Alma Mater Studiorum
Università di Bologna

Dottorato di Ricerca in

Scienze Statistiche

Ciclo XXXII

Settore Concorsuale: 13/D1
Settore Scientifico Disciplinare: SECS-S/01

A penalized likelihood-based framework for single and multiple-group factor analysis models

Presentata da:
Elena Geminiani

Coordinatore Dottorato:
Prof.ssa Alessandra Luati

Supervisore:
Prof.ssa Angela Montanari

Co-supervisori:
Prof.ssa Silvia Cagnone
Prof.ssa Irini Moustaki

Esame finale anno 2020

Abstract

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. Penalized models are generally less prone to instability in the estimation process and are easier to interpret and generalize than their unpenalized counterparts. However, sparse solutions and stable model selection procedures are only possible if the employed penalty is singular (non-differentiable) at the origin, which poses certain theoretical and computational challenges.

This thesis proposes a general penalized likelihood-based estimation approach for normal linear factor analysis models. The framework builds upon differentiable approximations of non-differentiable penalties and a theoretically founded definition of degrees of freedom. The employed optimization algorithm exploits second-order analytical derivative information and is integrated with an automatic tuning parameter selection procedure that finds the optimal value of the tuning without resorting to grid-searches. Some theoretical aspects of the penalized estimator are discussed. The proposed approach is evaluated in an extensive simulation study and illustrated using a psychometric data set.

As a meaningful addition, the illustrated framework is extended to multiple-group factor analysis models, which are commonly used in cross-national surveys. The employed penalty simultaneously induces sparsity and cross-group equality of loadings and intercepts. The automatic procedure proves particularly useful in this challenging context, as it allows for the estimation of the multiple tuning parameters that compose the penalty term in a fast, stable and efficient way. The merits of the proposed technique are demonstrated through numerical and empirical examples.

All the necessary routines are integrated into the `R` package `GJRM` to enhance reproducible research and transparent dissemination of results.

To my grandparents

Acknowledgements

I am very grateful to my supervisors Prof. Angela Montanari and Prof. Silvia Cagnone for suggesting to focus my research on such an exciting topic and their guidance throughout this project. They offered me the opportunity of spending visiting periods abroad, which made my PhD experience immensely enriching and stimulating.

My sincerest thanks go to Prof. Irini Moustaki for welcoming me at the London School of Economics and Political Science and her great support and encouragement. I am honoured that my work has benefited from her valuable suggestions. Her spirit of collaborative research brought me to a further visiting period at University College London, where my work has been enthusiastically guided by Prof. Giampiero Marra. He has provided me with innumerable worthy ideas on my project, and pushed me beyond my limits.

I acknowledge the financial support that I have received during my stays in London from the ERASMUS+ and Marco Polo programs.

Finally, an honourable mention goes to my family and Luca, for their ever-present encouragement, understanding and empowering support in completing this thesis.

There is an inherent implausibility
about assuming homogeneity in any human population
even if one does not wish to attribute
any heterogeneity to differences
in innate ability.

David J. Bartholomew

Table of Contents

Figures

Tables

Theorems

Propositions

CHAPTER 1

Introduction

Factor analysis has been extensively applied in the social, behavioral and natural sciences as a tool for summarizing the interrelationships among the observed variables into a smaller set of latent variables (factors). 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 the number of factors that are required to explain their interrelationships. 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 used for testing a hypothesis arising from past evidence and theory or after a preliminary EFA, so the number of latent variables and the observed variables that are used to measure them is known in advance. An intermediate step between the two that allows one to develop more realistic solutions while remaining in the CFA framework is E/CFA (Brown, 2014), which consists of a CFA model applying the same number of restrictions used in EFA (i.e., all factor loadings are estimated). In the same spirit, in exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) the CFA measurement model of a structural equation model (SEM) is replaced with an EFA.

In data reduction techniques such as factor analysis, the interest is in obtaining factor solutions that exhibit a "simple structure" (Thurstone, 1947), which are

particularly easy to interpret. Under simple structure, each factor is defined by the subset of the observed variables that load highly on the factor (referred to as pure measures), and each observed variable preferably has a high loading on one factor (referred to as primary loading) and close to zero loadings on the remaining factors (referred to as cross-loadings). 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 and E/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 singular at the origin (Fan & Li, 2001), so that it produces a sparse factor structure, that is, a loading matrix where the number of non-zero entries is much smaller than the total number of its elements. This definition does not impose any pattern on the non-zero 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 non-zero 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 enforced in the model. A grid-search over a range of tuning values is generally conducted, and the optimal

model 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") 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 reparameterized 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 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 (in press) 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) 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.

This thesis proposes 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. We also provide a theoretically founded definition of degrees of freedom (required when performing model selection), discuss the asymptotic properties of the penalized estimator 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 thesis is organized as follows. In the next chapter, we review the classical linear factor analysis model and illustrate the local approximation of several convex and non-convex penalties, including lasso, adaptive lasso, scad and mcp. The differentiable approximations of the penalties are motivated by the necessity of having a differentiable objective function, which is an indispensable prerequisite for the theoretical derivation of the degrees of freedom of the model and the computationally and theoretically founded estimation framework illustrated in Chapter 3. A separate section is devoted to the discussion of the asymptotic properties of the penalized estimator. In Chapter 4, we numerically and empirically evaluate the performances of the model and compare them to other penalized methods present in the literature through an extensive simulation study and a psychometric application. The extension of the model and the penalized estimation approach to the case of multiple groups are delineated in Chapter 5. In this challenging context, a suitable penalty function should simultaneously encourage sparsity and invariance in the factor loadings and intercepts. We then describe how the penalized estimation framework can be adapted in presence of the multiple tuning parameters that compose the penalty term. Numerical and empirical examples on the penalized multiple-group factor model are given in Chapter 6. The proposed methodology is integrated into the freely available `R` package `GJRM` (Marra & Radice, 2019b) to enhance reproducible research and transparent dissemination of results. For an overview of the main functions and a practical illustration of the analyses reported in this work, refer to Chapter 7. Finally, we present a general discussion and suggest directions for future research in Chapter 8.

Additional details on several topics (e.g., the single and multiple-group factor analysis model, the estimation framework, and the theoretical derivations and descriptions) are covered in Appendix A through Appendix F.

CHAPTER 2

Sparsity in the normal linear factor model

After a review of the normal linear factor analysis model (Section 2.1), we illustrate several well-known convex and non-convex penalties commonly used to introduce sparsity in a subset of the parameters (Section 2.2). These penalties, which include the lasso, alasso, scad and mcp, all belong to the L_1-type family and are thus singular at the origin, which is problematic for developing a coherent computational and theoretical inferential framework. To address this issue, we propose to replace the non-differentiable penalties with their differentiable counterparts obtained via local approximations (Section 2.3). An example clarifying the formulation of the employed penalties is provided in Section 2.3.1.

2.1 The normal linear factor analysis model

The classical linear factor analysis model takes the form:

$$\boldsymbol{x} = \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\epsilon}, \tag{2.1}$$

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 $\boldsymbol{\epsilon}$ is the $p \times 1$ vector of unique factors. It is assumed that $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi})$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi}$ usually a diagonal matrix (i.e., the observed variables are conditionally independent), and $\boldsymbol{f}$

is uncorrelated with $\boldsymbol{\epsilon}$. The factor loadings quantify the relationship between each observed variable and latent variable; in other words, how much each observed variable contributes to measuring the factor. The unique variances define the portions of variance in the observed variables not accounted for by the common factors. From the above assumptions, it follows that $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where the model-implied covariance matrix is $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$.

The common factors in expression (2.1) are allowed to covary, since there is in general no prior reason to expect substantively interesting latent variables to be uncorrelated (Bartholomew, Knott & Moustaki, 2011). In the social and behavioral sciences, latent factors are often intercorrelated. Typical examples are questionnaires whose latent structures entail several interrelated dimensions of broader constructs, mental disorders manifested by various clusters of interconnected symptoms or delinquency behaviors defined by various intertwined acts of misconduct. The estimation of the factor covariances also provides significant information, such as the existence of redundant factors or a potential higher-order structure (Brown, 2014). Lastly, even if the common factors are uncorrelated in the population, due to the practical necessity of sampling individuals from a population, it has been argued that it is always reasonable for common factors in a model to be correlated (McArdle, 2007).

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 $m \leq \min\left(N, \frac{p(p+1)}{2}\right)$ elements, with N the total sample size, constitute the free parameters, and are collected in the vector $\boldsymbol{\theta} = (\text{vec}(\boldsymbol{\Lambda})^T, \text{diag}(\boldsymbol{\Psi})^T, \text{vech}(\boldsymbol{\Phi})^T)^T$, where the vec$(\cdot)$ operator converts the enclosed matrix into a vector by stacking its columns, diag$(\cdot)$ extracts the diagonal elements of the enclosed square matrix, and 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.

The common factors are latent variables. As such, they are unobserved and thus have no defined metrics, which must be set by the researcher. This is usually

done in one of two ways. In the first method, the variance of the latent variable is fixed to a specific value, usually 1.0, which generates a standardized solution if the observed variables are standardized. This method is particularly useful in the following circumstances: as a parallel to traditional exploratory factor analysis; when the observed variables have been assessed on an arbitrary metric; and when the standardized solution is of more interest. In the second way, the researcher fixes the metric of the latent variable to be the same as one of its observed variables. The observed variable selected to pass its metric on to the factor is often referred to as a "marker" or "reference" variable. This model leads to an unstandardized solution, which is especially useful in tests of measurement invariance across groups and in evaluations of scale reliability. A third procedure, known as effects-coding, specifies the scale of a latent variable by constraining the corresponding set of loadings to average 1.0. However, it is not ideal in the presence of many cross-loadings among the observed variables (see Little, Slegers & Card, 2006 for details).

The way in which the scale of the latent variable is identified has no impact on overall goodness of fit (i.e., the above solutions produce identical goodness of fit indices), as each scale setting method is simply an alternative but equivalent parameterization of the same model. However, the standard errors are not invariant to the method used to define the scale of the latent variable. In other words, the magnitude of the standard error and the corresponding conclusions regarding the statistical significance of freely estimated parameters might vary based on the selection of the marker variable, or when the scale of the latent variable is defined by fixing its variance to 1.0 (Bollen, 1989).

In this work, we opted for the first approach and fixed the factor variances to unity, as it is common practice in single-group analyses.

The normal linear factor model is not identified because there is an infinite number of matrices $(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ that will reproduce the covariance structure $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$. Equation (2.1) is still satisfied if we replace $\boldsymbol{f}$ by $\boldsymbol{M}\boldsymbol{f}$, $\boldsymbol{\Lambda}$ by $\boldsymbol{\Lambda}\boldsymbol{M}^{-1}$ and $\boldsymbol{\Phi}$ by $\boldsymbol{M}\boldsymbol{\Phi}\boldsymbol{M}^T$, where $\boldsymbol{M}$ is any nonsingular orthogonal matrix of order r corresponding to a nonsingular transformation of the factors. This means

that the parameters in $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ are not independent of one another, and to make the estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ unique, we must impose (at least) r^2 constraints on the elements of $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$, since $\boldsymbol{M}$ has r^2 elements.

