PSYCHOMETRIKA
https://doi.org/10.1007/s11336-021-09751-8

SINGLE- AND MULTIPLE-GROUP PENALIZED FACTOR ANALYSIS: A TRUST-REGION ALGORITHM APPROACH WITH INTEGRATED AUTOMATIC MULTIPLE TUNING PARAMETER SELECTION

ELENA GEMINIANI

UNIVERSITY OF BOLOGNA

GIAMPIERO MARRA

UNIVERSITY COLLEGE LONDON

IRINI MOUSTAKI

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Penalized factor analysis is an efficient technique that produces a factor loading matrix with many zero elements thanks to the introduction of sparsity-inducing penalties within the estimation process. However, sparse solutions and stable model selection procedures are only possible if the employed penalty is non-differentiable, which poses certain theoretical and computational challenges. This article proposes a general penalized likelihood-based estimation approach for single- and multiple-group factor analysis models. The framework builds upon differentiable approximations of non-differentiable penalties, a theoretically founded definition of degrees of freedom, and an algorithm with integrated automatic multiple tuning parameter selection that exploits second-order analytical derivative information. The proposed approach is evaluated in two simulation studies and illustrated using a real data set. All the necessary routines are integrated into the `R` package `penfa`.

Key words: effective degrees of freedom, generalized information criterion, measurement invariance, penalized likelihood, simple structure.

1. Introduction

Factor analysis has been extensively applied in the social, behavioral and natural sciences as a data reduction method. For a given set of observed variables $x_1, \ldots, x_p$ one would like to find a set of latent factors $f_1, \ldots, f_r$, fewer in number than the observed variables ($r < p$), that contain essentially the same information. Factor analysis can be conducted in an exploratory (EFA; Mulaik, 2009) or confirmatory (CFA; Jöreskog, 1979) way. EFA analyzes a set of correlated observed variables without knowing in advance either the number of factors that are required to explain their interrelationships or their meaning. Depending on the r-factor model finally chosen (based on goodness-of-fit criteria and fit measures) as well as the rotation applied, an interpretation and labelling of the factors are given. CFA postulates certain relationships among the observed and latent variables by assuming a pre-specified pattern for the model parameters (factor loadings, structural parameters, unique variances). It is mainly concerned with testing hypotheses about the values of the factor loadings (usually, that some of them are zero).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-021-09751-8.

Correspondence should be made to Elena Geminiani, Department of Statistical Sciences, University of Bologna, Via Delle Belle Arti 41, 40126 Bologna, Italy. Email: elena.geminiani4@unibo.it; geminianielena@gmail.com

Published online: 26 March 2021

In data reduction techniques such as factor analysis, the interest is in obtaining factor solutions that exhibit a "simple structure" (Thurstone, 1947), that is, with many zero loadings and pure measures (i.e., each variable loads only on a single factor). In EFA this is accomplished with orthogonal or oblique factor rotations. However, rotations often do not generate loadings precisely equal to zero, so users have to manually set to zero those loadings that are smaller than a threshold (e.g., 0.30; Hair et al., 2010). Secondly, because each rotation is based on a specific optimization criterion, different rotations often lead to different factor structures which may all be far from "simple". In CFA, one usually resorts to modification indices (Chou & Huh, 2012) instead, but, if used extensively, they can lead to higher risks of capitalization on chance (MacCallum et al., 1992), and a lower probability of finding the best model specification (Chou & Bentler, 1990).

Penalized factor analysis is an alternative technique that produces parsimonious models using largely an automated procedure. The resulting models are less prone to instability in the estimation process and are easier to interpret and generalize than their unpenalized counterparts. It is based on the use of penalty functions that allow a subset of the model parameters (typically the factor loadings) to be automatically set to zero. The penalty is usually non-differentiable (Fan & Li, 2001), so that it produces a sparse factor structure, that is, a loading matrix where the number of nonzero entries is much smaller than the total number of its elements. This definition does not impose any pattern on the nonzero entries, so a simple structure is not enforced if it is not supported by the data. These sparsity-inducing penalties can reduce model complexity, enhance the interpretability of the results, and produce more stable parameter estimates. These benefits come, however, with a loss in model fit (i.e., a nonzero bias), so it is crucial to balance goodness of fit and sparsity appropriately. This can be achieved via the selection of a tuning parameter, which controls the amount of sparsity introduced in the model. A grid-search over a range of tuning values is generally conducted, and the optimal model is picked on the basis of information criteria or cross-validation.

In the last few years, several works have applied penalized estimation and regularization methods to models with latent variables. Choi, Oehlert and Zou (2010) used lasso ("least absolute shrinkage and selection operator"; Tibshirani, 1996) and adaptive lasso penalties in EFA. Since the lasso leads to biased estimates and overly dense factor structures, Hirose and Yamamoto (2014a; 2014b) employed non-convex penalties, such as the scad ("smoothly clipped absolute deviation") and the mcp ("minimax concave penalty"). Trendafilov, Fontanella and Adachi (2017) penalized a reparametrized loading matrix, whereas Jin, Moustaki and Yang-Wallentin (2018) considered a quadratic approximation of the objective function. Regularized methods have also been applied to structural equation models (SEM) for which CFA is a special case. Jacobucci, Grimm and McArdle (2016) developed the regularized SEM (RegSEM) using a reticular action model formulation and coordinate descent or general optimization routines. Huang, Chen and Weng (2017) and Huang (2020) examined the same problem of penalizing a SEM but employed a modification of the quasi-Newton algorithm.

Penalized estimation can be also extended to multiple-group analyses, such as cross-national surveys or cross-cultural assessments in psychological or educational testing. Recently, Huang (2018) and Lindstrøm and Dahl (2020) developed a penalized approach for multiple-group SEM, showing the benefits of using regularization techniques as alternatives to factorial invariance testing procedures (Meredith, 1993) to ascertain the differences and similarities of the parameter estimates across groups (see Bauer, Belzak & Cole, 2020 for a regularized approach for moderated non-linear factor analysis).

In this paper, we propose a penalized-estimation strategy for single- and multiple-group factor analysis models based on a carefully structured trust-region algorithm. The penalized optimization problem requires the availability of second-order analytical derivative information and thus twice-continuously differentiable functions. Because a sparse solution can be only achieved with non-differentiable penalties, we employ differentiable approximations of them. In particular, we locally approximate several convex and non-convex penalties, including lasso, adaptive lasso, scad and mcp. We also provide a theoretically founded definition of degrees of freedom (required when performing model selection) and present an efficient automatic procedure for the estimation of the tuning parameters, hence eliminating the need for computationally intensive grid searches as done in the literature. The proposed methodology is integrated into the `R` package `penfa` (a short form for *PENalized Factor Analysis*).

The paper is organized as follows. Sect. 2 briefly discusses the classical linear factor analysis model. In Sect. 3 we review and develop penalized likelihood estimation via locally approximated penalties. The extension of the model and the penalized approach for the case of multiple groups are described in Sects. 4 and 5, respectively. The derivation of the model degrees of freedom is presented in Sect. 6. Parameter estimation and the automatic selection of the tuning parameters are detailed in Sect. 7. The performance of the model is evaluated in two simulation studies (Sect. 8) and an empirical application (Sect. 9). Lastly, Sect. 10 concludes the paper and gives directions for future research. Additional details can be found in the Online Resources.

2. The Normal Linear Factor Analysis Model

The classical linear factor analysis model takes the form[1]:

$$\boldsymbol{x} = \boldsymbol{\Lambda f} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{x}$ is the $p \times 1$ vector of observed variables, $\boldsymbol{\Lambda}$ is the $p \times r$ factor loading matrix, $\boldsymbol{f}$ is the $r \times 1$ vector of common factors, and $\boldsymbol{\varepsilon}$ is the $p \times 1$ vector of unique factors. It is assumed that $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$, and $\boldsymbol{f}$ is independent of $\boldsymbol{\varepsilon}$. The observed variables are assumed to be conditionally independent (i.e., $\boldsymbol{\Psi}$ is a diagonal matrix), although this assumption can be relaxed if required. It then follows that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where the model-implied covariance matrix is $\boldsymbol{\Sigma} = \boldsymbol{\Lambda \Phi \Lambda}^T + \boldsymbol{\Psi}$.

It is possible to fix certain elements in $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ to zero based on a data generating hypothesis. The remaining $m \le \min\left(N, \frac{p(p+1)}{2}\right)$ elements, with N the total sample size, constitute the free parameters in $\mathrm{vec}(\boldsymbol{\Lambda})$, $\mathrm{diag}(\boldsymbol{\Psi})$, and $\mathrm{vech}(\boldsymbol{\Phi})$, and are collected in the vector $\boldsymbol{\theta}$, where the $\mathrm{vec}(\cdot)$ operator converts the enclosed matrix into a vector by stacking its columns, $\mathrm{diag}(\cdot)$ extracts the diagonal elements of the enclosed square matrix, and $\mathrm{vech}(\cdot)$ vectorizes the lower-diagonal part of the enclosed symmetric matrix. As it is common practice in these cases, we assume that the observed variables are measured as deviations from their means, so that the parameters only strive to reproduce the covariance matrix. As in Jöreskog (1979), we fix the variances of the common factors to unity for scale setting, and $r - 1$ elements of $\boldsymbol{\Lambda}$, in each column, to zero for uniqueness under factor rotation.

For a random sample of size N the log-likelihood is written as

$$\ell(\boldsymbol{\theta}) = -\frac{N}{2}\left\{\log|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{S\Sigma}^{-1}) + p\log(2\pi)\right\}, \tag{2}$$

where $\boldsymbol{S}$ is the sample covariance matrix. Since we are interested in introducing sparsity in the factor loading matrix, the estimation of the factor model will involve penalized likelihood procedures. The next section illustrates how such sparsity-inducing penalty functions can be specified and suitably approximated.

[1] An alternative factor model formulation would include the intercepts of the observed variables and the factor means. See the multiple-group extension in Sect. 4 for an example of a mean and covariance structure model.

3. Sparsity-Inducing Penalties

Since the primary interest of factor analysis is a sparse loading matrix, penalization is imposed on the factor loading matrix $\mathbf{\Lambda}$. Let us write the parameter vector as $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{q^\star}, \theta_{q^\star+1}, \dots, \theta_m)^T$, where the sub-vector $(\theta_1, \dots, \theta_{q^\star})^T$ collects the penalized parameters (i.e., the factor loadings), whereas $(\theta_{q^\star+1}, \dots, \theta_m)^T$ the unpenalized parameters (i.e., the free elements in $\mathbf{\Psi}$ and $\mathbf{\Phi}$). Because of the presence of fixed elements in $\mathbf{\Lambda}$ (Sect. 2), the number of penalized factor loadings $q^\star$ is smaller than $p \times r$. Define $\boldsymbol{R}_q = \text{diag}(0, 0, \dots, 0, 1, 0, \dots, 0)$ a diagonal matrix where the 1 on the $(q, q)^{\text{th}}$ entry of the matrix corresponds to the q^{th} parameter in $\boldsymbol{\theta}$, for $q = 1, \dots, q^\star$, and $\boldsymbol{R}_q = \boldsymbol{O}_{m \times m}$ for $q = q^\star + 1, \dots, m$. Let $\mathcal{P}_\eta(\boldsymbol{\theta})$ be a penalty function on the parameter vector $\boldsymbol{\theta}$, where $\eta \in [0, \infty)$ is a positive tuning parameter which determines the amount of shrinkage or penalization. The overall penalty is then given by the sum of the penalty terms for each parameter, that is, $\mathcal{P}_\eta(\boldsymbol{\theta}) = \sum_{q=1}^{m} \mathcal{P}_{\eta,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1)$, where $||\boldsymbol{R}_q\boldsymbol{\theta}||_1 = |\theta_q|$ if $q = 1, \dots, q^\star$, and zero otherwise. An example clarifying the formulation of this penalty is provided in Section B.1.1. One of the best-known penalties is the lasso (Tibshirani, 1996), which is defined as

$$\mathcal{P}_\eta^L(\boldsymbol{\theta}) = \eta \sum_{q=1}^{q^\star} |\theta_q|. \tag{3}$$

The potential problem with this penalty is that it penalizes all parameters equally, and thus can either select an overly complicated model or over-shrink large parameters. An ideal penalty should induce weak shrinkage on large effects and strong shrinkage on irrelevant effects (Tang, Shen, Zhang & Yi, 2017). To address this issue, alternative penalties have been developed, the most common ones being the adaptive lasso (alasso; Zou, 2006), scad (Fan & Li, 2001) and mcp (Zhang, 2010). These penalties give different amounts of shrinkage to each parameter, so each factor loading is weighted differently. Because of this, they lead to sparser solutions and enjoy the so-called "oracle" property, that is, when the true parameters have some zero loadings, they are estimated as zero with probability tending to one, and the nonzero loadings are estimated as well as when the correct submodel is known (Fan & Li, 2001). The alasso is defined as

$$\mathcal{P}_\eta^A(\boldsymbol{\theta}) = \eta \sum_{q=1}^{q^\star} w_q |\theta_q| = \eta \sum_{q=1}^{q^\star} \frac{|\theta_q|}{|\hat{\theta}_q|^a} \quad \text{for } a > 0. \tag{4}$$

It uses an adaptive weighting scheme based on a set of available weights $w_q = \dfrac{1}{|\hat{\theta}_q|^a}$ $(q = 1, \dots, q^\star)$, which are often taken to be the maximum likelihood estimates, that is, $w_q = \dfrac{1}{|\hat{\theta}_q^{\text{MLE}}|^a}$. As the exponent a gets larger, the relative strength of the penalization increases for smaller maximum likelihood estimates compared to larger maximum likelihood estimates.

