Fast M-Estimation of GLLVM in High Dimensions

Maria-Pia Victoria-Feser

Department of Statistics, University of Bologna
Research Center for Statistics, University of Geneva

joint work with
G. Blanc, University of Zurich, S. Guerrier, University of Geneva
& S. Cagnone, University of Bologna.

December 8, 2023

Introduction

- For fitting a GLLVM to the data, one relies on an approximation of the (marginal) MLE, since the likelihood function involves integrating over the latent variables, which does not admit, in general, a closed form.
- E.g. Expectation-Maximization (Dempster et al., 1977) with Monte Carlo simulations (Cappé and Moulines, 2009), Laplace approximation (Huber et al., 2004), adaptive quadrature (Cagnone and Monari, 2013), fully exponential Laplace approximation (Bianconcini and Cagnone, 2012), variational approximations (Hui et al., 2017; Niku et al., 2019).
- However, as either the number of observations n, the number of manifest variables p or the number of latent variables q increase, the computational speed deteriorates to the point of rendering the fitting impractical or even impossible.
- E.g. $n = 10,000$, $p = 2,000$, $q = 10$.

Introduction

- Moreover, the (approximated) MLE is computable only when $p < n$.
- A notable exception is the FA model, for which Robertson and Symons (2007) show that the MLE exists when $p > n$ if $q < n$.
  - Sundberg and Feldmann (2016) propose estimating equations and an iterative algorithm to compute the MLE when $p > n$, based on a rescaling of the data,
  - Dai et al. (2020) use a profile (marginal) likelihood (avoiding inverting the covariance matrix), implemeted in the `R` package `fad`.
- For Binary responses, the currently only available route is to use penalized (or regularized) estimators:
  - Chen et al. (2020) propose a Penalized Joint MLE, implemented in their `R` package `mirtjlm`,
  - Kidzinski et al. (2023), propose a type of Penalized Quasi-Likelihood Estimator, implemented in their `R` package `gmf`.

Introduction

- Alternative route: define inconsistent but numerically efficient estimators and define a bias reduction method.
- E.g. Guerrier et al. (2019).
- We choose the estimation framework of the extended log-likelihood function (Bjørnstad, 1996):
  - we use a profiled score function,
  - the latent scores are replaced by a (suitable) function of the data and the parameters,
  - we use a correction factor for consistency, leading to an M-estimator (Huber, 1964) and
  - we use a stochastic approximation algorithm (Kiefer and Wolfowitz, 1952; Blum, 1954; Fabian, 1968; Wei and Tanner, 1990) to compute it.
- It is a fast alternative to the adjusted profile h-likelihood (Lee and Nelder, 1996, 2001; Lee et al., 2006)
- It can be computed when $p > n$.

GLLVM

- Observations: $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$, a set of n independent p-dimensional response (or manifest) variables with correlated elements $Y_{i1}, \ldots, Y_{ip}, i = 1, \ldots, n$.
- Latent variables (for the correlation): $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$, $q \ll p$, independently distributed from a standard multivariate Gaussian distribution, with density denoted by $\phi(\mathbf{z}_i), i = 1, \ldots, n$.
- Covariates: $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, k-dimensional.
- Model: the conditional distribution of the manifest variables Y_{ij}, say F_{θ}, conditionally on the latent variables and covariates, is taken from the exponential family with
  - $\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{Z}_i^\top \boldsymbol{\lambda}_j, i = 1, \ldots, n, j = 1, \ldots, p$, the linear predictor,
  - $\mathbb{E}[Y_{ij}|\mathbf{Z}_i, \mathbf{x}_i, \beta_{0j}, \boldsymbol{\beta}_j, \boldsymbol{\lambda}_j] = g_j(\eta_{ij})$, where $g_j(\cdot), j = 1, \ldots, p$ is a known link function,
  - and response variable-specific dispersion parameter τ_j.

