New developments on pairwise likelihood estimation for latent variable models.

Irini Moustaki
London School of Economics & Political Science

collaborators: Chris Skinner, Haziq Jamil, Yunxiao Chen, Guiseppe Alfonzeti, Ruggero Bellio

Outline

- Brief introduction to latent variable models for categorical variables.
- Model framework.
- Estimation and inference framework: Pairwise Likelihood (PL)
- Topics that will be discussed:
    - Limited goodness-of-fit tests under SRS and complex sample designs
    - Stochastics optimization for reducing computational complexity

Latent variables and measurement

Using statistical models to understand constructs better: a question of **measurement**

- Many theories in behavioral and social sciences are formulated in terms of theoretical constructs that are not directly observed

  attitudes, opinions, abilities, motivations, etc.

- The measurement of a construct is achieved through one or more observable **indicators** (questionnaire **items**, tests).
- The purpose of a measurement model is to describe how well the observed indicators serve as a measurement instrument for the constructs, also known as **latent variables**.
- **Measurement models** often suggest ways in which the observed measurements can be improved.

Motivation of our work

- Improve the estimation in cases of intractable integrals and complex models.
- Provide an inferential framework for model testing and model selection.
- Improve the computational time and cost.

Notation

- $\mathbf{y}$: $p-$dimensional vector of the observed variables (binary, ordinal, continuous, mixed).
- $\mathbf{y}^\star$: $p-$dimensional vector of corresponding underlying continuous variables.
- The connection between y_i and $y_i^\star$ is

$$y_i = c_i \Longleftrightarrow \tau_{c_i-1}^{(y_i)} < y_i^\star < \tau_{c_i}^{(y_i)}, \tag{1}$$

$-\infty = \tau_0^{(y_i)} < \tau_1^{(y_i)} < \ldots < \tau_{m_i-1}^{(y_i)} < \tau_{m_i}^{(y_i)} = +\infty.$

- c : the c-th response category of variable y_i, $c = 1, \ldots, m_i$, $\tau_{i,c}$: the c-th threshold of variable y_i,
- In practice, $y_i^\star \sim N(0,1)$
- y_i is continuous: $y_i = y_i^\star$.

Structural Equation Model

Following Muthén (1984):

$$\mathbf{y}^{\star} = \boldsymbol{\nu} + \Lambda\boldsymbol{\eta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathrm{B}\boldsymbol{\eta} + \Gamma\mathbf{x} + \boldsymbol{\zeta}$$

$\boldsymbol{\eta}$: vector of latent variables, $q-$dimensional,
$\mathbf{x}$: vector of covariates,
$\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$: vectors of error terms, and
$\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$: vectors of intercepts.
Standard assumptions:

- $\boldsymbol{\eta}$, $\boldsymbol{\epsilon}$, $\boldsymbol{\zeta}$ follow multivariate normal distribution,
- $Cov\left(\boldsymbol{\eta}, \boldsymbol{\epsilon}\right) = Cov\left(\boldsymbol{\eta}, \boldsymbol{\zeta}\right) = Cov\left(\boldsymbol{\epsilon}, \boldsymbol{\zeta}\right) = \mathbf{0}$,
- $I - \mathrm{B}$ is non-singular, I the identity matrix.

Structural Equation Model

Based on the model:

$$\boldsymbol{\mu} \equiv E\left(\mathbf{y}^{\star}|\mathbf{x}\right) = \boldsymbol{\nu} + \Lambda\left(I - \mathrm{B}\right)^{-1}\left(\boldsymbol{\alpha} + \Gamma\mathbf{x}\right)$$

$$\Sigma \equiv Cov\left(\mathbf{y}^{\star}|\mathbf{x}\right) = \Lambda\left(I - \mathrm{B}\right)^{-1}\Psi\left[\left(I - \mathrm{B}\right)^{-1}\right]'\Lambda' + \Theta$$

Let $\boldsymbol{\theta}$ be the parameter vector of the model.

$$\boldsymbol{\theta}' = \left(\mathrm{vec}\left(\Lambda\right)', \mathrm{vec}\left(B\right)', \mathrm{vec}\left(\Gamma\right)', \mathrm{vech}\left(\Psi\right)', \mathrm{vech}\left(\Theta\right)', \boldsymbol{\alpha}', \boldsymbol{\nu}', \boldsymbol{\tau}'\right)$$

.

Likelihood Function

- Under the model,the probability of a response pattern r is:

$$\pi_r(\boldsymbol{\theta}) = \pi\left(y_1 = c_1, \ldots, y_p = c_p; \boldsymbol{\theta}\right) = \int \ldots \int \phi_p(\mathbf{y}^\star; \Sigma_{\mathbf{y}^\star}) d\mathbf{y}^\star\ , \tag{2}$$

  where $\phi_p(\mathbf{y}^\star; \Sigma_{\mathbf{y}^\star})$ is a p-dimensional normal density with zero mean, and correlation matrix $\Sigma_{\mathbf{y}^\star}$.
- The maximization of log-likelihood over the parameter vector $\boldsymbol{\theta}$ requires the evaluation of the p-dimensional integral which cannot be written in a closed form.
- Maximum likelihood infeasible for large number of observed variables.

Composite likelihood (1)

Review the composite likelihood setup:

- $\mathbf{y} = (y_1, \ldots, y_p)^\top$ with true density $p(\mathbf{y}; \theta_0)$, $\theta_0 \in \Theta \subseteq \mathbb{R}^d$;
- $p(\mathbf{y}; \theta_0)$ is unknown or too expensive to compute (e.g. large integrals involved).

- Define a set $\mathcal{A}$ of size K, made of marginal or conditional events for y.
- For each $A_k \in \mathcal{A}$, $k = 1, \ldots, K$, define a proper likelihood function $\mathcal{L}_k(\theta; \mathbf{y})$;
- Construct a composite likelihood with $\mathcal{L}_C(\theta; \mathbf{y}) = \prod_{k=1}^{K} \mathcal{L}_k(\theta; \mathbf{y})$.
- Let $c\ell(\theta; \mathbf{y})$ and $u(\theta; \mathbf{y})$ be respectively the composite log-likelihood and the composite score:

$$c\ell(\theta; \mathbf{y}) = \sum_{k=1}^{K} \ell_k(\theta; \mathbf{y}) \quad \text{and} \quad u(\theta; \mathbf{y}) = \sum_{k=1}^{K} \nabla \ell_k(\theta; \mathbf{y}).$$

Composite likelihood (2)

Finite sample quantities:

- Given a sample of size N, with $\mathbf{y}_{i.} = (y_{i1}, \ldots, y_{ip})$ for $i = 1, \ldots, n$, we can define

$$c\ell_n(\theta; \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \ell_k(\theta; \mathbf{y}_{i.}) \quad \text{and} \quad u_N(\theta; \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \nabla \ell_k(\theta; \mathbf{y}_{i.});$$

- Define the composite likelihood estimator θ_{CL} as the solution of $u_N(\theta_{CL}; \mathbf{y}) = 0$.