Similarly, the scad and mcp use a varying weighting scheme. The scad is defined as

$$\begin{aligned}\mathcal{P}_\eta^S(\boldsymbol{\theta}) = \sum_{q=1}^{q^\star} \Bigg\{ & \eta|\theta_q| \mathbb{1}(0 \leq |\theta_q| \leq \eta) - \left[\frac{\theta_q^2 + \eta^2 - 2\eta a|\theta_q|}{2(a-1)}\right] \mathbb{1}(\eta < |\theta_q| \leq a\eta) \\ & + \frac{\eta^2(a+1)}{2} \mathbb{1}(|\theta_q| > a\eta) \Bigg\} \text{ for } a > 2, \end{aligned} \tag{5}$$

and the mcp as

$$\mathcal{P}_{\eta}^{M}(\boldsymbol{\theta}) = \sum_{q=1}^{q^{\star}} \left\{ \left(\eta|\theta_q| - \frac{\theta_q^2}{2a} \right) \mathbb{1}(0 \leq |\theta_q^2| \leq a\eta) + \frac{\eta^2 a}{2} \mathbb{1}(|\theta_q| > a\eta) \right\} \text{ for } a > 1, \tag{6}$$

where a is an additional tuning parameter. The superscripts L, A, S, M in equations (3)-(6) refer to the lasso, alasso, scad and mcp, respectively. The derivations of expressions (3)-(6) can be found in Section B.1.2. While the lasso and alasso are convex penalties, the scad and mcp are non-convex and can, therefore, make the optimization problem non-convex. In fact, a challenge with non-convex penalties is to find a good balance between sparsity and stability. To this end, both scad and mcp have an extra tuning parameter (a) which regulates their concavity so that, when it exceeds a threshold, the optimization problem becomes convex.

The above penalties help to obtain sparse solutions, however, they are non-differentiable, which is problematic for developing a coherent computational and theoretical inferential framework. The next section addresses this issue by replacing the non-differentiable penalties with their differentiable counterparts obtained via local approximations.

3.1. *Locally Approximated Penalties*

Ulbricht (2010) pointed out that a good penalty function should satisfy the following properties, for $q = 1, \ldots, m$: (P.1) $\mathcal{P}_{\eta,q} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ and $\mathcal{P}_{\eta,q}(0) = 0$; (P.2) $\mathcal{P}_{\eta,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1)$ continuous and strictly monotone in $||\boldsymbol{R}_q\boldsymbol{\theta}||_1$; (P.3) $\mathcal{P}_{\eta,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1)$ continuously differentiable $\forall\, ||\boldsymbol{R}_q\boldsymbol{\theta}||_1 \neq 0$, such that $\dfrac{\partial \mathcal{P}_{\eta,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1)}{\partial ||\boldsymbol{R}_q\boldsymbol{\theta}||_1} > 0$. We develop differentiable approximations of the above penalties that satisfy these properties. These approximations make the objective function differentiable, which is an indispensable prerequisite for the theoretical derivation of the degrees of freedom of the model, and a computationally and theoretically founded estimation framework (Sects. 6–7). In the same spirit, as for instance, Filippou, Marra and Radice (2017), we locally approximate the non-differentiable L_1-norms in (3)-(6) at $||\boldsymbol{R}_q\boldsymbol{\theta}||_1 = 0$ and combine this with ideas by Fan and Li (2001) and Ulbricht (2010). Let $||\boldsymbol{R}_q\boldsymbol{\theta}||_1 = ||\boldsymbol{\xi}_q||_1$, where the q^{th} element in $\boldsymbol{\xi}_q = (0, \ldots, 0, \theta_q, 0, \ldots, 0)^T$ corresponds to the q^{th} parameter in $\boldsymbol{\theta}$. Assume that an approximation $\mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{A})$ of the L_1-norm $||\cdot||_1$ exists such that

$$||\boldsymbol{\xi}_q||_1 = \mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{B}) = \lim_{\mathcal{A} \rightarrow \mathcal{B}} \mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{A}),$$

where $\mathcal{A}$ represents a set of possible tuning parameters, $\mathcal{B}$ is the set of boundary values for $||\boldsymbol{\xi}_q||_1$ and $\mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{A})$ is at least twice differentiable. We use $||\boldsymbol{\xi}_q||_1 = \mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{A}) = (\boldsymbol{\xi}_q^T\boldsymbol{\xi}_q + \bar{c})^{\frac{1}{2}}$ (Koch, 1996), with $\bar{c}$ a small positive real number (e.g., 10^{-8}) which controls the closeness between the approximation and the exact function. For all $\boldsymbol{\xi}_q$ for which the derivative $\dfrac{\partial ||\boldsymbol{\xi}_q||_1}{\partial \boldsymbol{\xi}_q}$ is defined, we assume that

$$\frac{\partial ||\boldsymbol{\xi}_q||_1}{\partial \boldsymbol{\xi}_q} = \frac{\partial \mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{B})}{\partial \boldsymbol{\xi}_q} = \lim_{\mathcal{A} \rightarrow \mathcal{B}} \mathcal{D}_1(\boldsymbol{\xi}_q, \mathcal{A}),$$

where $\mathcal{D}_1(\boldsymbol{\xi}_q, \mathcal{A}) = \dfrac{\partial \mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{A})}{\partial \boldsymbol{\xi}_q}$, and that $\mathcal{D}_1(\mathbf{0}, \mathcal{A}) = \mathbf{0}$. Then, the first derivative $\mathcal{D}_1(\boldsymbol{\xi}_q, \mathcal{A}) = (\boldsymbol{\xi}_q^T\boldsymbol{\xi}_q + \bar{c})^{-\frac{1}{2}}\boldsymbol{\xi}_q$ is a continuous approximation of the first-order derivative of the L_1-norm. Notice that $\mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{A})$ deviates only slightly from $\mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{B})$: when $\boldsymbol{\xi}_q = \mathbf{0}$ the deviation is $\sqrt{\bar{c}}$, whereas for any other value of $\boldsymbol{\xi}_q$ the deviation is less than $\bar{c}$.

Penalty $\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta})$ for $\mathcal{T} = \{L, A, S, M\}$ can be locally approximated by a quadratic function as follows. Suppose that $\tilde{\boldsymbol{\theta}}$ is an initial value close to the true value of $\boldsymbol{\theta}$. Then, we approximate $\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta})$ by a Taylor expansion of order one at $\tilde{\boldsymbol{\theta}}$, that is,

$$\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta}) \approx \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}), \tag{7}$$

where $\nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \dfrac{\partial \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}$. As proved in Section B.1.3, $\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta})$ is approximated as

$$\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta}) \approx \frac{1}{2}\boldsymbol{\theta}^T \left\{ \sum_{q=1}^{m} \frac{\partial \mathcal{P}_{\eta,q}^{\mathcal{T}}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T \boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \boldsymbol{R}_q^T \boldsymbol{R}_q \right\} \boldsymbol{\theta} = \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}.$$

The penalty matrix $\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ is an $m \times m$ block diagonal matrix of the form:

$$\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} \boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}. \tag{8}$$

The first block is composed of the $q^\star \times q^\star$ diagonal matrix $\boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ and corresponds to the parameters to penalize, whereas the second block is an $(m - q^\star)$-dimensional null matrix relative to the parameters unaffected by the penalization. The matrix $\boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ is in turn a diagonal matrix whose entries $m_q^{\mathcal{T}} = \dfrac{\partial \mathcal{P}_{\eta,q}^{\mathcal{T}}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} \dfrac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T \boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}}$ (for $q = 1, \ldots, q^\star$) determine the amount of shrinkage on $\tilde{\theta}_q$ controlled by the tuning η and required by penalty $\mathcal{T}$. Their expressions for the lasso, alasso, scad and mcp are (see Section B.1.3.1)

$$\left[\boldsymbol{\mathcal{M}}_\eta^{L}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^L = \frac{\eta}{\sqrt{\tilde{\theta}_q^2 + \bar{c}}}, \tag{9}$$

$$\left[\boldsymbol{\mathcal{M}}_\eta^{A}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^A = \frac{\eta}{|\hat{\theta}_q|^a \sqrt{\tilde{\theta}_q^2 + \bar{c}}}, \tag{10}$$

$$\left[\boldsymbol{\mathcal{M}}_\eta^{S}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^S = \frac{\eta\left[\mathbb{1}(|\tilde{\theta}_q| \leq \eta) + \dfrac{\max(a\eta - |\tilde{\theta}_q|, 0)}{(a-1)\eta}\mathbb{1}(|\tilde{\theta}_q| > \eta)\right]}{\sqrt{\tilde{\theta}_q^2 + \bar{c}}}, \tag{11}$$

$$\left[\boldsymbol{\mathcal{M}}_{\eta}^{M}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^M = \frac{\left(\eta - \dfrac{|\tilde{\theta}_q|}{a}\right)\mathbb{1}(|\tilde{\theta}_q| < \eta a)}{\sqrt{\tilde{\theta}_q^2 + \bar{c}}}. \tag{12}$$

4. The Multiple-Group Factor Analysis Model

In studies of multiple groups of respondents, such as cross-national surveys and cross-cultural assessments in psychological or educational testing, the interest often lies in the comparisons of the groups with respect to their factor structures. In this case, the model becomes

$$\boldsymbol{x}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{f}_g + \boldsymbol{\varepsilon}_g \qquad \text{for } g = 1, \ldots, G, \tag{13}$$

where the subscript g denotes the group, and $\boldsymbol{\tau}_g$ the intercept terms. It is assumed that $\boldsymbol{f}_g \sim \mathcal{N}(\boldsymbol{\kappa}_g, \boldsymbol{\Phi}_g)$, $\boldsymbol{\varepsilon}_g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$, $\boldsymbol{f}_g$ is independent of $\boldsymbol{\varepsilon}_g$, and $\boldsymbol{\Psi}_g$ is a diagonal matrix. Then, it follows that $\boldsymbol{x}_g \sim \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where the model-implied moments are $\boldsymbol{\mu}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g\boldsymbol{\kappa}_g$ and $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Phi}_g\boldsymbol{\Lambda}_g^T + \boldsymbol{\Psi}_g$.

We set the metric of the factors and the necessary identification restrictions through the "marker-variable" approach (Little, Slegers & Card, 2006), which relies on the selection of a representative variable (marker) for each factor in each group. Then, we fix the intercepts of the markers to zero, the loadings on the "marked" factors to 1.0, and those on the remaining factors to zero. All of the other parameters are estimated. The choice of the markers is crucial and should be an accurate one (Millsap, 2001). Alternative identification methods are discussed in Millsap (2012).

The free parameters of each group appearing in $\text{vec}(\boldsymbol{\Lambda}_g)$, $\boldsymbol{\tau}_g$, $\text{diag}(\boldsymbol{\Psi}_g)$, $\text{vech}(\boldsymbol{\Phi}_g)$, and $\boldsymbol{\kappa}_g$ are collected in the m_g-dimensional vector $\boldsymbol{\theta}_g$, for $g = 1, \ldots, G$. Each group parameter vector is collected in the overall m-dimensional vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T, \ldots, \boldsymbol{\theta}_G^T)^T$, where $m = \sum_{g=1}^{G} m_g$. Assume for convenience that the same set of parameters is estimated in every group, which implies that the number of observed variables p and factors r is the same across groups, the fixed elements required for identification are placed in the same positions across groups, and that $m_1 = \ldots = m_G$, so that $m = m_1 G$. Given random samples of sizes $N_1, \ldots, N_G$, with $N = \sum_{g=1}^{G} N_g$ the total sample size across groups, the log-likelihood of the multiple-group factor model is

$$\ell(\boldsymbol{\theta}) = -\sum_{g=1}^{G} \frac{N_g}{2}\{\log|\boldsymbol{\Sigma}_g| + \text{tr}(\boldsymbol{W}_g\boldsymbol{\Sigma}_g^{-1}) + p\log(2\pi)\}, \tag{14}$$

where $\boldsymbol{W}_g = \boldsymbol{S}_g + (\bar{\boldsymbol{x}}_g - \boldsymbol{\mu}_g)(\bar{\boldsymbol{x}}_g - \boldsymbol{\mu}_g)^T$.

In multiple-group analyses, an important methodological consideration is the establishment of the comparability or "equivalence" of measurement across the groups (e.g., countries, socio-economical groups). Measurement (or factorial) invariance occurs when the factors have the same meaning in each group, which translates into equal measurement models (i.e., factor loadings, intercepts and unique variances) across groups (Millsap 2012). If non-equivalence of measurement exists, substantively interesting group comparisons may become distorted. Testing for measurement invariance in the parameters is, however, an intensive process. A sequence of nested tests is progressively conducted to establish the equivalence in the factor loadings, the intercepts, and optionally the unique variances (Vandenberg & Lance, 2000). The next section describes the penalty functions that can be incorporated into the multiple-group model to obtain a technique that automatically detects parameter equivalence across groups.

5. Sparsity and Invariance-Inducing Penalties

As in the single-group factor model, we can penalize the factor loadings to automatically obtain a sparse loading matrix in each of the groups. Define the diagonal matrix $\boldsymbol{R}_q = \text{diag}(0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 on the $(q, q)^{\text{th}}$ entry of the matrix corresponds to the q^{th} factor loading in $\boldsymbol{\theta}$, for $q = (g-1)m_1 + 1, \ldots, (g-1)m_1 + q^\star$ and $g = 1, \ldots, G$, and $\boldsymbol{R}_q = \boldsymbol{O}_{m \times m}$ for the remaining parameters. The quantity $q^\star$ represents the number of penalized loadings in each group. Then, the sparsity-inducing penalty on the factor loadings is $\mathcal{P}^{\mathcal{T}}_{\eta_1}(\boldsymbol{\theta}) = \sum_{q=1}^{m} \mathcal{P}^{\mathcal{T}}_{\eta_1,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1)$, where $\eta_1 \in [0, \infty)$ controls the overall amount of shrinkage.

