Pairwise Stochastic Approximation for Confirmatory Factor Analysis of Categorical Data

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

[1]University of Udine
[2]London School of Economics & Political Science

Abstract

Pairwise likelihood is a limited information method widely used to estimate latent variable models, including factor analysis of categorical data. It can often avoid evaluating high-dimensional integrals and, thus, is computationally more efficient than relying on the full likelihood. Despite its computational advantage, the pairwise likelihood approach can still be demanding for large-scale problems that involve many observed variables. We tackle this challenge by employing an approximation of the pairwise likelihood estimator, which is derived from an optimisation procedure relying on stochastic gradients. The stochastic gradients are constructed by subsampling the pairwise log-likelihood contributions, for which the subsampling scheme controls the per-iteration computational complexity. The stochastic estimator is shown to be asymptotically equivalent to the pairwise likelihood one. However, finite sample performances can be improved by compounding the sampling variability of the data with the uncertainty introduced by the subsampling scheme. We demonstrate the performance of the proposed method using simulation studies and two real data applications.

Keywords: Item factor analysis, structural equation models, composite likelihood, asymptotic normality, stochastic gradient descent

1 Introduction

Structural equation models (SEMs) – a general family of latent variable models – are widely used to analyze multiple observed variables from social surveys and administrative data. These models jointly treat observed variables as measures of unobserved (latent) constructs, where the constructs often receive substantive interpretations such as attitudes, beliefs, and abilities. When used in a confirmatory manner, statistical analysis based on SEMs can reveal the inter-relationships among latent constructs and observed variables, based on which specific hypotheses driven by social or economic theory can be tested. Factor models are a special case of SEMs and also serve as a building block for more general SEMs. With a confirmatory factor model, one can postulate certain relationships among the observed and latent variables by assuming a pre-specified pattern for certain model parameters (factor loadings). Researchers can test whether the postulated relationships exist by drawing statistical inference under this model, which is often known as confirmatory factor analysis (CFA). We refer the readers to Bartholomew, Steele, Moustaki, and Galbraith (2008) for a comprehensive review of SEM and CFA.

Questionnaire items in survey data are often categorical (ordinal or nominal). In the SEM literature, one common approach for analyzing categorical variables with factor models is the underlying variable approach, which assumes categorical variables to be generated by underlying continuous variables (e.g., see Jöreskog, 1990, 1994; Lee, Poon, and Bentler, 1990, 1992; Muthén, 1984). Under this modelling framework, full maximum likelihood is computationally challenging when there are many items because evaluating the likelihood function requires calculating integrals with respect to the respective underlying variables. Limited information estimation methods have been proposed in the literature to tackle this computational challenge,

such as the three-stage least squares estimation method (Jöreskog, 1990, 1994; Muthén, 1984) and composite likelihood methods. Three-stage methods often suffer from instability issues when the sample size is small, or the number of observed variables is large. This is due to the fact that statistical inference with these methods requires estimating a polychoric correlation matrix for the observed variables and further computing the asymptotic covariance matrix of the polychoric correlations. Such asymptotic covariance matrix, in fact, is often unstable.

Composite likelihood methods (see e.g., Besag, 1974; Lindsay, 1988; Cox and Reid, 2004; Varin, 2008; Varin, Reid, and Firth, 2011) are a general class of inference functions particularly suitable when the full likelihood is too expensive to compute. The main idea behind this approach is to construct a pseudo-likelihood from marginal or conditional distributions of lower-order margins of the data. The most popular approach for SEMs with categorical data is the pairwise likelihood function, which employs information from all possible bivariate margins of the data. Under the underlying variable framework, pairwise likelihood estimators have been proposed under different model settings, including the estimation of thresholds and polychoric correlations of ordinal data (de Leon, 2005), factor models for ordinal (Jöreskog and Moustaki, 2001; Katsikatsou, Moustaki, Yang-Wallentin, and Jöreskog, 2012) and mixed data (Katsikatsou, 2013), SEMs for longitudinal data (Vasdekis, Cagnone, and Moustaki, 2012) and random effects models (Vasdekis, Rizopoulos, and Moustaki, 2014).

Despite its advantages, pairwise likelihood estimation can still be computationally demanding for large-scale problems that involve many observed variables. An attempt has been made to reduce that computational burden by proposing a sampling method for selecting pairs based on their contribution to the total variance from all pairs (Papageorgiou and Moustaki, 2018). However, while the method manages to select relatively small subsets of important pairs (in terms of variance contribution), the computational saving is partially offset by the complexity of the sampling scheme, such that computation times are reported to improve only about 20% on average.

This paper tackles the computational challenge of estimating large-scale pairwise likelihood SEMs from a different perspective. Instead of sampling the pairs before the estimation as in Papageorgiou and Moustaki (2018), we iteratively subsample them along the optimisation as recently suggested in Alfonzetti, Bellio, Chen, and Moustaki (2023). In particular, such a strategy can be framed as a stochastic approximation method, where each gradient iteration is constructed stochastically based on a new small subset of the pairs. While the complexity-per-iteration drastically reduces, such an approach still allows to account for all the pairs as the optimisation proceeds. Differently from Alfonzetti et al. (2023), we show that our stochastic estimator is always asymptotically equivalent to the conventional pairwise likelihood estimator as long as both the number of iterations and the sample size go to infinity, regardless of their relative divergence rate. For ease of exposition, we focus on the CFA setting for ordinal data. However, our proposal can easily be generalised to other SEM settings, including those considered in the above-mentioned pairwise likelihood literature.

The paper is organised as follows. Section 2 reviews the confirmatory factor model for categorical data and statistical inference based on pairwise likelihood. Section 3 proposes the stochastic approximation method to tackle the computational challenge with large-scale pairwise likelihood estimation. We illustrate the capability of the proposed method using simulation studies in Section 4 and two real data applications in Section 5. We conclude with a discussion in Section 6.

2 Problem Setting

2.1 Confirmatory Factor Analysis of Categorical Data

Consider n respondents answering p ordinal or binary items. With the vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)^\top$ we denote the responses from a single respondent, where $Y_i \in \{1, ..., m_i\}$ has m_i categories, $i = 1, \ldots, p$. We say an item is binary if $m_i = 2$ and ordinal if $m_i > 2$. The underlying variable approach assumes the categorical variables in $\mathbf{Y}$ to be generated by the partial observation of a set of underlying continuous variables $\mathbf{Y}^* = (Y_1^*, ..., Y_p^*)^\top$, where the random vector $\mathbf{Y}^*$ is assumed normally distributed. More specifically, the connection between an

observed variable Y_i and the underlying continuous variable Y_i^* is

$$Y_i = c \text{ if and only if } \tau_{c-1}^{(i)} \leq Y_i^* < \tau_c^{(i)}, c = 1, \cdots, m_i, \tag{1}$$

where $\tau_k^{(i)}$ is the k^{th} threshold of variable Y_i, $k = 0, ..., m_i$, satisfying $-\infty = \tau_0^{(i)} < \tau_1^{(i)} < \ldots < \tau_{m_i-1}^{(i)} < \tau_{m_i}^{(i)} = +\infty$. Since only ordinal information is available, the distribution of Y_i^* is determined up to a linear transformation. To ensure model identifiability, it is convenient to assume each Y_i^* to follow a standard normal distribution.

Then, a factor model for categorical data $\mathbf{Y}$ is defined by a classical linear factor model for $\mathbf{Y}^*$, taking the form

$$\mathbf{Y}^* = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{2}$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})_{p\times q}$ is a $p \times q$ matrix of factor loadings, $\boldsymbol{\xi}$ is a $q \times 1$ vector of latent variables, and $\boldsymbol{\delta}$ is a p-dimensional vector of measurement errors. It is assumed that $\boldsymbol{\xi}$ follow a normal distribution $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}_\xi)$, where $\boldsymbol{\Sigma}_\xi$ is a $q \times q$ matrix. The diagonal entries of $\boldsymbol{\Sigma}_\xi$ are set to one to ensure that the scale of the latent variables is identified. In addition, the measurement error vector $\boldsymbol{\delta}$ is assumed to be independent of $\boldsymbol{\xi}$ and follows a normal distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\delta)$ with the covariance matrix $\boldsymbol{\Sigma}_\delta$ being a $p \times p$ diagonal matrix. The above model assumptions imply that $\text{Cov}(\boldsymbol{\xi}, \boldsymbol{\delta}) = \mathbf{0}$ and $\boldsymbol{\Sigma}_\delta = \boldsymbol{I}_p - \text{diag}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}^\top)$, where $\text{diag}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}^\top)$ denotes a diagonal matrix whose diagonal entries are the same as those of $\boldsymbol{\Lambda}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}^\top$ and $\boldsymbol{I}_p$ the p-dimensional identity matrix.

For CFA, zero constraints are imposed on the loading parameters in $\boldsymbol{\Lambda}$, typically determined by prior knowledge or hypothesis about the measured constructs. More specifically, λ_{ij} is set to zero if one postulates that item i does not directly measure latent variable j. A sensible CFA setting avoids the rotational indeterminacy of the latent variables, which further ensures the identification of the model (Anderson and Rubin, 1956).

For ease of exposition, in the rest of the paper, let us denote with Θ the parameter space, and $\boldsymbol{\theta} \in \Theta$ the generic vector of unknown model parameters collecting free non-redundant elements in $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_\xi$, as well as the free thresholds.

2.2 Inference Based on Pairwise Likelihood

Suppose that data from n respondents are collected, where data from observation l is denoted by $\mathbf{y}_l = (y_{l1}, ..., y_{lp})^\top$, $l = 1, ..., n$. The objective is to draw statistical inference on $\boldsymbol{\theta}^* \in \Theta$, the true parameter vector. Theoretically, it is a simple parametric inference problem falling under the classical maximum likelihood estimation framework. Unfortunately, maximum likelihood estimation is computationally very intensive. Let $\mathcal{R} = \{(c_1, ..., c_p)^\top : c_i \in \{1, ..., m_i\}\}$ be the set containing all the $R = \prod_{i=1}^p m_i$ possible response patterns. Then, the full log-likelihood takes the form $l(\boldsymbol{\theta}) = \sum_{\mathbf{c}\in\mathcal{R}} n_\mathbf{c} \log \pi_\mathbf{c}(\boldsymbol{\theta})$ with $n_\mathbf{c}$ the observed frequency of pattern $\mathbf{c} = (c_1, ..., c_p)^\top \in \mathcal{R}$ and

$$\pi_\mathbf{c}(\boldsymbol{\theta}) = P(Y_1 = c_1, \ldots, Y_p = c_p; \boldsymbol{\theta}) = \int_{\tau_{c_1-1}^{(1)}}^{\tau_{c_1}^{(1)}} \cdots \int_{\tau_{c_p-1}^{(p)}}^{\tau_{c_p}^{(p)}} \phi_p(\boldsymbol{y}^*; \boldsymbol{\Lambda}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}^\top + \boldsymbol{\Sigma}_\delta) d\boldsymbol{y}^*.$$

Thus, computing $l(\boldsymbol{\theta})$ requires evaluating potentially R p-dimensional integrals. It follows that, for a moderately large p, optimizing the full likelihood becomes computationally unaffordable due to the high complexity of evaluating each $\pi_c(\boldsymbol{\theta})$.