When we fix the scales of the latent variables, r constraints are imposed on either $\boldsymbol{\Phi}$ or $\boldsymbol{\Lambda}$. The remaining $r(r-1)$ constraints are imposed by requiring that certain elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ have values specified in advance. The most common method requires that at least $r-1$ elements of $\boldsymbol{\Lambda}$, in each column, are zero. Jöreskog (1979) showed that in case of an oblique solution, the following set of conditions is sufficient for uniqueness of $\boldsymbol{\Lambda}$:

1. Let $\boldsymbol{\Phi}$ be a symmetric positive definite matrix with $\text{diag}(\boldsymbol{\Phi}) = \boldsymbol{I}$;

2. Let $\boldsymbol{\Lambda}$ have at least $r-1$ fixed zeros in each column;

3. Let $\boldsymbol{\Lambda}_j$ have rank $r-1$, where $\boldsymbol{\Lambda}_j$, $j = 1, \ldots, r$ is the submatrix of $\boldsymbol{\Lambda}$, consisting of the rows of $\boldsymbol{\Lambda}$ which have fixed zero elements in the j^{th} column.

The fixed unities in the diagonal of $\boldsymbol{\Phi}$ set the unit of measurement of the factors. As previously mentioned, an alternative way of doing this is to fix one non-zero value in each column of $\boldsymbol{\Lambda}$ instead. Conditions 1–2 are therefore equivalent to requiring that $\boldsymbol{\Lambda}$ has at least $r-1$ fixed zeros in each column and one fixed non-zero value in each column, the latter values being in different rows. In this work, we impose for the normal linear factor model the set of restrictions illustrated by conditions 1–3.

It is important to notice that these conditions solve the "rotational uniqueness problem", but do not guarantee that the factor model is identified (Bollen & Jöreskog, 1985). The so-called "global identification" problem has only been solved for simple models, e.g., the congeneric model (Jöreskog, 1971), and no general necessary and sufficient rules exist for more complex models, like the ones with cross-loadings for all observed variables. In practice, software packages perform several empirical checks to test for "local identification". A more detailed treatment of these issues is provided in Bollen (1989) and Millsap (2012).

For a random sample of deviation scores $\boldsymbol{x}_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of size N from a

multivariate normal distribution, the likelihood function is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = \prod_{\alpha=1}^{N} f(\boldsymbol{x}_\alpha|\boldsymbol{\theta}) &= \prod_{\alpha=1}^{N} \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\boldsymbol{x}_\alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_\alpha\right\} \\ &= (2\pi)^{-\frac{N}{2}p}|\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{N}\boldsymbol{x}_\alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_\alpha\right\}.\end{aligned}$$

The log-likelihood, which is defined as the logarithm of $\mathcal{L}(\boldsymbol{\theta})$, takes the form (see Appendix A.1):

$$\ell(\boldsymbol{\theta}) \coloneqq \log \mathcal{L}(\boldsymbol{\theta}) = -\frac{N}{2}\left\{\log|\boldsymbol{\Sigma}| + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + p\log(2\pi)\right\}, \tag{2.2}$$

where $\boldsymbol{S}$ is the sample covariance matrix. The maximum likelihood estimator (MLE) is then defined as

$$\hat{\boldsymbol{\theta}}^{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}).$$

As noticed by Jöreskog (1967), the maximum likelihood estimator resulting from the maximization of the log-likelihood is equivalent to the one obtained by the minimization of the fit function

$$F = \log|\boldsymbol{\Sigma}| + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) - \log|\boldsymbol{S}| - p. \tag{2.3}$$

After discarding the numerical constants in (2.2) and (2.3), the expressions of the log-likelihood and the fit function differ by the multiplicative factor $-\frac{N}{2}$, which does not impact the optimization process and only produces a different value of the objective function.

From standard asymptotic theory (Anderson, 1989; Yuan & Bentler, 1997), $\hat{\boldsymbol{\theta}}^{\text{MLE}}$ is asymptotically consistent and efficient, and follows a multivariate normal distribution with covariance matrix obtained from the inverse of the expected Fisher information matrix $\boldsymbol{\mathcal{J}}$,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}^{\text{MLE}} - \boldsymbol{\theta}_0) \to \mathcal{N}\left(\mathbf{0}, \left[\frac{1}{N}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\right]^{-1}\right),$$

where $\boldsymbol{\theta}_0$ is the true parameter vector.

Let θ_q denote the q^{th} parameter from the m-dimensional vector $\boldsymbol{\theta}$. The propositions below enunciate the general expressions of the gradient of the log-likelihood $\boldsymbol{g}(\boldsymbol{\theta}) \coloneqq \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, the Hessian matrix of the second-order derivatives $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) \coloneqq \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$, and the expected Fisher information $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) \coloneqq \mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})^T] = -\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})\right]$ for the normal linear factor model.

Proposition 2.1 (Gradient of the normal linear factor model)**.** *The gradient of the log-likelihood of the normal linear factor analysis model in equation (2.1) with respect to an arbitrary scalar variable θ_q takes the form:*

$$[\boldsymbol{g}(\boldsymbol{\theta})]_q = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_q} = -\frac{N}{2} tr\left\{ \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q} \right\}. \tag{2.4}$$

Proof. See Appendix A.2.1.1. ■

Proposition 2.2 (Hessian of the normal linear factor model)**.** *The Hessian matrix of the normal linear factor analysis model in equation (2.1) with respect to two arbitrary scalar variables θ_q and $\theta_{q'}$ takes the form:*

$$\begin{aligned}
[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})]_{qq'} &= \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_q \partial \theta_{q'}} \\
&= -\frac{N}{2}\Bigg\{ tr\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}} \right) \\
&\qquad + tr\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1} \left(\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_q \partial \theta_{q'}} - 2\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}} \right) \right] \Bigg\}.
\end{aligned} \tag{2.5}$$

Proof. See Appendix A.2.2.1. ■

Proposition 2.3 (Expected Fisher information of the normal linear factor model)**.** *The expected Fisher information matrix of the normal linear factor analysis model in equation (2.1) with respect to two arbitrary scalar variables θ_q and $\theta_{q'}$ takes the form:*

$$[\boldsymbol{\mathcal{J}}(\boldsymbol{\theta})]_{qq'} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_q \partial \theta_{q'}} \right] = \frac{N}{2} tr\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}} \right). \tag{2.6}$$

Proof. The Fisher information is derived by noticing that $\mathbb{E}[\boldsymbol{S}] = \boldsymbol{\Sigma}$ as $N \to \infty$, and thus neglecting the second term in (2.5). ■

The specific forms of these derivatives with respect to each parameter matrix are given in propositions A.1-A.3 in Appendix A.2.

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 sections illustrate how such sparsity-inducing penalty functions can be specified (Section 2.2) and suitably approximated (Section 2.3).

2.2 Sparsity-inducing penalties

Given that the primary interest of factor analysis is a sparse loading matrix, penalization is imposed on the factor loading matrix $\boldsymbol{\Lambda}$. Let us write the parameter vector as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{q^\star}, \theta_{q^\star+1}, \ldots, \theta_m)^T$, where the sub-vector $(\theta_1, \ldots, \theta_{q^\star})^T$ collects the penalized parameters (i.e., the factor loadings), whereas $(\theta_{q^\star+1}, \ldots, \theta_m)^T$ the unpenalized parameters (i.e., the free elements in $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$). Because of the presence of fixed elements in $\boldsymbol{\Lambda}$ (Section 2.1), the number of penalized factor loadings $q^\star$ is smaller than $p \times r$. Define the diagonal matrix $\boldsymbol{R}_q = \text{diag}(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ where the 1 on the $(q, q)^{\text{th}}$ entry of the matrix corresponds to the q^{th} parameter in $\boldsymbol{\theta}$, for $q = 1, \ldots, q^\star$, and $\boldsymbol{R}_q = \boldsymbol{O}_{m \times m}$ for $q = q^\star + 1, \ldots, m$.

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 $q = 1, \ldots, q^\star$, and zero otherwise. One of the best-known penalties is the lasso (Tibshirani, 1996), which is defined as

$$\mathcal{P}^L_\eta(\boldsymbol{\theta}) = \eta \sum_{q=1}^{q^\star} |\theta_q|. \tag{2.7}$$

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 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, their estimator works as if the true non-zero parameters were known beforehand. The alasso is defined as

$$\mathcal{P}_\eta^A(\boldsymbol{\theta}) = \eta \sum_{q=1}^{q^\star} w_q |\theta_q| = \eta \sum_{q=1}^{q^\star} \frac{|\theta_q|}{|\hat{\theta}_q|^a} \quad \text{for } a > 0. \tag{2.8}$$

This penalty uses an adaptive weighting scheme based on a set of available weights $w_q = \dfrac{1}{|\hat{\theta}_q|^a}\,(q = 1, \ldots, q^\star)$, which are often taken to be the maximum likelihood estimates, that is, $w_q = \dfrac{1}{|\hat{\theta}_q^{\text{MLE}}|^a}$. The higher the exponent a, the more influential the weights, and in turn, the larger the penalization.

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

$$\begin{aligned}\mathcal{P}_\eta^S(\boldsymbol{\theta}) = \sum_{q=1}^{q^\star} \Bigg\{ & \eta|\theta_q| \mathbb{1}(0 \leq |\theta_q| \leq \eta) \\ & - \left[\frac{\theta_q^2 + \eta^2 - 2\eta a|\theta_q|}{2(a-1)}\right] \mathbb{1}(\eta < |\theta_q| \leq a\eta) \\ & + \frac{\eta^2(a+1)}{2} \mathbb{1}(|\theta_q| > a\eta) \Bigg\} \quad \text{for } a > 2,\end{aligned} \tag{2.9}$$

and the mcp as

$$\begin{aligned}\mathcal{P}_\eta^M(\boldsymbol{\theta}) = \sum_{q=1}^{q^\star} \Bigg\{ & \left(\eta|\theta_q| - \frac{\theta_q^2}{2a}\right) \mathbb{1}(0 \leq |\theta_q^2| \leq a\eta) \\ & + \frac{\eta^2 a}{2} \mathbb{1}(|\theta_q| > a\eta) \Bigg\} \quad \text{for } a > 1,\end{aligned} \tag{2.10}$$

where a is an additional tuning parameter. The superscripts L, A, S, M in equations (2.7)-(2.10) refer to the lasso, alasso, scad and mcp, respectively. The derivations of expressions (2.7)-(2.10) can be found in Appendix B.1.

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.