In the same spirit as factorial invariance, we can specify a penalty encouraging the equality of the loadings across groups. Conveniently, this can be achieved by shrinking the pairwise absolute differences of every factor loading across groups. Let $\boldsymbol{D}^{\boldsymbol{\Lambda}}_q$, for $q = 1, \ldots, q^\star$, be the matrix computing the differences of the factor loading pairs $(\theta_{(g-1)m_1+q}, \theta_{(g'-1)m_1+q})$ for $g < g'$, whereas for the other parameters $\boldsymbol{D}^{\boldsymbol{\Lambda}}_q = \boldsymbol{O}_{m_1\binom{G}{2}\times m}$. Then, the penalty inducing equal loadings across groups can be written as $\mathcal{P}^{\mathcal{T}}_{\eta_2}(\boldsymbol{\theta}) = \sum_{q=1}^{m} \mathcal{P}^{\mathcal{T}}_{\eta_2,q}(||\boldsymbol{D}^{\boldsymbol{\Lambda}}_q\boldsymbol{\theta}||_1)$, where $||\boldsymbol{D}^{\boldsymbol{\Lambda}}_q\boldsymbol{\theta}||_1 = \sum_{g<g'}|\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}|$ for $q = 1, \ldots, q^\star$, and zero otherwise. If $G = 2$, the absolute difference of the q^{th} loading across the two groups is expressed as $||\boldsymbol{D}^{\boldsymbol{\Lambda}}_q\boldsymbol{\theta}||_1 = |\theta_q - \theta_{m_1+q}|$, where $\boldsymbol{D}^{\boldsymbol{\Lambda}}_q = [\boldsymbol{R}_q, \ -\boldsymbol{R}_q]$. The expression of $\mathcal{P}^{\mathcal{T}}_{\eta_2}(\boldsymbol{\theta})$ for lasso, alasso, scad and mcp is given in Section B.2.1. The tuning parameter $\eta_2 \in [0, \infty)$ controls the amount of loading equality across groups. When the loadings are truly invariant and η_2 is properly chosen, the penalized group loading matrices "fuse", and share the same values.

Lastly, we can encourage the equality of the intercepts across groups by specifying a penalty shrinking their pairwise absolute group differences. Let $k^\star$ be the number of estimated intercepts in each group. Due to the presence of fixed elements in $\boldsymbol{\tau}_g$ for identification, $k^\star$ is smaller than p. Let $\boldsymbol{D}^{\boldsymbol{\tau}}_q$, for $q = (g-1)m_1+q^\star+1, \ldots, (g-1)m_1+q^\star+k^\star$, be a matrix of known constants computing the differences of the intercepts across groups, whereas for all of the other parameters (i.e., the loadings, the unique variances and the structural parameters) $\boldsymbol{D}^{\boldsymbol{\tau}}_q = \boldsymbol{O}_{m_1\binom{G}{2}\times m}$. The penalty inducing equal intercepts across groups is then written as $\mathcal{P}^{\mathcal{T}}_{\eta_3}(\boldsymbol{\theta}) = \sum_{q=1}^{m} \mathcal{P}^{\mathcal{T}}_{\eta_3,q}(||\boldsymbol{D}^{\boldsymbol{\tau}}_q\boldsymbol{\theta}||_1)$, where $\eta_3 \in [0, \infty)$ governs the amount of intercept invariance.

Optionally, one can encourage the invariance of the unique variances. However, as argued by Little et al. (2012), these quantities contain both random sources of errors, for which there is no theoretical reason to expect equality across groups, and item-specific components, which can vary as a function of various measurement factors. In light of this, we do not introduce a penalty on the unique variances, as their cross-group equivalence would not provide any additional evidence of comparability of the constructs because the important measurement parameters (i.e., the factor loadings and the intercepts) are already encouraged to be invariant by penalties $\mathcal{P}^{\mathcal{T}}_{\eta_2}$ and $\mathcal{P}^{\mathcal{T}}_{\eta_3}$.

The three penalties can be easily combined into a single penalty that simultaneously generates sparsity on the factor loading matrices and equivalent loadings and intercepts

$$\begin{aligned}\mathcal{P}^{\mathcal{T}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) &= \mathcal{P}^{\mathcal{T}}_{\eta_1}(\boldsymbol{\theta}) + \mathcal{P}^{\mathcal{T}}_{\eta_2}(\boldsymbol{\theta}) + \mathcal{P}^{\mathcal{T}}_{\eta_3}(\boldsymbol{\theta}) \\ &= \sum_{q=1}^{m} \left\{ \mathcal{P}^{\mathcal{T}}_{\eta_1,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1) + \mathcal{P}^{\mathcal{T}}_{\eta_2,q}(||\boldsymbol{D}^{\boldsymbol{\Lambda}}_q\boldsymbol{\theta}||_1) + \mathcal{P}^{\mathcal{T}}_{\eta_3,q}(||\boldsymbol{D}^{\boldsymbol{\tau}}_q\boldsymbol{\theta}||_1) \right\},\end{aligned} \tag{15}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)^T$ is the vector of the tuning parameters. Each penalty is controlled by its own tuning parameter, as we do not a priori expect these values to be equal. The penalties in (15) can be any of the functions illustrated in Sect. 3, including lasso, alasso, scad and mcp, and different penalty functions can be in principle combined. Suppose that the adaptive weights are available for the intercepts but not for the full loading matrices, possibly due to some inadmissible loading values. In this case, one can combine the alasso penalty for intercept similarity with the lasso (which also supports the automatic procedure, contrarily to the scad and mcp) for sparsity and loading equivalence. By following the rationale described in Sect. 3.1, we replace each non-differentiable penalty in (15) with its differentiable local approximation:

$$\mathcal{P}^{\mathcal{T}}_{\eta_1}(\boldsymbol{\theta}) \approx \frac{1}{2}\boldsymbol{\theta}^T \left\{ \sum_{q=1}^{m} \frac{\partial \mathcal{P}^{\mathcal{T}}_{\eta_1,q}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T \boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \boldsymbol{R}_q^T \boldsymbol{R}_q \right\} \boldsymbol{\theta} = \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{D}}^{\mathcal{T}}_{\eta_1}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta},$$

$$\mathcal{P}^{\mathcal{T}}_{\eta_2}(\boldsymbol{\theta}) \approx \frac{1}{2}\boldsymbol{\theta}^T \left\{ \sum_{q=1}^{m} \frac{\partial \mathcal{P}^{\mathcal{T}}_{\eta_2,q}(||\boldsymbol{D}^{\boldsymbol{\Lambda}}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{D}^{\boldsymbol{\Lambda}}_q\tilde{\boldsymbol{\theta}}||_1} \frac{1}{\sqrt{(\boldsymbol{D}^{\boldsymbol{\Lambda}}_q\tilde{\boldsymbol{\theta}})^T \boldsymbol{D}^{\boldsymbol{\Lambda}}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \boldsymbol{D}^{\boldsymbol{\Lambda}^T}_q \boldsymbol{D}^{\boldsymbol{\Lambda}}_q \right\} \boldsymbol{\theta} = \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{D}}^{\mathcal{T}}_{\eta_2}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta},$$

$$\mathcal{P}^{\mathcal{T}}_{\eta_3}(\boldsymbol{\theta}) \approx \frac{1}{2}\boldsymbol{\theta}^T \left\{ \sum_{q=1}^{m} \frac{\partial \mathcal{P}^{\mathcal{T}}_{\eta_3,q}(||\boldsymbol{D}^{\boldsymbol{\tau}}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{D}^{\boldsymbol{\tau}}_q\tilde{\boldsymbol{\theta}}||_1} \frac{1}{\sqrt{(\boldsymbol{D}^{\boldsymbol{\tau}}_q\tilde{\boldsymbol{\theta}})^T \boldsymbol{D}^{\boldsymbol{\tau}}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \boldsymbol{D}^{\boldsymbol{\tau}^T}_q \boldsymbol{D}^{\boldsymbol{\tau}}_q \right\} \boldsymbol{\theta} = \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{D}}^{\mathcal{T}}_{\eta_3}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta},$$

which leads to the following differentiable form of the combined penalty:

$$\mathcal{P}^{\mathcal{T}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^T \{\boldsymbol{\mathcal{D}}^{\mathcal{T}}_{\eta_1}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{\mathcal{D}}^{\mathcal{T}}_{\eta_2}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{\mathcal{D}}^{\mathcal{T}}_{\eta_3}(\tilde{\boldsymbol{\theta}})\}\boldsymbol{\theta} = \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}^{\mathcal{T}}_{\boldsymbol{\eta}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}. \tag{16}$$

Additional details on the structure of the matrix $\boldsymbol{\mathcal{D}}^{\mathcal{T}}_{\eta_2}(\tilde{\boldsymbol{\theta}})$ are given in Section B.2.1. For an example clarifying the formulation of these matrices for the multiple-group model, the reader is referred to Section B.2.2.

6. Generalized Information Criterion

The previously illustrated penalties can be directly introduced within the estimation process by means of penalized maximum likelihood estimation procedures. The penalized log-likelihood is

$$\ell_p(\boldsymbol{\theta}) := \sum_{\alpha=1}^{N} \left\{ \ell(\boldsymbol{x}_\alpha|\boldsymbol{\theta}) - \mathcal{P}^{\mathcal{T}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \right\} = \ell(\boldsymbol{\theta}) - N\,\mathcal{P}^{\mathcal{T}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}). \tag{17}$$

For the normal linear factor model, $\ell(\boldsymbol{\theta})$ is given in equation (2), $\mathcal{P}^{\mathcal{T}}_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ is one of the penalties of Sect. 3 generating a sparse factor solution, and the vector $\boldsymbol{\eta}$ reduces to the scalar η; for the multiple-group model, $\ell(\boldsymbol{\theta})$ is given in equation (14), $\mathcal{P}^{\mathcal{T}}_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ is one of the penalties of Sect. 5 inducing sparsity and invariant loadings and intercepts, and $\boldsymbol{\eta}$ is equal to the triplet $(\eta_1, \eta_2, \eta_3)^T$.

Simultaneous estimation of all parameters is achieved by maximizing the penalized log-likelihood in (17) and using a local approximation of $\mathcal{P}^{\mathcal{T}}_{\boldsymbol{\eta}}(\boldsymbol{\theta})$, that is,

$$\max_{\boldsymbol{\theta}} \ell_p(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \left\{ \ell(\boldsymbol{\theta}) - \frac{N}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}^{\mathcal{T}}_{\boldsymbol{\eta}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right\}, \tag{18}$$

where the function in brackets is now twice-continuously differentiable. The penalized maximum likelihood estimator (PMLE) is then defined as $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_p(\boldsymbol{\theta})$. Conveniently, the gradient of the penalized log-likelihood can be written as $\boldsymbol{g}_p(\boldsymbol{\theta}) := \frac{\partial \ell_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{g}(\boldsymbol{\theta}) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}$, where $\boldsymbol{g}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, the Hessian matrix of the second-order derivatives $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}) := \frac{\partial^2 \ell_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$, where $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$, and the expected Fisher information $\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}) := -\mathbb{E}\left[\frac{\partial^2 \ell_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] = \boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$, where $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]$.

A crucial aspect lies in the selection of $\boldsymbol{\eta}$, which controls the amount of penalization introduced in the model. To select $\boldsymbol{\eta}$, we elect to use the Generalized Information Criterion (GIC; Konishi & Kitagawa, 1996), which is based on a theoretically founded definition of degrees of freedom. Notice that this choice is possible because the quantities we are dealing with are twice-continuously differentiable. We resort to the general approach of the GIC because the penalized maximum likelihood estimator cannot be ascribed to the ordinary maximum likelihood framework postulated by the AIC, and not for relaxing the assumption $\mathbb{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] = \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right]$, which does hold true for the normal linear factor models considered in this paper. Let G be the true distribution function that generated the data $\boldsymbol{x}_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, which are realizations of the random vector $\boldsymbol{\mathcal{X}}_N = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)^T$. Let us express the parameter vector as $\boldsymbol{\theta} = \boldsymbol{T}(G)$, where $\boldsymbol{T}(G)$ is the m-dimensional functional vector of G defined as the solution of the implicit equations $\int \boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{T}(G)) dG(\boldsymbol{x}) = \boldsymbol{0}$, with $\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{T}(G)) = \frac{\partial}{\partial \boldsymbol{\theta}}\{\ell(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\}\Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)}$. The log-likelihood and the penalty matrix take different forms depending on whether we deal with a single- or multiple-group factor model. The GIC evaluating the goodness of fit of the penalized model, when used to predict independent future data z generated from the unknown distribution G, is (see Online Resource C)

$$\mathrm{GIC}(\boldsymbol{\mathcal{X}}_N; G) = -2\sum_{\alpha=1}^{N} \ell(\boldsymbol{x}_\alpha|\boldsymbol{\theta}) + 2\mathrm{tr}\{\boldsymbol{R}(\boldsymbol{\psi}, G)^{-1}\boldsymbol{Q}(G)\},$$

where

$$\boldsymbol{R}(\boldsymbol{\psi}, G) = -\int \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\left\{\ell(z|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right\}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \mathrm{d}G(z),$$
$$\boldsymbol{Q}(G) = -\int \frac{\partial^2 \ell(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \mathrm{d}G(z),$$

and $\boldsymbol{\eta}$ enters the penalty matrix $\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$. By considering the PMLE $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$, and replacing the unknown distribution G with its empirical counterpart $\hat{G}$ based on the data, we have

$$\mathrm{GIC}(\boldsymbol{\mathcal{X}}_N; \hat{G}) = -2\sum_{\alpha=1}^{N} \ell(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + 2\mathrm{tr}\{\boldsymbol{R}(\boldsymbol{\psi}, \hat{G})^{-1}\boldsymbol{Q}(\hat{G})\},$$

where

$$\begin{aligned}
\boldsymbol{R}(\boldsymbol{\psi}, \hat{G}) &= -\frac{1}{N}\sum_{\alpha=1}^{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\left\{\ell(\boldsymbol{x}_\alpha|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\boldsymbol{\theta})\boldsymbol{\theta}\right\}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\frac{1}{N}\{\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\}\\
&= -\frac{1}{N}\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}}),\\
\boldsymbol{Q}(\hat{G}) &= -\frac{1}{N}\sum_{\alpha=1}^{N}\frac{\partial^2\ell(\boldsymbol{x}_\alpha|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\frac{1}{N}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}).
\end{aligned}$$

The effective number or estimated degrees of freedom (edf) of the model is thus equal to edf $=$ tr$\left\{\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\}$. The formula for the edf is thus readily obtained by adapting the existing results for general likelihoods (of which the differentiable function in (18) is an example) to the penalized framework and assuming the usual regularity conditions. The GIC is an extension of the Akaike Information Criterion (AIC; Akaike, 1974), and as such, it may inherit the tendency of the latter to select overly complex models. To avoid this issue, we can change the constant 2 of the bias term to $\log(N)$ (used in the Bayesian Information Criterion; Schwarz, 1978). Then, given grid(s) of values, the optimal $\hat{\boldsymbol{\eta}}$ can be chosen using the following Generalized Bayesian Information Criterion (GBIC)

$$\text{GBIC}(\boldsymbol{\mathcal{X}}_N; \hat{G}) = -2\sum_{\alpha=1}^{N}\ell(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + \log(N)\text{tr}\{\boldsymbol{R}(\boldsymbol{\psi}, \hat{G})^{-1}\boldsymbol{Q}(\hat{G})\}. \tag{19}$$

The optimal penalized factor model is hence chosen to be the one with the lowest BIC, as this is the information criterion routinely employed in sparse settings. However, if researchers are more interested in accuracy and achieving minimum prediction error, then the AIC is to be preferred. In the presence of moderate sample size and many variables, the extended BIC (EBIC; Chen & Chen, 2008) may be more suitable.