A pairwise likelihood approach has been proposed in Katsikatsou et al. (2012) to draw statistical inference when the full likelihood is too expensive to compute. More specifically, the (unweighted) pairwise log-likelihood takes the form

$$\text{pl}(\boldsymbol{\theta}) = \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} n_{c_i c_j}^{(ij)} \log\{\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})\}, \tag{3}$$

where $n_{c_i c_j}^{(ij)} = \sum_{l=1}^{n} \mathbf{1}_{\{y_{li}=c_i, y_{lj}=c_j\}}$ is the observed frequency of the bivariate pattern given by responses c_i on Y_i and c_j on Y_j, while $\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})$ represents the model-implied probability of that specific response pattern when evaluated at parameters $\boldsymbol{\theta}$, namely

$$\begin{aligned}\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}) &= P\left(Y_i = c_i, Y_j = c_j; \boldsymbol{\theta}\right) \\ &= \Phi_2\left(\tau_{c_i}^{(i)}, \tau_{c_j}^{(j)}; \rho_{ij}(\boldsymbol{\theta})\right) - \Phi_2\left(\tau_{c_i}^{(i)}, \tau_{c_j-1}^{(j)}; \rho_{ij}(\boldsymbol{\theta})\right) \\ &- \Phi_2\left(\tau_{c_i-1}^{(i)}, \tau_{c_j}^{(j)}; \rho_{ij}(\boldsymbol{\theta})\right) + \Phi_2\left(\tau_{c_i-1}^{(i)}, \tau_{c_j-1}^{(j)}; \rho_{ij}(\boldsymbol{\theta})\right), \end{aligned} \tag{4}$$

The notation $\Phi_2\left(a, b; \rho\right)$ refers to the zero-mean bivariate cumulative normal distribution with correlation ρ evaluated at the point (a, b), while

$$\rho_{ij}(\boldsymbol{\theta}) = \left(\lambda_{i1}, ..., \lambda_{iq}\right) \boldsymbol{\Sigma}_{\xi} \left(\lambda_{i1}, ..., \lambda_{iq}\right)^{\top}.$$

Recall that λ_{ik} is a loading parameter in $\boldsymbol{\Lambda}$, and $\boldsymbol{\Sigma}_{\xi}$ is a $q \times q$ covariance matrix of the latent variables. Note that the pairwise likelihood only involves two-dimensional integrals. Thus, it is computationally more feasible than the full likelihood. The maximum pairwise likelihood estimator is then defined as

$$\hat{\boldsymbol{\theta}}_{PML} = \arg\max_{\boldsymbol{\theta}} \ \mathrm{pl}(\boldsymbol{\theta}).$$

Following the theory for composite likelihood (Varin et al., 2011), $\hat{\boldsymbol{\theta}}_{PML}$ is consistent and asymptotically normal. That is, as the sample size n goes to infinity, the PML estimator converges in distribution to a multivariate normal random variable centered in $\boldsymbol{\theta}^*$, with asymptotic covariance matrix depending on $\boldsymbol{H} = E_{\boldsymbol{Y}}\left\{-\nabla^2 \mathrm{pl}(\boldsymbol{\theta}^*)/n\right\}$ and $\boldsymbol{J} = \mathrm{Var}_{\boldsymbol{Y}}\left\{\nabla \mathrm{pl}(\boldsymbol{\theta}^*)/n\right\}$. Namely,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{PML} - \boldsymbol{\theta}^*\right) \xrightarrow[n]{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}\right). \tag{5}$$

More generally, one can add a pre-specified non-negative weight w_{ij} to each item pair (i, j) to obtain a weighted pairwise log-likelihood; see Varin et al. (2011) for more details. Furthermore, we refer the readers to Katsikatsou and Moustaki (2016) for more results on estimating and testing SEMs with pairwise likelihood.

Finally, note that the original formulation of the PML estimator for ordinal factor models proposed in Katsikatsou et al. (2012) directly estimates the entries of $\boldsymbol{\Sigma}_{\xi}$, without ensuring it to be a proper correlation matrix. We refer to this unconstrained version of the PML estimator as the UPML estimator. This paper adopts a suitable reparameterisation of $\boldsymbol{\Sigma}_{\xi}$ based on the transformation proposed in Lewandowski, Kurowicka, and Joe (2009), which ensures $\boldsymbol{\Sigma}_{\xi}$ to be symmetric and strictly positive definite. Such an approach maps $\boldsymbol{\Sigma}_{\xi}$ to a set of free $q(q-1)/2$ parameters to be estimated without explicit constraints. Accounting for the specific nature of $\boldsymbol{\Sigma}_{\xi}$ via reparameterisation is needed because the UPML estimate $\hat{\boldsymbol{\Sigma}}_{\xi}$ may not be a proper correlation matrix, and consequently, the dependence relationships among the latent variables, which are often of substantive interest, cannot be assessed. The mathematical details about the reparameterisation are left in Appendix B.1.

3 Inference based on Pairwise Stochastic Approximation

Although the pairwise likelihood is computationally more feasible than the full likelihood, it still suffers from a high computational cost when the number of variables, p, is large. This is because the complexity of the pairwise log-likelihood is dominated by the number of item pairs, which is $P = p(p-1)/2$. When p is too large, optimising the pairwise log-likelihood can become computationally unaffordable with a gradient-based optimisation routine. For example, with $p = 50$ binary items, the number of bivariate probabilities involved in a single evaluation of the score function $\nabla \mathrm{pl}(\boldsymbol{\theta})$ becomes as high as $4P = 4,900$, which is already computationally demanding and, overall, needs to be carried out at each iteration of the optimisation.

To tackle this computational issue, we propose a stochastic estimator that replaces $\nabla \text{pl}(\boldsymbol{\theta})$ in the optimisation with a computationally convenient stochastic approximation by subsampling the pairwise components in (3), while being asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{\text{PML}}$. Let us define the $p \times p$-dimensional matrix $\boldsymbol{W}$, with generic element $W_{ij} \in \{0, 1\}$. Each element W_{ij} corresponds to a specific pair of items. Suppose now to draw ν out of the possible P pairs of items without replacement. If the pair (i, j) has been drawn, its corresponding weight is set to one. Otherwise, its weight is set to zero. Then, we can construct an unbiased approximation of $\nabla \text{pl}(\boldsymbol{\theta})$ via

$$S(\boldsymbol{\theta}; \boldsymbol{W}) = \frac{P}{\nu} \sum_{i<j} W_{ij} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} \frac{n_{c_i c_j}^{(ij)}}{\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta})} \nabla \pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}). \tag{6}$$

Note that $E_{\boldsymbol{W}} \{S(\boldsymbol{\theta}; \boldsymbol{W})\} = \nabla \text{pl}(\boldsymbol{\theta})$, since $E_{\boldsymbol{W}} (W_{ij}) = \nu / P$ for all is and js. Thus, $S(\boldsymbol{\theta}; \boldsymbol{W})$ can be safely used as a stochastic gradient in a stochastic approximation algorithm (Robbins and Monro, 1951; Polyak and Juditsky, 1992; Ruppert, 1988).

Let T_n be the number of iterations in the algorithm, which diverges as n goes to infinity, and B an initial burn-in period. Let $\eta_0 > 0$ be the initial choice of the stepsize, with decreasing scheduling $\eta_t = \eta_0 (1 + a \eta_0 t)^{-3/4}$, $a > 0$, in accordance with Xu (2011). We construct an estimator $\bar{\boldsymbol{\theta}}$ with the following algorithm:

1. Given ν, η_0, T_n, B, $\boldsymbol{\theta}_0$, for $t = 1, \ldots, T_n$, alternate:
   (a) *Sampling Step:* Sample $\boldsymbol{W}_t$ by drawing ν out of P pairwise components;
   (b) *Approximation Step:* Compute $\boldsymbol{S}_t = S(\boldsymbol{\theta}; \boldsymbol{W}_t)$ via (6);
   (c) *Update Step:* Update the current estimate via $\tilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_{t-1} - \eta_t \boldsymbol{S}_t$;
   (d) *Projection Step:* Ensure the estimate to be in Θ with $\boldsymbol{\theta}_t = \prod(\tilde{\boldsymbol{\theta}}_t)$.
2. *Trajectories averaging:* Compute $\bar{\boldsymbol{\theta}} = \frac{1}{T_n - B} \sum_{t=B+1}^{T_n} \boldsymbol{\theta}_t$.

At each iteration t, step (a) samples a new set of weights $\boldsymbol{W}_t$. Then, with the subset of pairs identified by $\boldsymbol{W}_t$, step (b) constructs the new stochastic gradient $\boldsymbol{S}_t$. Successively, step (c) updates the parameter estimates, and step (d) ensures they remain within the parameter space. The step (d) is needed to guarantee that $\rho(\boldsymbol{\theta})_{ij}$ in (4) is a proper correlation. In practice, it consists of adequately rescaling the loadings in order to constrain their scale. See Appendix B for further details. Finally, after T_n iterations, the algorithm computes the final estimate $\bar{\boldsymbol{\theta}}$ by averaging the trajectories of the estimates along the optimisation, ignoring an initial burn-in period aimed at limiting the influence of $\boldsymbol{\theta}_0$ in the computation of $\bar{\boldsymbol{\theta}}$. Such technique is also known as Ruppert-Polyak averaging, from Polyak and Juditsky (1992) and Ruppert (1988).

This final averaging step plays a crucial role in defining the asymptotic behaviour of the stochastic estimator because it allows a central limit theorem to characterise the demeanour of $\bar{\boldsymbol{\theta}}$ as n and T_n diverge. In particular, it can be shown that $\bar{\boldsymbol{\theta}}$ is consistent and asymptotically normally distributed, i.e.

$$\boldsymbol{\Omega}_n^{-1/2} (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow[n]{d} \mathcal{N}(\mathbf{0}; \boldsymbol{I}_d), \tag{7}$$

with an asymptotic covariance matrix defined by

$$\boldsymbol{\Omega}_n = \frac{1}{n} \boldsymbol{H}^{-1} \boldsymbol{J} \boldsymbol{H}^{-1} + \frac{1}{T_n} \boldsymbol{H}^{-1} \boldsymbol{V}_n \boldsymbol{H}^{-1}. \tag{8}$$

The decomposition of $\boldsymbol{\Omega}_n$ in two terms, as outlined in (8), is particularly illustrative to understand how $\bar{\boldsymbol{\theta}}$ behaves asymptotically. The first term on the right-hand-side in (8) denotes the sampling variability of the data, and the sample size scales it. It coincides with the conventional variance of $\hat{\boldsymbol{\theta}}_{\text{PML}}$ outlined in (5). Thus, it is the component of the variance that diminishes as the sample size grows. The second term, instead, is scaled by T_n and represents the additional variability due to the randomness of the optimisation. As reasonable, it vanishes as the optimisation proceeds. At the core of this second term, we find the matrix

$\boldsymbol{V}_n$, which is connected to the variance of the stochastic gradient in (6) and, thus, to the distribution of the weight matrix $\boldsymbol{W}$.