In the expressions of the penalties $\mathcal{P}_\eta^A(\boldsymbol{\theta}), \mathcal{P}_\eta^S(\boldsymbol{\theta}), \mathcal{P}_\eta^M(\boldsymbol{\theta})$, we did not stress their dependence on the additional tuning parameter a because this quantity is implicitly assumed to be fixed, for instance, it has been determined from prior trials. Common values of the shape parameter of the scad range between 2.5 and 4.5 (Huang et al., 2017), with 3.7 being the conventional level employed in the literature and suggested by Fan and Li (2001). For the mcp, values of a between 1.5 and 3.5 are often considered (Huang, 2018), whereas the exponent of the alasso does not typically exceed 2 (Zou, 2006).

Simplified examples of the shapes of the illustrated penalties are shown in Figure 2.1. For all penalties $\eta = 1$, whereas the shape parameter for the scad is $a = 3.7$, for the mcp is $a = 3$, and the exponent of the alasso is $a = 1$. All of the four penalties belong to the L_1-type family and are singular at $\theta = 0$. Contrarily to the lasso and alasso, the depicted scad and mcp penalties are concave functions.

Figure 2.2 represents the surface plot of the alasso penalty by varying the values of the parameter θ and the estimate $\hat{\theta}$ appearing in the adaptive weight (equation (2.8)). For fixed θ, the penalty has a V-shape and increases as the value of $\hat{\theta}$ gets larger, with the magnitude of the penalization being inversely related to the size of θ. As a consequence, the amount of penalization on $\hat{\theta}$ increases as θ approaches zero.

Figure 2.3 illustrates the shapes of the alasso, scad and mcp by varying the value of their additional tuning parameter a. The exponent in the expression of

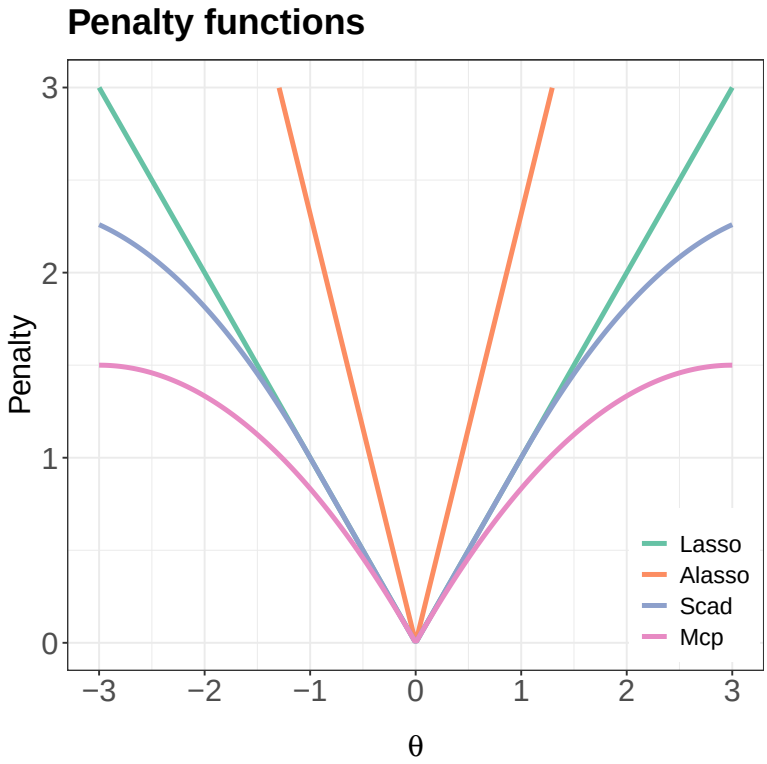


Figure 2.1: Shapes of the lasso, alasso ($a = 1$), scad ($a = 3.7$) and mcp ($a = 3$) penalty functions for $\eta = 1$.

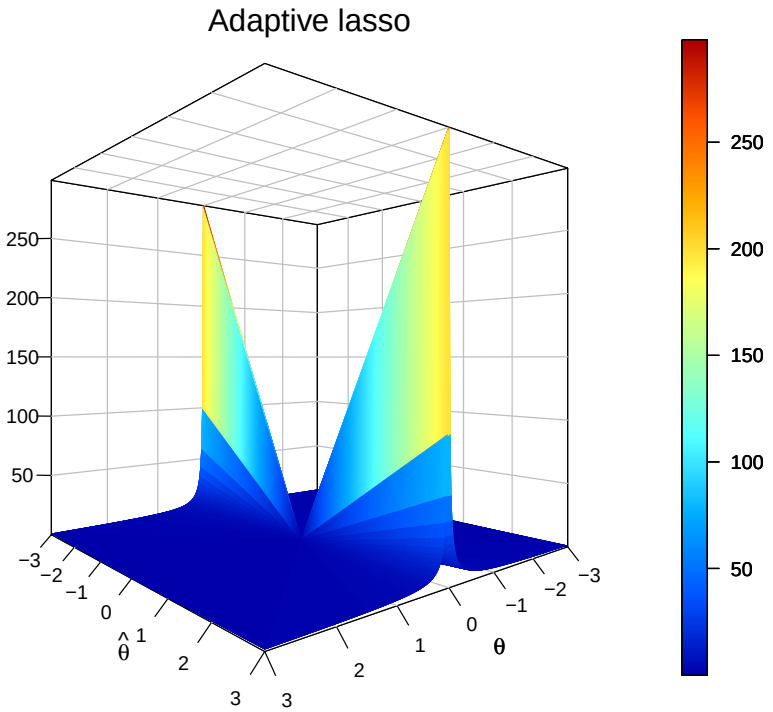


Figure 2.2: Three-dimensional surface plot of the alasso penalty ($\eta = 1, a = 1$) by varying the parameter θ and the estimate $\hat{\theta}$.

the alasso controls the importance given to the adaptive weights. As a gets higher, the magnitude of the penalization progressively increases for small values of $\hat{\theta}$, and decreases for large values. The shapes of the scad and mcp are similar, with their degree of concavity decreasing as the shape parameter a increases. When $a \to \infty$ (see for instance, $a = 50$), the two penalties converge to the lasso.

The above penalties help to obtain sparse solutions, however, they are non-differentiable at the origin, 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.

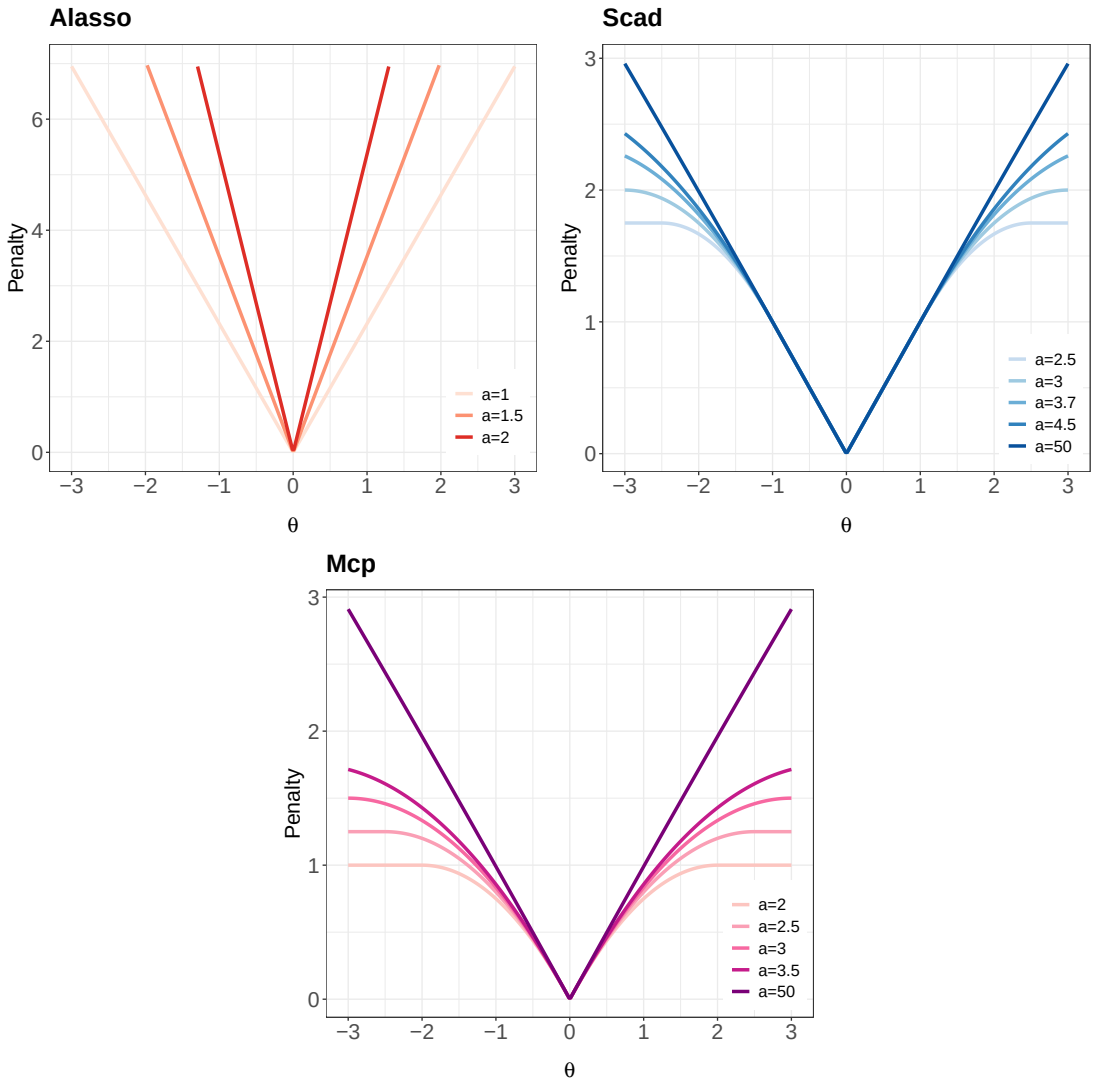


Figure 2.3: The alasso, scad and mcp penalties by varying the value of their additional tuning parameter a.