Pairwise likelihood estimation

Following Cox & Reid (2004), the composite-likelihood could be modified as follows:

$$c\ell_n(\theta;\mathbf{y}) = \sum_{i<j} \ln L\left(\boldsymbol{\theta};(y_i,y_j)\right) - ap\sum_{i} \ln L\left(\boldsymbol{\theta};y_i\right) \ ,$$

where c is a constant to be chosen for optimal efficiency.

Trying different values of a so that the value of ap ranges from 0 to 1, and conducting some small scale simulation studies, our results indicate that, practically, the sum of univariate log-likelihoods affect neither the accuracy nor the efficiency of estimation.

Pairwise likelihood for SEM

Basic assumption:

$$\begin{pmatrix} y_i^\star \\ y_j^\star \end{pmatrix} \Bigg| \mathbf{x} \sim N_2 \left(\begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}, \begin{pmatrix} \sigma_{ii} & \\ \sigma_{ji} & \sigma_{jj} \end{pmatrix} \right)$$

The pl for N independent observations[1]:

$$pl\left(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}\right) = \sum_{n=1}^{N} \sum_{i<i'} \ln L\left(\boldsymbol{\theta}; (y_{in}, y_{i'n})|\mathbf{x}\right) .$$

The specific form of $\ln L\left(\boldsymbol{\theta}; (y_{in}, y_{i'n})|\mathbf{x}\right)$ depends on the type of the observed variables (binary/ ordinal, continuous).

[1] Myrsini Katsikatsou et al. "Pairwise likelihood estimation for factor analysis models with ordinal data". In: *Computational Statistics & Data Analysis* 56.12 (2012), pp. 4243–4258.

Pairwise Likelihood Estimation for Binary Responses (1) - no covariates

- For a pair of variables y_i and y_j. The basic pairwise log-likelihood takes the form

$$\sum_{i<j} \sum_{c_i=0}^{1} \sum_{c_j=0}^{1} n_{c_i c_j}^{(y_i y_j)} \ln \pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta}) \tag{3}$$

  where $n_{c_i c_j}$ is the observed frequency of sample units with $y_i = c_i$ and $y_j = c_j$.
- To accommodate complex sampling, the PL becomes:

$$pl(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i<j} \sum_{c_i=0}^{1} \sum_{c_j=0}^{1} p_{c_i c_j}^{(y_i y_j)} \ln \pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta}) \ , \tag{4}$$

  where $p_{c_i c_j} = \sum_{h \in s} w_h I(y_i^{(h)} = c_i, y_j^{(h)} = c_j) / \sum_{h \in s} w_h$.

Pairwise Likelihood Estimation for Binary Responses (2)

The score function

$$\nabla pl(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i<j} \sum_{c_i=0}^{1} \sum_{c_j=0}^{1} p_{c_i c_j}^{(y_i y_j)} (\pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta}))^{-1} \frac{\partial \pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{5}$$

Using Taylor expansion, we may write

$$\hat{\boldsymbol{\theta}}_{PL} = \boldsymbol{\theta} + H(\boldsymbol{\theta})^{-1} \nabla pl(\boldsymbol{\theta}; \mathbf{y}) + o_p(N^{-1/2}) \tag{6}$$

where $H(\boldsymbol{\theta})$ is the sensitivity matrix, $H(\boldsymbol{\theta}) = E\left\{-\nabla^2 pl(\boldsymbol{\theta}; \mathbf{y})\right\}$. It follows that

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}\right) \xrightarrow{d} N_t\left(0, H(\boldsymbol{\theta}) J^{-1}(\boldsymbol{\theta}) H(\boldsymbol{\theta})\right) ,$$

where t is the dimension of $\boldsymbol{\theta}$, and $J(\boldsymbol{\theta})$ is the variability matrix, $J(\boldsymbol{\theta}) = Var\left\{\sqrt{N} \nabla pl(\boldsymbol{\theta}; \mathbf{y})\right\}$.

Finite-sample properties of PL estimation

For factor analysis models with categorical data (Katsikatsou et al., 2012):

- PL estimates and standard errors present a close-to-zero bias and mean squared error (MSE).
- PL performs very similarly to three-stage least squares methods and maximum likelihood as implemented in the GLLVM approach.

Model fit and model selection

Katsikatsou and Moustaki, 2016 (Psychometrika).

- Pairwise Likelihood Ratio Test (PLRT) for overall fit
- Pairwise Likelihood Ratio Test for comparing models (e.g. equality constraints)
- Model selection criteria: PL versions of AIC and BIC
- The PLRT statistic performs in accordance with the asymptotic results at 5% and 1% significance levels for $N = 500, 1000$ but not satisfactorily for $N = 200$.
- Both adjusted AIC and BIC criteria perform very well with a minimum rate of success 82.9%.

Software

In the **R package** `lavaan`

PL is available for fitting and testing factor analysis models or SEMs where

- all observed variables are binary or ordinal, and
- the standard parametrization for the underlying variables is used (zero means and unit variances)
- Multigroup analysis is also possible.
- Handling MAR and Non ignorable missigness.

Current work

- Limited information test statistics under SRS and complex designs (with Skinner and Jamil).
- Methods for reducing the computational complexity of pairwise estimation
  - Employ sampling methodology for selecting pairs (Papageorgiou and Moustaki, 2019)
  - Stochastic optimization (with Alfonzetti, Chen, and Bellio)

Limited Information Test Statistics for PL estimators

Overall goodness-of-fit tests, simple hypothesis

- Let us denote with $\mathbf{p}$ the $2^p \times 1$ vector of sample proportions corresponding to the vector of population proportions $\boldsymbol{\pi}$. Assuming i.i.d, it is known that:

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} N(0, \Sigma), \tag{7}$$

- where $\Sigma = D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ and N is the sample size.
- Under complex sampling design, the vector $\mathbf{p}$ becomes the weighted vector of proportions $\mathbf{p}$ with elements $\sum_{h \in s} w_h I(\mathbf{y}^{(h)} = \mathbf{y}_r) / \sum_{h \in s} w_h$.
- Under suitable conditions (e.g. Fuller, 2009, sect. 1.3.2) we still have a central limit theorem, where the covariance matrix Σ need now not take a multinomial form.

Fit on the Lower order margins

- Let $\dot{\boldsymbol{\pi}}_1 = (P(y_1 = 1), P(y_2 = 1), \ldots, P(y_p = 1))'$ be the $p \times 1$ vector that contains all univariate probabilities of a positive response to an item.
- Let $\dot{\boldsymbol{\pi}}_2$ be the $\binom{p}{2} \times 1$ vector of bivariate probabilities with elements, $\dot{\pi}_{ij} = P(y_i = 1, y_j = 1), j < i$.
- Let $\boldsymbol{\pi}_2$ be the vector that contains both these univariate and bivariate probabilities with dimension $s = p + \binom{p}{2} = p(p+1)/2$.
- We also define an $s \times 2^p$ indicator matrix T_2 of rank s such that $\boldsymbol{\pi}_2 = T_2\boldsymbol{\pi}$.