The edf of an unpenalized model ($\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}} = \boldsymbol{O}_{m\times m}$) coincide with the dimension of the parameter vector $\boldsymbol{\theta}$, since tr$\left\{\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\}$ $=$ tr$\left\{\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\}$ $=$ tr$(\boldsymbol{I}_m) = m$, where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. For a penalized model edf $=$ tr$\left\{\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\}$ $=$ $m-$tr$\left\{[-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})]^{-1}N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right\}$. This shows that edf $\to m$ as $\boldsymbol{\eta} \to \boldsymbol{0}$, and edf $\to m - r^\star$ as $\boldsymbol{\eta} \to \infty$, where $r^\star$ is the number of penalized elements (equal to $q^\star$ for the factor model, and $G(q^\star + k^\star)$ for the multiple-group extension). When $\boldsymbol{0} < \boldsymbol{\eta} < \infty$, the edf $\in [m - r^\star, m]$. The overall edf of a fitted model is given by the sum of the edf for each parameter; each single edf takes a value in the range [0, 1] and quantifies precisely the extent to which each coefficient is penalized.

The existing penalized factor models (Choi et al., 2010; Hirose & Yamamoto, 2014a; Jacobucci et al., 2016; Huang et al., 2017; Huang, 2018; Jin et al., 2018) compute the degrees of freedom as the number of nonzero parameters (referred in the following as dof), by advocating the fact that the number of nonzero coefficients in a lasso-penalized linear model gives an unbiased estimate of the total degrees of freedom (Zou et al., 2007). This way of estimating the degrees of freedom implies that each dof can be either 0 if its parameter has been shrunken to zero, or 1 otherwise. On the contrary, the edf can take any value in [0, 1]. This suggests that, while the definitions of dof and edf may produce equivalent results (for penalties enjoying the oracle property, as the alasso, scad and mcp), in practical situations using edf is expected to yield better-calibrated degrees of freedom. The proposed method also treats the estimated edf as they are. Importantly, the definition of edf directly stems from the estimated bias term of the GIC, which gives it a theoretically founded basis.

7. Penalized Maximum Likelihood Estimation

For any given set of values of $\boldsymbol{\eta}$ in the penalty matrix, which is hence denoted in the following as $\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$, we minimize $-\ell_p(\boldsymbol{\theta})$ via a trust-region algorithm (Nocedal & Wright, 2006). At iteration $\mathfrak{t}$, a "model function" $\mathcal{Q}_p^{[\mathfrak{t}]}$ is constructed, whose behavior near the current point $\boldsymbol{\theta}^{[\mathfrak{t}]}$ is similar to that of the actual objective function. The model function is usually a quadratic approximation of $-\ell_p$ at $\boldsymbol{\theta}^{[\mathfrak{t}]}$:

$$\mathcal{Q}_p^{[\mathfrak{t}]}(\boldsymbol{s}) = -\left\{\ell_p\left(\boldsymbol{\theta}^{[\mathfrak{t}]}\right) + \boldsymbol{s}^T\boldsymbol{g}_p\left(\boldsymbol{\theta}^{[\mathfrak{t}]}\right) + \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\mathcal{H}}_p\left(\boldsymbol{\theta}^{[\mathfrak{t}]}\right)\boldsymbol{s}\right\},$$

where $\boldsymbol{s}$ is the trial step vector aiming at reducing the model function, $\boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathfrak{t}]})$ the penalized score function, and $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}^{[\mathfrak{t}]})$ the penalized Hessian matrix. For the normal linear factor model, the derivation of the second-order derivatives is a tedious and lengthy process; however, the availability of these quantities guarantees a better accuracy of the algorithm since no numerical approximation is employed. Because the Hessian requires computing many elements, we also considered the Fisher information matrix. If the elements of $(\hat{\boldsymbol{\Sigma}} - \boldsymbol{S})$ are small and the second derivatives not too large, which is often the case, the information matrix is very close to the true Hessian. For the multiple-group model, due to the presence of the parameters for the mean structure besides those for the covariance structure, we only considered the information matrix as it exhibited similar numerical performances to the Hessian at a reduced computational cost. The analytical expressions of these derivatives for the single- and multiple-group model are given in Geminiani (2020, Appendices A, F, respectively).

The search for a minimizer of $\mathcal{Q}_p^{[\mathfrak{t}]}$ is restricted to some region around $\boldsymbol{\theta}^{[\mathfrak{t}]}$, which is usually the ball $||\boldsymbol{s}||_2 < \Delta^{[\mathfrak{t}]}$, where $||\cdot||_2$ is the Euclidean norm, and $\Delta^{[\mathfrak{t}]} > 0$ the trust-region radius at iteration $\mathfrak{t}$. The size of the trust region is critical to the effectiveness of each step: if it is too small, the algorithm may miss the opportunity to take a step that moves it closer to the minimizer of the objective function; if it is too large, the minimizer of the model may be far from the one of the objective function in the region, so it may be necessary to reduce the region size and repeat the process. Each iteration of the trust-region algorithm solves the subproblem:

$$\boldsymbol{s}^{[\mathfrak{t}]} = \arg\min_{\boldsymbol{s}\in\mathbb{R}^m} \mathcal{Q}_p^{[\mathfrak{t}]}(\boldsymbol{s}) \qquad \text{subject to } ||\boldsymbol{s}||_2 \leq \Delta^{[\mathfrak{t}]}, \tag{20}$$

$$\boldsymbol{\theta}^{[\mathfrak{t}+1]} = \boldsymbol{\theta}^{[\mathfrak{t}]} + \boldsymbol{s}^{[\mathfrak{t}]}, \tag{21}$$

where the current iteration $\boldsymbol{\theta}^{[\mathfrak{t}]}$ is updated with $\boldsymbol{s}^{[\mathfrak{t}]}$ if this step produces an improvement over the objective function. The size of the region is chosen by measuring the agreement between the model function and the objective function at previous iterations through the ratio:

$$r^{[\mathfrak{t}]} = \frac{-\left[\ell_p(\boldsymbol{\theta}^{[\mathfrak{t}]}) - \ell_p(\boldsymbol{\theta}^{[\mathfrak{t}]} + \boldsymbol{s}^{[\mathfrak{t}]})\right]}{\mathcal{Q}_p^{[\mathfrak{t}]}(\boldsymbol{0}) - \mathcal{Q}_p^{[\mathfrak{t}]}(\boldsymbol{s}^{[\mathfrak{t}]})}. \tag{22}$$

Algorithm 1
Trust-region algorithm

```
Require: $\Delta_{\max} > 0$, $\Delta_0 \in (0, \Delta_{\max})$, $\boldsymbol{\theta}^{[0]}$
Compute $\ell_p(\boldsymbol{\theta}^{[0]})$, $\boldsymbol{g}_p(\boldsymbol{\theta}^{[0]})$, $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}^{[0]})$
Set $\epsilon = \texttt{.Machine\$double.eps}^{\frac{1}{2}} = 1.490116 \times 10^{-8}$
while t $\leq$ 1000 or $\left| -\left[\ell_p(\boldsymbol{\theta}^{[t]}) - \ell_p(\boldsymbol{\theta}^{[t+1]})\right] \right| < \epsilon$ do
    $\boldsymbol{s}^{[t]} = \arg\min_{\boldsymbol{s}:||\boldsymbol{s}||_2 \leq \Delta^{[t]}} \mathcal{Q}_p^{[t]}(\boldsymbol{s})$
    $r^{[t]} = \dfrac{-\left[\ell_p(\boldsymbol{\theta}^{[t]}) - \ell_p(\boldsymbol{\theta}^{[t]} + \boldsymbol{s}^{[t]})\right]}{\mathcal{Q}_p^{[t]}(\mathbf{0}) - \mathcal{Q}_p^{[t]}(\boldsymbol{s}^{[t]})}$
    if $r^{[t]} < \frac{1}{4}$ then
        $\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]}$
        $\Delta^{[t+1]} = \dfrac{||\boldsymbol{s}^{[t]}||_2}{4}$
    else
        $\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} + \boldsymbol{s}^{[t]}$
        if $r^{[t]} > \frac{3}{4}$ and $||\boldsymbol{s}^{[t]}||_2 = \Delta^{[t]}$ then
            $\Delta^{[t+1]} = \min(2\Delta^{[t]}, \Delta_{\max})$
        else
            $\Delta^{[t+1]} = \Delta^{[t]}$
        end if
    end if
end while
```

The numerator quantifies the actual reduction, and the denominator the predicted reduction. If $r^{[t]}$ is negative, the model is an inadequate representation of the objective function over the current trust region, so the step $\boldsymbol{s}^{[t]}$ is rejected, and the new problem is solved with a smaller region. If $r^{[t]}$ is close to 1, there is good agreement between $\mathcal{Q}_p^{[t]}$ and $-\ell_p$ over $\boldsymbol{s}^{[t]}$, so the model can accurately predict the behavior of the objective function along that step, and the trust region is enlarged for the next iteration. If $r^{[t]}$ is positive, but not close to 1, the trust region is not altered, unless it is close to zero or negative, in which case it is shrunken. Algorithm 1 describes the process. The term $\Delta_{\max}$ represents an overall bound on the step lengths. Trust-region algorithms never run too far from the current iteration as the points outside the trust region are not considered. For this reason, they were shown to be more stable and faster than line search methods, particularly for functions that are non-concave and/or exhibit regions close to flat (Radice, Marra & Wojtyś, 2016).

An alternative proposal to using a grid-search combined with GBIC is to estimate $\boldsymbol{\eta}$ automatically and in a data-driven fashion, a development that has not been so far considered in penalized factor analysis. To this end, we propose adapting to the current context the automatic multiple tuning (a.k.a smoothing) parameter selection of Marra and Radice (2020, see also references therein), which is based on an approximate AIC.

Assume that, near the solution, the trust-region method behaves like a classic unconstrained Newton-Raphson algorithm (Nocedal & Wright, 2006). Suppose also that $\boldsymbol{\theta}^{[\mathrm{t}+1]}$ is the "true" parameter value, and thus $\boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}+1]}) = \boldsymbol{0}$. By using a first-order Taylor expansion of $\boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}+1]})$ at $\boldsymbol{\theta}^{[\mathrm{t}]}$ it follows that

$$\boldsymbol{0} = \boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}+1]}) \approx \boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}]}) + \boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}^{[\mathrm{t}]})(\boldsymbol{\theta}^{[\mathrm{t}+1]} - \boldsymbol{\theta}^{[\mathrm{t}]}).$$

Solving for $\boldsymbol{\theta}^{[\mathrm{t}]}$ yields, after some manipulation (see Section D.1),

$$\boldsymbol{\theta}^{[t+1]} = \left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}^{[\mathrm{t}]})\right]^{-1}\sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]})}\boldsymbol{K}^{[\mathrm{t}]}, \tag{23}$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]}) = -\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}^{[\mathrm{t}]})$, $\boldsymbol{K}^{[\mathrm{t}]} = \boldsymbol{\mu}_{\boldsymbol{K}}^{[\mathrm{t}]} + \boldsymbol{\vartheta}^{[\mathrm{t}]}$ with $\boldsymbol{\mu}_{\boldsymbol{K}}^{[\mathrm{t}]} = \sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]})}\boldsymbol{\theta}^{[\mathrm{t}]}$ and $\boldsymbol{\vartheta}^{[\mathrm{t}]} = \sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]})}^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{[\mathrm{t}]})$. The square root of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]})$ and its inverse are obtained by eigenvalue decomposition. If they are not positive-definite, they are corrected as described in Section D.2. From standard likelihood theory, we have that $\boldsymbol{\vartheta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$ and $\boldsymbol{K} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{K}}, \boldsymbol{I}_m)$, where $\boldsymbol{\mu}_{\boldsymbol{K}} = \sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)}\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}_0$ the true parameter vector. Let $\hat{\boldsymbol{\mu}}_{\boldsymbol{K}}$ be the predicted value vector for $\boldsymbol{K}$ defined as

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{K}} = \sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}\hat{\boldsymbol{\theta}} = \sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}\left[\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right]^{-1}\sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}\boldsymbol{K} = \boldsymbol{A}_{\hat{\boldsymbol{\eta}}}^{\mathcal{T}}\boldsymbol{K},$$

where $\boldsymbol{A}_{\hat{\boldsymbol{\eta}}}^{\mathcal{T}} = \sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}\left[\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right]^{-1}\sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}$ is the influence (or hat) matrix of the fitting problem and depends on the tuning parameters. The quantity $\hat{\boldsymbol{\theta}} = \left[\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right]^{-1}\sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}\boldsymbol{K}$ denotes the PMLE. Ideally, the estimation of the tuning parameters should suppress the model complexity unsupported by the data. This can be achieved by minimizing the expected mean squared error of $\hat{\boldsymbol{\mu}}_{\boldsymbol{K}}$ from its expectation $\boldsymbol{\mu}_{\boldsymbol{K}}$ (Section D.3):

$$\mathbb{E}\left[\frac{1}{N}||\boldsymbol{\mu}_{\boldsymbol{K}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}||_2^2\right] = \frac{1}{N}\mathbb{E}\left[||\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}\boldsymbol{K}||_2^2\right] + \frac{2}{N}\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}) - 1. \tag{24}$$