2.3 Locally approximated penalties

Ulbricht (2010) pointed out that a good penalty function should satisfy the following properties, for $q = 1, \ldots, m$:

(P.1) $\mathcal{P}_{\eta,q} : \mathbb{R}^+ \to \mathbb{R}^+$ 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 $\dfrac{\partial \mathcal{P}_{\eta,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1)}{\partial ||\boldsymbol{R}_q\boldsymbol{\theta}||_1} > 0.$

However, the lasso, alasso, scad and mcp are all singular at $\theta_q = 0$. To address this issue, in the same spirit as for instance Filippou, Marra and Radice (2017), we locally approximate the non-differentiable L_1-norms in (2.7)-(2.10) at their critical point $||\boldsymbol{R}_q\boldsymbol{\theta}||_1 = 0$ and combine this with ideas by Fan and Li (2001) and Ulbricht (2010). Let $||\boldsymbol{R}_q\boldsymbol{\theta}||_1 = ||\boldsymbol{\xi}_q||_1$, and assume that an approximation $\mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{A})$ of the L_1-norm $||\cdot||_1$ exists such that

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

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. As in Koch (1996), we use $||\boldsymbol{\xi}_q||_1 = \mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{A}) = (\boldsymbol{\xi}_q^T\boldsymbol{\xi}_q + \bar{c})^{\frac{1}{2}}$, 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

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

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 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,

$$\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta}) \approx \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}), \tag{2.11}$$

where $\nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \dfrac{\partial \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}$. By applying the chain rule, the penalty $\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta})$ can be written as

$$\begin{aligned}
\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta}) &\approx \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\
&\approx \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \frac{\partial \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})^T}{\partial \tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\
&\approx \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \sum_{q=1}^{m}\left[\frac{\partial \mathcal{P}_{\eta,q}^{\mathcal{T}}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial \tilde{\boldsymbol{\theta}}}\right]^T(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\
&\approx \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \sum_{q=1}^{m}\left[\frac{\partial \mathcal{P}_{\eta,q}^{\mathcal{T}}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\right]^T \cdot \left[\frac{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}{\partial \boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}\right]^T \cdot \left[\frac{\partial \boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}{\partial \tilde{\boldsymbol{\theta}}}\right]^T(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}).
\end{aligned} \tag{2.12}$$

Let us examine the quantities that make up each addend of expression (2.12). The first factor represents the derivative of $\mathcal{P}_{\eta,q}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ with respect to the L_1 norm of its argument $\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}$. Because the expression depends on the specific form of the penalty $\mathcal{T}$, it is separately computed for each of the examined penalties in Appendix B.2. The second factor denotes the derivative of the L_1-norm with respect to its argument, and is equal for all penalties to

$$\begin{aligned}
\frac{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}{\partial \boldsymbol{R}_q\tilde{\boldsymbol{\theta}}} &= \frac{\partial}{\partial \boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}\left\{\sum_{s=1}^{m}\left[(\boldsymbol{R}_s\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_s\tilde{\boldsymbol{\theta}}\right]^{\frac{1}{2}}\right\} = \frac{\partial}{\partial \boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}\left[(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}\right]^{\frac{1}{2}} \\
&= \frac{1}{2}\left[(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}\right]^{-\frac{1}{2}} \cdot 2\boldsymbol{R}_q\tilde{\boldsymbol{\theta}} = [(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}]^{-\frac{1}{2}}\boldsymbol{R}_q\tilde{\boldsymbol{\theta}} \\
&\approx \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}}\boldsymbol{R}_q\tilde{\boldsymbol{\theta}},
\end{aligned}$$

where the denominator is approximated by $\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+\bar{c}}$ to allow for the case of $\tilde{\boldsymbol{\theta}}=\mathbf{0}$. Finally, the third factor is simply $\dfrac{\partial \boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}{\partial\tilde{\boldsymbol{\theta}}}=\boldsymbol{R}_q$.

By combining the local approximation $(\boldsymbol{R}_q\boldsymbol{\theta})\approx(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})$ (Fan & Li, 2001) with the following approximation introduced in Ulbricht (2010):

$$\begin{aligned}(\boldsymbol{R}_q\boldsymbol{\theta})^T\boldsymbol{R}_q(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}) &= (\boldsymbol{R}_q\boldsymbol{\theta})^T\boldsymbol{R}_q\boldsymbol{\theta}-(\boldsymbol{R}_q\boldsymbol{\theta})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}\\ &= \frac{1}{2}\left\{(\boldsymbol{R}_q\boldsymbol{\theta})^T\boldsymbol{R}_q\boldsymbol{\theta}-2(\boldsymbol{R}_q\boldsymbol{\theta})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}\right\}\\ &\quad+\frac{1}{2}\left\{(\boldsymbol{R}_q\boldsymbol{\theta})^T\boldsymbol{R}_q\boldsymbol{\theta}-(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}\right\}\\ &= \frac{1}{2}\left\{(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q^T\boldsymbol{R}_q(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})\right\}+\frac{1}{2}\left\{(\boldsymbol{R}_q\boldsymbol{\theta})^T\boldsymbol{R}_q\boldsymbol{\theta}-(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}\right\}\\ &\approx \frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{R}_q^T\boldsymbol{R}_q\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}^T\boldsymbol{R}_q^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}\right),\end{aligned}$$

we have that

$$\begin{aligned}&\left[\frac{\partial\mathcal{P}^{\mathcal{T}}_{\eta,q}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\right]^T\cdot\left[\frac{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}{\partial\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}\right]^T\cdot\left[\frac{\partial\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}{\partial\tilde{\boldsymbol{\theta}}}\right]^T(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})\\ =\;&\frac{\partial\mathcal{P}^{\mathcal{T}}_{\eta,q}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\cdot\left[\frac{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}{\partial\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}\right]^T\cdot\frac{\partial\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}{\partial\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})\\ =\;&\frac{\partial\mathcal{P}^{\mathcal{T}}_{\eta,q}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\cdot\frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+\bar{c}}}(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\cdot\boldsymbol{R}_q(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})\\ \approx\;&\frac{\partial\mathcal{P}^{\mathcal{T}}_{\eta,q}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\cdot\frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+\bar{c}}}\frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{R}_q^T\boldsymbol{R}_q\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}^T\boldsymbol{R}_q^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}\right)\\ =\;&\frac{1}{2}\boldsymbol{\theta}^T\left\{\frac{\partial\mathcal{P}^{\mathcal{T}}_{\eta,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}\tilde{\boldsymbol{\theta}}^T\left\{\frac{\partial\mathcal{P}^{\mathcal{T}}_{\eta,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\}\tilde{\boldsymbol{\theta}}\\ =\;&\frac{1}{2}\left[\boldsymbol{\theta}^T\boldsymbol{\mathcal{S}}^{\mathcal{T}}_{\eta,q}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}^T\boldsymbol{\mathcal{S}}^{\mathcal{T}}_{\eta,q}(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\theta}}\right],\end{aligned}$$

where $\boldsymbol{\mathcal{S}}^{\mathcal{T}}_{\eta,q}(\tilde{\boldsymbol{\theta}})=\dfrac{\partial\mathcal{P}^{\mathcal{T}}_{\eta,q}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\dfrac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+\bar{c}}}\boldsymbol{R}_q^T\boldsymbol{R}_q$.

Let us denote $\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \sum_{q=1}^m \boldsymbol{\mathcal{S}}_{\eta,q}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$. Then, equation (2.12) can be rewritten as

$$\begin{aligned}
\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta}) &\approx \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \sum_{q=1}^m \left[\frac{\partial \mathcal{P}_{\eta,q}^{\mathcal{T}}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\right]^T \left[\frac{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}{\partial \boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}\right]^T \left[\frac{\partial \boldsymbol{R}_q\tilde{\boldsymbol{\theta}}}{\partial \tilde{\boldsymbol{\theta}}}\right]^T (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\
&= \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \sum_{q=1}^m \frac{1}{2}\left[\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_{\eta,q}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}^T \boldsymbol{\mathcal{S}}_{\eta,q}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\theta}}\right] \\
&= \mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} - \frac{1}{2}\tilde{\boldsymbol{\theta}}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\theta}}.
\end{aligned}$$

We can ignore the constant terms that do not depend on $\boldsymbol{\theta}$, namely, $\mathcal{P}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ and $\frac{1}{2}\tilde{\boldsymbol{\theta}}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\theta}}$. Then, the differentiable local approximation of the penalty $\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta})$ is

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

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

$$\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} \boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}. \tag{2.13}$$

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

$$m_q^{\mathcal{T}} = \frac{\partial \mathcal{P}_{\eta,q}^{\mathcal{T}}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T \boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \qquad \text{for } q = 1, \ldots, q^\star,$$

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 Appendix B.2)

$$\left[\boldsymbol{\mathcal{M}}_{\eta}^{L}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^L = \frac{\eta}{\sqrt{\tilde{\theta}_q^2 + \bar{c}}}, \tag{2.14}$$

$$\left[\boldsymbol{\mathcal{M}}_{\eta}^{A}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^A = \frac{\eta}{|\hat{\theta}_q|^a \sqrt{\tilde{\theta}_q^2 + \bar{c}}}, \tag{2.15}$$

$$\left[\boldsymbol{\mathcal{M}}_{\eta}^{S}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^S = \frac{\eta\left[\mathbb{1}(|\tilde{\theta}_q| \leq \eta) + \frac{\max(a\eta - |\tilde{\theta}_q|, 0)}{(a-1)\eta}\mathbb{1}(|\tilde{\theta}_q| > \eta)\right]}{\sqrt{\tilde{\theta}_q^2 + \bar{c}}}, \tag{2.16}$$

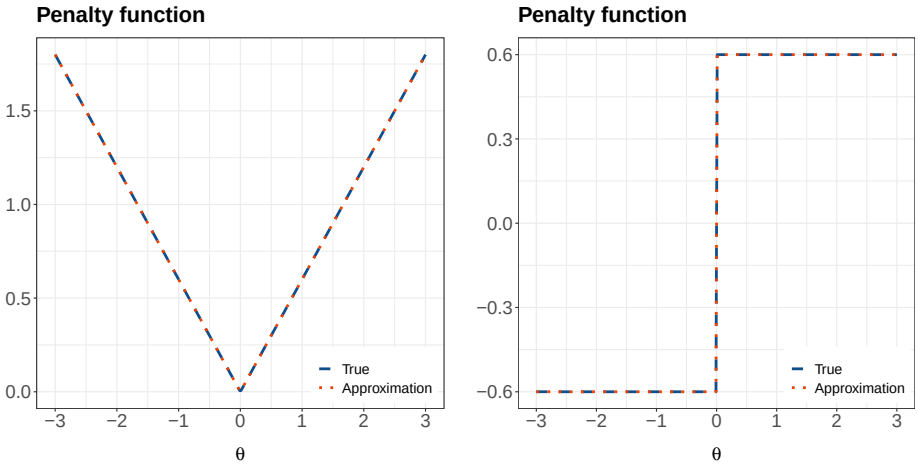
$$\left[\boldsymbol{\mathcal{M}}_{\eta}^{M}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^M = \frac{\left(\eta - \frac{|\tilde{\theta}_q|}{a}\right)\mathbb{1}(|\tilde{\theta}_q| < \eta a)}{\sqrt{\tilde{\theta}_q^2 + \bar{c}}}. \tag{2.17}$$

Figure 2.4 shows a graphical representation of the examined penalties and their first derivatives. On the left-hand side we have the penalty functions and their local approximations, whereas the right-hand side reports the original discontinuous derivatives and the continuous derivatives resulting from the local approximation ($\bar{c} = 10^{-8}$). The plots for the alasso are not presented as the shape of this penalty is proportional to the one of the lasso.