Limited information goodness-of-fit tests

Reiser (1996, 2008), Bartholomew and Leung (2002), Maydey-Olivares and Joe (2005, 2006) Cagnone and Mignani (2007).

The test statistics developed are based on marginal distributions rather than on the whole response pattern.

1. $H_o : \boldsymbol{\pi}_2 = \boldsymbol{\pi}_2(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus $H_1 : \boldsymbol{\pi}_2 \neq \boldsymbol{\pi}_2(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$.
2. Construct test statistics based upon the residual vector $\hat{\mathbf{e}}_2 = \mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}_{PL})$ derived from the bivariate marginal distributions of $\mathbf{y}$ and with $\boldsymbol{\theta}_{PL}$.
3. We first derive the asymptotic distribution of $\hat{\mathbf{e}}_2$.

Distribution of residuals (1)

- Following earlier notation, we can write $s \times 1$ vectors: $\boldsymbol{\pi}_2(\boldsymbol{\theta}) = T_2\boldsymbol{\pi}(\boldsymbol{\theta})$ and $\mathbf{p}_2 = T_2\mathbf{p}$.
- It follows that:

$$\sqrt{n}(\mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta})) \xrightarrow{d} N(0, \Sigma_2), \tag{8}$$

  where $\Sigma_2 = T_2 \Sigma T_2'$.
- Because T_2 is of full rank s, Σ_2 is also of full rank s.

Distribution of residuals (2)

Noting that $\boldsymbol{\pi}_2(\boldsymbol{\theta}) = T_2\boldsymbol{\pi}(\boldsymbol{\theta})$, a Taylor series expansion gives:

$$\boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}_{PL}) = \boldsymbol{\pi}_2(\boldsymbol{\theta}) + T_2\Delta(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}) + o_p(N^{-1/2}), \tag{9}$$

where $\Delta = \dfrac{\partial\boldsymbol{\pi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$
Hence, using

$$\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta} = H(\boldsymbol{\theta})^{-1}\nabla pl(\boldsymbol{\theta}; \mathbf{y}) + o_p(N^{-1/2})$$

we have

$$\hat{\mathbf{e}}_2 = \mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}_{PL}) = \mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta}) - T_2\Delta H(\boldsymbol{\theta})^{-1}\nabla pl(\boldsymbol{\theta}; \mathbf{y}) + o_p(N^{-1/2}). \tag{10}$$

Finally we need to express $\nabla pl(\boldsymbol{\theta}; \mathbf{y})$ in terms of $\mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta})$

Distribution of residuals (3)

Hence, there is a $t \times s$ matrix $B(\boldsymbol{\theta})$ such that

$$\nabla pl(\boldsymbol{\theta}; \mathbf{y}) = B(\boldsymbol{\theta})(\mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta})) \tag{11}$$

Hence, from (10)

$$\hat{\mathbf{e}}_2 = (I - T_2 \Delta H(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}))(\mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta})) + o_p(n^{-1/2}) \tag{12}$$

So from (8), we have under H_0 that:

$$\sqrt{N} \hat{\mathbf{e}}_2 \xrightarrow{d} N(0, \Omega). \tag{13}$$

where $\Omega = (I - T_2 \Delta H(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}))\Sigma_2 (I - T_2 \Delta H(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}))'$.

Distribution of residuals (4)

To estimate the asymptotic covariance matrix of $\hat{\mathrm{e}}_2$, we evaluate $\frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ at the PL estimate $\hat{\boldsymbol{\theta}}_{PL}$ to obtain $\hat{\Delta}$ and set:

$$\hat{\Omega} = (I - T_2 \hat{\Delta} \hat{H}(\hat{\boldsymbol{\theta}}_{PL})^{-1} B(\hat{\boldsymbol{\theta}}_{PL})) \hat{\Sigma}_2 (I - T_2 \hat{\Delta} \hat{H}(\hat{\boldsymbol{\theta}}_{PL})^{-1} B(\hat{\boldsymbol{\theta}}_{PL}))',$$

where $\hat{\Sigma}_2 = T_2 \hat{\Sigma} T_2'$.

- In the case of iid observations with a multinomial covariance matrix, we may set $\hat{\Sigma} = D(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}})) - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\boldsymbol{\pi}(\hat{\boldsymbol{\theta}})'$.
- In the case of a complex sample design we need to derive a consistent estimator for Σ

Proposed test statistics

Wald test type statistics

A Wald test statistic is given by:

- $$L_2 = N(\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}_{PL}))'\hat{\Omega}^{+}(\mathbf{p_2} - \boldsymbol{\pi_2}(\hat{\boldsymbol{\theta}}_{\mathbf{PL}})), \tag{14}$$
    - $\hat{\Omega}^{+}$ is the Moore-Penrose inverse of $\hat{\Omega}$.
    - Under H_0, this test statistic is asymptotically distributed as χ^2 with degrees of freedom equal to the rank of $\hat{\Omega}^{+}$, which is between $s - t$ and s.
- An alternative Wald test: $\hat{\boldsymbol{\Xi}}_2 = \mathrm{diag}(\hat{\boldsymbol{\Omega}}_2)^{-1}$ is used instead of the pseudoinverse of $\boldsymbol{\Omega}_2$. We refer to this *Diagonal Wald test*, (Wald v2). Its distribution needs to be determined using moment-matching procedures. We employ a three moment adjustment.
- The estimation of Ω_2 can be computationally involved in some cases (large models).
- The rank of Ω_2 cannot be determined a priori instead one needs to inspect the eigen values of $\hat{\Omega}_2$.

Variance-covariance free Wald test, Wald v3

Maydeu-Olivares and Joe (2005, 2006) suggested using a weight matrix $\boldsymbol{\Xi}$ such that $\boldsymbol{\Omega}_2$ is a generalized inverse of $\boldsymbol{\Xi}$, i.e. $\boldsymbol{\Xi} = \boldsymbol{\Xi}\boldsymbol{\Omega}_2\boldsymbol{\Xi}$.
The test statistic proposed:

$$X^2 = n\hat{\mathbf{e}}_2^\top \hat{\boldsymbol{\Xi}} \hat{\mathbf{e}}_2 = n\hat{\mathbf{e}}_2^\top \hat{\boldsymbol{\Delta}}_2^\perp \left((\hat{\boldsymbol{\Delta}}_2^\perp)^\top \hat{\boldsymbol{\Sigma}}_2 \hat{\boldsymbol{\Delta}}_2^\perp \right)^{-1} (\hat{\boldsymbol{\Delta}}_2^\perp)^\top \hat{\mathbf{e}}_2$$

- where $\boldsymbol{\Delta}_2^\perp$ is an $S \times (S - m)$ orthogonal complement to $\boldsymbol{\Delta}_2$, i.e. it satisfies $(\boldsymbol{\Delta}_2^\perp)^\top \boldsymbol{\Delta}_2 = \mathbf{0}$.
- It converges in distribution to a χ^2_{S-m} variate as $n \to \infty$.