To get an intuition about the behaviour of $\boldsymbol{V}_n$, let us focus on the role of the sufficiency data reduction in characterising the distribution of $\boldsymbol{W}$ in (6). Having $\boldsymbol{\theta}$ depending on the data only through the set of $n_{c_i c_j}^{(ij)}$ allows the stochastic weights W_{ij} not to be indexed in $l = 1, \ldots, n$, because computing the gradient on a single observation has the same cost of computing it on the whole sample. While it might seem trivial, such a structure has deep consequences on the behaviour of the optimisation noise. It follows that the stochastic gradient $S(\boldsymbol{\theta}; \boldsymbol{W})$ in (6), while considering only a subset of the available pairs, still accounts for the full sample at each iteration. Hence, the amount of information $S(\boldsymbol{\theta}; \boldsymbol{W})$ accounts for at each iteration grows with growing sample size, and thus its average variability decreases as n diverges. It can be shown that, with the stochastic gradient defined in (6), the matrix $\boldsymbol{V}_n$ in (8) takes the form

$$\boldsymbol{V}_n = \frac{1}{n} \left\{ \frac{P(P-\nu)}{\nu(P-1)} \boldsymbol{H} - \frac{P-\nu}{\nu(P-1)} \boldsymbol{J} \right\}. \tag{9}$$

The variability injected in the optimisation by the randomness of $\boldsymbol{W}$ corresponds to a weighted average of the matrices $\boldsymbol{H}$ and $\boldsymbol{J}$, with weights determined by the values of ν and P. When $\nu = P$, no randomness is left in $\boldsymbol{W}$, and the weights of $\boldsymbol{H}$ and $\boldsymbol{J}$ in (9) collapse to zero. However, regardless of the choice of ν, $\boldsymbol{V}_n$ vanishes naturally asymptotically because of the $1/n$ factor implied by the sufficiency reduction. It follows that the second term on the right-hand-side in (8) decreases faster than the first because both n and T_n scale it.

Therefore, it can be stated that $\bar{\boldsymbol{\theta}}$ in (7) is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{\text{PML}}$ in (5). Nevertheless, in finite sample settings, the optimisation noise may still be non-negligible, such that it is always beneficial to account for all the terms in (8). Note that the variance component in (8) referring to the optimisation noise only demands an estimate of H and J to be evaluated as the sampling variability. In Section 4, we show that considering the optimisation randomness improves finite sample performance of inference with $\bar{\boldsymbol{\theta}}$.

For a detailed derivation of (7) through (9), see Appendix C. Note that the technical proof borrows from the results in Alfonzetti et al. (2023). However, their setting assumes the weights used to construct $S(\boldsymbol{\theta}; \boldsymbol{W})$ to be indexed both in the pairs (i, j) and in the observations $l = 1, \ldots, n$. In other words, by fixing ν, their set of weights only accounts for a fixed amount of statistical information, while in the case of (6), this amount grows with n, as discussed previously. Such a difference results in our optimisation noise in (8) disappearing much faster than what happens in Alfonzetti et al. (2023) because of being simultaneously scaled by T_n and n. The main implication is that, in our case, the stochastic estimator $\bar{\boldsymbol{\theta}}$ is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{\text{PML}}$, whatever the relative divergence rate of n and T_n, while in their case this happens only if $n/(T_n + n)$ goes to 0 as n goes to infinity.

4 Simulation experiments

4.1 Setup

A simulation study is conducted to examine the performance of the proposed estimator. Let the number of ordinal items be fixed to $p = 40$, while the number of latent variables, $q \in \{4, 8\}$, and the sample size, $n \in \{1000, 2000, 5000\}$ vary across the experimental settings. The simulated data consists of ordinal variables with $m_i = 4$ categories, and the true thresholds for all variables are fixed at $\tau_i = (-1.2, 0, 1.2)$, with $i = 1, \ldots, p$. Each combination of p, q and n accounts for $S = 1000$ replications. The true loading matrices are generated to have a simple structure (one loading per item) with additional $q - 1$ cross-loadings. For illustration purposes, the estimation proceeds by drawing only $\nu = 8$ pairs per iteration out of the possible $P = 780$. The initial stepsize η_0 used in the stochastic updates is fixed at $\eta_0 = 10^{-2}$. An initial burn-in period runs until the iteration n, so only subsequent iterations directly enter the computation of $\bar{\boldsymbol{\theta}}$. See Appendix C for further details about the simulation settings and the setup of the stochastic algorithm.

4.2 Performance criteria

We are interested in evaluating the properties of the proposed estimator both in terms of pointwise convergence to $\boldsymbol{\theta}^*$ and variability around its target. Since the number of parameters is large, instead of showing results for each parameter estimate, we average performance criteria across all estimated parameters of the same type (factor loadings and factor correlations). For brevity reasons, we omit results for thresholds since they are typically of less interest to practitioners.

The convergence of the estimator is evaluated by computing the average mean squared error (MSE) along the optimisation for $t = 1, \ldots, T_n$, with

$$MSE_t = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{k} \sum_{j=1}^{k} \left(\bar{\boldsymbol{\theta}}_{sjt} - \boldsymbol{\theta}^* \right)^2,$$

where S here is the number of replicates, k is the total number of parameters of the same type, $\bar{\boldsymbol{\theta}}_{sjt}$ is the estimate of the j^{th} parameter for the replication s at the iteration t. Computational times are reported in Appendix C.3, while additional sensitivity experiments can be found in Appendix C.4. Similarly, we want to investigate how inference performance about $\boldsymbol{\theta}^*$ evolves along the optimisation. To this end, we compute confidence intervals for $\bar{\boldsymbol{\theta}}$ and their empirical coverage levels for different T_n. In particular, we rely on the asymptotic covariance matrix in (7) assuming $\boldsymbol{V}_n = \mathbf{0}$ and its finite sample counterpart accounting for both the terms outlined in (8). To estimate the matrices $\boldsymbol{H}$ and $\boldsymbol{J}$, we use their conventional empirical estimators (e.g., see Section 5.1 in Varin et al., 2011).

4.3 Results

First, let us start by showing the convergence of $\bar{\boldsymbol{\theta}}$ towards $\boldsymbol{\theta}^*$ when both n and T_n diverge. As expected, Figure 1 outlines that the average MSE performance for correlations and loadings improves both with n and T_n increasing. The MSE typically drops rapidly at the beginning of the optimisation and then slows down its decrease during the remaining iterations. Furthermore, with n and T_n fixed, results are better when $q = 4$ rather than $q = 8$. This is expected because of the larger number of parameters involved in the latter case. As a benchmark, we also compute $\hat{\boldsymbol{\theta}}_{\text{PML}}$, whose MSE performance is denoted by horizontal dashed lines in Figure 1. In most cases, $\bar{\boldsymbol{\theta}}$ almost overlaps $\hat{\boldsymbol{\theta}}_{\text{PML}}$ in terms of MSE even before reaching $2n$ iterations.

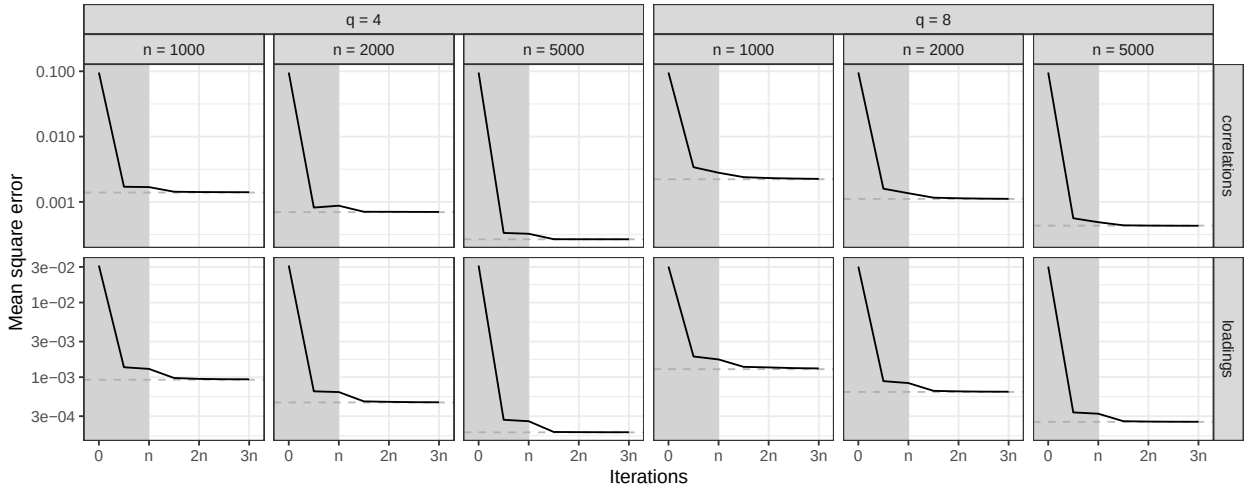

Figure 1: Average Mean Squared Error performance along the optimisation of loadings and latent correlations for $p = 40$, $q \in \{4, 8\}$, and $n \in \{1000, 2000, 5000\}$. Grey areas denote the initial burn-in period. Horizontal dashed lines refer to $\hat{\boldsymbol{\theta}}_{\text{PML}}$ performance.

Second, we investigate the performance of confidence intervals built from (7), i.e. we track their observed coverage levels when both the sample size and the number of iterations increase. Figure 2 shows the boxplots of the empirical coverage grouped by parameter type. Furthermore, it also compares the performance obtained by accounting only for the sampling variability of the data, the first term on the right-hand-side in (8), with the one corrected by the optimisation noise, the last summand in (8). Results show that differences between the two methods taper both when the sample size increases and when the number of iterations diverges. Such behaviour is expected since, as seen in Section 3, the optimisation noise vanishes simultaneously with n and T_n. However, the advantage of computing confidence intervals corrected for the optimisation noise is evident when n or T_n are small, that is when the variance of the stochastic gradients is still non-negligible. As in the case of Figure 1, coverage levels are typically better when $q = 4$ rather than $q = 8$. Nevertheless, both cases tend to align to the nominal level when n and T_n increase.

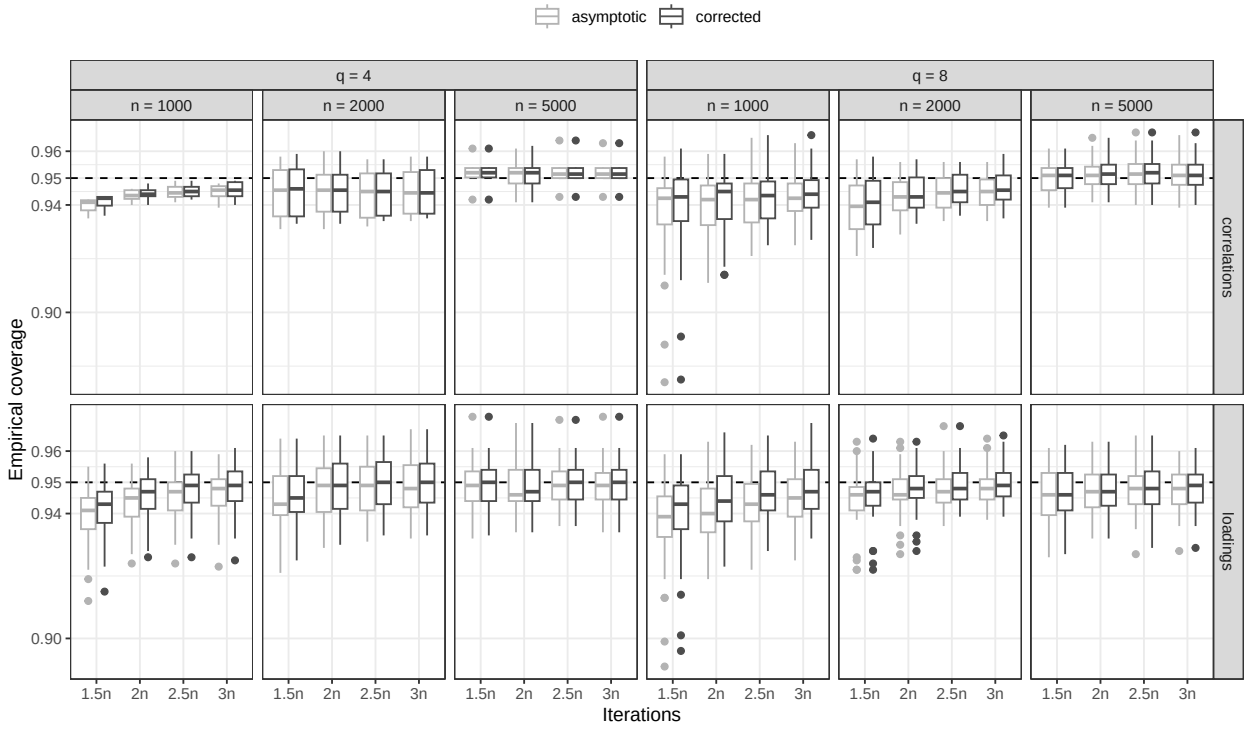

Figure 2: Empirical coverage levels for confidence intervals for $p = 40$, $q \in \{4, 8\}$, $n \in \{1000, 2000, 5000\}$ and $T_n \in \{1.5n, 2n, 2.5n, 3n\}$. The horizontal dashed line denotes the target nominal coverage level, i.e. 95%. The gray scale distinguishes confidence intervals constructed using only the first term in (8) (`asymptotic`) from the ones considering both terms in (8) (`corrected`).