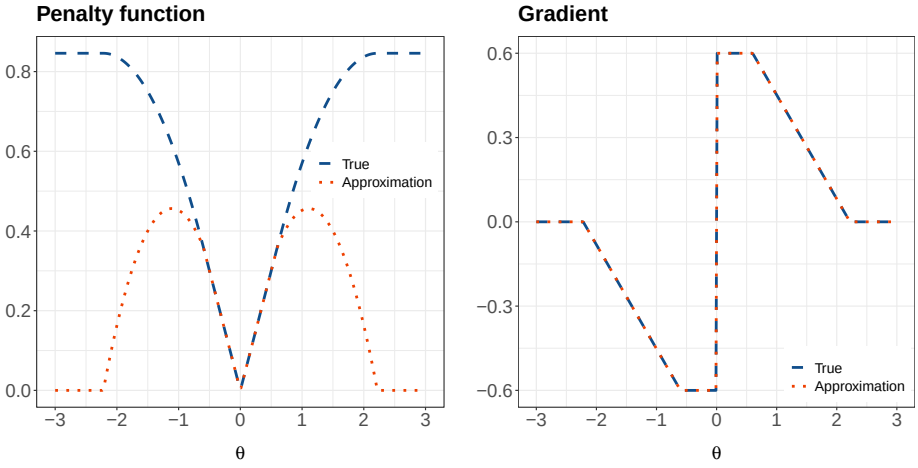
Figures 2.5 and 2.6 extend the bi-dimensional plots to three-dimensional surfaces. On the left-hand side of each figure we find the true penalty functions (or their first derivatives), whereas their local approximations are depicted on the right-hand side.

Although one could employ linear rather than quadratic approximations of the penalties (see e.g., Jin et al., 2018 for a local linear approximation of the scad in EFA), the presented method performs well in our studies, hence we keep this possible modification as a future task to explore.

(a) *Lasso*, $\eta = 0.6$.



(b) *Scad*, $\eta = 0.6, a = 3.7$.



(c) *Mcp*, $\eta = 0.6, a = 3$.

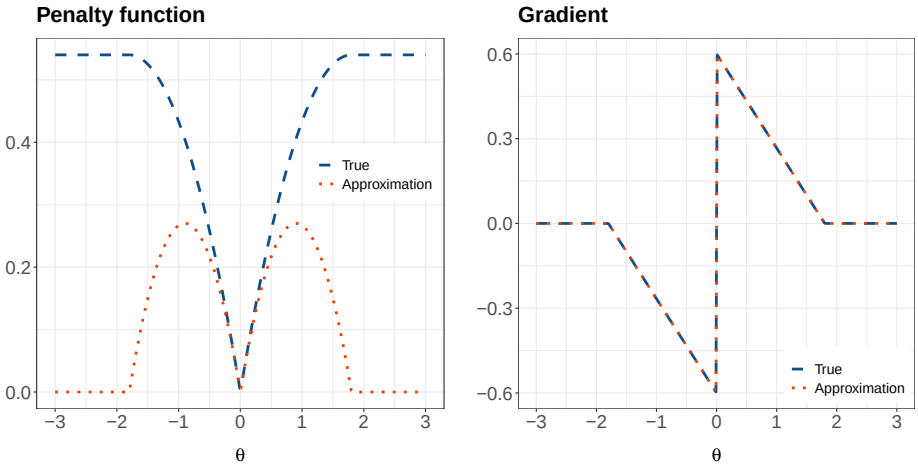


Figure 2.4: The true penalty functions (left-hand side) and their first derivatives (right-hand side) with the local approximations superimposed ($\bar{c} = 10^{-8}$).

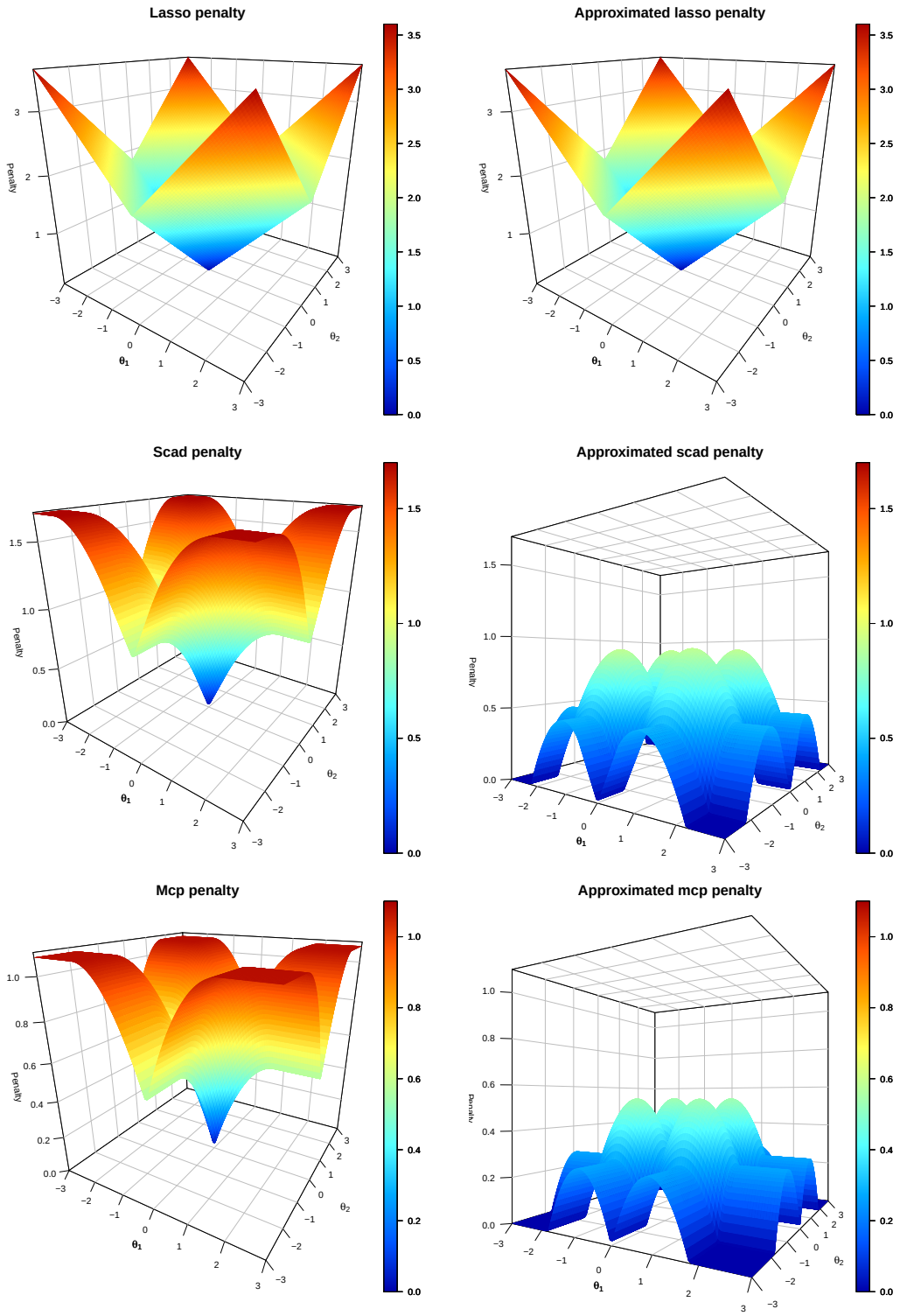


Figure 2.5: The lasso, scad and mcp penalty functions (left-hand side) and their local approximations (right-hand side; $\bar{c} = 10^{-8}$). The tuning parameter $\eta = 0.6$; for the scad $a = 3.7$ and for the mcp $a = 3$.

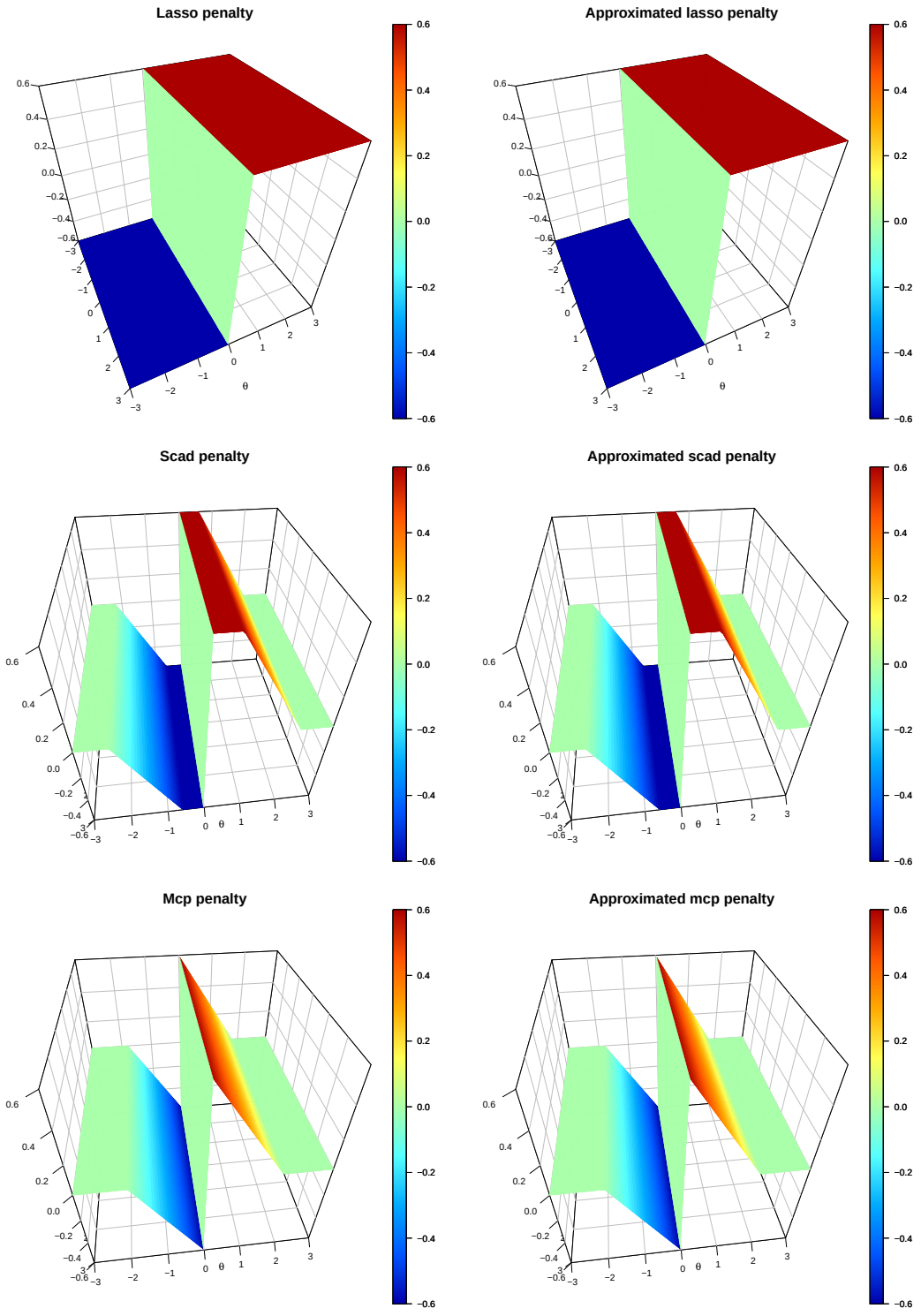


Figure 2.6: The first derivatives of the lasso, scad and mcp penalties (left-hand side) and their local approximations (right-hand side; $\bar{c} = 10^{-8}$). The tuning parameter $\eta = 0.6$, for the scad $a = 3.7$ and for the mcp $a = 3$.