GLLVM

- Parameters: $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^r$, collects
  - $\beta_{0j} \in \mathbb{R}, j = 1, \ldots, p$,
  - $\boldsymbol{\beta}_j \in \mathbb{R}^k, j = 1, \ldots, p$ the fixed-effect coefficients,
  - $\boldsymbol{\lambda}_j \in \mathbb{R}^q, j = 1, \ldots, p$ the factor loadings,
  - response variable-specific dispersion parameter τ_j, with $\boldsymbol{\tau} = [\tau_j]_{j=1,\ldots,p}$.
- Hence, $r = (k + 1 + q + 1)p$, e.g. if $p = 10,000$, $k = 10$ and $q = 10$, we have $r = 220,000$ parameters...
- Note that the $\boldsymbol{\lambda}_i$ are identifiable up to a rotation.

Log-Likelihood Functions

- The conditional (on the latent variables) joint density $f_j(y_{ij}|\mathbf{Z}_i, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau})$ of the sample of observations $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, under the assumption of conditional independence, is

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}_i|\mathbf{Z}_i, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau}) = \prod_{j=1}^{p} \exp\left(\frac{y_{ij}\eta_{ij} - b_j(\eta_{ij})}{\tau_j} + c_j(y_{ij}, \tau_j)\right).$$

  for known response-specific functions b_j and c_j.

- Extended log-likelihood :

$$l(\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{z}|\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} \log\left(f_{\mathbf{Y}_i|\mathbf{Z}_i}(\mathbf{y}_i|\mathbf{z}_i, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau})\phi(\mathbf{z}_i)\right).$$

- Maximizing the extended log-likelihood leads to trivial solutions (e.g. $\mathbf{z} = 0$)...

Log-Likelihood Functions

- Profile log-likelihood function, for a suitable function $\mathbf{z}^* := m(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\tau})$, also called the profile *h*-likelihood (Lee and Nelder, 1996, 2001):

$$l(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} \log \left(f_{\mathbf{Y}_i|\mathbf{z}_i}(\mathbf{y}_i|\mathbf{z}_i^*, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau}) \phi(\mathbf{z}_i^*) \right).$$

- Again, the profile log-likelihood can lead to trivial solutions (Lee and Nelder, 2009)...
- Adjusted profile *h*-likelihood:

$$l(\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{z}|\mathbf{y}, \mathbf{x}) - h(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}, \hat{\mathbf{z}})$$

  for a suitable function h and where $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\tau}}$ and $\hat{\mathbf{z}}$ are the maximizers of $l(\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{z}|\mathbf{y}, \mathbf{x})$.
- Wu and Bentler (2012): GLLVM with binary manifest variables.

PRoflle M-Estimator (PRIME)

- Consider a (suitable) function $\mathbf{z}^* = m(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\tau})$, explicit or implicit, and the profiled score functions

$$\begin{aligned}\Psi(\mathbf{y}_i|\mathbf{z}_i^*, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau}) &= \begin{bmatrix} \Psi_1(\mathbf{y}_i|\mathbf{z}_i^*, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau}) \\ \Psi_2(\mathbf{y}_i|\mathbf{z}_i^*, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau}) \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j=1}^{p} \frac{\partial \eta_{ij}}{\partial \boldsymbol{\theta}} \left(\frac{y_{ij} - b_j'(\eta_{ij})}{\tau_j} \right) \\ -\sum_{j=1}^{p} \left(y_{ij}\eta_{ij} - b_j(\eta_{ij}) \right) / \tau_j^2 + c'(y_{ij}, \tau_j) \end{bmatrix},\end{aligned}$$

- we propose to center the score function to its expectation, therefore defining a Fisher-consistent M-estimator $\hat{\boldsymbol{\theta}}$ (and $\hat{\boldsymbol{\tau}}$) through

$$\frac{1}{n} \sum_{i=1}^{n} \left(\Psi(\mathbf{y}_i|\mathbf{z}_i^*, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau}) - \mathbb{E}_{\mathbf{Y}} \left[\Psi(\mathbf{Y}|\mathbf{z}_i^*, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau}) \right] \right) = \mathbf{0}.$$

PRoflle M-Estimator (PRIME)

- For $\mathbf{z}_i^*$, we propose to use the (estimated) mode of $f_{\mathbf{Z}_i|\mathbf{Y}_i}(\mathbf{z}_i|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\tau})$, also called the maximum a posteriori (MAP).
- Other choices are possible and do not impact the consistency of the PRIME.
- The form of the PRIME (M-estimator with Fisher consistency centering) allows to compute it by means of a stochastic approximation algorithm.