The quantity $\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}) = \mathrm{tr}\left\{\left[\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right]^{-1}\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})\right\}$ can be interpreted as the edf of the penalized model, and is equivalent to the expression of the bias term of the GBIC. The right-hand side of (24) depends on the tuning parameters through $\boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}$, whereas $\boldsymbol{K}$ is linked to the unpenalized part of the model. The tuning parameters are estimated by minimizing an estimate of (24):

$$\mathcal{V}(\boldsymbol{\eta}) = \frac{1}{N}||\widehat{\boldsymbol{\mu}_{\boldsymbol{K}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}}||_2^2 = \frac{1}{N}||\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}\boldsymbol{K}||_2^2 + \frac{2}{N}\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}) - 1. \tag{25}$$

This is equivalent to the Un-Biased Risk Estimator (UBRE; Wood, 2017, Ch. 6) and an approximate AIC (Section D.4), which means that $\boldsymbol{\eta}$ is estimated by minimizing what is effectively the AIC with number of parameters given by $\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}})$. In practice, given $\boldsymbol{\theta}^{[\mathrm{t}+1]}$, the estimation problem is expressed as

$$\boldsymbol{\eta}^{[\mathrm{t}+1]} = \arg\min_{\boldsymbol{\eta}} \mathcal{V}^{[\mathrm{t}+1]}(\boldsymbol{\eta}) = \arg\min_{\boldsymbol{\eta}}\left\{\frac{1}{N}||\boldsymbol{K}^{[\mathrm{t}+1]} - \boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}^{[\mathrm{t}+1]}}\boldsymbol{K}^{[\mathrm{t}+1]}||_2^2 + \frac{2}{N}\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}^{[\mathrm{t}+1]}}) - 1\right\},$$

and solved by adapting the approach by Wood (2004) to the current context. This approach is based on Newton's method and can evaluate in a stable and efficient way the components in $\mathcal{V}(\boldsymbol{\eta})$ and their derivatives with respect to $\log(\boldsymbol{\eta})$ (since the tuning parameters can only take positive values). The two steps, one for the estimation of $\boldsymbol{\theta}$ and the other for $\boldsymbol{\eta}$, are iterated until the algorithm satisfies the stopping criterion $\dfrac{|\ell(\boldsymbol{\theta}^{[t+1]}) - \ell(\boldsymbol{\theta}^{[t]})|}{0.1 + |\ell(\boldsymbol{\theta}^{[t+1]})|} < 10^{-7}$.

Sometimes the final model could be overly dense and sparser solutions may be desired. One way to achieve this systematically is to increase the amount that each model edf counts, in the UBRE score, by a factor $\gamma \geq 1$, called "influence factor" (Wood, 2017). The slightly modified tuning criterion then is

$$\mathcal{V}(\boldsymbol{\eta}) = \frac{1}{N}||\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}\boldsymbol{K}||_2^2 + \frac{2}{N}\gamma \operatorname{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}) - 1. \tag{26}$$

For smoothing spline regression models, Kim and Gu (2004) found that $\gamma = 1.4$ can correct the tendency to over-fitting of prediction error criteria. However, this work deals with different models, and our focus is not only on fit but also on the recovery of sparse structures, thus higher values may be more appropriate.

The automatic procedure described above is general and easy to implement, but it may occasionally suffer at small sample sizes since it finds its justification asymptotically when the dependence of the Hessian on the tuning parameter(s) vanishes. As argued by Wood (2017), at small sample sizes, it would in principle be more reliable to select the tuning parameter(s) based on a non-approximate function, such as the GBIC and grid-searches, although implementing such an approach in the multiple-group context would introduce further complications and possibly new computational problems and instabilities. Notice also that the automatic procedure relies on the separability of the penalty matrix from the tuning parameter(s). This requirement is satisfied by the lasso and alasso (thus, $\mathcal{T} = \{L, A\}$), but not by the scad and mcp which are therefore confined to the grid-search approach. However, this is not problematic because in our simulation experiments and empirical application the alasso generally represented the most convenient choice of penalty based on a number of criteria.

At convergence, the covariance matrix of $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}} = \mathcal{J}_p(\hat{\boldsymbol{\theta}})^{-1}\mathcal{J}(\hat{\boldsymbol{\theta}})\mathcal{J}_p(\hat{\boldsymbol{\theta}})^{-1}$. However, instead of $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$, for practical purposes, it is more convenient to employ at convergence the alternative Bayesian result $\boldsymbol{V}_{\boldsymbol{\theta}} = \mathcal{J}_p(\hat{\boldsymbol{\theta}})^{-1}$ (Marra & Wood, 2012). The goodness of fit of the penalized model can then be evaluated through confidence intervals, which are available for each model parameter, obtained from the posterior distribution $\boldsymbol{\theta}|\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \boldsymbol{\eta} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \boldsymbol{V}_{\boldsymbol{\theta}})$ (Section D.5). Notice that the proposed approach can be regarded as a Bayesian method with the exponential prior $\exp\left\{-\frac{N}{2}\boldsymbol{\theta}^T\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right\}$ on the penalty function. The process of determining the optimal loading pattern can indeed be formulated as a Bayesian variable selection problem (Lu, Chow & Loken, 2016). For instance, Bayesian Structural Equation Modeling (BSEM; Muthén & Asparouhov, 2012)—in which the elements that would be fixed to zero in a confirmatory analysis (usually the cross-loadings) are replaced with approximate zeros based on informative, small-variance priors—is a particular case where the shrinkage is achieved through an informative ridge prior. With the proposed method, users can rely on the automatic procedure for recovering optimally sparse factor solutions without manually specifying the variance of the Bayesian prior employed in BSEM.

The presented modeling framework has been implemented in the freely available `R` package `penfa` and we refer the reader to Online Resource F for a brief description of the software and practical illustrations.

8. Simulation Studies

The performances of the proposed PMLE were evaluated and compared to the penalized methods by Jacobucci et al. (2016, `R` package `regsem`) and Huang (2018, `R` package `lslx`) in two extensive simulation studies, one for the normal linear factor model and the other for its multiple-group extension. Despite the presence of other penalized factor analysis techniques (Choi et al., 2010; Hirose & Yamamoto 2014b, 2014a; Trendafilov et al., 2017; Jin et al., 2018), our choice fell on `regsem` and `lslx` because they allow the specification of fixed, free and penalized parameters, as well the estimation of the structural model.

8.1. Simulation Study I

The first simulation evaluates the performances of the proposed technique in a single-group factor analysis model. We evaluate the impact of several conditions, including the sample size, the penalty function, the type of second-order derivative information used in the trust-region algorithm, the strategy for the choice of the tuning parameter, the magnitude of the influence factor and—for some of the penalties—the value of the additional tuning parameter. The simulation was partly inspired by the empirical application (Sect. 9), therefore the number of variables ($p = 9$) and of factors ($r = 3$) exactly match those of the real data analysis. The conditions that were varied are:

- Sample size: 300, 500, and 1000 observations. These values are in line with those investigated in similar simulation studies (Huang et al., 2017; Jacobucci et al., 2016; Jin et al., 2018; Hirose & Yamamoto, 2014b) and include two moderate sample sizes (which are commonly found in psychometric applications) and a large one (to mimic asymptotic behavior). Note that 300 is close to the number of observations in the empirical example;
- Penalty function: lasso, alasso, scad, and mcp were examined in their ability to shrink to zero small loadings without possibly affecting the remaining ones;
- Information matrix: either the Hessian or the Fisher information matrix was used in the optimization process (see Sect. 7);
- Shrinkage parameter selection: this was achieved either by a grid-search or through the automatic procedure. The grid-search was conducted over 200 distinct values of η and for all four penalty types, with the optimal model being the one with the lowest GBIC. The elements of the grid were adapted based on the specific combination of penalty type and sample size. The automatic procedure was used with lasso and alasso;
- Influence factor: informed by the values that performed well in the application, we investigated different values for the influence factor, namely, $\gamma = \{1, 1.4, 2, 2.5, 3, 3.5, 4, 4.5\}$;
- Additional tuning parameter: we tested different values of the additional tuning parameter of the alasso, scad and mcp. For the alasso $a = \{1, 2\}$, for the scad $a = \{2.5, 3, 3.7, 4.5\}$ (with 3.7 being the conventional level employed in the literature and suggested by Fan & Li, 2001), and for the mcp $a = \{2.5, 3, 3.5\}$.

The population parameters complied to the following structure:

$$\boldsymbol{\Lambda}^T = \begin{bmatrix} 0.85 & 0.75 & 0.65 & \underline{\mathit{0}} & 0 & 0 & \underline{\mathit{0}} & 0 & 0.30 \\ \underline{\mathit{0}} & 0 & 0.30 & 0.85 & 0.75 & 0.65 & \underline{\mathit{0}} & 0 & 0 \\ \underline{\mathit{0}} & 0 & 0 & \underline{\mathit{0}} & 0 & 0.30 & 0.85 & 0.75 & 0.65 \end{bmatrix}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} \underline{\mathit{1}} & 0.3 & 0.3 \\ & \underline{\mathit{1}} & 0.3 \\ & & \underline{\mathit{1}} \end{bmatrix}$$

with $\boldsymbol{\Psi} = \boldsymbol{I}_p - \boldsymbol{\Lambda\Phi\Lambda}^T$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Elements in italic and underlined were fixed for scale setting and identification purposes. The specific values of the factor loadings were inspired by the numerical example in Huang et al. (2017). As it is common in many factor analysis applications, a subset of the observed variables does not load only on one factor but also presents a cross-loading.

All of the factor loadings were penalized for assessing the effectiveness of the proposed method in recovering the underlying factor structure and not erroneously shrinking the small cross-loadings to zero. Based on results from previous studies (see for instance Choi et al., 2010 for the alasso, and Hirose & Yamamoto, 2014b and Huang et al., 2017 for the mcp), the alasso and the non-convex penalties are expected to outperform the lasso, which is known to be biased due to its tendency to overly shrink nonzero parameters. Concerning the influence factor, higher values favor sparsity at the expense of an increase in bias, whereas lower values favor goodness of fit.

Data were simulated in `R` (R Core Team, 2018) according to the population parameters. The resulting data matrix was then analyzed in `penfa`, `regsem` (Jacobucci et al., 2019) and `lslx` (Huang & Hu, 2019) by estimating a factor model with the correct number of factors, the specified fixed elements, and all of the free loadings were penalized. Common factors were estimated to be correlated and with fixed unit variance. Whenever present, sign reversal of the factors was accounted for to ensure that the sign of the primary loadings corresponded to that of the corresponding population parameters. Based on the availability of the respective software implementations,[2] lasso, alasso, scad and mcp were tried for `regsem`, and lasso and mcp for `lslx`. For each scenario, we generated $L = 1000$ replications for which the unpenalized factor model produced admissible[3] solutions.

We evaluated the performance of the methods according to the criteria illustrated in Huang et al. (2017), which are briefly mentioned here. The overall estimation quality was assessed using the estimated mean squared error (MSE):

$$\widehat{\text{MSE}}(\hat{\boldsymbol{\theta}}) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\boldsymbol{\theta}}^{(l)} - \boldsymbol{\theta}_0)^T(\hat{\boldsymbol{\theta}}^{(l)} - \boldsymbol{\theta}_0), \tag{27}$$

where $\hat{\boldsymbol{\theta}}^{(l)} = (\hat{\theta}_1^{(l)}, \ldots, \hat{\theta}_m^{(l)})^T$ denotes the vector of estimated parameters in replicate l, $\boldsymbol{\theta}_0$ the true parameter vector, and L the number of replications. The degree of bias of each estimator was evaluated by the estimated squared bias (SB):

$$\widehat{\text{SB}}(\hat{\boldsymbol{\theta}}) = (\bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0)^T(\bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0), \tag{28}$$

where $\bar{\hat{\boldsymbol{\theta}}} = \frac{1}{L}\sum_{l=1}^{L}\hat{\boldsymbol{\theta}}^{(l)}$ represents the empirical mean of $\hat{\boldsymbol{\theta}}$. Let $\mathcal{F} = \{q \mid \theta_{0q} \neq 0\,\&\,\hat{\theta}_q \text{ penalized}\}$ indicate the set of indices associated to the true nonzero parameters that have been penalized (i.e., the penalized nonzero factor loadings) and $|\mathcal{F}|$ the cardinality of $\mathcal{F}$, which in the simulation is equal to 12. The chance of correctly identifying the true nonzero parameters was evaluated via the estimated true positive rate (TPR):

[2] For `regsem`, we used the default optimizer `Rsolnp` for lasso and alasso, and coordinate descent for scad and mcp. The additional tuning parameters of the penalties were kept to the values specified in the package, that is, $a = 1$ for alasso, and $a = 3.7$ for scad and mcp. For `lslx`, as per software implementations, the shape parameter of the mcp was internally selected over a varying three-dimensional grid of values adapted for each replicate. Because of the specificities of each package implementation, a customized and sensible grid of η values was considered for each technique.