Model modifications

When a global goodness-of-fit test statistic or local-fit indices (Bollen & Long, 1993) indicate lack of fit, modification indices are used to suggest ways for model improvement (e.g., by estimating the loadings erroneously fixed to zero). Modification indices are univariate statistics for each fixed parameter quantifying the minimum decrease in the overall chi-square value that would be achieved if that parameter was freely estimated. However, since model modifications are largely guided by the results obtained from fitting an initial model to a particular sample, they tend to capitalize on chance and yield inflated type I errors.

The presented penalized-likelihood approach bypasses the need for model modifications by automatically recovering an optimally sparse factor structure.

2.3.1 An example

For notational clarity, we illustrate the aforementioned penalties in a simple example. Consider the following normal linear factor analysis model with $p = 6$ observed variables and $r = 2$ common factors:

$$\boldsymbol{x} = \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\epsilon},$$

where it is assumed that $\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi}$ a diagonal matrix, and $\boldsymbol{f}$ is uncorrelated with $\boldsymbol{\epsilon}$. The population parameters are as follows:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \underline{0} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \underline{0} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \\ \lambda_{61} & \lambda_{62} \end{bmatrix} \quad \boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & 0 & 0 & 0 & 0 & 0 \\ & \psi_{22} & 0 & 0 & 0 & 0 \\ & & \psi_{33} & 0 & 0 & 0 \\ & & & \psi_{44} & 0 & 0 \\ & & & & \psi_{55} & 0 \\ & & & & & \psi_{66} \end{bmatrix} \quad \boldsymbol{\Phi} = \begin{bmatrix} \underline{1} & \phi_{12} \\ & \underline{1} \end{bmatrix},$$

where the elements in italic and underlined were fixed for scale setting and identification purposes, as illustrated in Section 2.1. The parameter vector $\boldsymbol{\theta}$ collecting

the free elements of the parameter matrices can be written as

$$\begin{aligned}\boldsymbol{\theta} &= (\text{vec}(\boldsymbol{\Lambda})^T, \text{diag}(\boldsymbol{\Psi})^T, \text{vech}(\boldsymbol{\Phi})^T)^T \\ &= (\lambda_{11}, \lambda_{21}, \lambda_{31}, \lambda_{51}, \lambda_{61}, \lambda_{22}, \lambda_{32}, \lambda_{42}, \lambda_{52}, \lambda_{62}, \psi_{11}, \psi_{22}, \psi_{33}, \psi_{44}, \psi_{55}, \psi_{66}, \phi_{12})^T.\end{aligned}$$

Conveniently, the parameter vector can be rewritten as

$$\boldsymbol{\theta} = (\underbrace{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}}_{\textit{Factor loadings}}, \theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{16}, \theta_{17})^T,$$

where the sub-vector $(\theta_1, \ldots, \theta_{10})^T$ collects the parameters that are being penalized (i.e., the factor loadings), whereas $(\theta_{11}, \ldots, \theta_{17})^T$ the unpenalized parameters (i.e., the free elements in $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$). Let $q^\star = 10$ be the number of penalized parameters, and $m = 17$ the total number of parameters. Define

$$\boldsymbol{R}_q = \begin{array}{c} \\ 1 \\ \\ q \\ \\ \\ 17 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & & q & & & 17 \end{array} \\ \begin{bmatrix} 0 & \ldots & 0 & \ldots & \ldots & 0 \\ \vdots & \ddots & \vdots & & & \vdots \\ 0 & \ldots & 1 & \ldots & \ldots & 0 \\ \vdots & & \vdots & \ddots & & \vdots \\ \vdots & & \vdots & & \ddots & \vdots \\ 0 & \ldots & 0 & \ldots & \ldots & 0 \end{bmatrix} \end{array} \quad \text{for } q = 1, \ldots, 10,$$

and $\boldsymbol{R}_q = \boldsymbol{O}_{17\times 17}$ for $q = 11, \ldots, 17$. Then, the sparsity-inducing penalty is expressed as

$$\mathcal{P}_\eta(\boldsymbol{\theta}) = \sum_{q=1}^{17} \mathcal{P}_{\eta,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1),$$

where $||\boldsymbol{R}_q\boldsymbol{\theta}||_1 = |\theta_q|$ for $q = 1, \ldots, 10$, and 0 for $q = 11, \ldots, 17$.

CHAPTER 3

Penalized estimation framework and theoretical aspects

In Section 3.1, we illustrate how simultaneous estimation of the model parameters is achieved using a carefully structured trust-region algorithm. We then describe two possible approaches for the determination of the tuning parameter of the penalized model. The first solution is based on a grid-search over a range of tuning values, and picks the optimal model on the basis of a generalized information criterion (Section 3.2). Alternatively, we propose an automatic tuning parameter selection procedure, which finds the optimal amount of sparsity without resorting to grid-searches (Section 3.3). The chapter concludes with a discussion of the theoretical properties of the proposed estimator (Section 3.4).

3.1 Penalized maximum likelihood estimation

The penalty functions illustrated in Chapter 2 can be directly introduced within the estimation process by means of penalized maximum likelihood estimation procedures. The penalized log-likelihood is given by

$$\ell_p(\boldsymbol{\theta}) \coloneqq \sum_{\alpha=1}^{N} \ell(\boldsymbol{x}_\alpha|\boldsymbol{\theta}) - \sum_{\alpha=1}^{N} \mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - N\,\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta}), \tag{3.1}$$

where $\ell(\boldsymbol{\theta})$ is given in equation (2.2), and $\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta})$ is one of the penalties of Section 2.2 generating a sparse factor solution.

Simultaneous estimation of all parameters is achieved by maximizing the penalized log-likelihood in (3.1) and using a local approximation of $\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta})$ (Section 2.3), that is,

$$\max_{\boldsymbol{\theta}} \ell_p(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \left\{ \ell(\boldsymbol{\theta}) - \frac{N}{2} \boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\theta} \right\}, \tag{3.2}$$

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, the Hessian matrix of the second-order derivatives and the expected Fisher information matrix can be written as

$$\begin{aligned}
\boldsymbol{g}_p(\boldsymbol{\theta}) &\coloneqq \frac{\partial \ell_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{g}(\boldsymbol{\theta}) - N \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\theta}, \\
\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}) &\coloneqq \frac{\partial^2 \ell_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) - N \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}), \\
\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}) &\coloneqq -\mathbb{E}\left[\frac{\partial^2 \ell_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right] = \boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) + N \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}).
\end{aligned}$$

For a given value of η in the penalty matrix, which is hence denoted in the following as $\boldsymbol{\mathcal{S}}_{\hat{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$, we seek to minimize the negative penalized log-likelihood $-\ell_p(\boldsymbol{\theta})$. This can be done via a trust-region algorithm (Conn, Gould & Toint, 2000). According to this strategy, at iteration $\mathtt{t}$, the information gathered around $-\ell_p$ is used to construct a "model function" $\mathcal{Q}_p^{[\mathtt{t}]}$ whose behavior near the current point $\boldsymbol{\theta}^{[\mathtt{t}]}$ is similar to that of the actual objective function $-\ell_p$. Because the model $\mathcal{Q}_p^{[\mathtt{t}]}$ may not be a good approximation of $-\ell_p$ when $\boldsymbol{\theta}$ is far away from $\boldsymbol{\theta}^{[\mathtt{t}]}$, the search for a minimizer of $\mathcal{Q}_p^{[\mathtt{t}]}$ is restricted to some region $\mathcal{R}^{[\mathtt{t}]}$ around $\boldsymbol{\theta}^{[\mathtt{t}]}$. This region is usually the ball $||\boldsymbol{s}||_2 < \Delta$, where $||\cdot||_2$ is the Euclidean norm, $\boldsymbol{s}$ the trial step vector aiming at reducing the model function, and the scalar $\Delta > 0$ the trust-region radius. 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.

The model $\mathcal{Q}_p^{[t]}$ is usually a quadratic function of the form:

$$\mathcal{Q}_p^{[t]}(\boldsymbol{s}) = -\left\{\ell_p(\boldsymbol{\theta}^{[t]}) + \boldsymbol{s}^T\boldsymbol{g}_p(\boldsymbol{\theta}^{[t]}) + \frac{1}{2}\boldsymbol{s}^T\boldsymbol{B}(\boldsymbol{\theta}^{[t]})\boldsymbol{s}\right\}, \tag{3.3}$$

where $\boldsymbol{g}_p(\boldsymbol{\theta}^{[t]}) = \boldsymbol{g}(\boldsymbol{\theta}^{[t]}) - N\boldsymbol{\mathcal{S}}_{\hat{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}^{[t]})\boldsymbol{\theta}^{[t]}$ is the penalized score function. The matrix $\boldsymbol{B}(\boldsymbol{\theta}^{[t]})$ can be the penalized Hessian $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}^{[t]}) = \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}^{[t]}) - N\boldsymbol{\mathcal{S}}_{\hat{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}^{[t]})$, or some approximation thereof, as $\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}^{[t]}) = -\mathbb{E}\left[\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}^{[t]})\right]$. If $\boldsymbol{B}(\boldsymbol{\theta}^{[t]})$ is equal to the penalized Hessian, $\mathcal{Q}_p^{[t]}$ agrees with the Taylor-series expansion of $-\ell_p$ around $\boldsymbol{\theta}^{[t]}$ to the first three terms, otherwise the agreement between the two functions is to the first two terms. The derivation of the first and 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 for the normal linear factor model requires computing many elements (see Appendix A.2), the Fisher information matrix is particularly convenient. 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.