5 Real applications

We use two datasets to demonstrate the benefits of combining pairwise likelihood and stochastic optimisation. The first is from the Big Five personality test, and the second is on learning capabilities of targeted customer-supplier relationships. In the first case, we aim to present a real-data application where MPL is not a computationally viable alternative. In such a scenario, $\bar{\boldsymbol{\theta}}$ is a fast and reliable choice to estimate the model. In the second application, we investigate a dataset with more contained dimensions. Although computing $\hat{\boldsymbol{\theta}}_{\text{PML}}$ is easily feasible, we show $\bar{\boldsymbol{\theta}}$ is still reliable in terms of both pointwise and inferential performance. All

experiments have been carried out on commodity hardware[1].

5.1 Real application 1: The Big Five personality test

The dataset[2] consists of responses obtained from a web-based survey designed to measure five dimensions of human personality: Neuroticism (N), Agreeableness (A), Extraversion (E), Openness to experience (O), and Conscientiousness (C). The dataset includes answers to $p = 120$ items on a 5-grade scale, totalling $619,150$ observations. It is worth noting that the dataset contains $367,593$ missing answers, which may require appropriate handling or imputation techniques during the analysis (see, e.g. Katsikatsou, Moustaki, and Jamil, 2022). However, given the dimension of the survey, we only retain the ones with complete records, leading to a sample size of $n = 410,376$.

As outlined in Johnson (2014), each of the five personality traits can be further split into six facets measured by four items each, for a total of $q = 30$ latent traits to account for potentially mutually correlated variables. We refer the reader to Johnson (2014) for a detailed inventory of each facet interpretation and label. The following refers to the facets and items based on their respective area. For instance, facets `N1` through `N6` correspond to the Neuroticism (N) trait, while `n1` through `n24` to the related items. Similarly, the survey is structured such that non-overlapping groups of four items measure each of the q latent dimensions. Therefore, we assume that the positions of the zeros in the loading matrix $\mathbf{\Lambda}$ are known. It is worth noting that all loadings are expected to be positive since any "negatively worded" items have been appropriately re-coded.

The total number of parameters to estimate is 1035, namely 480 thresholds, 120 loadings, and 435 latent correlations. The dimensions of the problem prevent numerical optimisation from being a viable option (we manually interrupted the numerical optimiser after twelve hours of running). Nevertheless, the flexibility of the proposed stochastic optimiser allows for overcoming such a challenge. Given the high computational cost of the problem, we set $\nu = 1$. In other words, the algorithm only draws one random pair per iteration and uses it to update the stochastic estimates (however, almost identical results have been obtained using higher values of ν). The initial stepsize is set at $\eta_0 = 10^{-4}$. Convergence is checked by monitoring the complete negative pairwise likelihood on a validation portion of the data. The algorithm checks the validation likelihood every five thousand consecutive iterations. The estimation stops when the difference between successive checks drops under a certain tolerance level. In particular, we use 60% of the dataset as a training partition and the remaining 40% as the validation set. Convergence is reached in about forty-five thousand iterations, including an initial burn-in period of five thousand iterations, with a total computing time of around twenty minutes.

Final results are shown in Figure 3 while complete parameter trajectories are provided in the Appendix. On the left-hand side, Figure 3 reports estimates for the free loadings. As expected, all loadings are positive. Furthermore, most items have similar importance in measuring the latent facets, given the comparable magnitude of the estimated loadings. However, two questions appear less relevant in estimating their respective latent traits. Specifically, `e16`, "Like to take it easy", for Extraversion, and `a20`, "Boast about my virtues", for Agreeableness.

Consistently with the psychometric design of the personality test, the 30 latent facets cluster into five blocks with high within-block positive correlations, as reported on the right-hand side of Figure 3. Note that considering all thirty latent personality facets in the model simultaneously, the estimation uncovers complex correlation patterns that might not be evident when fitting sub-models or analysing individual traits separately. For instance, the results reveal negative correlations between the Neuroticism and Conscientiousness traits, indicating an inverse relationship between these personality dimensions. Additionally, we observe interesting correlation structures between the Agreeableness and Extraversion traits. Some facets within these traits exhibit positive correlations, while others do not. For instance, consider facet `A5` from the Agreeableness trait, "Modesty". While it correlates positively with the other Agreeableness facets, its correlation pattern with the rest of the latent variables in the model is almost opposite to the overall behaviour within its block.

[1] Specifications: AMD Ryzen 7 4800H 16 x 2.9 GHz, R 4.1.2, gcc 11.4.0, Ubuntu 22.04

[2] Downloadable at `https://osf.io/tbmh5/`

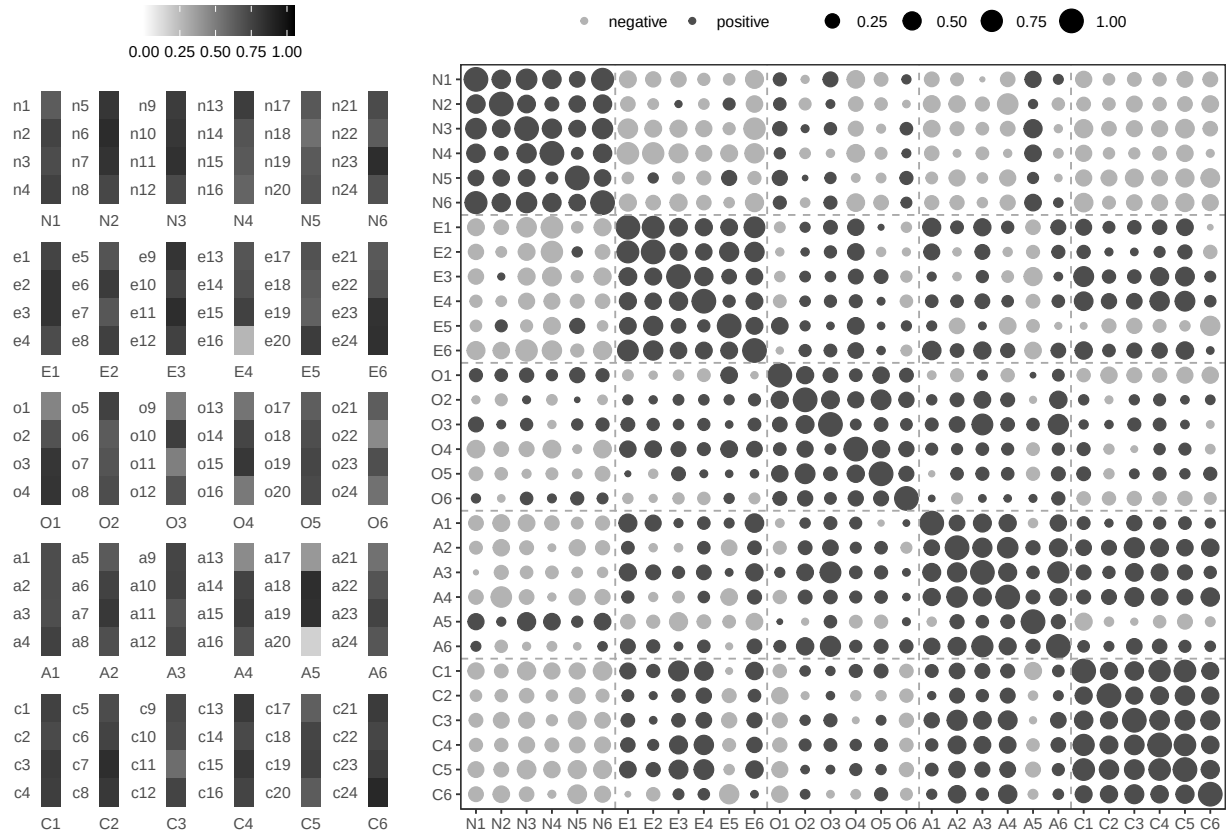

Figure 3: Estimated parameters on the Big Five dataset. Free loadings on the left-hand side. Latent correlations on the right-hand side.

5.2 Real application 2: Learning capabilities

As a second example, we present a smaller application to benchmark the inferential performance of the proposed stochastic estimator on a dataset where numerical pairwise likelihood is computationally viable. Additionally, we also highlight how the numerical estimates are affected by properly accounting for the constraints on $\boldsymbol{\Sigma}_\xi$. Recall that we refer to UPML as the unconstrained version of PML described in Section 2.2, which directly estimates the non-diagonal entries of $\boldsymbol{\Sigma}_\xi$ without imposing constraints other than symmetry. We remind the reader that, as mentioned in Section 2.2 and detailed in Appendix B.1, both $\hat{\boldsymbol{\theta}}_{\text{PML}}$ and $\bar{\boldsymbol{\theta}}$ adopt a suitable reparameterisation of $\boldsymbol{\Sigma}_\xi$ that ensures it to be a proper correlation matrix.

We use the same dataset that was analysed in Katsikatsou et al. (2012) using the UPML estimator. Originally, the data were collected by Selnes and Sallis (2003), who aimed to study whether specific factors affect the learning capabilities of targeted customer-supplier relationships. The eighteen items considered serve as indicators of four factors, which are interpreted as collaborative commitment (ξ_1), internal complexity (ξ_2), relational trust (ξ_3), and environmental uncertainty (ξ_4). See the Appendix for more details about the items considered. All indicators were measured on a seven-point scale. The sample size is 286 after listwise deletion. The matrix $\boldsymbol{\Lambda}$ is assumed to have a known simple structure (i.e. each item is loaded on just one of the four factors). There are a total of 132 free parameters to be estimated, namely 108 thresholds, 18 loadings and 6 latent correlations.

The estimation via $\bar{\boldsymbol{\theta}}$ is carried out with $\nu = 8$ and $\eta_0 = 0.05$. Stopped after 2,500 iterations by monitoring the complete pairwise likelihood on the data every 500 updates after a burn-in period of 500 iterations. Table 1

Table 1: Relationship learning data: Estimated factor loadings and correlations with $\bar{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}_{\text{PML}}$ and $\hat{\boldsymbol{\theta}}_{\text{UPML}}$. Standard errors in brackets. Results for $\hat{\boldsymbol{\theta}}_{\text{UPML}}$ are taken from Katsikatsou et al. (2012).