[3] A solution is considered admissible if it does not present Heywood cases (negative unique variances), the covariance matrices of the unique factors and common factors are positive-definite, the factor loading matrix is of full column rank and does not contain any null rows (Jöreskog & Sörbom, 1996).

$$\widehat{\mathrm{TPR}}(\hat{\boldsymbol{\theta}}) = \frac{1}{L}\sum_{l=1}^{L} \frac{\sum_{q\in\mathcal{F}} \mathbb{1}\left(\hat{\theta}_q^{(l)} \neq 0\right)}{|\mathcal{F}|}. \tag{29}$$

Denote as $\mathcal{F}^c = \{q \mid \theta_{0q} = 0 \,\&\, \hat{\theta}_q \text{ penalized}\}$ the set collecting the indices of the true zero parameters that have been penalized (i.e., the penalized zero factor loadings), with $|\mathcal{F}^c|$ equal to 9. The estimated false positive rate (FPR) examined the degree to which the true zero parameters were incorrectly identified as nonzero:

$$\widehat{\mathrm{FPR}}(\hat{\boldsymbol{\theta}}) = \frac{1}{L}\sum_{l=1}^{L} \frac{\sum_{q\in\mathcal{F}^c} \mathbb{1}\left(\hat{\theta}_q^{(l)} \neq 0\right)}{|\mathcal{F}^c|}. \tag{30}$$

Lastly, selection consistency was assessed via the proportion of times the true model—for which all the true zero and nonzero factor loadings were correctly identified as equal to zero and different from zero, respectively—was chosen over the replicates (proportion choosing the true model; PCTM):

$$\widehat{\mathrm{PCTM}}(\hat{\boldsymbol{\theta}}) = \frac{1}{L}\sum_{l=1}^{L} \frac{\sum_{q\in\mathcal{F}} \mathbb{1}\left(\hat{\theta}_q^{(l)} \neq 0\right) + \sum_{q\in\mathcal{F}^c} \mathbb{1}\left(\hat{\theta}_q^{(l)} = 0\right)}{|\mathcal{F}| + |\mathcal{F}^c|}, \tag{31}$$

where $|\mathcal{F}| + |\mathcal{F}^c| = q^\star$. For the computation of PCTM and FPR,[4] the parameter estimates were rounded to one decimal digit[5] for all models. For the sake of clarity, we report a selection of the most relevant results for the configurations of `penfa-alasso` ($a = 2, \gamma = 4.5$), `penfa-scad` ($a = 3$) and `penfa-mcp` ($a = 3$) that produced the best models in terms of the aforementioned performance criteria. Due to its generally higher numerical stability in comparison to the Hessian, only `penfa` models estimated with the Fisher information matrix are presented. The results for `penfa-lasso` are described in Online Resource A. In the same spirit, the results of `regsem` and `lslx` are presented for their best performing models (i.e., with the mcp for both of them). All other results can be requested from the corresponding author.

Overall, the low values for MSE, the bias, and FPR which are very close to zero, together with high PCTM and excellent TPR show that the examined penalized techniques possess very good empirical performances and outperform the unpenalized (MLE) model (Table 1). The MSE of all penalized methods are very similar to each other and improve as the sample size increased. The results with the lower bias were associated with the use of non-convex penalties, although the bias of `penfa-alasso` very quickly converged to zero when the sample size increased, and hence the impact of the penalty decreased. The true positive rates were always equal to 1.0, which showed that the inspected methods never suppressed the nonzero penalized parameters (i.e., the primary loadings and the cross-loadings). In terms of both false positive rates and selection consistency, `penfa-alasso` with automatic tuning parameter selection presented by far the best performances for all the sample sizes. The coverage probabilities for the parameters of all fitted models (Section A.1) were generally close to their true nominal level for all penalty functions.

[4] No rounding was necessary for TPR because, based on the simulation design, the estimates were never mistakenly estimated close to zero.

[5] This choice was made on practical grounds, deeming estimates in absolute value below 0.05 to be "practically" and "substantively" equal to zero. The use of tighter rounding thresholds may worsen FPR and PCTM. For `penfa`, numerical experiments showed that this was the case when using the grid-search, whereas the models with the automatic procedure exhibited fairly comparable FPR and PCTM even after two or three decimal digits roundings.

TABLE 1.
Performance measures of the examined models in simulation study I by varying the sample size N.

	Unpenalized (MLE)	`penfa`				`lslx`	`regsem`
		ALASSO		SCAD	MCP	MCP	MCP
		Grid	Auto	Grid	Grid	Grid	Grid
MSE							
$N = 300$	0.108	0.073	0.075	0.074	0.074	0.075	0.071
	(0.04–0.41)	(0.02–0.19)	(0.02–0.08)	(0.02–0.17)	(0.02–0.17)	(0.02–0.20)	(0.02–0.33)
$N = 500$	0.064	0.041	0.041	0.042	0.042	0.042	0.041
	(0.02–0.23)	(0.01–0.12)	(0.01–0.12)	(0.01–0.12)	(0.01–0.12)	(0.01–0.13)	(0.01–0.12)
$N = 1000$	0.031	0.020	0.020	0.020	0.020	0.020	0.020
	(0.01–0.08)	(0.01–0.06)	(0.01–0.05)	(0.01–0.06)	(0.01–0.06)	(0.01–0.06)	(0.01–0.06)
SB							
$N = 300$	0.001	0.003	0.004	0.002	0.002	0.003	0.000
$N = 500$	0.000	0.001	0.001	0.001	0.001	0.001	0.000
$N = 1000$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TPR							
$N = 300$	1	1	1	1	1	1	1
						(0.92–1)	(0.92–1)
$N = 500$	1	1	1	1	1	1	1
$N = 1000$	1	1	1	1	1	1	1
FPR							
$N = 300$	0.390	0.022	0.008	0.016	0.019	0.036	0.018
	(0.00–0.89)	(0.00–0.33)	(0.00–0.22)	(0.00–0.22)	(0.00–0.33)	(0.00–0.44)	(0.00–0.22)
$N = 500$	0.264	0.012	0.004	0.007	0.008	0.016	0.012
	(0.00–0.89)	(0.00–0.22)	(0.000–0.11)	(0.00–0.22)	(0.00–0.22)	(0.00–0.33)	(0.00–0.22)
$N = 1000$	0.113	0.003	0.001	0.002	0.002	0.004	0.009
	(0.00–0.56)	(0.00–0.11)	(0.00–0.11)	(0.00–0.11)	(0.00–0.11)	(0.00–0.22)	(0.00–0.22)
PCTM							
$N = 300$	0.009	0.820	0.932	0.871	0.843	0.743	0.848
$N = 500$	0.073	0.898	0.962	0.936	0.925	0.877	0.897
$N = 1000$	0.356	0.974	0.991	0.982	0.979	0.966	0.923

MSE mean-squared error, *SB* squared bias, *TPR* true positive rate, *FPR* false positive rate, *PCTM* proportion choosing the true model. In brackets, the ranges of MSE, TPR, and FPR across replicates.

The mean squared error and bias of penfa-alasso with automatic tuning parameter selection were similar to those obtained with the same penalty and grid-search, but the false positives and PCTM were markedly lower and higher, respectively. This is due to the way the optimal penalized model is picked. With the automatic procedure, the optimal model is the one whose tuning parameter minimizes the criterion in (26); with the grid-search, the optimal model minimizes the GBIC in (19). However refined, a grid-search cannot compete with an approach that looks for the optimal tuning parameter on the positive real line. In addition, the presence of a sparsity-inducing quantity (the influence factor) in the optimization criterion helped the model obtain a nicer tradeoff between goodness of fit and model complexity. With reference to the exponent a in the expression of the alasso, as this quantity increased the weights became more influential, and we observed a general improvement in all the performance measures. The best results were obtained for $a = 2$.

By comparing the quality measures of the three methods for the same penalty function (i.e., the mcp), we notice that penfa outperformed lslx and was generally close to regsem for MSE and SB and superior for FPR and PCTM. This might be due to several aspects, e.g., the optimization algorithm, the internal software package implementations, the formulation of the degrees of freedom, and possibly the approximation of the penalty.

The examined performance criteria explored different conflicting objectives. Ideally, one desires a model with low bias and little complexity (i.e., a sparse solution), but the two measures cannot be minimized simultaneously. This can be seen by looking at the performances of the penfa-alasso model for extreme values of the influence factor (i.e., $\gamma = 4.5$ in Table 1 and $\gamma = 1$ in Table A.2 in Section A.2). The higher value of γ produced sparser solutions (i.e., smaller FPR and larger PCTM), at the cost of a larger bias. As the sample size increased, the discrepancies in the performances of the models with different values of γ diminished though.

The models fitted through the automatic tuning parameter procedure exhibited markedly shorter computational times[6] than grid-search methods. Specifically, the average median elapsed times were 17 sec for penfa-alasso with grid (1-dim. grid for η; $a = 2$) and 0.3 sec with automatic procedure ($a = 2$; $\gamma = 4.5$), 21.1 sec for penfa-scad (1-dim. grid for η; $a = 3$), 20.7 sec for penfa-mcp (1-dim. grid for η; $a = 3$), 6.6 sec for lslx-mcp (2-dim. grid for η and a) and 42.2 sec for regsem-mcp (1-dim grid for η, $a = 3.7$ as per default software implementations). The penfa models with the automatic procedure exhibited the lowest computational times, which is also merit of the stability of the trust-region optimizer, whose parameter updates only involve the points within a proper trust-region. The computational times of lslx are lower than those of the other grid-search techniques because the underlying optimizer is implemented in C++, which significantly boosts the computations with respect to base R routines.

8.2. *Simulation Study II*

The second simulation evaluates the ability of the proposed technique in identifying the pattern of partial invariance in a multiple-group factor model as a function of the sample size, the size of the generated difference in the group-specific loadings and intercepts, the magnitude of the influence factor and the value of the additional tuning parameter. Since the current implementation of regsem does not allow for multiple-group analyses, our method is only compared with lslx.

We consider a population multiple-group factor model with $p = 12$ variables, $r = 3$ factors and $G = 2$ groups. We explore a range of conditions, under which the factor loading matrices and intercepts are either invariant or non-invariant, with the level of non-invariance becoming progressively larger. Based on the findings from Simulation study I, we employ the alasso penalty for inducing sparsity and invariant loadings and intercepts, that is, $\mathcal{S}^{A}_{\boldsymbol{\eta}}(\tilde{\boldsymbol{\theta}}) =$

[6] All computations were carried out on a machine with Intel(R) Core(TM) i7-5600U 2.60GHz (quad-core) processor and 16GB of RAM.

TABLE 2.
The factor loading matrices and intercepts of the two groups under each difference scenario of simulation study II.

	Group 1			Group 2								
	All conditions			Small			Medium			Large		
	$\boldsymbol{\Lambda}_1$		$\boldsymbol{\tau}_1$	$\boldsymbol{\Lambda}_2$		$\boldsymbol{\tau}_2$	$\boldsymbol{\Lambda}_2$		$\boldsymbol{\tau}_2$	$\boldsymbol{\Lambda}_2$		$\boldsymbol{\tau}_2$
x_1	*0.85*	*0*	*0*	*0.85*	*0*	*0*	*0.85*	*0*	*0*	*0.85*	*0*	*0*
x_2	0.85	0	0	0.85	0	0	0.85	0	0	0.85	0	0
x_3	0.85	0	0	0.85	0	0	0.85	0	0	0.85	0	0
x_4	0.75	0	0	0.75	0	0	0.75	0	0	0.75	0	0
x_5	0.75	0	0	0.75	0	0	0.75	0	0	0.75	0	0
x_6	0.75	0	0	0.65	0	− 0.1	0.55	0	− 0.2	0.45	0	− 0.3
x_7	*0*	*0.85*	*0*	*0*	*0.85*	*0*	*0*	*0.85*	*0*	*0*	*0.85*	*0*
x_8	0	0.85	0	0	0.85	0	0	0.85	0	0	0.85	0
x_9	0	0.85	0	0	0.85	0	0	0.85	0	0	0.85	0
x_{10}	0	0.75	0	0	0.75	0	0	0.75	0	0	0.75	0
x_{11}	0	0.75	0	0	0.75	0	0	0.75	0	0	0.75	0
x_{12}	0	0.75	0	0	0.65	− 0.1	0	0.55	− 0.2	0	0.45	− 0.3

Elements fixed for origin and scale setting and identification purposes are italic and underlined. Under the null condition, the parameters of Group 2 coincide with those of Group 1.

$\mathcal{D}^{A}_{\eta_1}(\tilde{\boldsymbol{\theta}})+\mathcal{D}^{A}_{\eta_2}(\tilde{\boldsymbol{\theta}})+\mathcal{D}^{A}_{\eta_3}(\tilde{\boldsymbol{\theta}})$. The three tuning parameters $(\eta_1, \eta_2, \eta_3)^T$ in $\boldsymbol{\eta}$ are estimated alongside the model parameters through the automatic multiple tuning parameter procedure. For `lslx` we used the mcp penalty, which had better performances than the lasso. The optimization technique currently employed in `lslx` makes use of a single penalty for both shrinking the parameters and their differences across groups. Therefore, there is only one shrinkage parameter η, whose optimal value is determined through a grid-search. For `lslx-mcp`, we carried out a grid-search over 200 values of the shrinkage parameter η and 4 of the shape parameter a. The conditions that were varied are:

- Sample size: 300, 500, and 1000 observations evenly split between the two groups, with 300 being close to the number of observations in the empirical example;
- Difference size: either null, small, medium or large group differences in the primary loadings and the intercepts of two variables were created (details are given below). This condition was partly inspired by the simulation conducted by Huang (2018);
- Influence factor: informed by the values that performed well in Simulation study I, we investigated three values of the influence factor, namely, $\gamma = \{3.5, 4, 4.5\}$;
- Additional tuning parameter: two values were tested for the exponent in the expression of the alasso, namely $a = \{1, 2\}$.

The factor loading matrix and the vector of intercepts of Group 1 are reported on the left-hand side of Table 2, and are the same under every difference scenario. Elements in italic and underlined are fixed for metric setting and identification purposes. The factor loadings and intercepts of Group 2 are presented by difference scenario on the right-hand side of Table 2. In case of a null difference, the two groups share the same parameter matrices. Under the small, medium and large scenarios, the primary loadings and the intercepts of two variables (i.e., x_6 and x_{12}) in Group 1 differ from the corresponding parameters in Group 2 by a size of 0.1, 0.2, and 0.3, respectively. Under all conditions, the structural parameters are assumed to be invariant across groups, that is, $\text{vech}(\boldsymbol{\Phi}_1) = \text{vech}(\boldsymbol{\Phi}_2) = \text{vech}(\boldsymbol{\Phi}) = (1, 0.3, 1)^T$ and $\boldsymbol{\kappa}_1 = \boldsymbol{\kappa}_2 = (0, 0)^T$, whereas $\boldsymbol{\Psi}_g = \boldsymbol{I}_p - \boldsymbol{\Lambda}_g \boldsymbol{\Phi} \boldsymbol{\Lambda}_g^T$, for $g = 1, 2$. The factor loadings and the intercepts are penalized in the way described in Sect. 5 (i.e., shrinkage of the loadings and of the pairwise group differences of loadings and intercepts), whereas the remaining model parameters are estimated without penalization. For each scenario, we generated $L = 1000$ replications for which the unpenalized multiple-group model produced admissible solutions, and analyzed them as in simulation study I.