Each iteration of the trust-region algorithm solves the sub-problem:

$$\boldsymbol{s}^{[t]} = \arg\min_{\boldsymbol{s}\in\mathbb{R}^m} \mathcal{Q}_p^{[t]}(\boldsymbol{s}) \qquad \text{subject to } ||\boldsymbol{s}||_2 \leq \Delta^{[t]}, \tag{3.4}$$

$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} + \boldsymbol{s}^{[t]}, \tag{3.5}$$

where the current iteration $\boldsymbol{\theta}^{[t]}$ is updated with $\boldsymbol{s}^{[t]}$ if this step produces an improvement over the objective function. In practice, the size of the region is chosen according to the performance of the algorithm during previous iterations, and specifically, to the agreement between the model function and the objective function at previous iterations. Given a step $\boldsymbol{s}^{[t]}$, define the ratio

$$r^{[t]} = \frac{-\left[\ell_p(\boldsymbol{\theta}^{[t]}) - \ell_p(\boldsymbol{\theta}^{[t]} + \boldsymbol{s}^{[t]})\right]}{\mathcal{Q}_p^{[t]}(\mathbf{0}) - \mathcal{Q}_p^{[t]}(\boldsymbol{s}^{[t]})}; \tag{3.6}$$

the numerator is called the actual reduction, whereas the denominator is the predicted reduction. If $r^{[t]}$ is negative, the new objective value $-\ell_p(\boldsymbol{\theta}^{[t]} + \boldsymbol{s}^{[t]})$ is greater than the current value $-\ell_p(\boldsymbol{\theta}^{[t]})$, which means that 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 the model $\mathcal{Q}_p^{[t]}$ and the function $-\ell_p$ over this step. This means that the model can accurately predict the behavior of the objective function along the step $\boldsymbol{s}^{[t]}$, so 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. The starting values of the model parameters in $\boldsymbol{\theta}^{[0]}$ are inspired by the values used by established software for latent variable analyses, such as the `R` package `lavaan` (Rosseel, 2012) and the commercial software `Mplus` (L. Muthén & Muthén, 2020). Specifically, the starting values of the factor loadings are computed through instrumental variables methods (Hägglund, 1982), the factor variances and covariances are initialized at 0.05 and zero, respectively, whereas the unique variances at half the variances of the observed variables in the data set. These initial values can be replaced with informative user-defined values (see Chapter 7 for additional details). The solution resulting from the optimization process undergoes admissibility checks. 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).

It should be noticed that the trust-region radius is increased only if $||\boldsymbol{s}^{[t]}||_2$ reaches the boundary of the region. If the step stays strictly inside the region, we can conclude that the current $\Delta^{[t]}$ is not interfering with the progress of the algorithm, so its value is left unchanged for the following iteration. The trust-region algorithm is implemented in the `R` package `trust`.

Algorithm 1 Trust-region algorithm

Require: $\Delta_{\max} > 0, \Delta_0 \in (0, \Delta_{\max}), \boldsymbol{\theta}^{[0]}$

Compute $\ell_p(\boldsymbol{\theta}^{[0]}), \boldsymbol{g}_p(\boldsymbol{\theta}^{[0]}), \boldsymbol{B}(\boldsymbol{\theta}^{[0]})$

Set $\epsilon =$ `.Machine$double.eps`$^{\frac{1}{2}} = 1.490116 \times 10^{-8}$

while $\mathrm{t} \leq 1000$ or $\left| - \left[\ell_p(\boldsymbol{\theta}^{[\mathrm{t}]}) - \ell_p(\boldsymbol{\theta}^{[\mathrm{t}+1]}) \right] \right| < \epsilon$ **do**

  $\boldsymbol{s}^{[\mathrm{t}]} = \arg\min_{\boldsymbol{s}:||\boldsymbol{s}||_2 \leq \Delta^{[\mathrm{t}]}} \mathcal{Q}_p^{[\mathrm{t}]}(\boldsymbol{s})$

  $r^{[\mathrm{t}]} = \dfrac{- \left[\ell_p(\boldsymbol{\theta}^{[\mathrm{t}]}) - \ell_p(\boldsymbol{\theta}^{[\mathrm{t}]} + \boldsymbol{s}^{[\mathrm{t}]}) \right]}{\mathcal{Q}_p^{[\mathrm{t}]}(\mathbf{0}) - \mathcal{Q}_p^{[\mathrm{t}]}(\boldsymbol{s}^{[\mathrm{t}]})}$

  if $r^{[\mathrm{t}]} < \frac{1}{4}$ **then**

    $\boldsymbol{\theta}^{[\mathrm{t}+1]} = \boldsymbol{\theta}^{[\mathrm{t}]}$

    $\Delta^{[\mathrm{t}+1]} = \dfrac{||\boldsymbol{s}^{[\mathrm{t}]}||_2}{4}$

  else

    $\boldsymbol{\theta}^{[\mathrm{t}+1]} = \boldsymbol{\theta}^{[\mathrm{t}]} + \boldsymbol{s}^{[\mathrm{t}]}$

    if $r^{[\mathrm{t}]} > \frac{3}{4}$ and $||\boldsymbol{s}^{[\mathrm{t}]}||_2 = \Delta^{[\mathrm{t}]}$ **then**

      $\Delta^{[\mathrm{t}+1]} = \min(2\Delta^{[\mathrm{t}]}, \Delta_{\max})$

    else

      $\Delta^{[\mathrm{t}+1]} = \Delta^{[\mathrm{t}]}$

    end if

  end if

end while

3.1.1 Comparison to line search methods

Line search algorithms choose a direction $\boldsymbol{s}^{[t]}$ and then search along this direction for a new iterate with a value of the objective function lower than the one at the previous iteration. The distance to move along $\boldsymbol{s}^{[t]}$ is determined by $\rho^{[t]}$, a positive scalar referred to as step length. The line search algorithm solves the problem

$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} + \rho^{[t]}\boldsymbol{s}^{[t]}. \tag{3.7}$$

Typically, the search direction is a descent direction (like steepest-descent or Newton's direction), whereas the step length is chosen through an inexact line search that identifies the value among a sequence of candidate trials achieving adequate reductions in $-\ell_p$ at a minimal cost.

Line search and trust-region methods differ in the order in which they choose the direction and distance of the move to the next iterate. Line search methods first fix the direction $\boldsymbol{s}^{[t]}$ and then identify an appropriate distance (the step length $\rho^{[t]}$). In trust-region methods, a maximum distance (the radius $\Delta^{[t]}$) is first chosen and then a direction and step that attain the best improvement subject to this distance constraint.

If the objective function is non-convex, line search algorithms may search far away from $\boldsymbol{\theta}^{[t]}$, but still choose $\boldsymbol{\theta}^{[t+1]}$ to be close to $\boldsymbol{\theta}^{[t]}$. In some cases, the function can be evaluated so far away from $\boldsymbol{\theta}^{[t]}$ that it is not finite and the algorithm fails. On the contrary, trust-region methods never run too far from the current iteration as the points outside the trust region are not considered. Trust-region algorithms 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). A detailed exposition of trust-region and line search techniques can be found in Nocedal and Wright (2006, Ch. 3–4).

A crucial aspect of penalized models lies in the selection of the tuning parameter, which controls the amount of sparsity introduced in the model. The next sections propose two approaches for the selection of the tuning parameter of the penalized model.

3.2 Generalized Information Criterion

To select η, we elect to use the Generalized Information Criterion (GIC; Konishi & Kitagawa, 1996), which is an extension of the Akaike Information Criterion (AIC; Akaike, 1974) to the case where the estimation is not conducted through ordinary maximum likelihood and 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.

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 $\boldsymbol{\mathcal{X}}_N = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)^T$. Assume that the distribution that generated the data is included in the class of parametric models $\{f(\boldsymbol{x}|\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m\}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$ is the m-dimensional vector of unknown parameters and $\boldsymbol{\Theta}$ an open subset of $\mathbb{R}^m$. A statistical model $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ is then obtained by replacing the parameter vector $\boldsymbol{\theta}$ with the PMLE $\hat{\boldsymbol{\theta}}$. 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

$$\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{T}(G)) = \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\theta} \right\} \Bigg|_{\boldsymbol{\theta} = \boldsymbol{T}(G)}.$$

The GIC evaluating the goodness of fit of the model, when used to predict independent future data $\boldsymbol{z}$ generated from the unknown distribution G, is

$$GIC(\boldsymbol{\mathcal{X}}_N; \hat{G}) = -2 \sum_{\alpha=1}^{N} \log f(\boldsymbol{X}_\alpha|\hat{\boldsymbol{\theta}}) + 2N\, b(\hat{G}), \tag{3.8}$$

where $\hat{G}$ is the empirical distribution function based on the data, and $b(\hat{G})$ the bias estimate arising from using the data twice for estimating the model and the evaluation measure of the goodness of the estimated model (details in Appendix C). Konishi and Kitagawa (1996) showed that the asymptotic bias of the log-likelihood can be represented as the integral of the product of the influence function of the employed estimator and the score function of the probability model, i.e.,

$$b(G) = \frac{1}{N} b_1(G) + o\left(\frac{1}{N}\right),$$

where

$$b_1(G) = \operatorname{tr}\left\{ \int \boldsymbol{T}^{(1)}(\boldsymbol{z}; G) \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \right\}. \tag{3.9}$$

The quantity $\boldsymbol{T}^{(1)}(\boldsymbol{z}; G)$ is the influence function of the functional $\boldsymbol{T}(G)$ at the true distribution G, and describes the effect of an infinitesimal contamination at $\boldsymbol{z}$. The influence function that defines the PMLE is given by (see Appendix C)

$$\boldsymbol{T}^{(1)}(\boldsymbol{z}; G) = \boldsymbol{R}(\boldsymbol{\psi}, G)^{-1} \boldsymbol{\psi}(\boldsymbol{z}; \boldsymbol{T}(G)), \tag{3.10}$$

where $\boldsymbol{R}(\boldsymbol{\psi}, G)$ is an $m \times m$ matrix defined as

$$\begin{aligned}
\boldsymbol{R}(\boldsymbol{\psi}, G) &= -\int \frac{\partial \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\
&= -\int \frac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\
&\qquad + \int \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \left(\frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\theta} \right) \bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}).
\end{aligned}$$

If we denote $\boldsymbol{\theta} = (\boldsymbol{\theta}^\star, \check{\boldsymbol{\theta}})^T$, where $\boldsymbol{\theta}^\star$ collects the penalized parameters and $\check{\boldsymbol{\theta}}$ the unpenalized parameters, we have that

$$\frac{\partial \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \dfrac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\star \partial \boldsymbol{\theta}^{\star T}} - \boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) & \dfrac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\star \partial \check{\boldsymbol{\theta}}^T} \\ \dfrac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \check{\boldsymbol{\theta}} \partial \boldsymbol{\theta}^{\star T}} & \dfrac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \check{\boldsymbol{\theta}} \partial \check{\boldsymbol{\theta}}^T} \end{bmatrix},$$

where $\boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ is the sub-matrix of $\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ corresponding to the penalized parameters defined in Section 2.3. By substituting the expression of the influence function of the PMLE into equation (3.9), we get the following expression of the bias

$$\begin{aligned}
b_1(G) &= \operatorname{tr}\left\{ \boldsymbol{R}(\boldsymbol{\psi}, G)^{-1} \int \boldsymbol{\psi}(\boldsymbol{z}; \boldsymbol{T}(G)) \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \right\} \\
&= \operatorname{tr}\left\{ \boldsymbol{R}(\boldsymbol{\psi}, G)^{-1} \boldsymbol{Q}(\boldsymbol{\psi}, G) \right\},
\end{aligned}$$

where $\boldsymbol{Q}(\boldsymbol{\psi}, G)$ is an $m \times m$ matrix defined as

$$\begin{aligned}
\boldsymbol{Q}(\boldsymbol{\psi}, G) &= \int \boldsymbol{\psi}(\boldsymbol{z}; \boldsymbol{T}(G)) \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(\boldsymbol{z}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right\}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \\
&\qquad \times \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\
&= \int \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\
&= -\int \frac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) = \boldsymbol{Q}(G).
\end{aligned}$$

