

with $n_{c_i c_j}^{(ij)} = \sum_{l=1}^{n} \mathbf{1}_{\{y_{li}=c_i, y_{lj}=c_j\}}$ and $\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}) = P\left(Y_i = c_i, Y_j = c_j; \boldsymbol{\theta}\right)$.

- ✔ It only involves **bidimensional integrals**
- ✔ $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{PML} - \boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}\right)$, with $\boldsymbol{H} = E\{-\nabla^2 p\ell(\boldsymbol{\theta}^*)/n\}$ and $\boldsymbol{J} = \text{Var}\{\nabla p\ell(\boldsymbol{\theta}^*)/n\}$.

[2] M. Katsikatsou et al. "Pairwise likelihood estimation for factor analysis models with ordinal data". In: *Computational Statistics and Data Analysis* 56 (2012), pp. 4243–4258.

Pairwise Likelihood inference[2] (i)

Pairwise Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_{PML} = \arg\max_{\boldsymbol{\theta}} \; p\ell(\boldsymbol{\theta}), \quad \text{where} \quad p\ell(\boldsymbol{\theta}) = \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} n_{c_i c_j}^{(ij)} \log\left\{\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})\right\},$$

with $n_{c_i c_j}^{(ij)} = \sum_{l=1}^{n} \mathbf{1}_{\{y_{li}=c_i, y_{lj}=c_j\}}$ and $\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}) = P\left(Y_i = c_i, Y_j = c_j; \boldsymbol{\theta}\right)$.

- ✔ It only involves **bidimensional integrals**
- ✔ $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{PML} - \boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}\right)$, with $\boldsymbol{H} = E\{-\nabla^2 p\ell(\boldsymbol{\theta}^*)/n\}$ and $\boldsymbol{J} = \mathrm{Var}\{\nabla p\ell(\boldsymbol{\theta}^*)/n\}$.
- ✘ The number of summands grows with $\boldsymbol{O(K)}$, where K is the number of pairs, i.e. $O(p^2)$.

[2] M. Katsikatsou et al. “Pairwise likelihood estimation for factor analysis models with ordinal data”. In: *Computational Statistics and Data Analysis* 56 (2012), pp. 4243–4258.

Do we need all the pairs?

Stochastic approximations - Gradients

Pairwise gradient: $$\nabla p\ell(\boldsymbol{\theta}) = \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} \frac{n_{c_i c_j}^{(ij)}}{\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})} \nabla \pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}).$$

Stochastic approximations - Gradients

Pairwise gradient: $$\nabla p\ell(\boldsymbol{\theta}) = \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} \frac{n^{(ij)}_{c_i c_j}}{\pi^{(ij)}_{c_i c_j}(\boldsymbol{\theta})} \nabla \pi^{(ij)}_{c_i c_j}(\boldsymbol{\theta}).$$

Consider a $p \times p$ matrix $\boldsymbol{W}$, with generic element $W_{ij} \in \{0, 1\}$ and let $E\{W_{ij}\} = \nu/K$.

Stochastic approximations - Gradients

Pairwise gradient: $$\nabla p\ell(\boldsymbol{\theta}) = \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} \frac{n_{c_i c_j}^{(ij)}}{\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})} \nabla \pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}).$$

Consider a $p \times p$ matrix $\boldsymbol{W}$, with generic element $W_{ij} \in \{0,1\}$ and let $E\{W_{ij}\} = \nu / K$.

Stochastic gradient: $$S(\boldsymbol{\theta}; \boldsymbol{W}) = \frac{\boldsymbol{K}}{\boldsymbol{\nu}} \sum_{i<j} \boldsymbol{W_{ij}} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} \frac{n_{c_i c_j}^{(ij)}}{\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})} \nabla \pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}).$$

The key feature is $E_{\boldsymbol{W}}\left\{S(\boldsymbol{\theta}; \boldsymbol{W})\right\} = \nabla p\ell(\boldsymbol{\theta})$, where $S(\boldsymbol{\theta}; \boldsymbol{W})$ accounts for ν pairs on average.

Stochastic optimization:[3,4]

Given a suitable decreasing scheduling for η_t, at each iteration $t = 1, \dots, T_n$ repeat

1. **Sample** $\boldsymbol{W}_t$ by drawing ν out of K pairwise components;
2. **Update** the current estimate via $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \frac{\eta_t}{n} S(\boldsymbol{\theta}_{t-1}; \boldsymbol{W}_t)$;

Trajectories averaging via $\bar{\boldsymbol{\theta}} = \frac{1}{T_n} \sum_{t=1}^{T_n} \boldsymbol{\theta}_t$.