PRofIle *M*-Estimator (PRIME)

- Moreover, for numerical efficiency, without loosing consistency, the score functions can be further simplified to

$$\widetilde{\Psi}_1(\mathbf{y}_i|\mathbf{z}_i,\mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\tau}) = \sum_{j=1}^{p}\frac{\partial\eta_{ij}}{\partial\boldsymbol{\theta}}\frac{y_{ij}}{\tau_j},$$

$$\widetilde{\Psi}_2(\mathbf{y}_i|\mathbf{z}_i,\mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\tau}) = -2\sum_{j=1}^{p}\frac{\partial}{\partial\boldsymbol{\tau}}\left(\frac{y_{ij}^2}{\tau_j}\right),$$

  which we call the Simplified PRIME (SPRIME) $\widetilde{\boldsymbol{\theta}}$.
- Property 1: In the Gaussian case, i.e. FA, the PRIME and SPRIME are equivalent to the MLE!
- This suggests that, for numerical efficiency, a FA model in large dimensions should be estimated via the PRIME (or SPRIME), which avoids the computation of the inverse of the covariance matrix...

PRoflle *M*-Estimator (PRIME)

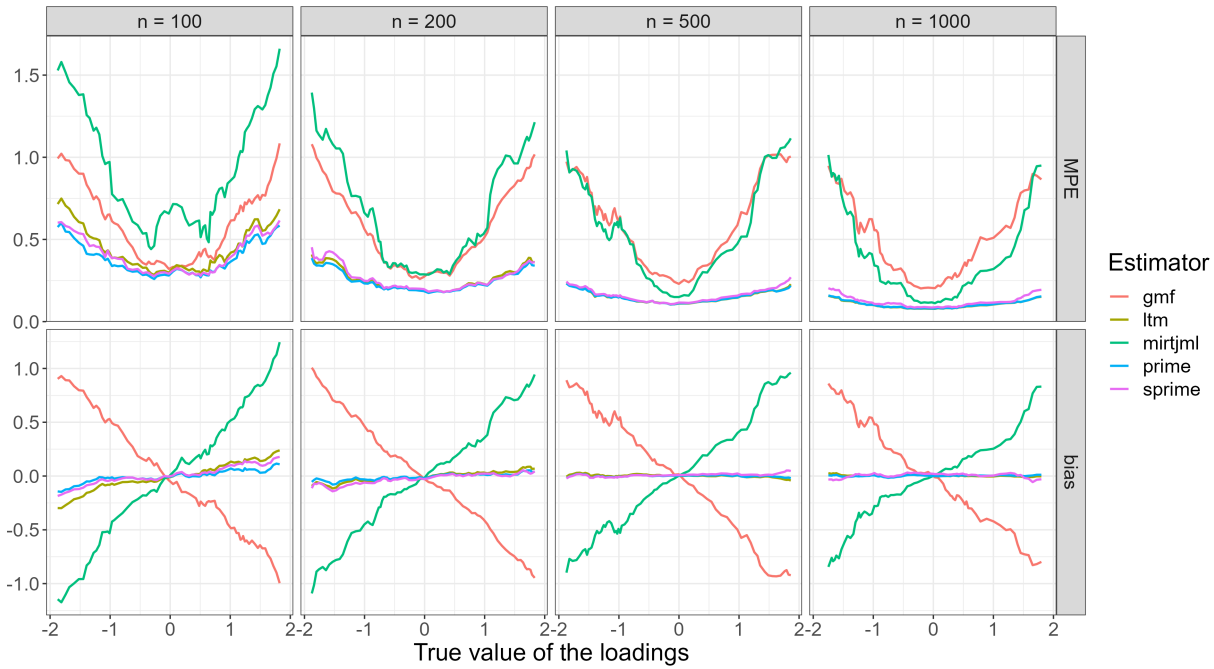
- Property 2: In the absence of latent variables ($\mathbf{z}_i^* = \mathbf{z}_i = \mathbf{0}$), i.e. supposing uncorrelated responses, the estimating equations of the SPRIME are the score functions of a GLM for each response variable $\mathbf{Y}_j$, and therefore defining the MLE.
- Property 3: Asymptotic normality of $\hat{\boldsymbol{\theta}}$ (and $\widetilde{\boldsymbol{\theta}}$), with covariance matrix