The performances of the penalized models are evaluated through the criteria (27)-(31) used in simulation study I. For the sake of conciseness, we report the results for the `penfa-alasso` model ($a = 2, \gamma = 4.5$) that produced the best solution in terms of these performance criteria. All other results can be requested from the corresponding author. Overall, the low values of MSE, SB, FPR, high PCTM and excellent TPR show that the penalized techniques possess very good empirical performances, with all measures improving as the sample size increased (Table 3). Higher difference sizes were associated with higher MSE and squared bias, with the lower values generally occurring for `penfa-alasso`. We separately computed these measures for each parameter matrix (that is, $\boldsymbol{\Lambda}_g, \boldsymbol{\tau}_g, \boldsymbol{\Psi}_g, \boldsymbol{\Phi}_g, \boldsymbol{\kappa}_g$, for $g = 1, 2$) produced by `penfa-alasso`. The largest MSE were observed for the factor variances and covariances, followed by the factor loadings. The bias tended to increase for the penalized parameters (factor loadings and intercepts) across the difference conditions, while remaining almost unaltered for the unique variances and the structural parameters. The squared bias quickly converged towards zero in all difference scenarios as the sample size increased. The TPR were always equal to 1.0, which showed that the examined methods never suppressed the nonzero penalized parameters.

Whereas under the null and small scenarios the two methods produced similar measures, `penfa-alasso` markedly outperformed `lslx-mcp` under the medium and large conditions, especially in terms of selection consistency at the smallest sample size. On top of that, whereas these performance measures for `lslx` noticeably degraded as the difference size increased, they remained fairly stable for `penfa-alasso`; even with the smallest sample size, `penfa-alasso` identified the true heterogeneity pattern more than 90% of the times. Thanks to the use of the automatic multiple tuning parameter procedure, the average median computational time to fit a `penfa-alasso` model with 3 tuning parameters (3.2 seconds) was much lower than the one necessary to fit an `lslx-mcp` model with a single shrinkage parameter η and the associated shape parameter a selected through a grid-search (45 seconds).

9. Empirical Application

The Holzinger & Swineford data set (Holzinger & Swineford, 1939; Kelley, 2019) is a classical psychometric application containing the responses of $N = 301$ students on some psychological tests. This data set (or subsets of it) has been often used to demonstrate CFA (Jöreskog, 1979), EFA (Browne, 2001; Jöreskog & Sörbom, 1993) and various penalized factor analysis techniques (Trendafilov et al., 2017; Jacobucci et al., 2016; Huang et al., 2017; Jin et al., 2018). For space constraints, the description of the data set is reported in Online Resource E.

9.1. Normal Linear Factor Model

Following Jacobucci et al. (2016) and Huang et al. (2017), to illustrate the proposed method in the normal linear factor model, we use a subset of nine mental tests (VISUAL, CUBES, FLAGS, PARAGRAP, SENTENCE, WORDM, ADDITION, COUNTING, and STRAIGHT) underlying three latent factors. The data set was column-wise centered since the model in equation (1) assumes that the observed variables have zero means and scaled as described in Yuan and Bentler (2006). The inspection at the covariance matrix of the observed variables revealed the presence of relationships between tests designed to measure distinct mental abilities. The CFA model assuming a simple structure showed a poor fit to the data (p-value of the chi-square goodness of fit test < 0.001), which suggested the multi-dimensionality of some of the tests. In these

TABLE 3.
Performance measures of `penfa-alasso` ($a = 2$, $\gamma = 4.5$) and `lslx-mcp` models by sample size and difference scenario.

Difference scenario	Null		Small		Medium		Large	
	`penfa`	`lslx`	`penfa`	`lslx`	`penfa`	`lslx`	`penfa`	`lslx`
MSE								
$N = 300$	0.275	0.279	0.303	0.307	0.356	0.372	0.385	0.416
	(0.11–0.65)	(0.10–0.75)	(0.12–0.76)	(0.11–0.79)	(0.15–0.83)	(0.19–0.86)	(0.15–0.90)	(0.14–1.04)
$N = 500$	0.165	0.164	0.189	0.189	0.220	0.239	0.221	0.235
	(0.07–0.50)	(0.06–0.47)	(0.08–0.54)	(0.08–0.51)	(0.11–0.60)	(0.12–0.58)	(0.09–0.60)	(0.08–0.52)
$N = 1000$	0.083	0.082	0.102	0.104	0.105	0.115	0.103	0.101
	(0.04–0.20)	(0.04–0.20)	(0.05–0.22)	(0.05–0.22)	(0.04–0.25)	(0.04–0.26)	(0.04–0.26)	(0.04–0.23)
SB								
$N = 300$	0.003	0.002	0.020	0.021	0.046	0.062	0.043	0.050
$N = 500$	0.001	0.001	0.017	0.020	0.026	0.042	0.018	0.012
$N = 1000$	0.000	0.000	0.012	0.018	0.007	0.007	0.005	0.001
TPR								
$N = 300$	1	1	1	1	1	1	1	1
$N = 500$	1	1	1	1	1	1	1	1
$N = 1000$	1	1	1	1	1	1	1	1
FPR								
$N = 300$	0.006	0.010	0.005	0.012	0.005	0.019	0.004	0.035
	(0.00–0.40)	(0.00–0.30)	(0.00–0.20)	(0.00–0.30)	(0.00–0.20)	(0.000–0.30)	(0.00–0.20)	(0.00–0.40)
$N = 500$	0.004	0.004	0.005	0.005	0.004	0.014	0.003	0.020
	(0.00–0.20)	(0.00–0.20)	(0.00–0.30)	(0.00–0.20)	(0.00–0.20)	(0.000–0.25)	(0.00–0.20)	(0.00–0.25)
$N = 1000$	0.002	0.001	0.002	0.002	0.002	0.005	0.001	0.003
	(0.00–0.15)	(0.00–0.10)	(0.00–0.20)	(0.00–0.15)	(0.00–0.20)	(0.000–0.15)	(0.00–0.20)	(0.00–0.15)
PCTM								
$N = 300$	0.935	0.890	0.945	0.880	0.933	0.820	0.948	0.677
$N = 500$	0.951	0.956	0.948	0.949	0.947	0.854	0.967	0.781
$N = 1000$	0.980	0.991	0.969	0.977	0.976	0.930	0.984	0.958

MSE mean-squared error, *SB* squared bias, *TPR* true positive rate, *FPR* false positive rate, *PCTM* proportion choosing the true model. In brackets, the ranges of MSE and FPR across replicates; TPR were always equal to one.

TABLE 4.
BIC of the best configurations of the fitted models.

Method	Penalty	BIC
penfa	ALASSO	7558.03
penfa	MCP	7561.57
penfa	SCAD	7561.68
penfa	LASSO	7562.94
CFA		7595.34
Unpenalized		7601.42

For `penfa-alasso` (automatic procedure) $a = 1$ and $\gamma = 4.5$, for `penfa-scad` $a = 4.5$, for `penfa-mcp` $a = 1.5$, and for `penfa-lasso` (automatic procedure) $\gamma = 4.5$. For all models the Fisher information was used.

circumstances where it may be difficult to specify the correct sparsity pattern of the loading matrix in advance, it is beneficial to resort to penalized techniques to explore and unveil the underlying loading pattern. We hence penalize all of the factor loadings and freely estimate the remaining model parameters. Factor variances are fixed to one for scale setting and some elements of the loading matrix to zero for identification purposes. Even if the proposed method does allow us to obtain sparsity, we should acknowledge that its achievement also depends on the features of the statistical model under investigation and the amount of information carried by the data. Concerning the former, as pointed out by Trendafilov et al. (2017), inducing sparsity in a factor model, and even more so one with correlated factors, is more complicated than for other types of models (e.g. principal component analysis) due to the presence of other parameters (unique variances and factor variances and covariances) affecting the overall model fit. As a result, if too large a value for the tuning parameter is chosen, an excessive number of loadings is shrunken, and the remaining parameters are forced to explode to compensate for this lack of fit. This issue can be avoided if the appropriate amount of sparsity is introduced into the model, which in turn is only possible if the tuning parameter governing the amount of sparsity is selected according to a valid procedure, such as the one introduced in the paper.

We fitted a large number of models involving all four penalties. For grid-search, 200 models corresponding to varying levels of the tuning parameter were fitted. We also tried a sequence of values for the additional tuning parameter of the alasso ($a = \{1, 1.5, 2\}$), scad ($a = \{2.5, 3.7, 4.5\}$), mcp ($a = \{1.5, 2, 2.5, 3, 3.5\}$), and for the influence factor ($\gamma = \{1, 1.4, 2, 2.5, 3, 3.5, 4, 4.5\}$). The GBIC[7] values were calculated for each of the fitted `penfa` models and are ranked in Table 4 for the best model configurations. In particular, the alasso (automatic procedure, $a = 1$, $\gamma = 4.5$) presented the lowest BIC, closely followed by the mcp ($a = 1.5$) and scad ($a = 4.5$). Interestingly, the BIC of `penfa-lasso` with grid-search (7567.62) markedly decreased when the model was fitted through the automatic procedure with an influence factor of 4.5 (7562.94). Notice that both the CFA and the unpenalized solution (corresponding to the factor analysis model in equation (1) with the minimum identification restrictions) resulted in worse fits than the ones of the penalized models, probably because of the strict assumption of no cross-loadings of the former, and the unnecessary complexity of the latter. This indicates that the analysis benefited from the introduction of sparsity.

Table 5 (left-hand side) reports the parameter estimates of the unpenalized model and the best `penfa-alasso` model. A blank cell in the factor loading matrix indicates that the corresponding

[7] We used the BIC as a criterion for model comparisons due to its widespread use in sparse settings, but different evaluation measures can be employed depending on the research question.

estimate was zero after one decimal digit rounding.[8] The unpenalized model presented various cross-loadings, which resulted in a complex model. For penfa, only four secondary loadings ($\hat{\lambda}_{51}$, $\hat{\lambda}_{81}$, $\hat{\lambda}_{91}$, $\hat{\lambda}_{32}$) were identified as nonzero. If a sparser loading matrix is desired, users can increase the value of the exponent a of the alasso and/or the influence factor γ in the automatic procedure. For instance, a penfa-alasso model (BIC = 7565.39) with $a = 2$ and $\gamma = 5.5$ (Table 5, right-hand side) produced a sparser factor solution with a single cross-loading ($\hat{\lambda}_{91}$). The data analysis was also conducted for regsem and lslx using the available penalties (i.e., lasso, alasso, scad, and mcp for the former, and lasso and mcp for the latter) and is presented in Online Resource E. The factor structures of the penalized models looked similar, but the proposed method reported the lowest BIC values, showing the potential of the presented procedure. As argued by Huang et al. (2017), this example shows that complex models do not necessarily outperform simpler ones when model complexity is also taken into account in the model selection criterion.

9.2. Multiple-Group Factor Model

Besides considering the sample of the students as a whole, we divided it into two groups ($N_1 = 156$, $N_2 = 145$) based on the attended school, and then conducted a multiple-group analysis. One school (Pasteur) included students with parents who immigrated from Europe, whereas the other (Grant-White) was composed of students coming from middle-income American white families. Following Huang (2018), we considered the 19 mental tests and standardized the data to handle the scaling effect.

The traditional approach consists of the estimation of an unpenalized multiple-group CFA in which the tests are assumed to be pure measures, followed by factorial invariance testing procedures. The model assuming equal loadings across groups shows an adequate fit to the data (p-value of the chi-square goodness of fit test = 0.266), which, however, significantly worsens when the intercepts are also equated across groups (p-value of the likelihood ratio test comparing the model with invariant loadings and intercepts versus the one with only invariant loadings $<$ 0.001). Model modifications are typically conducted to determine and freely estimate the non-invariant elements.

Alternatively, the invariance pattern can be explored via penalized techniques employing penalties that combine sparsity and cross-group equivalence of loadings and intercepts. In light of its superior performance in the single-group analysis and simulation, we employed the alasso with the automatic multiple tuning parameter procedure, and tested various values of the influence factor ($\gamma = \{1, 2, 3, 3.5, 4, 4.5\}$) and the exponent ($a = \{1, 2\}$). The tests VISUAL, WORDM, COUNTING and NUMBERR are assumed to be the markers, and thus have fixed loadings and intercepts. The data analysis was also conducted in lslx with the mcp (see Table E.4), but not in regsem as its current implementation does not allow for multiple-group analyses. Note that lslx uses only one penalty for shrinking both the parameters and their differences, hence it has a single tuning parameter η.

The parameter estimates of penfa-alasso are reported in Table 6. The better fit of penfa-alasso (BIC = 14658) as compared to lslx-mcp (BIC = 14697.75) is also merit of the greater flexibility of the former, which employs three distinct penalties having their own tuning parameters, with respect to the latter, where a single tuning has to take care of the shrinkage of the parameters as well as their cross-group differences. penfa-alasso produces sparse loading matrices with many zero-entries, but the presence of a couple of nonzero cross-loadings demonstrates that the structure hypothesized by a multiple-group CFA is too restrictive. The factor loading matrices of penfa-alasso are also fully equivalent, in agreement to the results of invariance testing. Conversely, the intercepts are not fully invariant, which is again in line with

[8] This was done for a neater visual illustration; in the penfa package, no internal rounding is implemented, and the estimates are shown as returned by the trust-region optimizer.

TABLE 5.
Parameter estimates of the nine mental tests from the Holzinger & Swineford data set for the unpenalized model, and `penfa-alasso` with automatic procedure (on the left-hand side, $\hat{\eta} = 0.017$, $a = 1$ and $\gamma = 4.5$; on the right-hand side, $\hat{\eta} = 0.011$, $a = 2$ and $\gamma = 5.5$).