Pearson Chi-square Test Statistic

- Let D_2 be the $s \times s$ matrix $D_2 = diag(\pi_2(\boldsymbol{\theta}))$ and let $\hat{D}_2 = diag(\pi_2(\hat{\boldsymbol{\theta}}_{PL}))$.
- The Pearson test statistic is given by

$$X_P^2 = n\hat{\mathbf{e}}_2'\hat{D}_2^{-1}\hat{\mathbf{e}}_2 = n(\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}_{PL}))'\hat{D}_2^{-1}(\mathbf{p_2} - \boldsymbol{\pi_2}(\hat{\boldsymbol{\theta}}_{\mathbf{PL}})). \tag{15}$$

- The limiting distribution of $\sqrt{n}\hat{D}_2^{-0.5}\hat{\mathbf{e}}_2$ under the hypothesis that the model is correct is given by $N(0, {D_2}^{-0.5}\Omega_2 {D_2}^{-0.5})$.
- Hence X_P^2 has the limiting distribution of $\sum \delta_i W_i$, where the δ_i are eigenvalues of ${D_2}^{-0.5}\Omega_2 {D_2}^{-0.5}$ and the W_i are independent chi-square random variables, each with one degree of freedom.
- These eigenvalues can be estimated by the eigenvalues of $\hat{D}_2^{-0.5}\hat{\Omega}_2\hat{D}_2^{-0.5}$.
- A first and a second order Rao-Scott type test can be obtained.

Estimation of the covariance matrix under complex sampling

Estimation of the covariance matrix under complex sampling: stratified multistage sampling (1)

- $$\begin{aligned} \Sigma &= limvar\{\sqrt{N}(\mathbf{p} - \boldsymbol{\pi})\} \\ &= limvar\{\sqrt{N}(\frac{\sum_{h \in s} w_h \mathbf{y}^{(h)}}{\sum_{h \in s} w_h} - \boldsymbol{\pi})\} \end{aligned}$$

  where $limvar$ denotes the asymptotic covariance matrix.
- Using a usual linearization argument for a ratio:

$$\Sigma = limvar\{\sqrt{N}\frac{\sum_{h \in s} w_h (\mathbf{y}^{(h)} - \boldsymbol{\pi})}{E(\sum_{h \in s} w_h)}\}. \tag{16}$$

Estimation of the covariance matrix: stratified multistage sampling (2)

- Strata are labelled a and the primary sampling units are labelled $b = 1, \ldots, N_a$, where N_a is the number of primary sampling units selected in stratum a.
- Then we write

$$[\sum_{h \in s} w_h(\mathbf{y}^{(h)} - \boldsymbol{\pi})]/[E(\sum_{h \in s} w_h)] = \sum_a \sum_b \tilde{\mathbf{u}}_{ab}, \tag{17}$$

- where $\tilde{\mathbf{u}}_{ab} = \sum_{h \in s_{ab}} w_h(\mathbf{y}^{(h)} - \boldsymbol{\pi})/[E(\sum_{h \in s} w_h)]$ and s_{ab} is the set of sample units contained within primary sampling unit b within stratum a. So

$$\Sigma = limvar\{\sqrt{N} \sum_a \sum_b \tilde{\mathbf{u}}_{ab}\}. \tag{18}$$

Estimation of the covariance matrix: stratified multistage sampling (3)

- A standard estimator of $N^{-1}\Sigma$ is then given by

$$N^{-1}\hat{\Sigma} = \sum_a \frac{N_a}{N_a - 1} \sum_b (\mathbf{u}_{ab} - \bar{\mathbf{u}}_a)(\mathbf{u}_{ab} - \bar{\mathbf{u}}_a)' \qquad (19)$$

- where $\mathbf{u}_{ab} = \sum_{h \in s_{ab}} w_h(\mathbf{y}^{(h)} - \mathbf{p})/(\sum_{h \in s} w_h)$ and $\bar{\mathbf{u}}_a = N_a^{-1} \sum_b \mathbf{u}_{ab}$

Estimation of the covariance matrix under complex sampling (4)

- In order to compute the Wald and Pearson test statistic, we only require $\hat{\Sigma}_2 = T_2 \hat{\Sigma} T_2'$.

$$N^{-1}\hat{\Sigma}_2 = \sum_a \frac{N_a}{N_a - 1} \sum_b (\mathbf{v}_{ab} - \bar{\mathbf{v}}_a)(\mathbf{v}_{ab} - \bar{\mathbf{v}}_a)' \tag{20}$$

where $\mathbf{v}_{ab} = \sum_{h \in s_{ab}} w_h (\mathbf{y}_2^{(h)} - \mathbf{p}_2)/(\sum_{h \in s} w_h)$, $\bar{\mathbf{v}}_a = N_a^{-1} \sum_b \mathbf{v}_{ab}$ and $\mathbf{y}_2^{(h)} = T_2 \mathbf{y}^{(h)}$ is the $s \times 1$ vector containing indicator values $I(y_i^{(h)} = 1)$ and $I(y_i^{(h)} = y_j^{(h)} = 1)$ for different values of i and j.

Simulation study

Simulation A: data generated under SRS

- Four sample sizes $(n = 500, 1000, 2000, 3000)$.
  1. $p = 5$ and $q = 1$ (1F 5V)
  2. $p = 8$ and $q = 1$ (1F 8V)
  3. $p = 15$ and $q = 1$ (1F 15V)
  4. $p = 10$ and $q = 2$, 5 indicators per factor (2F 10V)
  5. $p = 15$ and $q = 3$, 5 indicators per factor (3F 15V)
- Models 4 and 5 are confirmatory factor analysis models.
- The number of replications within each condition is 1000.
- Power analysis: a latent variable $z \sim N(0, 1)$ added to the data generating model.

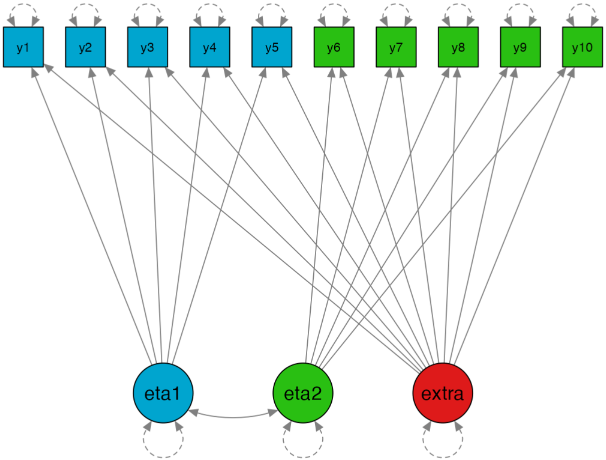
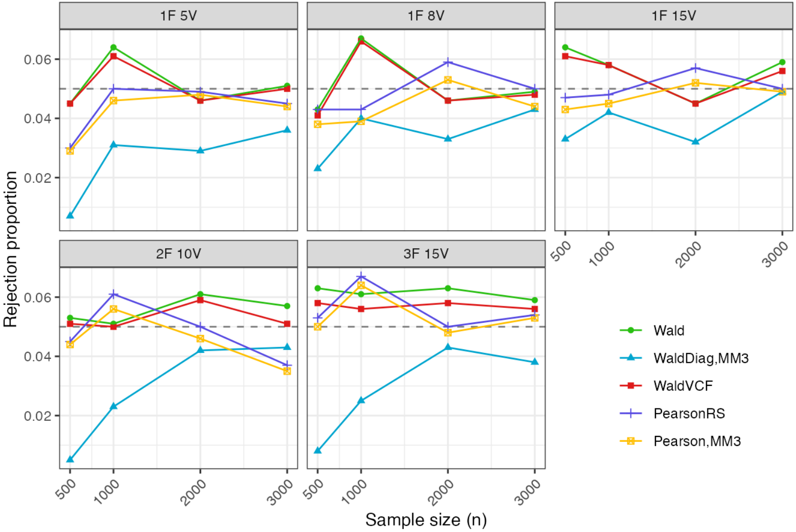