Factors	$\boldsymbol{\theta}$	$\bar{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\theta}}_{\text{PML}}$	$\hat{\boldsymbol{\theta}}_{\text{UPML}}$
Collaborative Commitment (ξ_1)	$\lambda_{1,1}$	0.877 (0.025)	0.883 (0.023)	0.882 (0.023)
	$\lambda_{2,1}$	0.885 (0.019)	0.891 (0.017)	0.891 (0.017)
	$\lambda_{3,1}$	0.875 (0.021)	0.884 (0.018)	0.884 (0.018)
	$\lambda_{4,1}$	0.887 (0.018)	0.898 (0.016)	0.897 (0.016)
	$\lambda_{5,1}$	0.869 (0.022)	0.875 (0.020)	0.875 (0.020)
Internal Complexity (ξ_2)	$\lambda_{6,2}$	0.624 (0.077)	0.626 (0.077)	0.622 (0.079)
	$\lambda_{7,2}$	0.827 (0.065)	0.820 (0.063)	0.821 (0.065)
	$\lambda_{8,2}$	0.782 (0.068)	0.788 (0.068)	0.784 (0.069)
Relational Trust (ξ_3)	$\lambda_{9,3}$	0.808 (0.027)	0.808 (0.027)	0.808 (0.027)
	$\lambda_{10,3}$	0.863 (0.023)	0.865 (0.023)	0.866 (0.023)
	$\lambda_{11,3}$	0.867 (0.024)	0.867 (0.024)	0.867 (0.024)
	$\lambda_{12,3}$	0.907 (0.016)	0.908 (0.016)	0.908 (0.016)
	$\lambda_{13,3}$	0.872 (0.020)	0.871 (0.020)	0.871 (0.020)
Environmental Uncertainty (ξ_4)	$\lambda_{14,4}$	0.766 (0.033)	0.765 (0.033)	0.767 (0.033)
	$\lambda_{15,4}$	0.852 (0.027)	0.852 (0.027)	0.854 (0.027)
	$\lambda_{16,4}$	0.750 (0.040)	0.750 (0.040)	0.752 (0.040)
	$\lambda_{17,4}$	0.701 (0.044)	0.703 (0.044)	0.697 (0.044)
	$\lambda_{18,4}$	0.704 (0.043)	0.704 (0.042)	0.705 (0.042)
ξ_1, ξ_2	$\boldsymbol{\Sigma}_\xi$	0.257 (0.083)	0.255 (0.082)	0.255 (0.083)
ξ_1, ξ_3		0.631 (0.044)	0.629 (0.044)	0.627 (0.044)
ξ_1, ξ_4		0.662 (0.047)	0.661 (0.047)	0.658 (0.047)
ξ_2, ξ_3		0.128 (0.073)	0.126 (0.073)	0.125 (0.073)
ξ_2, ξ_4		0.194 (0.079)	0.198 (0.079)	0.197 (0.079)
ξ_3, ξ_4		0.650 (0.049)	0.648 (0.049)	0.651 (0.048)

reports the estimated latent correlations, loadings and their standard errors. Variances for both $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\text{PML}}$ have been computed via multivariate delta method in order to be comparable with the parameterisation used by $\hat{\boldsymbol{\theta}}_{\text{UPML}}$. As expected, pointwise estimates provided by $\bar{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}_{\text{PML}}$ and $\hat{\boldsymbol{\theta}}_{\text{UPML}}$ are all quite similar. Nevertheless, $\hat{\boldsymbol{\theta}}_{\text{UPML}}$ sometimes exhibits slightly larger standard errors than its constrained counterpart $\hat{\boldsymbol{\theta}}_{\text{PML}}$, both regarding loadings and latent correlations. At the same time, while $\bar{\boldsymbol{\theta}}$ closely approximates $\hat{\boldsymbol{\theta}}_{\text{PML}}$ pointwise, it exhibits higher variability when accounting for the optimisation noise. Nevertheless, its variability is still comparable to the other two methods and sometimes lower than the one characterising UPML estimates.

6 Conclusions

This paper proposes a new estimator based on stochastic approximation to tackle the computational issue with the traditional pairwise likelihood for the confirmatory factor analysis of high-dimensional categorical data. The new estimator can provide asymptotically consistent point estimation and valid statistical inference (e.g., confidence intervals) under any reasonably small computational budget constraint at the price of a small sacrifice in statistical efficiency. The sacrifice in statistical efficiency becomes negligible as the sample size grows to infinity. The key to the proposed estimator is the stochastic approximation technique, which uses an optimisation procedure relying on stochastic gradients to give an approximation to the conventional pairwise maximum likelihood estimator. As a byproduct of this research, a positive definite constraint, which is not imposed in the previous estimators based on the pairwise likelihood, is introduced on the correlation matrix for the latent variables via a suitable reparameterisation.

Nevertheless, the idea of sampling pairs to reduce the computational burden of pairwise likelihood estimation of factor models is not completely new to the literature and has already been discussed in Papageorgiou and Moustaki (2018). However, their proposed method has a different rationale from our stochastic estimator

because it requires first selecting a subset of the pairs and then estimating the model on the selected pairs only. Their estimators have been seen to reduce the computational time of pairwise maximum likelihood while maintaining good statistical properties in terms of bias and mean squared error. Still, the inference is based solely on pairwise likelihood estimation without considering the initial sampling of pairs. In contrast, our method fully incorporates pair sampling in the inference procedure while allowing for the inclusion of all pairs in the estimation by selecting new subsets of pairs at each algorithm iteration.

The effectiveness of our proposal is shown via simulation studies and two real-data applications. The experiments highlight that the stochastic estimator is comparable to the conventional pairwise likelihood estimator in terms of mean squared error and stress how inferential performance improves when the optimisation noise is taken into account when drawing inference about the true parameters. In addition, we provide two real data examples of very different dimensions to underline the flexibility of the proposed estimator and its inferential reliability.

The current research can be extended in several directions. First, while we focus on the confirmatory factor analysis setting, the proposed method can be easily applied to the exploratory factor analysis setting after imposing certain minimum constraints to ensure model identifiability. Second, our procedure can be modified to estimate mixture models (Ranalli and Rocci, 2016) or polychoric correlation matrices without a factor model structure (Drasgow, 2004) based on similar pairwise likelihood approaches. Finally, sparsity-inducing penalties, such as the lasso penalty (Tibshirani, 1996), may be needed in high-dimensional item factor analysis (Chen, Li, Liu, and Ying, 2023). By incorporating techniques based on the stochastic proximal gradient (Zhang and Chen, 2022), the current method may be extended to solve statistical inference problems based on penalised pairwise likelihoods.

Reproducibility

The code to reproduce the simulated experiments and the real-data applications is available at `https://anonymous.4open.science/r/experimentsSTOPLFA`.

References

Alfonzetti, G., R. Bellio, Y. Chen, and I. Moustaki (2023). When composite likelihood meets stochastic approximation. *arXiv:2310.04165*.

Anderson, T. W. and H. Rubin (1956). Statistical inference in factor analysis. *Third Berkeley Symposium of Mathematical Statistics and Probability 5*, 111–150.

Bartholomew, D., F. Steele, I. Moustaki, and J. Galbraith (2008). *Analysis of Multivariate Social Science Data* (2nd ed.). Chapman and Hall/CRC.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Blum, J. R., H. Chernoff, M. Rosenblatt, and H. Teicher (1958). Central Limit Theorems for Interchangeable Processes. *Canadian Journal of Mathematics 10*, 222–229.

Chen, Y., X. Li, J. Liu, and Z. Ying (2023). Item response theory–a statistical framework for educational and psychological measurement. *Statistical Science*. In press.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika 91*, 729–737.

de Leon, A. R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters 75*, 49–57.

Dillon, W. R., A. Kumar, and N. Mulani (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to heywood cases. *Psychological Bulletin 101*(1), 126.

Drasgow, F. (2004). Polychoric and polyserial correlations. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, and N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*. New York, NY: John Wiley.

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society, A 134*, 486–510.

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality 51*, 78–89.

Jöreskog, K. G. (1990). New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity 24*, 387–404.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika 59*, 381–389.

Jöreskog, K. G. and I. Moustaki (2001). Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research 36*, 347–387.

Katsikatsou, M. (2013). *Composite Likelihood Estimation for Latent Variable Models with Ordinal and Continuous or Ranking Variables*. Ph. D. thesis, Uppsala University.

Katsikatsou, M. and I. Moustaki (2016). Pairwise likelihood ratio tests and model selection criteria for structural equation models with ordinal variables. *Psychometrika 81*, 1046–1068.

Katsikatsou, M., I. Moustaki, and H. Jamil (2022). Pairwise likelihood estimation for confirmatory factor analysis models with categorical variables and data that are missing at random. *British Journal of Mathematical and Statistical Psychology 75*(1), 23–45.

Katsikatsou, M., I. Moustaki, F. Yang-Wallentin, and K. G. Jöreskog (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis 56*, 4243–4258.

Lee, S. Y., W. Y. Poon, and P. M. Bentler (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics and Probability letters 9*, 91–97.

Lee, S. Y., W. Y. Poon, and P. M. Bentler (1992). Structural equation models with continuous and polytomous variables. *Psychometrika 57*, 89–105.

Lewandowski, D., D. Kurowicka, and H. Joe (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis 100*(9), 1989–2001.

Lindsay, B. (1988). Composite likelihood methods. In N. U. Prabhu (Ed.), *Statistical Inference from Stochastic Processes*, pp. 221–239. Providence, RI: American Mathematical Society.

Muthén, B. (1984). A general structural model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika 49*(1), 115–132.

Papageorgiou, I. and I. Moustaki (2018). Samplng of pairs in pairwise likelihood estimation for latent variable models with categorical observed variables. *Statistics and Computing 29*, 351–365.

Polyak, B. T. and A. B. Juditsky (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization 30*(4), 838–855.

Ranalli, M. and R. Rocci (2016). Mixture models for ordinal data: a pairwise likelihood approach. *Statistics and Computing 26*(1-2), 529–547.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*(3), 400–407.

Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Selnes, F. and J. Sallis (2003). Promoting relationship learning. *Journal of Marketing 67*, 80–95.

Stan Development Team (2022). The Stan Core Library. Version 2.33.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 58*(1), 267–288.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis 92*, 1–28.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*, 5–42.

Vasdekis, V., S. Cagnone, and I. Moustaki (2012). A pairwise likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika 77*, 425–441.

Vasdekis, V., D. Rizopoulos, and I. Moustaki (2014). Weighted pairwise likelihood estimation for a general class of random effects models. *Biostatistics 15*, 677–689.

Xu, W. (2011). Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv:1107.2490*.

Zhang, S. and Y. Chen (2022). Computation for latent variable model estimation: A unified stochastic proximal framework. *Psychometrika 87*(4), 1473–1502.

Appendix A Asymptotics

The proposed estimator, $\bar{\boldsymbol{\theta}}$, works in a double asymptotic regime where both the sample size and the number of iterations diverge to infinity. In the following, we show that:

i) The error $\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ is asymptotically normal;

ii) Its asymptotic covariance is stated in (8).

The part of the proof related to i) follows the derivation in Appendix B in Alfonzetti et al. (2023), so here we give an intuition about how the proof works, and we refer the readers to their work for a detailed derivation. The proof of ii) is peculiar to the current setting of factor models with ordinal data, since it takes advantage of the data reduction by sufficiency via bivariate frequencies.