$$\begin{aligned}\text{Acov}(\hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\Psi_1(\mathbf{y}_i|\mathbf{z}_i^*,\mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\tau})\right]^{-1}\mathbb{E}\left[\Psi_1(\mathbf{y}_i|\ldots)\Psi_1(\mathbf{y}_i|\ldots)^T\right]\\ &\quad\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\Psi_1(\mathbf{y}_i|\mathbf{z}_i^*,\mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\tau})\right]^{-T}\end{aligned}$$

- The Acov can be easily obtained using a single run of a Monte Carlo simulation.
- The efficiency loss of $\hat{\boldsymbol{\theta}}$ with respect to the (true) MLE is about 2% (based on simulations).

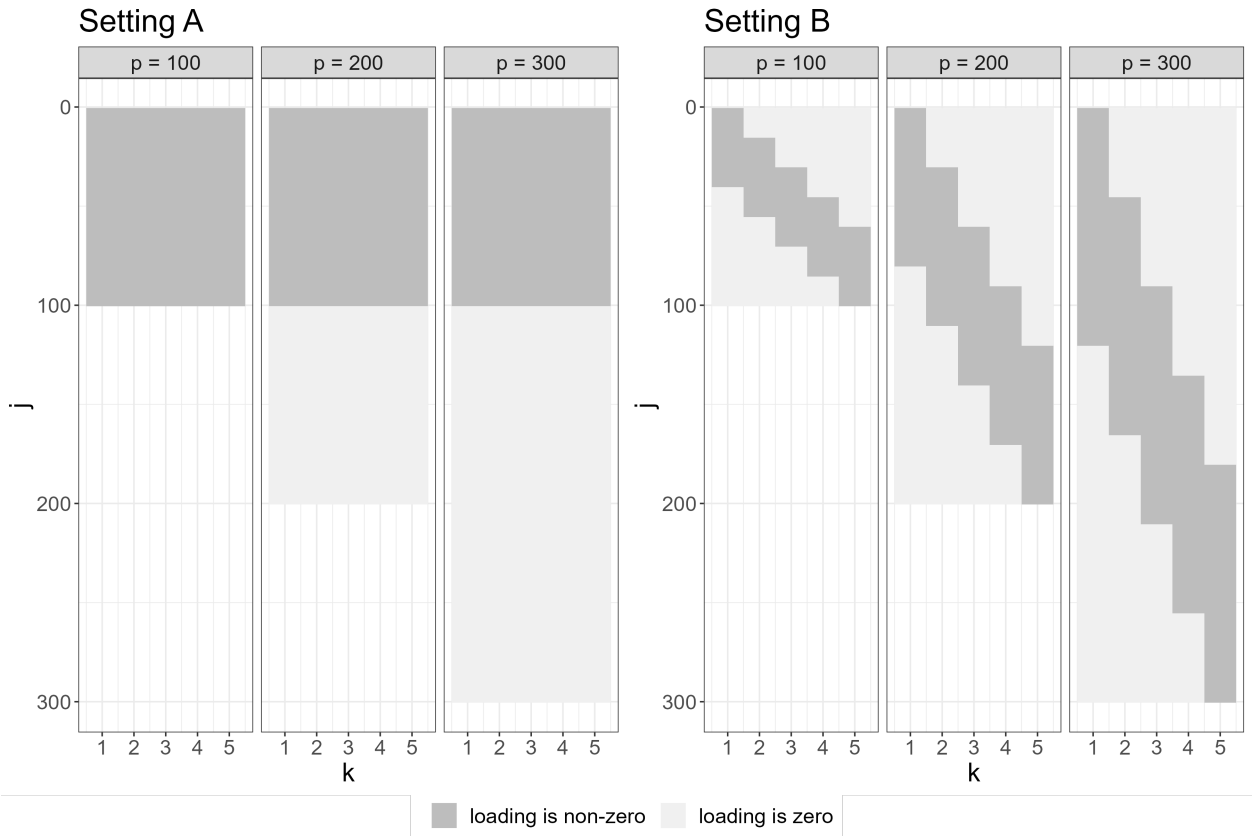
Simulations I: low p Binary