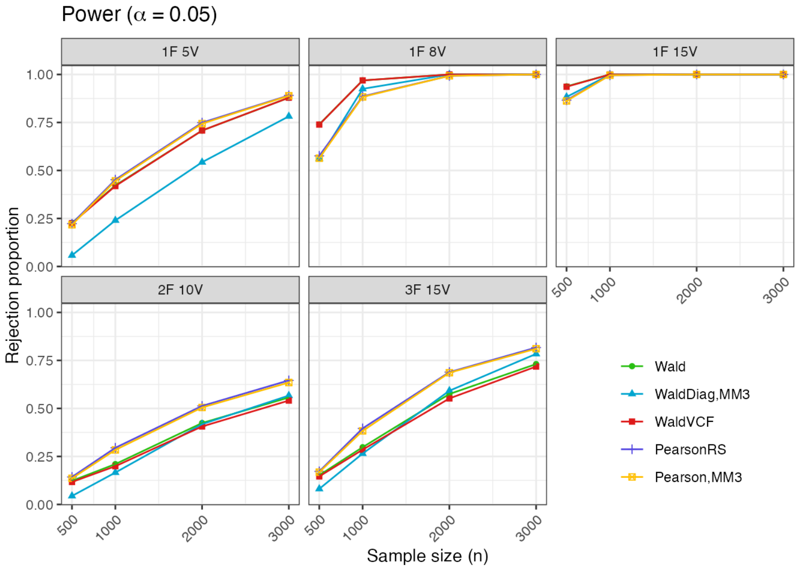
Figure: Model 4: Confirmatory factor analysis model

Simulation A: Test statistics computed

- The Wald test.
- The Wald v2 test (diagonal).
- The Wald v3 test (otrhogonal components)
- The Pearson test (PearsonRS).
- The first-and-second-moment adjusted (FSMadj) Pearson test statistic.

Type I errors ($\alpha = 0.05$)




Power (α = 0.05)
1F 5V
1F 8V
1F 15V
2F 10V
3F 15V
Rejection proportion
0.00
0.25
0.50
0.75
1.00
500
1000
2000
3000
Sample size (n)
Wald
WaldDiag,MM3
WaldVCF
PearsonRS
Pearson,MM3

Simulation A: Results

- The Wald v2 has the poorest performance. Both Pearson test statistics performed satisfactorily at all three significance levels $\alpha = 0.01, 0.05, 0.10$ and improved with the increase of the sample size.
- The power of all tests increases with the sample size but stayed at lower levels in the case of two and three-factor models.

Simulation B: data generated under complex sampling

- Four sample sizes ($n = 500, 1000, 2000, 3000$).
- We generate data for an entire population inspired by a sampling design used in large scale assessment surveys.
- The population consists of $2,000$ schools (Primary Sampling Units, PSU) of three types: "A" (400 units), "B" (1000 units), and "C" (600 units). The school type correlates with the average abilities of its students (stratification factor).
- Each school is assigned a random number of students from the normal distribution $N(500, 125^2)$ (the number then rounded down to a whole number).
- Students are then assigned randomly into classes of average sizes 15, 25 and 20 respectively for each school type A, B and C.
- The total population size is roughly 1 million students.

Simulation B: Sampling designs (1)

1. Stratified sampling: From each school type (strata), select 1000 students (PSU) using SRS. Let N_a be the total number of students in stratum $a \in \{1,2,3\}$. Probability of selection of a student in stratum a is $\Pr(\text{selection}) = \frac{1000}{N_a}$. The total sample size is $n = 3 \times 1000 = 3000$.

2. Two-stage cluster sampling: Select 140 schools (PSU; clusters) using probability proportional to size (PPS). For each school, select one class by SRS, and all students in that class. The probability of selection of a student in PSU $b = 1, \dots, 2000$:

$$\Pr(\text{selection}) = \Pr(\text{weighted school selection}) \times \frac{1}{\text{\# classes in school } b}.$$

   The total sample size will vary from sample to sample, but on average will be $n = 140 \times 21.5 = 3010$, where 21.5 is the average class size per school.

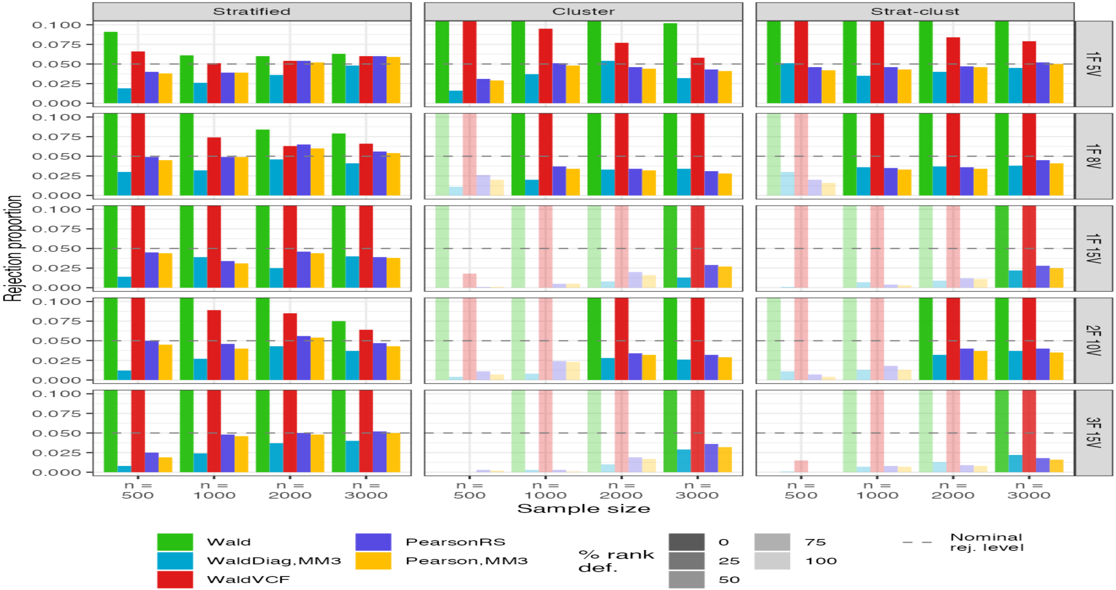
Simulation B: Sampling designs (2)