A.1 Asymptotic normality

The interest lies on the distribution of $\bar{\boldsymbol{\theta}}$ or, better, its distance from the true parameter $\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. Thanks to the linearity of the sum, we can rearrange it with

$$\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \frac{1}{T_n}\sum_{t=1}^{T_n} \boldsymbol{\theta}_t - \boldsymbol{\theta}^* = \frac{1}{T_n}\sum_{t=1}^{T_n} (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) = \bar{\boldsymbol{\Delta}}_{T_n},$$

where, thus, $\bar{\boldsymbol{\Delta}}_{T_n}$ is the average error of the algorithm. Following Proposition 2 in Appendix B in Alfonzetti et al. (2023), we can characterise the behaviour of $\bar{\boldsymbol{\Delta}}_{T_n}$ with

$$\sqrt{T_n + n}\bar{\boldsymbol{\Delta}}_{T_n} = \frac{\sqrt{T_n + n}}{nT_n}\boldsymbol{H}^{-1}\sum_{t=1}^{T_n}\sum_{l=1}^{n} \boldsymbol{S}^*_{l,t} + o_p(1), \tag{10}$$

where $\boldsymbol{S}^*_{l,t}$ is the contribution of unit l at iteration t, namely

$$\boldsymbol{S}^*_{l,t} = \frac{P}{\nu}\sum_{i<j} W_{ijt} \sum_{c_i=1}^{m_i}\sum_{c_j=1}^{m_j} \frac{\mathbf{1}_{\{y_{li}=c_i, y_{lj}=c_j\}}}{\pi^{(ij)}_{c_i c_j}(\boldsymbol{\theta}^*)} \nabla \pi^{(ij)}_{c_i c_j}(\boldsymbol{\theta}^*).$$

Thus, the asymptotic distribution of $\bar{\boldsymbol{\Delta}}_{T_n}$ depends on the average behaviour of $\boldsymbol{S}^*_{l,t}$ along the iterations and the sample. Let us define $\bar{\boldsymbol{S}}^* = n^{-1}\sum_l \bar{\boldsymbol{S}}^*_l = n^{-1}T_n^{-1}\sum_t\sum_l \boldsymbol{S}^*_{l,t}$. Note that, since the random variables $\bar{\boldsymbol{S}}_1, \ldots, \bar{\boldsymbol{S}}_n$ are independent conditioned on $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{T_n}$, we can combine the Central Limit Theorem for exchangeable random variables in Blum, Chernoff, Rosenblatt, and Teicher (1958) and the Cramér-Wold device to prove the asymptotic multivariate normality of $\bar{\boldsymbol{S}}^*$, like done in Alfonzetti et al. (2023). Finally, after (10), the asymptotic normality of $\bar{\boldsymbol{S}}^*$ guarantees the asymptotic normality of $\bar{\boldsymbol{\Delta}}_{T_n}$. Note that Alfonzetti et al. (2023) characterise (10) according to three asymptotic regimes depending on the relative divergence rate of n and T_n. In the case of factor models for ordinal data, there is no difference between them regarding the asymptotic distribution of $\bar{\boldsymbol{\theta}}$ because of the data reduction by sufficiency. Thus, we only report (10) for illustration purposes, but the same asymptotic result holds under the other regimes considered in their work.

A.2 Asymptotic covariance

Theorem 2 in Blum et al. (1958), jointly with the Cramér-Wold device, identifies the asymptotic covariance matrix of $\bar{\boldsymbol{S}}^*$ with $n^{-1}\text{Var}(\bar{\boldsymbol{S}}_l^*)$. By using the law of total variance, we can write

$$n^{-1}\text{Var}(\bar{\boldsymbol{S}}_l^*) = \frac{1}{nT_n} E_{\boldsymbol{Y}} \text{Var}_{\boldsymbol{W}|\boldsymbol{Y}}(\boldsymbol{S}_{l,t}^*) + \frac{1}{n} \text{Var}_{\boldsymbol{Y}} E_{\boldsymbol{W}|\boldsymbol{Y}}(\bar{\boldsymbol{S}}_l^*).$$

While the second term on the right-hand side coincides with $\boldsymbol{J}$, as in Alfonzetti et al. (2023), the first term behaves differently because of the specific structure of $S(\boldsymbol{\theta}; \boldsymbol{W})$, where the data reduction by sufficiency strongly affects the distribution of $\boldsymbol{W}$ and, thus, the variability of the stochastic gradients.

Let us call $\boldsymbol{V}_n = n^{-1} E_{\boldsymbol{Y}} \text{Var}_{\boldsymbol{W}|\boldsymbol{Y}}(\boldsymbol{S}_{l,t}^*)$. Then, it can be shown that

$$\begin{aligned} \boldsymbol{V}_n &= \psi_1^{-2} n^{-1} (\psi_1 - \psi_2) \boldsymbol{H} + n^{-1} \left(\psi_1^{-2} \psi_2 - 1 \right) \boldsymbol{J} \\ &= \frac{1}{n} \left\{ \frac{P(P-\nu)}{\nu(P-1)} \boldsymbol{H} - \frac{P-\nu}{\nu(P-1)} \boldsymbol{J} \right\}, \end{aligned}$$

where $\psi_1 = E_{\boldsymbol{W}}(W_{ijt}^2) = E_{\boldsymbol{W}}(W_{ijt}) = \nu/P$ and $\psi_2 = E_{\boldsymbol{W}}(W_{ijt} W_{i'j't}) = \{\nu(\nu-1)\}/\{P(P-1)\}$. Differently from Alfonzetti et al. (2023), the moments of our random weights do not depend on n, because of the sufficiency reduction of the data. Such property implies that $\boldsymbol{V}_n$ is an $o(1)$. That is, it vanishes as the sample size grows. In Alfonzetti et al. (2023), instead, it is an $O(1)$ and remains constant when n goes to infinity.

It follows that the asymptotic covariance matrix of $\bar{\boldsymbol{S}}^*$ can be written as $n^{-1}\text{Var}(\bar{\boldsymbol{S}}_l^*) = T_n^{-1} \boldsymbol{V}_n + n^{-1} \boldsymbol{J}$. Finally, after (10), the asymptotic distribution of the average error of the algorithm, $\bar{\boldsymbol{\Delta}}_{T_n}$, is characterised by the covariance matrix

$$\boldsymbol{\Omega}_n = \frac{1}{nT_n} \boldsymbol{H}^{-1} \left\{ \frac{P(P-\nu)}{\nu(P-1)} \boldsymbol{H} - \frac{P-\nu}{\nu(P-1)} \boldsymbol{J} \right\} \boldsymbol{H}^{-1} + \frac{1}{n} \boldsymbol{H}^{-1} \boldsymbol{J} \boldsymbol{H}^{-1}, \tag{11}$$

as reported in Section 3. Therefore, it holds that $\boldsymbol{\Omega}_n^{-1/2} \bar{\boldsymbol{\Delta}}_{T_n} \xrightarrow[n]{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. As discussed in Section 3, the first term on the right-hand side of (11) is scaled both by n and T_n. Hence, it decreases faster than the second term on the right-hand side of (11), whatever the relative divergence rate between n and T_n. It follows that, with n and T_n going to infinity simultaneously, $\bar{\boldsymbol{\theta}}$ is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{\text{PML}}$, because the second term on the right-hand side of (11) is the dominant part of the asymptotic covariance matrix. Finally, note that such result implies that differences among the three asymptotic scenarios outlined in Alfonzetti et al. (2023) are not relevant in the current setting since the data reduction by sufficiency always allows $\bar{\boldsymbol{\theta}}$ to match the statistical efficiency of $\hat{\boldsymbol{\theta}}_{\text{PML}}$ asymptotically.

Appendix B Model constraints

The model outlined in Section 2 implies the parameter space Θ to be affected by some sets of parametric constraints. More specifically, three types of constraints can be identified:

1. *Thresholds*: for $i = 1, \ldots, p$ it needs to hold that $-\infty = \tau_0^{(i)} < \tau_1^{(i)} < \ldots < \tau_{m_i - 1}^{(i)} < \tau_{m_i}^{(i)} = +\infty$ in order to model the ordinal nature of the responses correctly.

2. *Latent variables*: the latent covariance matrix $\boldsymbol{\Sigma}_\xi$ needs to be constrained to be a correlation matrix to fix the scale of the latent variables.

3. *Model implied correlations*: Similarly, to fix the scale of the underlying variables $Y_1^*, \ldots, Y_p^*$, we need to ensure that, for all is and js, $\boldsymbol{\lambda}_i^\top \boldsymbol{\Sigma}_\xi \boldsymbol{\lambda}_j = \rho(\boldsymbol{\theta})_{ij}$ are proper correlations, where $\boldsymbol{\lambda}_i^\top$ is the generic i-th row of $\boldsymbol{\Lambda}$.

The first set of constraints is the most straightforward to manage. One can account for it by directly reparametrising the thresholds (while fixing one threshold per item) with $\Delta_{m_i}^{(i)} = \exp\left\{\tau_{m_i}^{(i)} - \tau_{m_i-1}^{(i)}\right\}$, and thus estimating thresholds positive differences rather than thresholds per se. However, in practice, it is sufficient to initialise the thresholds far enough from each other in order to avoid their overlapping without relying on the above-mentioned reparameterisation.

The second and third sets of constraints are instead more involved to consider. We investigate them in detail in the following paragraphs.

B.1 Latent variables correlations

The latent covariance matrix $\boldsymbol{\Sigma}_\xi$ is constrained to be symmetric and positive semi-definite, so it is useful to parametrise it via its Cholesky decomposition $\boldsymbol{\Sigma}_\xi = \boldsymbol{U}^\top \boldsymbol{U}$, with $\boldsymbol{U}$ the upper triangular Cholesky factor. However, the scale of the latent variables is fixed to one to allow the identification of the parameter vector. Thus, $\boldsymbol{U}$ must be constrained to be the upper triangular Cholesky factor of a correlation matrix, which translates into imposing $\|\boldsymbol{u}_s\| = 1$, where $\boldsymbol{u}_s$ is the s-th column of $\boldsymbol{U}$. To do this, we rely on the transformation proposed in Lewandowski et al. (2009), similar to state-of-the-art statistical software like `Stan`[3] (Stan Development Team, 2022). Hence, the generic element of the matrix $\boldsymbol{U}$ is defined recursively via

$$U_{r,s} = \begin{cases} 0 & \text{if } r > s; \\ 1 & \text{if } r = s = 1; \\ z_{r,s} & \text{if } 1 = r < s; \\ \frac{z_{rs}}{z_{r-1,s}} U_{r-1,s}(1 - z_{r-1,s}^2) & \text{if } 1 < r < s; \\ \frac{U_{r-1,s}}{z_{r-1,s}}(1 - z_{r-1,s}^2) & \text{if } 1 < r = s, \end{cases}$$

where $z_{r,s}$ is the Fisher's transformation of an unconstrained parameter $h_{r,s}$, namely $z_{r,s} = \frac{\exp(2h_{r,s})-1}{\exp(2h_{r,s})+1} = \tanh(h_{r,s})$. Thus, from the $q(q-1)/2$ unconstrained parameters $h_{r,s}$, the transformation recovers $\boldsymbol{\Sigma}_\xi$ ensuring it is a proper correlation matrix.

B.2 Model implied correlations

Without explicit constraints on the loadings, it is possible in principle to visit along the optimisation a state of $\boldsymbol{\theta}$ such that $|\rho_{ij}(\boldsymbol{\theta})| = |\boldsymbol{\lambda}_i^\top \boldsymbol{\Sigma}_\xi \lambda_j| > 1$ for some pair of items (i, j). In this case $\rho_{ij}(\boldsymbol{\theta})$ is not a valid correlation and, hence, $\nabla \text{pl}(\boldsymbol{\theta})$ is not computable. In fact, while the parameterisation ensures $\boldsymbol{\Lambda}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}^\top$ is a valid covariance matrix, it does not explicitly constrain its scale. By the Cauchy-Schwarz inequality, it always holds that $|\rho_{ij}(\boldsymbol{\theta})|^2 \leq \rho_{ii}(\boldsymbol{\theta})\rho_{jj}(\boldsymbol{\theta})$. Thus, by imposing $\rho_{ii}(\boldsymbol{\theta}) \leq 1$ for all $i = 1, \ldots, p$, we ensure $|\boldsymbol{\lambda}_i^\top \boldsymbol{\Sigma}_\xi \lambda_j| \leq 1$ for all is and js.