Objective: compare Mean Procruste Errors (MPE) and Estimation Bias (ES) of different estimators for a Binary GLLVM, with $p = 40$, $q = 2$ and $n = 100, 200, 500, 1000$. The estimators are the two regularized ones (gmf and mirtjml), the PRIME and SPRIME, and the MLE based on Gaussian Quadrature approximations implemented in the `R` package `ltm`.



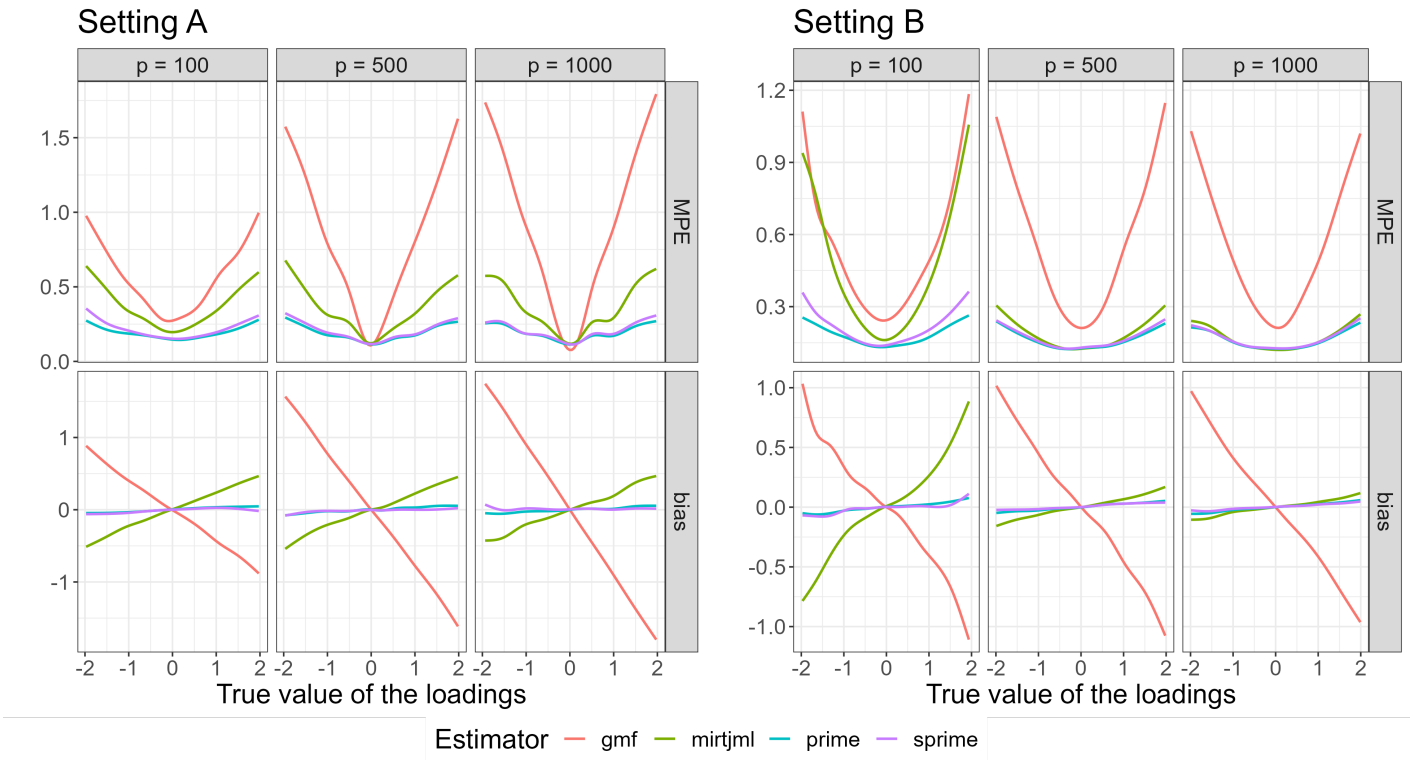
Simulations II: high p Binary

Objective: compare Mean Procruste Errors (MPE) and Estimation Bias (ES) of different estimators for a Binary GLLVM, with $p = 100$ to $p = 1000$, $q = 5$ and $n = 500$. The estimators are the two regularized ones (gmf and mirtjml), the PRIME and SPRIME. We consider two settings for the type of sparcity.