[3] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[4] Boris T Polyak and Anatoli B Juditsky. "Acceleration of stochastic approximation by averaging". In: *SIAM journal on control and optimization* 30.4 (1992), pp. 838–855.

Stochastic approximation - Asymptotics (i)

Let T_n diverge with $n \to \infty$. Then, Let T_n diverge with $n \to \infty$. Then,

Consistency and normality:

$$\boldsymbol{\Omega}_n^{-1/2}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow[n]{d} \mathcal{N}\left(\mathbf{0}; \boldsymbol{I}_d\right);$$

Stochastic approximation - Asymptotics (i)

Let T_n diverge with $n \to \infty$. Then,

Consistency and normality:

$$\boldsymbol{\Omega}_n^{-1/2}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow[n]{d} \mathcal{N}\left(\mathbf{0}; \boldsymbol{I}_d\right);$$

Covariance matrix:

$$\boldsymbol{\Omega}_n = \underbrace{\frac{1}{n}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}}_{\text{same as } \hat{\boldsymbol{\theta}}_{\text{PML}}} + \frac{\textbf{NOISE}}{\boldsymbol{T_n}}.$$

The **NOISE** depends on $\text{Var}\{n^{-1}S(\boldsymbol{\theta}; \boldsymbol{W})\}$; thus on the distribution of $\boldsymbol{W}$ and the chosen ν.

Stochastic approximation - Asymptotics (ii)

Recall that $$\frac{1}{n}S(\boldsymbol{\theta};\boldsymbol{W}) = \frac{K}{n\nu}\sum_{i<j} W_{ij}\sum_{c_i=1}^{m_i}\sum_{c_j=1}^{m_j}\frac{n^{(ij)}_{c_ic_j}}{\pi^{(ij)}_{c_ic_j}(\boldsymbol{\theta})}\nabla\pi^{(ij)}_{c_ic_j}(\boldsymbol{\theta}).$$

Stochastic approximation - Asymptotics (ii)

Recall that $$\frac{1}{n}S(\boldsymbol{\theta};\boldsymbol{W}) = \frac{K}{n\nu}\sum_{i<j} W_{ij} \sum_{c_i=1}^{m_i}\sum_{c_j=1}^{m_j} \frac{n_{c_ic_j}^{(ij)}}{\pi_{c_ic_j}^{(ij)}(\boldsymbol{\theta})}\nabla\pi_{c_ic_j}^{(ij)}(\boldsymbol{\theta}).$$

When $W_{ij} = 1$, we account for all n observations on the (i, j) pair. It can be shown that such special behaviour implies that **NOISE** $= O(1/n)$.

Stochastic approximation - Asymptotics (ii)

$$\text{Recall that} \quad \frac{1}{n}S(\boldsymbol{\theta};\boldsymbol{W}) = \frac{K}{n\nu}\sum_{i<j} W_{ij} \sum_{c_i=1}^{m_i}\sum_{c_j=1}^{m_j} \frac{n_{c_ic_j}^{(ij)}}{\pi_{c_ic_j}^{(ij)}(\boldsymbol{\theta})}\nabla\pi_{c_ic_j}^{(ij)}(\boldsymbol{\theta}).$$

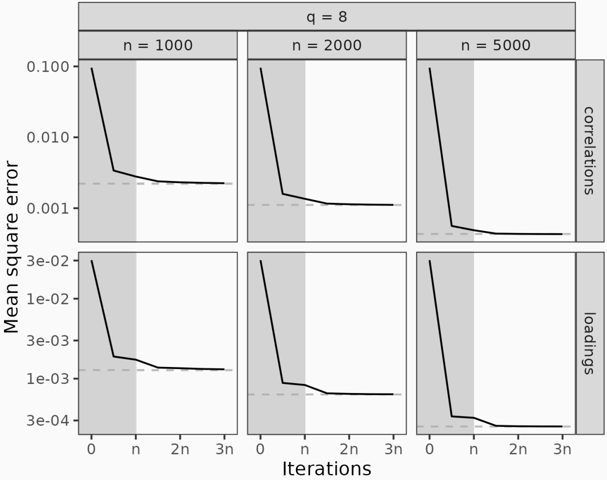
When $W_{ij} = 1$, we account for all n observations on the (i,j) pair. It can be shown that such special behaviour implies that **NOISE** $= O(1/n)$.