Measurement model	Unpenalized model				`penfa-alasso` ($a = 1, \gamma = 4.5$)				`penfa-alasso` ($a = 2, \gamma = 5.5$)			
	Spatial	Verbal	Speed	$\boldsymbol{\Psi}$	Spatial	Verbal	Speed	$\boldsymbol{\Psi}$	Spatial	Verbal	Speed	$\boldsymbol{\Psi}$
VISUAL	0.81	$\underline{0}$	$\underline{0}$	0.70	0.83	$\underline{0}$	$\underline{0}$	0.63	0.85	$\underline{0}$	$\underline{0}$	0.59
CUBES	0.65	−0.12	−0.16	1.03	0.49			1.11	0.46			1.13
FLAGS	0.91	−0.33		0.69	0.76	−0.16		0.75	0.66			0.82
PARAGRAP	$\underline{0}$	0.99	$\underline{0}$	0.38	$\underline{0}$	0.96	$\underline{0}$	0.38	$\underline{0}$	0.96	$\underline{0}$	0.37
SENTENCE	−0.13	1.19		0.40	−0.06	1.11		0.42		1.08		0.44
WORDM	0.07	0.87		0.37		0.89		0.36		0.89		0.36
ADDITION	$\underline{0}$	$\underline{0}$	0.77	0.59	$\underline{0}$	$\underline{0}$	0.70	0.67	$\underline{0}$	$\underline{0}$	0.62	0.76
COUNTING	0.30	−0.16	0.68	0.48	0.12		0.70	0.44			0.79	0.36
STRAIGHT	0.54	−0.14	0.43	0.55	0.41		0.42	0.56	0.36		0.39	0.58

TABLE 5.
continued

Structural model	Spatial	Verbal	Speed	Spatial	Verbal	Speed	Spatial	Verbal	Speed
Spatial	*1*	0.59	0.17	*1*	0.48	0.20	*1*	0.45	0.31
Verbal	–	*1*	0.22	–	*1*	0.16	–	*1*	0.19
Speed	–	–	*1*	–	–	*1*	–	–	*1*

Fixed parameters are italic and underlined. A blank cell indicates that the corresponding estimate is zero.

TABLE 6.
Parameter estimates of the 19 mental tests from the Holzinger & Swineford data set for `penfa-alasso` (automatic procedure, $\hat{\boldsymbol{\eta}} = (0.006, 16221.852, 0.013)^T$, $a = 1$, $\gamma = 4$)

Measurement model	`penfa - alasso`											
	PASTEUR SCHOOL						GRANT- WHITE SCHOOL					
	$\boldsymbol{\tau}_1$	Spatial	Verbal	Speed	Memory	$\boldsymbol{\Psi}_1$	$\boldsymbol{\tau}_2$	Spatial	Verbal	Speed	Memory	$\boldsymbol{\Psi}_2$
VISUAL	*0*	*1*	*0*	*0*	*0*	0.44	*0*	*1*	*0*	*0*	*0*	0.43
CUBES	0.01	0.58				0.89	0.01	0.58				0.68
PAPER	0	0.62				0.81	0	0.62				0.71
FLAGS	0.14⋆	0.86	− 0.09			0.61	− 0.16⋆	0.86	− 0.09			0.47
GENERAL	− 0.01		1.02		− 0.11	0.26	− 0.01		1.02		− 0.11	0.31
PARAGRAP	− 0.01		0.96			0.35	−0.01		0.96			0.31
SENTENCE	− 0.01	− 0.12	1.08			0.25	− 0.01	− 0.12	1.08			0.22
WORDC	− 0.08⋆		0.84			0.41	0.07⋆		0.84			0.45
WORDM	*0*	*0*	*1*	*0*	*0*	0.23	*0*	*0*	*1*	*0*	*0*	0.35
ADDITION	0.14⋆	− 0.40	0.14	0.99	0.15	0.52	− 0.18⋆	− 0.40	0.14	0.99	0.15	0.34
CODE	0		0.17	0.74	0.27	0.44	0		0.17	0.74	0.27	0.61
COUNTING	*0*	*0*	*0*	*1*	*0*	0.54	*0*	*0*	*0*	*1*	*0*	0.44
STRAIGHT	0	0.40		0.68		0.62	0	0.40		0.68		0.44
WORDR	*0*	*0*	*0*	*0*	*1*	0.58	*0*	*0*	*0*	*0*	*1*	0.56
NUMBERR	0		− 0.14		0.84	0.68	0		− 0.14		0.84	0.67
FIGURER	0.02	0.37			0.63	0.73	0.02	0.37			0.63	0.47
OBJECT	0.16⋆	− 0.23		0.32	0.87	0.63	− 0.19⋆	− 0.23		0.32	0.87	0.46
NUMBERF	0			0.25	0.65	0.78	0			0.25	0.65	0.65
FIGUREW	− 0.20⋆	0.06		0.09	0.53	0.85	0.24⋆	0.06		0.09	0.53	0.60
Spatial	− 0.02	0.59	0.28	0.16	0.17		0.02	0.60	0.36	0.29	0.24	
Verbal	− 0.26	–	0.66	0.19	0.10		0.29	–	0.62	0.23	0.26	
Speed	0.09	–	–	0.44	0.07		− 0.09	–	–	0.63	0.16	
Memory	− 0.05	–	–	–	0.52		0.05	–	–	–	0.42	

Fixed parameters are italic and underlined. A blank cell in the factor loading matrix indicates that the corresponding estimate is zero. Non-invariant parameters across groups are starred (⋆).

the findings from factorial invariance testing. This example clearly shows the benefits of using properly designed penalized techniques to explore the non-equivalence pattern of the parameter matrices in a multiple-group factor model.

10. Discussion

Penalized factor analysis is an efficient estimation technique that produces a factor loading matrix with many zero elements thanks to the introduction of sparsity-inducing penalty functions within the estimation process. In order to achieve sparse solutions and stable model selection procedures, the penalty functions must be non-differentiable. In this work, we adopted suitable local approximations of them. In this way, it was possible to employ in the optimization process a trust-region algorithm, which required analytical information on the score vector and the Hessian matrix (or a good approximation thereof). The use of differentiable penalties allowed us to recast the problem in a theoretically founded framework, where a precise definition of effective degrees of freedom was obtained, based on the bias term of the Generalized Information Criterion, or equivalently, the influence matrix of the model. This represents a novelty, as the existing proposals compute the degrees of freedom of a penalized factor model as the number of nonzero parameters. As an alternative to the usually time-consuming grid-searches, we also illustrated an efficient automatic technique for the estimation of the tuning parameter alongside the parameters of the factor model. The asymptotic properties of the penalized estimator can be established along the lines of Filippou et al. (2017) and Fan and Li (2001).

The simulations showed that the proposed approach produced trustworthy models with high accuracy, selection consistency, low bias and false positives. This indicates that the method is a valuable alternative to the existing techniques. Furthermore, it often generated the best tradeoff between goodness of fit and model complexity when compared to such models, as in the empirical application. As a result of this delicate balance, the proposed method may not necessarily provide the sparsest factor solution. Numerical experiments, however, confirm that the proposed method can produce very good results even if the penalized parameters are estimated just close enough to zero. This is because the edf are also being estimated close to zero, and we would actually need a considerable number of coefficients to see a substantive impact on the total edf and the GBIC. Still, if researchers desire more sparsity, they can manually and subjectively increase the value of the tuning parameter or the influence factor for the automatic procedure.

Notably, we extended the illustrated framework to multiple-group factor models by employing a penalty that simultaneously induced sparsity and cross-group equality of loadings and intercepts. As such, it revealed as a worthy alternative to invariance testing procedures. In this context, the automatic procedure proved particularly useful as it allowed for the estimation of the multiple tuning parameters composing the penalty term in a fast, stable and efficient way.

The presented framework allows one to easily and efficiently combine multiple penalty terms (like in the multiple-group model), as the automatic procedure scales well with the number of tuning parameters. In the empirical application, the alasso penalty was considered for all three penalty terms, but different penalty functions can also be combined if desired.

Another interesting modification pertains to the type of parameters that are penalized. Given the general estimation framework proposed in this work, also residual covariances (i.e., the off-diagonal elements of the covariance matrix of the unique factors) can be penalized to examine the assumption of conditional independence (that is, detect which pairs of variables are conditionally dependent). This model is known in the econometric literature as "sparse approximate factor model" (Bai & Liao, 2016).

We envisage several interesting lines of future research. Firstly, the proposed approach can be applied to structural equation models in which, in addition to the measurement model, a structural model (usually a mediation model for the factors) is tested. Secondly, the results described in this work were derived under the $N > p$ scenario, as it is the case for many applications from the social and behavioral sciences. However, penalized techniques can also be extremely useful in the high-dimensional case, where maximum likelihood estimation is not feasible. It would hence be interesting to review the presented methodology in this demanding set-up. Future research may also evaluate the impact of messy data and larger model sizes on the penalized estimation framework. Finally, the observed variables were assumed to follow a multivariate normal distribution. When this is not reasonable, one can resort to pseudo maximum likelihood (Arminger & Schoenberg, 1989) or, for categorical data, pairwise maximum likelihood (Katsikatsou et al., 2012). Further studies are needed to extend this work to the non-normal case, as this setting poses additional challenges since the asymptotic covariance matrix of the PMLE is no longer consistently estimated by the inverse Fisher information but by a "sandwich-type" covariance matrix (Yuan & Bentler 1997).

Funding Open access funding provided by Alma Mater Studiorum - Universitá di Bologna within the CRUI-CARE Agreement.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Arminger, G., & Schoenberg, R. J. (1989). Pseudo maximum likelihood estimation and a test for misspecification in mean and covariance structure models. *Psychometrika*, *54*(3), 409–425.

Bai, J., & Liao, Y. (2016). Efficient estimation of approximate factor models via penalized maximum likelihood. *Journal of Econometrics*, *191*(1), 1–18.

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55.

Browne, M. W. (2001). An overviewof analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Choi, J., Oehlert, G., & Zou, H. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and its Interface*, *3*(4), 429–436.

Chou, C., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research*, *25*(1), 115–136.

Chou, C., & Huh, J. (2012). Model modification in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 232–246). New York, NY: The Guilford Press.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Filippou, P., Marra, G., & Radice, R. (2017). Penalized likelihood estimation of a trivariate additive probit model. *Biostatistics*, *18*(3), 569–585.

Geminiani, E. (2020). *A penalized likelihood-based framework for single and multiple-group factor analysis models (Doctoral dissertation, University of Bologna)*. Retrieved from http://amsdottorato.unibo.it/9355/.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Hirose, K., & Yamamoto, M. (2014a). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, *79*, 120–132.

Hirose, K., & Yamamoto, M. (2014b). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, *25*(5), 863–875.

Holzinger, K. J. & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. Supplementary Educational Monographs, 48, University of Chicago.

Huang, P. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 499–522.

Huang, P. (2020). lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *Journal of Statistical Software*, *93*(7), 1–37.

Huang, P. & Hu, W. (2019). lslx: Semi-confirmatory structural equation modeling via penalized likelihood [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=lslx (R package version 0.6.8).

Huang, P., Chen, H., & Weng, L. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, *82*(2), 329–354.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566.

Jacobucci, R., Grimm, K. J., Brandmaier, A. M., Serang, S., Kievit, R. A. & Scharf, F. (2019). regsem: Regularized Structural Equation Modeling [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=regsem (R package version 1.3.2).

Jin, S., Moustaki, I., & Yang-Wallentin, F. (2018). Approximated penalized maximum likelihood for exploratory factor analysis: An orthogonal case. *Psychometrika*, *83*(3), 628–649.

Jöreskog, K. G. (1979). A general approach to confirmatory maximum likelihood factor analysis with addendum. In K. G. Jöreskog, D. Sörbom, & J. Magidson (Eds.), *Advances in factor analysis and structural equation models* (pp. 21–43). Cambridge, MA: Abt Books.

Jöreskog, K. G. & Sörbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software. International.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258.

Kelley, K. (2019). MBESS: The MBESS R package [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=MBESS.

Kim, Y., & Gu, C. (2004). Smoothing spline gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(2), 337–356.

Koch, I. (1996). On the asymptotic performance of median smoothers in image analysis and nonparametric regression. *The Annals of Statistics*, *24*(4), 1648–1666.

Konishi, S., & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, *83*(4), 875–890.

Lindstrøm, J. C., & Dahl, F. A. (2020). Model selection with lasso in multi-group structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 33–42.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(1), 59–72.

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2012). Representing contextual effects in multiple-group MACS models. In T. D. Little, J. A. Bovaird, N. A. Card, et al. (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 121–147). NY: Routledge New York.

Lu, Z., Chow, S., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, *51*(4), 519–539.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504.

Marra, G., & Radice, R. (2020). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, *115*(530), 886–895.

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, *39*(1), 53–74.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(1), 1–17.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Abingdon: Routledge.

Mulaik, S. A. (2009). *Foundations of factor analysis*. Boca Raton: Chapman and Hall/CRC.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335.

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Berlin: Springer Science & Business Media.

R Core Team. (2018). R:Alanguage and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/.

Radice, R., Marra, G., & Wojtyś, M. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, *26*(5), 981–995.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Tang, Z., Shen, Y., Zhang, X., & Yi, N. (2017). The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, *205*(1), 77–88.

Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of mind*. Chicago, IL: University of Chicago Press.

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

Trendafilov, N. T., Fontanella, S., & Adachi, K. (2017). Sparse exploratory factor analysis. *Psychometrika*, *82*(3), 778–794.

Ulbricht, J. (2010). Variable selection in generalized linear models. (Doctoral dissertation, Ludwig-Maximilians-Universität München). Verlag Dr. Hut.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Boca Raton: Chapman and Hall/CRC.

Yuan, K., & Bentler, P. M. (1997). Improving parameter tests in covariance structure analysis. *Computational Statistics & Data Analysis*, *26*(2), 177–198.

Yuan, K. & Bentler, P. M. (2006). Structural equation modeling. In C. Rao & S. Sinharay (Eds.), Handbook of Statistics (Vol. 10, pp. 297-358). Elsevier.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics*, *35*(5), 2173–2192.

Manuscript Received: 11 MAY 2020
Final Version Received: 18 FEB 2021
Accepted: 22 FEB 2021