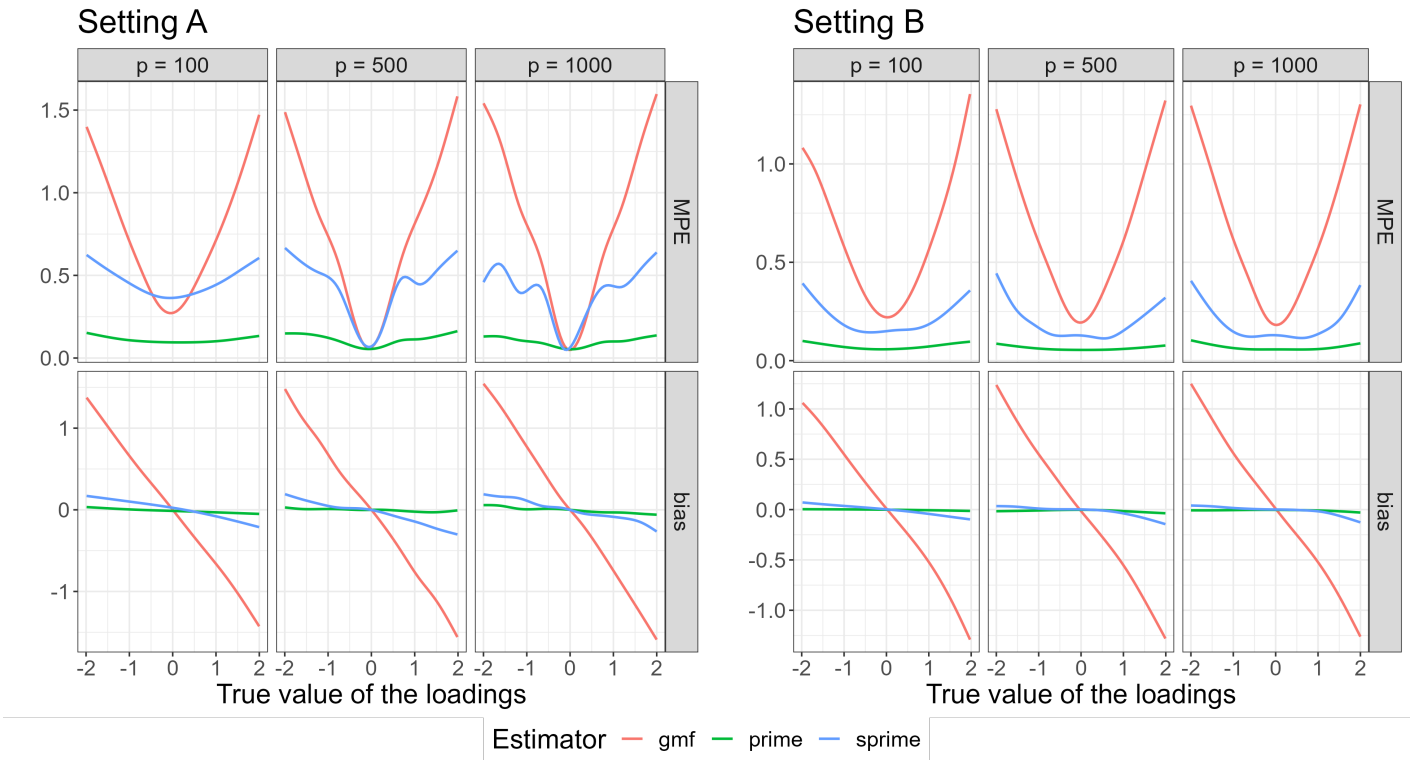
Simulations II: high p Binary

Objective: compare Mean Procruste Errors (MPE) and Estimation Bias (ES) of different estimators for a Binary GLLVM, with $p = 100$ to $p = 1000$, $q = 5$ and $n = 500$. The estimators are the two regularized ones (gmf and mirtjml), the PRIME and SPRIME. We consider two settings for the type of sparcity.