1. Two-stage stratified cluster sampling: For each school type (strata), select 50 schools using SRS. Then, within each school, select 1 class by SRS, and all students in that class are selected to the sample. The probability of selection of a student in PSU b from school type a is

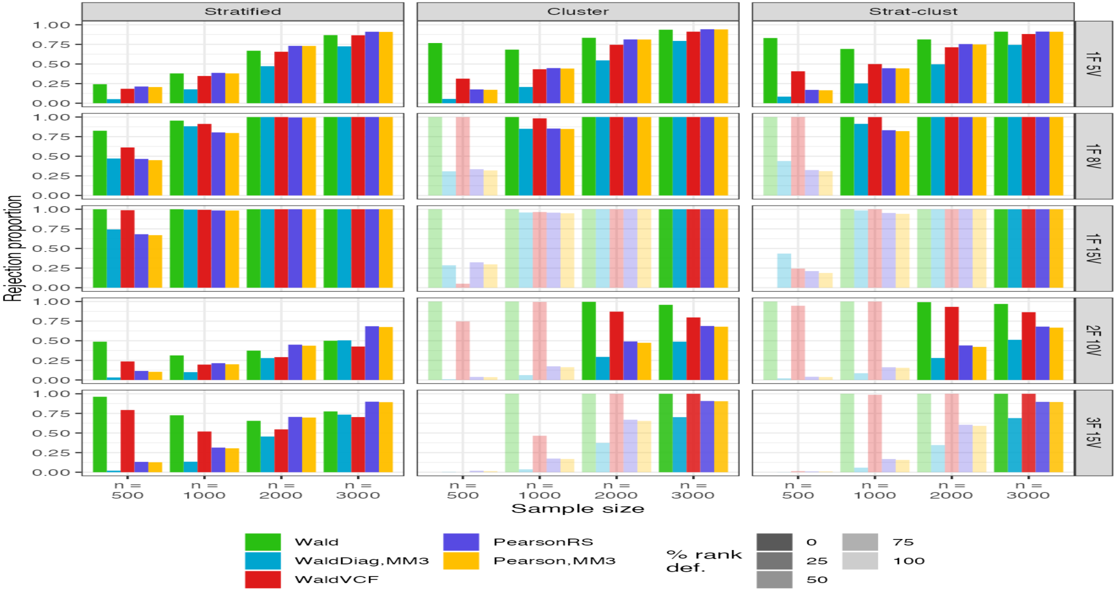
$$\Pr(\text{selection}) = \frac{50}{\#\ \text{schools of type}\ a} \times \frac{1}{\#\ \text{classes in school}\ b}.$$

Here, the expected sample size is $n = 50 \times (15 + 25 + 20) = 3000$.

Type I errors ($\alpha = 0.05$)



Power ($\alpha = 0.05$)



Simulation B: Results

- Type I error rates: Both Pearson tests performed satisfactorily under stratified sampling.
- In the cluster sampling and stratified cluster sampling and in samples sizes of 500 and 1000 we had a large proportion of rank deficiency issues with the estimated covariance matrix.
- The power of the test in the one-factor models and stratified sampling increased to 1 with the increase of the sample size.

Stochastic gradient descent

Composite likelihood

Finite sample quantities:

- Given a sample of size N, with $\mathbf{y}_{i.} = (y_{i1}, \dots, y_{ip})$ for $i = 1, \dots, N$, we can define

$$c\ell_n(\boldsymbol{\theta}; \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \ell_k(\boldsymbol{\theta}; \mathbf{y}_{i.}) \quad \text{and} \quad u_N(\boldsymbol{\theta}; \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \nabla \ell_k(\boldsymbol{\theta}; \mathbf{y}_{i.});$$

- Define the composite likelihood estimator $\boldsymbol{\theta}_{CL}$ as the solution of $u_N(\boldsymbol{\theta}_{CL}; \mathbf{y}) = 0$.

Notation consideration:

The value $\boldsymbol{\theta}_{CL}$ is the theoretical optimiser of $c\ell_n(\boldsymbol{\theta}; \mathbf{y})$ but, typically, we can't compute it exactly. We use $\hat{\boldsymbol{\theta}}_{CL}$ to refer to the output of a generic optimisation algorithm applied on $c\ell_n(\boldsymbol{\theta}; \mathbf{y})$. Otherwise stated, $\hat{\boldsymbol{\theta}}_{CL}$ is a numerical approximation of $\boldsymbol{\theta}_{CL}$.

Computational considerations

The computational bottleneck shifts from the intractability of $p(\mathbf{y};\boldsymbol{\theta}_0)$ to the number of components K to account for in $\mathcal{L}_C$. A numerical optimisation algorithm needs to re-evaluate $u_N(\boldsymbol{\theta};\mathbf{y})$ at each iteration, which has a complexity $O(NK)$.

Average stochastic gradient descent (1)

Problem setup:

- The target of the approximation is $\boldsymbol{\theta}^*$, such that $E_\Gamma\left\{u(\boldsymbol{\theta}^*;\mathbf{y})\right\}=0$
  - In an online setting, Γ is the true density of the data, and $\boldsymbol{\theta}^* \equiv \boldsymbol{\theta}_0$.
  - In an finite-sample setting, Γ is the data empirical distribution, and $\theta^* \equiv \theta_{CL}$.

The finite-sample setting:[2]

- The data are fixed at $\mathbf{y}$.
- Since data are fixed, stochastic gradients are based on an auxiliary random variable ζ.
- Define $U = U(\boldsymbol{\theta};\zeta \mid \mathbf{y})$, such that $E_\zeta\left\{U\right\} = u_N(\boldsymbol{\theta};\mathbf{y})$

[2] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". en. In: *The Annals of Mathematical Statistics* 22.3 (Sept. 1951), pp. 400–407.

Average stochastic gradient descent (2)

A generic SGD algorithm:

Given a starting value $\boldsymbol{\theta}^0$ and a decreasing scheduling for the stepsize $\eta^{(t)}$, $t = 1, \ldots, T$:

1. At the the generic t-th iteration, alternate:
   - Compute $U^{(t)}$;
   - Update the parameter state with $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \eta^{(t)} U^{(t)}$.
2. Return $\bar{\boldsymbol{\theta}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\theta}^{(t)}$.

Why averaging?[3],[4]

- Asymptotic normality: $\sqrt{T}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_{CL}) | \boldsymbol{\theta}_{CL} \xrightarrow{\mathrm{d}} \mathcal{N}_d \left\{0, \Omega_{\zeta|\mathbf{y}}\right\}$ with $\Omega_{\zeta|\mathbf{y}} = A^{-1} S A^{-1}$;
  - $A = A(\boldsymbol{\theta}_{CL}) = -\nabla u_n(\boldsymbol{\theta}_{CL}; \mathbf{y})$;
  - $S = S(\boldsymbol{\theta}_{CL}) = \mathrm{Var}_{\zeta|\mathbf{y}} \left\{U(\boldsymbol{\theta}_{CL}; \zeta|\mathbf{y})\right\}$.