Asymptotic equivalence between $\bar{\theta}$ and $\hat{\theta}_{\text{PML}}$

$$\boldsymbol{\Omega}_n = \frac{1}{n}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1} + \boldsymbol{O}\left(\frac{\mathbf{1}}{\boldsymbol{nT_n}}\right) \approx \frac{1}{n}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}.$$

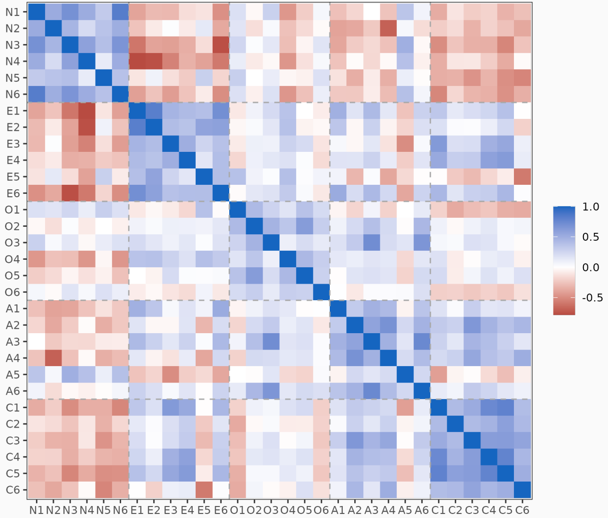
Simulation experiments

- $p = 40$ items, $q = 8$ latent traits with simple loading structure augmented with $q - 1$ cross-loadings;
- $m = 4$ categories per item;
- $n \in \{1000, 2000, 5000\}$;
- $\nu = 8$ i.e. 8 out of $K = 780$ pairs per iteration;
- Replications $= 1000$.



Big Five Application

- $p = 120$ items, $m = 5$ categories per item;
- $q = 30$ latent traits that we expect to be grouped in **N**euroticism, **A**greeableness, **E**xtraversion, **O**penness to experience, **C**onscientiousness;
- $n = 410,376$. The 60% is used for training;
- $\nu = 1$ out of $K = 7140$ pairs per iteration.



Thank you for your attention!

giuseppe.alfonzetti@uniud.it

giuseppealfonzetti

giuseppealfonzetti.github.io

Alfonzetti et al., "Pairwise stochastic approximation for confirmatory factor analysis of categorical data", *BJMSP* (2024)

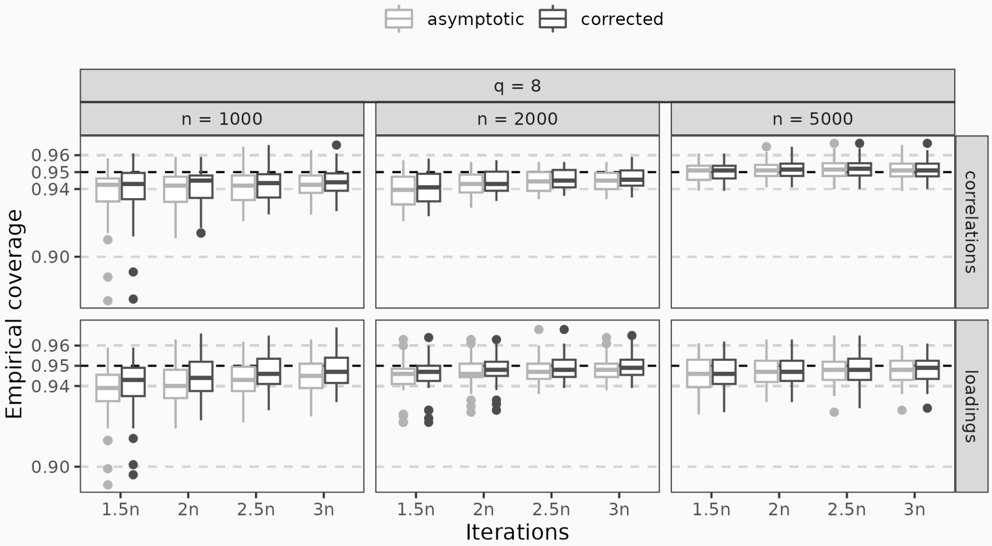
Finanziato dall'Unione europea NextGenerationEU

MUR Ministero dell'Università e della Ricerca

References

- Katsikatsou, M. et al. “Pairwise likelihood estimation for factor analysis models with ordinal data”. In: *Computational Statistics and Data Analysis* 56 (2012), pp. 4243–4258.
- Muthén, B. “A general structural model with dichotomous, ordered categorical and continuous latent variable indicators”. In: *Psychometrika* 49.1 (1984), pp. 115–132.
- Polyak, Boris T and Anatoli B Juditsky. “Acceleration of stochastic approximation by averaging”. In: *SIAM journal on control and optimization* 30.4 (1992), pp. 838–855.
- Robbins, Herbert and Sutton Monro. “A stochastic approximation method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

Simulation experiments - Empirical coverage



Big Five - Estimation trajectories