In other words, we need to impose a constraint on the loadings to ensure the residual variance estimates stay away from the negative domain since $\boldsymbol{\Sigma}_\delta = \boldsymbol{I}_p - \text{diag}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}^\top)$. Consider $\boldsymbol{\nu}_i = \boldsymbol{U}\boldsymbol{\lambda}_i$, where $\boldsymbol{U}$ is the upper Cholesky factor of $\boldsymbol{\Sigma}_\xi$. Then, the residual variance constraint is satisfied by imposing $\|\boldsymbol{\nu}_i\|_2 \leq 1$ for $i = 1, \ldots, p$. The projector operator on the L_2 unit ball has the well-known analytical form $\prod(\boldsymbol{\nu}_i) = \boldsymbol{\nu}_i / \max\{1, \|\boldsymbol{\nu}_i\|_2\}$. Thus, for each item i, the loadings are orthogonally projected into the parameter space with

$$\prod(\boldsymbol{\lambda}_i) = \frac{\boldsymbol{\lambda}_i}{\max\left\{1, \sqrt{\boldsymbol{\lambda}_i^\top \boldsymbol{\Sigma}_\xi \boldsymbol{\lambda}_i}\right\}}, \tag{12}$$

which guarantees $\|\boldsymbol{\nu}_i\|_2 \leq 1$. Such constraint is connected to what the factor analysis literature usually refers to as a *Heywood case* (Heywood, 1931), which denotes the occurrence of estimated non-positive residual

[3] `https://mc-stan.org/docs/reference-manual/correlation-matrix-transform.html`

variances. Typically, the observation of a Heywood case suggests two possible issues (Dillon, Kumar, and Mulani, 1987). On one hand, it can be related to the sampling variability of the data. The low statistical information provided by small sample sizes might lead the (pairwise) maximum likelihood estimator to wander outside the parameter space. On the other hand, it might be a symptom of model misspecification, thus represents an indicator suggesting practitioners to revise their models.

With the availability of ever larger datasets, the former interpretation has slowly lost relevance, such that Heywood cases are primarily interpreted as a sign of misspecification. However, instability issues related to data variability are particularly relevant in the case of stochastic procedures. Since only small subsets of the data are considered at each iteration, the higher variability of $\boldsymbol{S}_t$ compared to $\nabla \text{pl}(\boldsymbol{\theta})$, might lead to $|\boldsymbol{\lambda}_i^\top \boldsymbol{\Sigma}_\xi \lambda_j| > 1$ for some is and js and, thus, interrupting the optimisation. The projection in (12) completely avoids this scenario, pulling the loadings back within Θ whenever they push to escape. In practice, if the model is correctly specified, the projection only affects the iterations at the beginning of the optimisation when the estimate is far from the target and the optimisation noise is still large.

As a final consideration, note that, while (12) avoids the occurrence of $|\boldsymbol{\lambda}_i^\top \boldsymbol{\Sigma}_\xi \lambda_j| > 1$, it does not entirely write off the eventuality of Heywood cases, since it still allows for null residual variances (i.e. $\boldsymbol{\lambda}_i^\top \boldsymbol{\Sigma}_\xi \lambda_i = 1$). Thus, practitioners retain the possibility to interpret such boundary cases as a sign of model misspecification.

Appendix C Supplementary material for simulation experiments.

C.1 True parameter vector

In accordance with the simulations Katsikatsou et al. (2012), the ordinal items have been generated on a four-grade scale, with thresholds set at $\tau_i = (-1.2, 0, 1.2)$, with $i = 1, \ldots, p$. Since larger models are considered in the simulations here, differently from Katsikatsou et al. (2012) we generate true loadings and latent correlations randomly. In particular, true loadings are drawn uniformly at random in $[0.2, 0.8]$. We avoid generating negative loadings to dodge sign-flipping problems during the estimation. Instead, latent correlations are generated through the parameterisation outlined in Appendix B.1 to ensure that the true $\boldsymbol{\Sigma}_\xi$ is a proper correlation matrix. The unconstrained parameters are drawn uniformly at random in $[-0.8, 0.8]$, which typically leads to small-to-moderate values of latent correlations, both positive and negative. Finally, to ensure that the true models are non-degenerate, we further pass through the projection step in (12) for $i = 1, \ldots, p$. The two $\boldsymbol{\Lambda}$ matrices used for the settings with $p = 40$ and $q \in \{4, 8\}$ are reported in Figure 4, while the latent matrices $\boldsymbol{\Sigma}_\xi$ are plotted in Figure 5.

C.2 Stochastic estimator setup

As reported in Section 4, the choice of drawing $\nu = 8$ pairs per iteration is purely illustrative. Lower values of ν lead to larger optimisation noise and, thus, highlight even more the need for a correction in the asymptotic covariance matrix when constructing confidence intervals. However, the more considerable the noise, the higher the number of iterations needed to converge. On the contrary, larger values of ν drastically lower the optimisation noise, which helps stabilise parameter trajectories along the optimisation. However, while estimates would converge earlier in terms of iterations, the computational cost of the stochastic gradients grows linearly with ν, so it is non-trivial to anticipate which choice of ν would lead to the lowest total computational time. For this reason, $\nu = 8$ has been shown as an illustrative example without any claim of being an optimal choice.

The value of the initial stepsize $\eta_0 = 10^{-2}$ has been chosen manually by minimising the average MSE of the stochastic estimator (with $\nu = 8$) on the most challenging setting (i.e. $p = 40$, $q = 8$, $n = 1000$) on a grid of possible values. For a detailed discussion about the effects of the choice of η_0 on the inferential performance of $\bar{\boldsymbol{\theta}}$, we refer the readers to Alfonzetti et al. (2023) and in particular to their Appendix C. Note that the hyperparameter a, in the stepsize scheduling, is fixed arbitrarily low at $a = 10^{-3}$, since, following Xu (2011), it should be chosen as the lowest eigenvalue of $\boldsymbol{H}$, but it is unknown in practice.

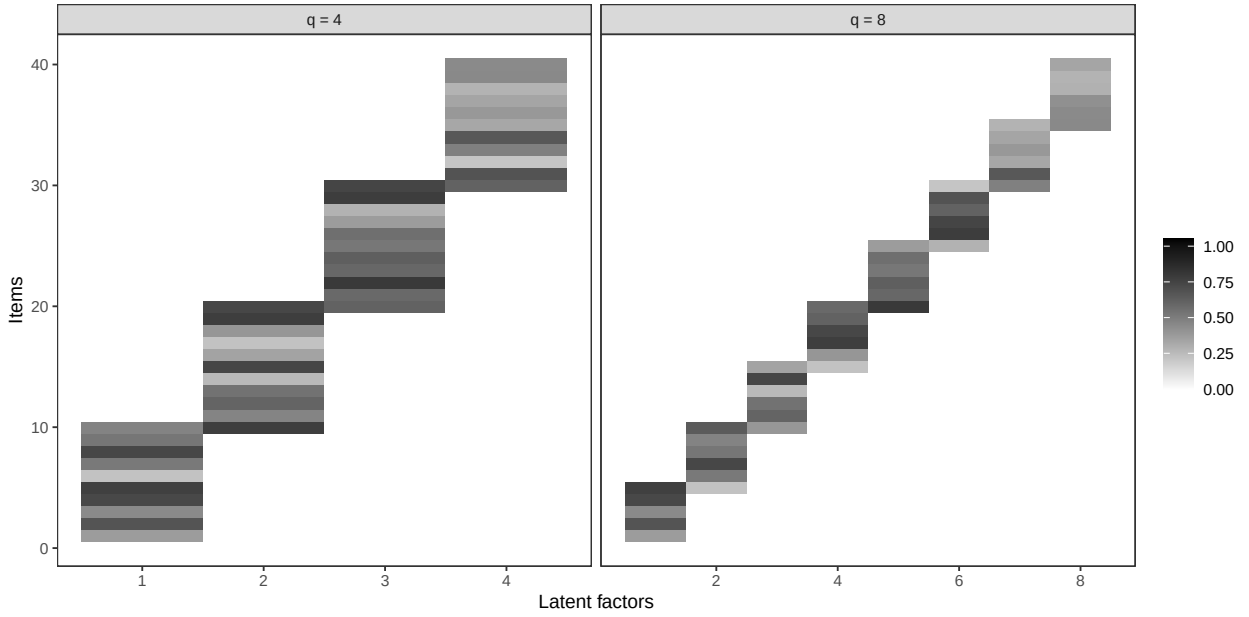

Figure 4: True loading matrices used in simulations in Section 4.

As a final note, all the $S = 1000$ replications are initialised at the same starting point. Namely, thresholds start from $(-1, 0, 1)$; all loadings are initialised at 0.5 since they are expected to be positive, while the unconstrained reparameterisation of the latent correlation matrix starts from the null vector. To get rid of the influence of the starting point in the computation of $\bar{\boldsymbol{\theta}}$, we add a burn-in period of length B. In particular, we let the algorithm discard the first $B = n$ iterations, such that the Ruppert-Polyak averaging used to compute $\bar{\boldsymbol{\theta}}$ starts from $t = n + 1$. While such an approach is highly beneficial in practice, its only consequence on the asymptotic theory outlined in Section 3 is substituting T_n with the effective averaging range, namely $T_n - B$.

C.3 Computational times

In Table 2 we report the computational times recorded for the experiments presented in Section 4.

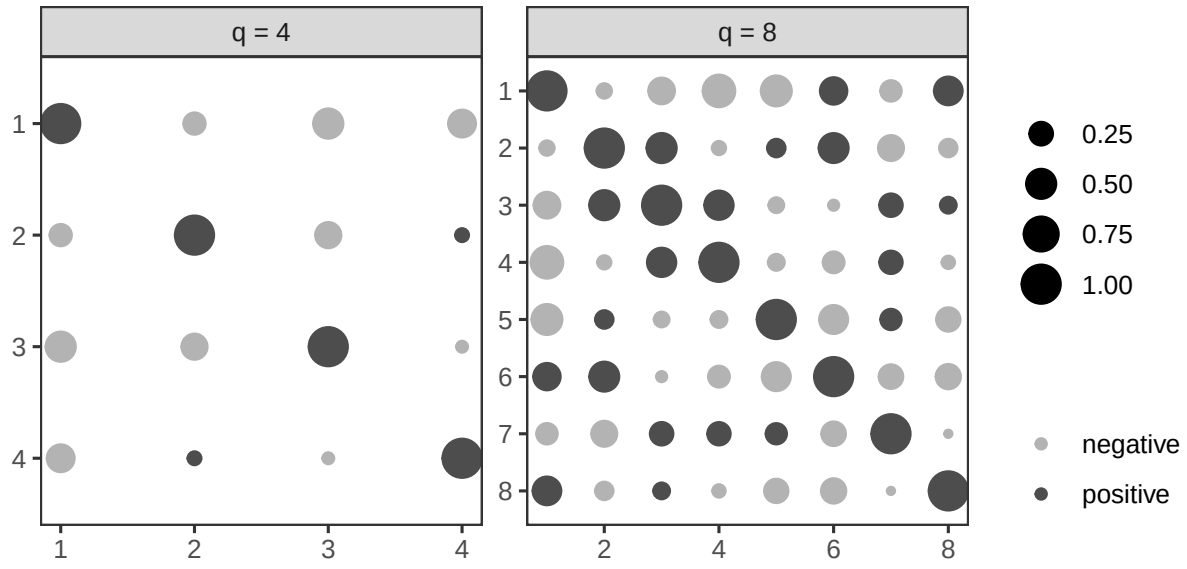

Figure 5: True latent correlation matrices used in simulations in Section 4.