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

[4] David Ruppert. *Efficient estimations from a slowly convergent Robbins-Monro process*. Tech. rep. Cornell University Operations Research and Industrial Engineering, 1988.

Average stochastic gradient descent (3)

A popular example of SGD:

- In most applications, stochastic gradients are constructed by considering a random subset of observations at each iteration.
- Namely, $U(\boldsymbol{\theta}; \zeta|\mathbf{y}) \propto \sum_i \zeta_i u(\boldsymbol{\theta}; \mathbf{y}_i)$, where $\zeta = (\zeta_1, \ldots, \zeta_N)$ follows a different distribution according to (1) how many observations to consider and (2) whether the sampling is chosen with or without replacement.
- We refer to this class of algorithms as observations-based SGD (or OSGD), to stress they represent a specific case of SGD.

CSGD - Composite Stochastic Gradient Descent

- Takes advantage of the peculiar structure of the composite likelihood;
- More computationally flexible than OSGD;
- Possibility for more efficient stochastic gradients than OSGD.

CSGD - What's new about it?

More flexible stochastic approximation of the composite score defined by

$$U_{\mathcal{P}} = U(\boldsymbol{\theta}; \mathbf{y}, W, \mathcal{P}) = c_{\mathcal{P}} \sum_{i=1}^{N} \sum_{k=1}^{K} W_{ik} \nabla \ell_k(\boldsymbol{\theta}; \mathbf{y}_{i\cdot}),$$

where $c_{\mathcal{P}}$ is a scaling constant that guarantees

$$E_{W|\mathbf{y}} \left\{ U(\boldsymbol{\theta}; \mathbf{y}, W, \mathcal{P}) \right\} = u_N(\boldsymbol{\theta}; \mathbf{y}), \quad \boldsymbol{\theta} \in \Theta,$$

and W is a random weighting matrix defined on some probability space $\mathcal{P}$ with realisation w.

$$W = \left.\overbrace{\boxed{\begin{matrix} W_{11} & W_{12} & \cdots & W_{1K} \\ W_{21} & W_{22} & \cdots & W_{2K} \\ \vdots & \vdots & & \vdots \\ W_{N1} & W_{N2} & \cdots & W_{NK} \end{matrix}}}^{K}\right\} N$$

Figure: The generic weighting matrix of the stochastic composite score.

CSGD - The algorithm

CSGD algorithm:

Given $\theta^{(0)}$, $\mathcal{P}$, $c_{\mathcal{P}}$, η, T, B;

1. For $t = 1, \dots, T$:
   - Sampling step: Draw a new $w^{(t)}$ according to $\mathcal{P}$;
   - Approximation step: Compute $U_{\mathcal{P}}^{(t)} = U(\boldsymbol{\theta}^{(t-1)}; \mathbf{y}, w^{(t)}, \mathcal{P})$;
   - Update: Compute $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \eta^{(t)} U_{\mathcal{P}}^{(t)}$, where $\eta^{(t)} = \eta t^{-\epsilon}$, with $\epsilon \in (1/2, 1]$.
2. Trajectories averaging: Return

$$\bar{\boldsymbol{\theta}}_{\mathcal{P}} = \frac{1}{T-B} \sum_{t=B+1}^{T} \boldsymbol{\theta}^{(t)},$$

where B is an initial burn-in period.

CSGD - Choosing the probability space

OSGD ($\mathcal{P}'$):

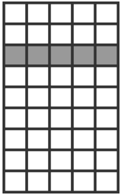
- $W_{i1} = \cdots = W_{iK}$ for $i = 1, \ldots, N$, with $(W_{11}, \ldots, W_{N1}) \sim$ Multi$\{1, (1/N, \ldots, 1/N)\}$

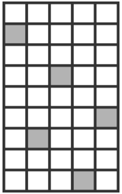
Bernoulli CSGD ($\mathcal{P}^*$):

- $W_{ik} \overset{iid}{\sim}$ Bernoulli$(1/N)$, for $i = 1, \ldots, N$ and $k = 1, \ldots, K$.

$$U_{\mathcal{P}'} = \sum_{i=1}^{N} W_{i1} c\ell(\boldsymbol{\theta}; \mathbf{y}_{i\cdot})$$

$$U_{\mathcal{P}^*} = \sum_{i=1}^{N} \sum_{k=1}^{K} W_{ik} \nabla \ell_k(\boldsymbol{\theta}; \mathbf{y}_{i\cdot}).$$





CSGD - Efficiency of the estimates

	$\mathcal{P}'$	$\mathcal{P}^*$
Stochastic gradient $(U_{\mathcal{P}})$	$U_{\mathcal{P}'} = \sum_{i=1}^{N} W_{i1} c\ell(\theta; y_{i.})$	$U_{\mathcal{P}*} = \sum_{i=1}^{N} \sum_{k=1}^{K} W_{ik} \nabla \ell_k(\theta; y_{i.})$
Computational budget	$O(K)$	$O(K)$
$S = \mathit{Var}_{W\|y}(U_{\mathcal{P}})$	$\hat{J}(\theta_{CL})$	$\hat{H}(\theta_{CL})$
$A = -\nabla u_N(\theta_{CL}; y)$	$\hat{H}(\theta_{CL})$	$\hat{H}(\theta_{CL})$
$\Omega_{W\|y} = A^{-1} S A^{-1}$	$\hat{H}^{-1} \hat{J} \hat{H}^{-1} = \hat{\Omega}$	$\hat{H}^{-1} \hat{H} \hat{H}^{-1} = \hat{H}^{-1}$
Asymptotic distribution:	$\sqrt{T}(\bar{\theta}_{\mathcal{P}'} - \theta_{CL}) \| \theta_{CL} \xrightarrow{\text{d}} \mathcal{N}_d \left\{0, \hat{\Omega}\right\}$	$\sqrt{T}(\bar{\theta}_{\mathcal{P}*} - \theta_{CL}) \| \theta_{CL} \xrightarrow{\text{d}} \mathcal{N}_d \left\{0, \hat{H}^{-1}\right\}$

Table: Effects of the choice of $\mathcal{P}$ on the efficiency of CSGD estimates.

Only conditional inference is available!

- We have the asymptotic distribution for both $\sqrt{T}(\bar{\theta}_{\mathcal{P}} - \theta_{CL})|\theta_{CL}$ and $\sqrt{N}(\theta_{CL} - \theta_0)$; ... What about $(\bar{\theta}_{\mathcal{P}} - \theta_0)$?
- What happens if the CSGD algorithm is stopped too early, when $(\bar{\theta}_{\mathcal{P}} - \theta_{CL})|\theta_{CL}$ is still large?