Simulations III: high p Poisson

Objective: compare Mean Procruste Errors (MPE) and Estimation Bias (ES) of different estimators for a Poisson GLLVM, with $p = 100$ to $p = 1000$, $q = 5$ and $n = 500$. The estimators are the regularized gmf, the PRIME and SPRIME. We consider two settings for the type of sparcity.



Extensions

- In principle, the PRIME and SPRIME can be used for any latent variable model, for which the marginalization of the likelihood function implies approximations that render the numerical aspects infeasible.
- In particular, they can be used for constrained GLLVM (i.e. confirmatory analysis).
- They can also be used for panel data for which the dimension of the parameters increase with time (under investigation).
- An `R` package is under construction...

Thank you very much for your attention!

Any questions?

More info...

- Maria-Pia.VictoriaFeser@unige.ch
- guillaume.blanc@jacobscenter.uzh.ch
- Stephane.Guerrier@unige.ch
- silvia.cagnone@unibo.it

S. Bianconcini and S. Cagnone. Estimation of generalized linear latent variable models via fully exponential Laplace approximation. *Journal of Multivariate Analysis*, 112:183–193, 2012.

J. F. Bjørnstad. On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91:791–806, 1996.

J. R. Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

S. Cagnone and P. Monari. Latent variable models for ordinal data by using the adaptive quadrature approximation. *Computational Statistics*, 28(2):597–619, 2013.

O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B*, 71(3):593–613, 2009.

Y. Chen, X. Li, and S. Zhang. Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115:1756–1770, 2020.

F. Dai, S. Dutta, and R. Maitra. A matrix-free likelihood method for exploratory factor analysis of high-dimensional gaussian data. *Journal of Computational and Graphical Statistics*, 29:675–680, 2020.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–22, 1977.

V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

S. Guerrier, E. Dupuis-Lozeron, Y. Ma, and M.-P. Victoria-Feser. Simulation-based bias correction methods for complex models. *Journal of the American Statistical Association*, 114:146–157, 2019.

P. Huber, E. Ronchetti, and M.-P. Victoria-Feser. Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B*, 66:893–908, 2004.

P. J. Huber. Robust estimation of a location parameter. *The Annals Mathematics Statistics*, 35:73–10, 1964.

F. K. C. Hui, D. I. Warton, J. T. Ormerod, V. Haapaniemi, and S. Taskinen. Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, 26:35–43, 2017.

L. Kidzinski, F. K. C. Hui, D. I. Warton, and T. J. Hastie. Generalized matrix factorization: efficient algorithms for fitting generalized linear latent variable models to large data arrays. *Journal of Machine Learning Research*, 2023. To appear.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

Y. Lee and J. A. Nelder. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, 58:619–678, 1996.

Y. Lee and J. A. Nelder. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88:987–1006, 2001.

Y. Lee and J. A. Nelder. Likelihood inference for models with unobservables: Another view. *Statistical Science*, 24:255–269, 2009.

Y. Lee, J. A. Nelder, and Y. Pawitan. *Generalized linear models with random effects: Unified analysis via H-likelihood*. Chapman and Hall, 2006.

J. Niku, W. Brooks, R. Herliansyah, F. K. C. Hui, S. Taskinen, and D. I. Warton. Efficient estimation of generalized linear latent variable models. *PloS one*, 14(5):e0216129, 2019.

D. Robertson and J. Symons. Maximum likelihood factor analysis with rank-deficient sample covariance matrices. *Journal of Multivariate Analysis*, 98:813–828, 2007.

R. Sundberg and U. Feldmann. Exploratory factor analysis—parameter estimation and scores prediction with high-dimensional data. *Journal of Multivariate Analysis*, 148:49–59, 2016.

G. C. G. Wei and M. A. Tanner. A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85:699–704, 1990.

J. Wu and P. M. Bentler. Application of h-likelihood to factor analysis models with binary response data. *Journal of Multivariate Analysis*, 106:72–79, 2012.