Table 2: Computational times (s) for $\hat{\boldsymbol{\theta}}_{\mathrm{PML}}$ and $\bar{\boldsymbol{\theta}}$ evaluated at $T_n = 2n$.

n	q	$\hat{\boldsymbol{\theta}}_{\mathrm{PML}}$	$\bar{\boldsymbol{\theta}}$
1000		35.74	3.59
2000	4	34.41	7.16
5000		33.90	17.33
1000		105.74	4.65
2000	8	114.74	9.07
5000		108.03	22.05

C.4 Sensitivity checks for starting points and stepsize

In this section, we present some additional simulation experiments to assess the sensitivity of the stochastic estimator setup outlined in Appendix C.2. In particular, we assess the mean squared error performance of the algorithm for different choices of the threshold starting points and the initial stepsize η_0. Concerning thresholds initialisation, two additional starting points have been tested under the setting with $q = 8$ and $n \in \{1000, 2000, 5000\}$, namely $(-2, 0, 2)$ and $(-0.5, 0, 0.5)$. Figure 6 presents the mean squared error performance averaged across 1000 simulations for the two starting points previously mentioned, together with the one used Section 4, namely $(-1, 0, 1)$. As apparent from the plots, the trajectories stemming from the three starting points almost overlap since the early stages of the optimisation, namely even before the end of the burn-in phase. This suggests that the burn-in period is successfully limiting the influence of the starting point in the averaging process.

Regarding the sequence of steps η_t, a correct setup of the decreasing scheduling is of key importance for efficient computations of stochastic approximations. See Xu (2011) for further details. In particular, the parameter η_0, determining the length of the initial step, needs to be set according to the data. In practical applications, η_0 should be chosen by monitoring the likelihood of the model after a small number of iterations. Figure 7 reports the average mean squared error performance across 1000 simulations for iteratively halved values of η_0 at the end of the burn-in period. Results show that the mean squared error performance is robust

to underspecification of η_0. When the value chosen for η_0 is too large, instead, the trajectories diverge. A closer inspection of the experiments revealed the reason for such behaviour to be threshold trajectories breaking their order constraints. As Appendix B outlines, the current parameterisation does not incorporate the explicit ordering of thresholds. Thus, while setting the η_0 to reasonably low values is enough to obtain good performance, a suitable reparameterisation of the model thresholds, as discussed in Appendix B, would potentially allow for larger steps and more robustness to η_0 misspecification.

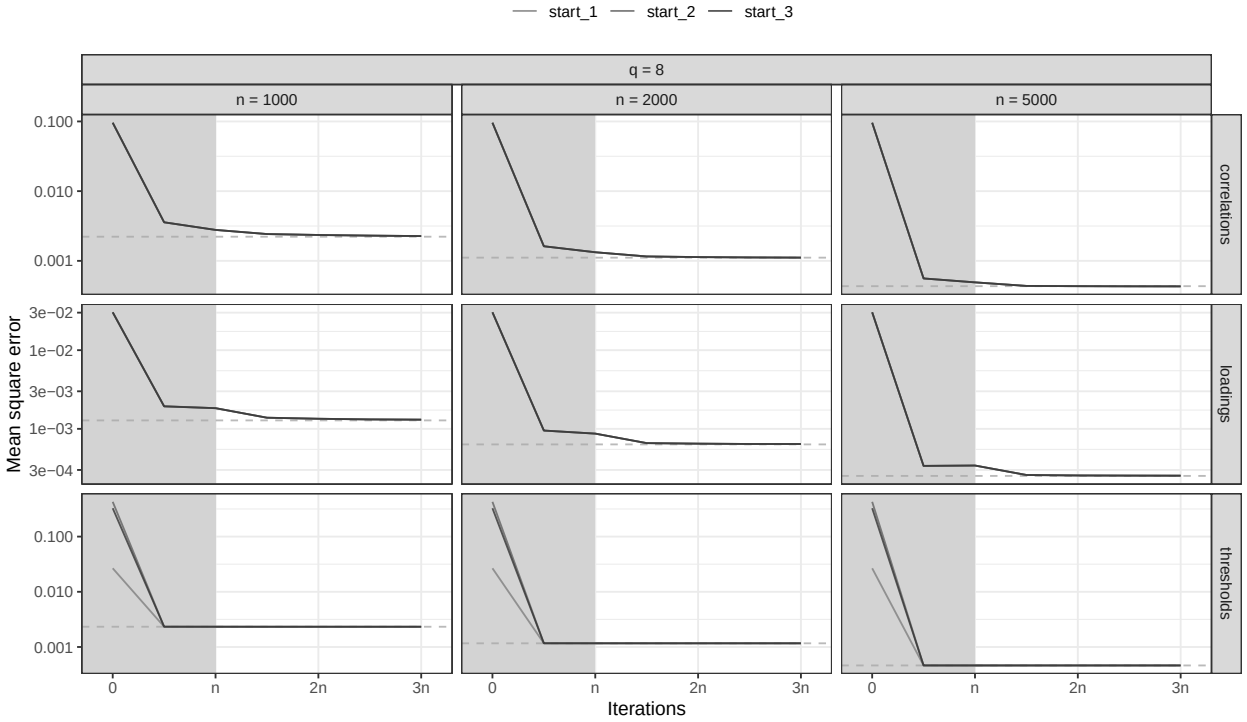

Figure 6: mean squared error performance from three different starting points for the thresholds: `start_1` (-1,0,1), `start_2` (-2,0,2), `start_3` (-0.5, 0, 0.5). Grey areas denote the burn-in period $B = n$.

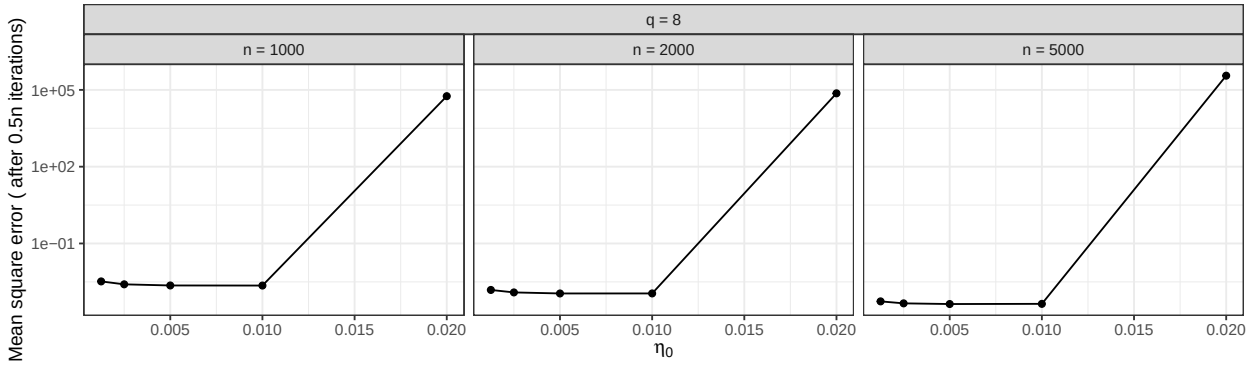

Figure 7: Average mean squared error performance across 1000 replications at the end of the burn-in period ($B = n$) for different values of η_0.

Appendix D Supplementary material for Big Five application

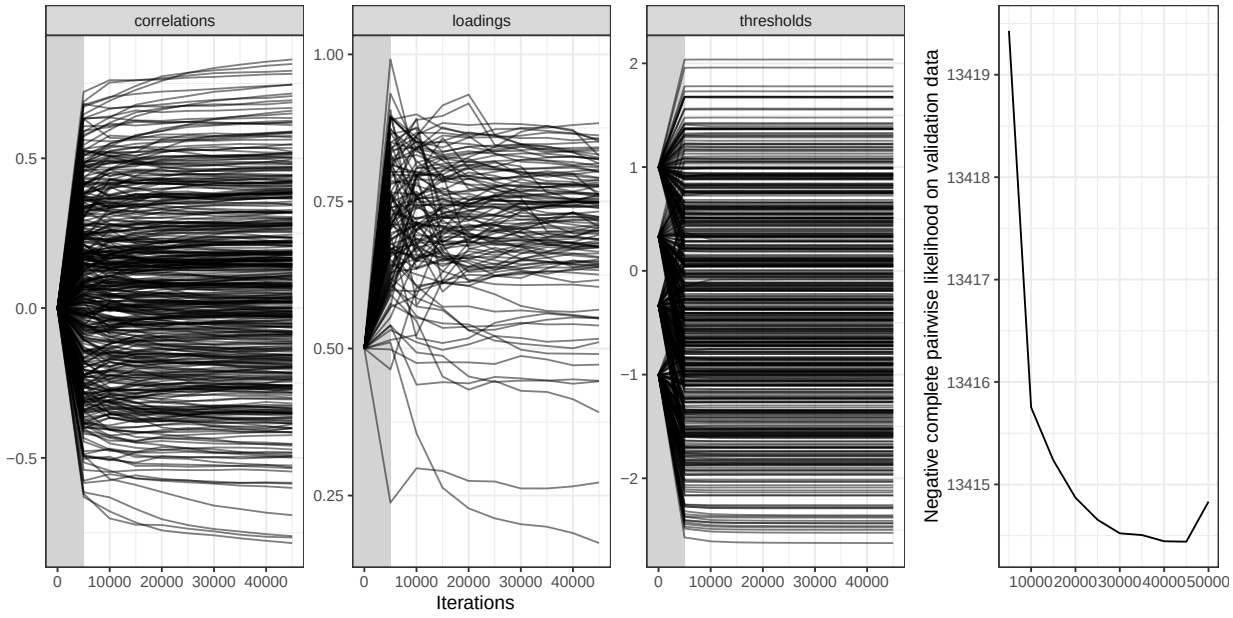

Figure 8: Trajectories along the optimisation of parameter estimates and validation objective function on the Big Five dataset. The grey areas denote the initial burn-in period.

Appendix E The indicators for the Relationship Learning data

Collaborative Commitment

cc1 To what degree do you discuss company goals with the other party in this relationship?

cc2 To what degree are these goals developed through joint analysis of potentials?

cc3 To what degree are these goals formalized in a joint agreement or contract?

cc4 To what degree are these goals implemented in day-to-day work?

cc5 To what degree have you developed measures that capture performance related to these goals?

Internal Complexity

ic1 The products we exchange are generally very complex.

ic2 There are many operating units involved from both organizations.

ic3 There are many contract points between different departments or professions between the two organizations.

Relational Trust

rt1 I believe the other organization will respond with understanding in the event of problems.

rt2 I trust that the other organization is able to fulfill contractual agreements.

rt3 We trust that the other organization is competent at what they are doing.

rt4 There is a general agreement in my organization that the other organization is trustworthy.

rt5 There is a general agreement in my organization that the contact people on the other organization are trustworthy.

Environmental Uncertainty

eu1 End-users needs and preferences change rapidly in our industry.

eu2 The competitors in our industry frequently make aggressive moves to capture market share.

eu3 Crises have caused some of our competitors to shut down or radically change the way they operate.

eu4 It is very difficult to forecast where the technology will be in the next 2-3 years in our industry.

eu5 In recent years, a large number of new product ideas have been made possible through technological breakthroughs in our industry.