CSGD - Three asymptotic regimes

Theorem: Asymptotic distribution for $\bar{\theta}_{\mathcal{P}}$

Consider $N/(T_N + N) \to \alpha$ with $0 \le \alpha \le 1$

- **Regime 1.** $\alpha = 0$:

$$\sqrt{N}(\bar{\theta}_{\mathcal{P}} - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}_d \left\{0, \Omega\right\}.$$

- **Regime 2.** $\alpha = 1$:

$$\sqrt{T_N}(\bar{\theta}_{\mathcal{P}} - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}_d \left\{0, E_Y\left(\Omega_{W\,|\,y}\right)\right\}.$$

- **Regime 3.** $0 < \alpha < 1$:

$$\sqrt{T_N + N}(\bar{\theta}_{\mathcal{P}} - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}_d \left\{0, \frac{E_Y\left(\Omega_{W\,|\,y}\right)}{1-\alpha} + \frac{\Omega}{\alpha}\right\}$$

Heuristic about total variability:

$$\begin{aligned}\mathsf{Var}_{W,Y}(\bar{\theta}_{\mathcal{P}}) &= E_Y\left\{\mathsf{Var}_{W\,|\,y}(\bar{\theta}_{\mathcal{P}})\right\} + \\ &+ \mathsf{Var}_Y\left\{E_Y(\bar{\theta}_{\mathcal{P}})\right\} \\ &\approx \frac{1}{T}E_Y\left(\Omega_{W\,|\,y}\right) + \frac{1}{N}\Omega.\end{aligned}$$

Factor analysis for ordinal data

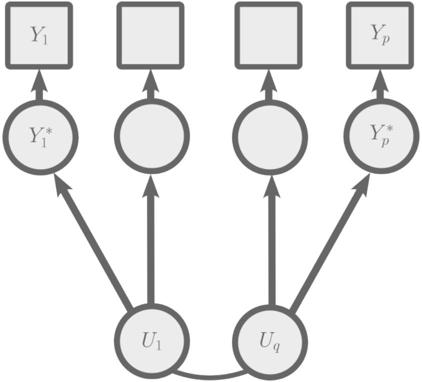


Figure: Example of ordinal factor model with simple loading structure.

Model setup:

- Data are assumed to be ordinal, $y_i = c_i \in \{0, \dots, m_i - 1\}$.

$$y_i = c_i \iff \tau^{(j)}_{c_i - 1} < y_i^* < \tau^{(i)}_{c_i},$$

- Underlying linear factor model:

$$y^* = \Lambda\eta + \epsilon,$$

where $\epsilon \sim \mathcal{N}_p(0, \Sigma_\epsilon)$ and $\Sigma_\epsilon = I_p - \text{diag}(\Lambda\Sigma_\eta\Lambda^T)$.

- $\theta = \Lambda, \Sigma_\eta, \tau$, where
  - Λ is the $p \times q$ loadings matrix $\Lambda = (\lambda_1^T, \dots, \lambda_p^T)$
  - Thresholds $\tau = (\tau^{(1)T}, \dots, \tau^{(p)T})^T$

Factor analysis for ordinal data - What's special?

Some considerations:

- Data reduced by sufficiency;
- The computational cost of $u_N(\boldsymbol{\theta}; \mathbf{y})$ is already $O(K)$ and does not depend on N;
- No way to use OSGD if $O(K)$ is still too expensive!
- We can adapt CSGD by collapsing the weighting matrix W onto a vector;

$$U(\theta; W; \mathbf{y}, \mathcal{P}) = \frac{1}{N} \sum_{j<j'} W_{jj'} \sum_{s_j, s_{j'}} \frac{n^{jj'}_{s_j s_{j'}}}{\pi^{jj'}_{s_j s_{j'}}} \nabla \pi^{jj'}_{s_j s_{j'}}$$

- We can arbitrarily choose how many sub-likelihoods to draw at each iteration (i.e. iteration complexity as low as $O(1)$).

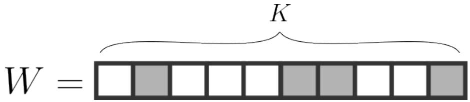


Figure: CSGD weighting vector for ordinal factor models.

The Big Five dataset

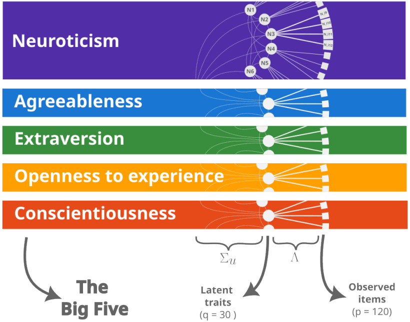


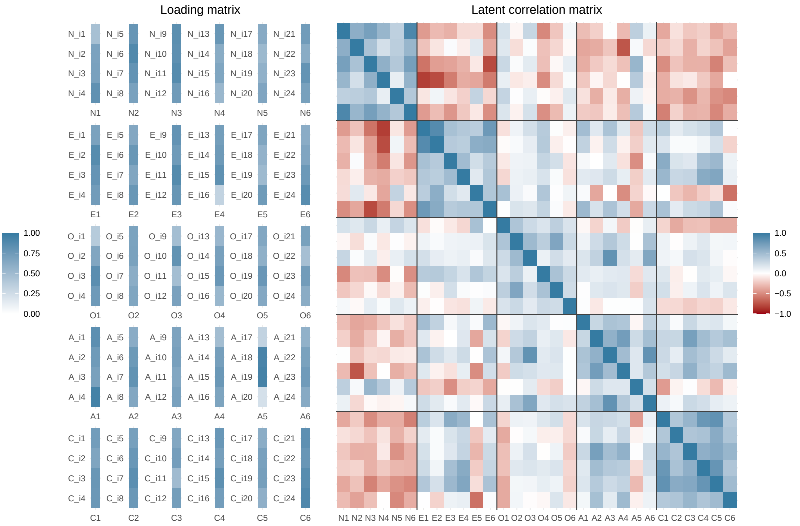
Figure: Structure of the Big Five factor model.

- Large-scale web-based test designed to measure 5 personality areas: **Neuroticism** (N), **Agreeableness** (A), **Extraversion**(E), **Openness to experience** (O) and **Conscientiousness** (C).
- Each area can be further split in 6 personality facets, for a total of **30 latent traits** to account for, potentially mutually correlated.
- The dataset consists of answers to **120 items** on a 5-point scale observed on more than **600 thousands units**.

The Big Five dataset - Results

Estimation details

- Confirmatory loading matrix with simple structure;
- Loadings and correlations initialized at 0.
- Sampling on average 16 pairs per iteration ($\approx 0.22\%$).
- Burn-in period of 2500 iterations.
- Convergence check on $\frac{|\theta^{(t)}-\theta^{(t-1)}|}{|\theta^{(t)}|}$. Tolerance set at 50 consecutive iterations below 5×10^{-5}.
- Convergence after 8311 iterations (≈ 955 seconds on single core, included frequencies computation).



Thank you for your attention!