Let $b_1(\hat{G})$ be a bias estimate obtained by replacing the unknown distribution G with the empirical distribution $\hat{G}$ based on the data:

$$b_1(\hat{G}) = \operatorname{tr}\left\{ \boldsymbol{R}(\boldsymbol{\psi}, \hat{G})^{-1} \boldsymbol{Q}(\hat{G}) \right\}, \tag{3.11}$$

where

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

$$\boldsymbol{Q}(\hat{G}) = -\frac{1}{N}\sum_{\alpha=1}^{N} \frac{\partial^2 \log f(\boldsymbol{x}_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(\hat{G})} = -\frac{1}{N}\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(\hat{G})} = -\frac{1}{N}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}).$$

The estimated bias $b_1(\hat{G})$ is an estimate of the effective number or estimated degrees of freedom (*edf*) of the penalized model, that is,

$$edf = b_1(\hat{G}) = \operatorname{tr}\left\{ \left[-\frac{1}{N}\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}}) \right]^{-1} \left[-\frac{1}{N}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) \right] \right\} = \operatorname{tr}\left\{ \boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) \right\}. \tag{3.12}$$

By substituting the asymptotic bias estimate in equation (3.12) into the expression (3.8) of the GIC, one obtains:

$$\begin{aligned} GIC(\boldsymbol{\mathcal{X}}_N; \hat{G}) &= -2N \left\{ \frac{1}{N} \sum_{\alpha=1}^{N} \log f(\boldsymbol{x}_\alpha | \hat{\boldsymbol{\theta}}) - \frac{1}{N} b_1(\hat{G}) \right\} \\ &= -2 \sum_{\alpha=1}^{N} \log f(\boldsymbol{x}_\alpha | \hat{\boldsymbol{\theta}}) + 2 \operatorname{tr}\{ \boldsymbol{R}(\boldsymbol{\psi}, \hat{G})^{-1} \boldsymbol{Q}(\hat{G}) \} \\ &= -2\, \ell(\hat{\boldsymbol{\theta}}) + 2 \operatorname{tr} \left\{ \boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) \right\}. \end{aligned} \tag{3.13}$$

The GIC is an extension of the AIC, 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) and obtain the following Generalized Bayesian Information Criterion (GBIC):

$$\begin{aligned} GBIC(\boldsymbol{\mathcal{X}}_N; \hat{G}) &= -2 \sum_{\alpha=1}^{N} \log f(\boldsymbol{x}_\alpha | \hat{\boldsymbol{\theta}}) + \log(N) \operatorname{tr}\{ \boldsymbol{R}(\boldsymbol{\psi}, \hat{G})^{-1} \boldsymbol{Q}(\hat{G}) \} \\ &= -2\, \ell(\hat{\boldsymbol{\theta}}) + \log(N) \operatorname{tr} \left\{ \boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) \right\}. \end{aligned} \tag{3.14}$$

The tuning parameter η enters through the penalty matrix, which is included in $\boldsymbol{\mathcal{H}}_p$. The determination of the tuning parameter can be viewed as a model selection and evaluation problem. Therefore, information criteria evaluating a penalized model can be used as tuning parameter selectors. By evaluating statistical models determined according to a grid of values of η, we take the optimal value of the tuning parameter $\hat{\eta}$ to be the one minimizing the value of the GBIC (since the BIC generally selects more sparse models than does the AIC), that is,

$$\hat{\eta} = \arg\min_{\eta} GBIC(\boldsymbol{\mathcal{X}}_N; \hat{G}).$$

The optimal penalized factor model is hence chosen to be the one with the lowest BIC, which 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, and hence expression (3.13) 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.

Degrees of freedom

The *edf* of an unpenalized model ($\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}} = \boldsymbol{O}_{m\times m}$) coincide with the dimension of the parameter vector $\boldsymbol{\theta}$, since $\text{tr}\left\{\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\} = \text{tr}\left\{\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\} = \text{tr}(\boldsymbol{I}_m) = m$, where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. For a penalized model $edf = \text{tr}\left\{\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\} = m - \text{tr}\left\{[-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\hat{\boldsymbol{\theta}})]^{-1}N\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right\}$. This shows that $edf \to m$ as $\eta \to 0$, and $edf \to m - q^\star$ as $\eta \to \infty$, where $q^\star$ is the number of penalized elements. When $0 < \eta < \infty$, the $edf \in [m - q^\star, m]$. 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.

Non-zero parameters

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 non-zero parameters (referred in the following as *dof*), by advocating the fact that the number of non-zero 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. Importantly, the definition of *edf* directly stems from the estimated bias term of the GIC, which gives it a theoretically founded basis.

3.3 Automatic tuning parameter selection

An alternative proposal to using a grid-search combined with GBIC is to estimate η 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 (2019a, 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 $\boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}+1]}) = \mathbf{0}$. By using a first-order Taylor expansion of $\boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}+1]})$ at $\boldsymbol{\theta}^{[\mathrm{t}]}$ it follows that

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

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

$$\boldsymbol{\theta}^{[t+1]} = \left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]}) + N\boldsymbol{\mathcal{S}}_{\hat{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}^{[\mathrm{t}]})\right]^{-1} \sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]})}\boldsymbol{K}^{[\mathrm{t}]}, \tag{3.15}$$

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 Appendix D.2. From standard likelihood theory, we have that $\boldsymbol{\vartheta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ 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

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

where $\boldsymbol{A}_{\hat{\eta}}^{\mathcal{T}} = \sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}\left[\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\hat{\eta}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right]^{-1}\sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}$ is the influence (or hat) matrix of the fitting problem and depends on the tuning parameter. The quantity

$\hat{\boldsymbol{\theta}} = \left[\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\hat{\eta}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right]^{-1}\sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}\boldsymbol{K}$ denotes the PMLE. Ideally, the estimation of the tuning parameter 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 $\boldsymbol{\mu}_{\boldsymbol{K}}$ (Appendix D.3):

$$\mathbb{E}\left[\frac{1}{N}||\boldsymbol{\mu}_{\boldsymbol{K}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}||_2^2\right] = \frac{1}{N}\mathbb{E}\left[||\boldsymbol{K} - \boldsymbol{A}_{\eta}^{\mathcal{T}}\boldsymbol{K}||_2^2\right] + \frac{2}{N}\text{tr}(\boldsymbol{A}_{\eta}^{\mathcal{T}}) - 1, \tag{3.16}$$

where $||\cdot||_2^2$ is the squared Euclidean norm. The quantity

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

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 (3.16) depends on the tuning parameter through $\boldsymbol{A}_{\eta}^{\mathcal{T}}$, whereas $\boldsymbol{K}$ is linked to the unpenalized part of the model. The tuning parameter is estimated by minimizing an estimate of (3.16):

$$\mathcal{V}(\eta) = \frac{1}{N}||\widehat{\boldsymbol{\mu}_{\boldsymbol{K}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}}||_2^2 = \frac{1}{N}||\boldsymbol{K} - \boldsymbol{A}_{\eta}^{\mathcal{T}}\boldsymbol{K}||_2^2 + \frac{2}{N}\text{tr}(\boldsymbol{A}_{\eta}^{\mathcal{T}}) - 1. \tag{3.17}$$

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

$$\begin{aligned}\eta^{[\text{t}+1]} &= \arg\min_{\eta}\mathcal{V}^{[\text{t}+1]}(\eta) \\ &= \arg\min_{\eta}\left\{\frac{1}{N}||\boldsymbol{K}^{[\text{t}+1]} - \boldsymbol{A}_{\eta}^{\mathcal{T}^{[\text{t}+1]}}\boldsymbol{K}^{[\text{t}+1]}||_2^2 + \frac{2}{N}\text{tr}(\boldsymbol{A}_{\eta}^{\mathcal{T}^{[\text{t}+1]}}) - 1\right\},\end{aligned} \tag{3.18}$$

and solved by adapting the approach by Wood (2004) to the current context (Appendix D.5). This approach is based on Newton's method and can evaluate in a stable and efficient way $\mathcal{V}(\eta)$ and its derivative with respect to $\log(\eta)$ (since the tuning parameter can only take positive values). The two steps, one for the

estimation of $\boldsymbol{\theta}$ and the other for η, are iterated until the algorithm satisfies the stopping criterion

$$\frac{|\ell(\boldsymbol{\theta}^{[\mathrm{t}+1]}) - \ell(\boldsymbol{\theta}^{[\mathrm{t}]})|}{0.1 + |\ell(\boldsymbol{\theta}^{[\mathrm{t}+1]})|} < 10^{-7}.$$

Influence factor

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

$$\mathcal{V}(\eta) = \frac{1}{N}||\boldsymbol{K} - \boldsymbol{A}_\eta^{\mathcal{T}}\boldsymbol{K}||_2^2 + \frac{2}{N}\gamma \operatorname{tr}(\boldsymbol{A}_\eta^{\mathcal{T}}) - 1. \tag{3.19}$$

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.

It is important to notice that the implementation of the automatic procedure described above relies on the separability of the penalty matrix from the tuning parameter. 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 the simulation experiments and the empirical application (see Chapter 4) the alasso generally represented the most convenient choice of penalty based on a number of criteria.

The presented modeling framework has been implemented in the `R` package `GJRM` (Marra & Radice, 2019b) and we refer the reader to Chapter 7 for a brief description of the software and practical illustrations.

3.4 Theoretical aspects of the PMLE

This section discusses some asymptotic properties of the PMLE. For notational convenience, let $\boldsymbol{\mathcal{S}}_\eta$ be the shorthand for $\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}$, for $\mathcal{T} = \{L, A, S, M\}$, and $\boldsymbol{\theta}_0$ the true parameter vector. The following results were derived under the regularity conditions reported in Appendix E.1.

Theorem 3.1 (Asymptotic distribution of the PMLE (I))**.** *Under certain regularity conditions, the PMLE has the following asymptotic distribution:*

$$\sqrt{N}\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)\left\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 + \boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}N\boldsymbol{\mathcal{S}}_\eta(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right\} \xrightarrow{d} \mathcal{N}(\mathbf{0}, N\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)),$$

and thus the asymptotic bias of $\hat{\boldsymbol{\theta}}$ *is equal to* $-\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}N\boldsymbol{\mathcal{S}}_\eta(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0$*, and the asymptotic covariance* $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}$*, where* $\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0) = \boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_\eta(\boldsymbol{\theta}_0)$.

Proof. See Appendix E.2 ■

Furthermore (see Appendix E.3 for the derivation of these results),

$$\begin{aligned}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 &= \mathcal{O}_P\left(N^{-\frac{1}{2}}\right),\\ \text{Bias}(\hat{\boldsymbol{\theta}}) &= o(N^{-\frac{1}{2}}),\\ \text{Cov}(\hat{\boldsymbol{\theta}}) &= \mathcal{O}(N^{-1}).\end{aligned}$$

The next theorem states that the asymptotic distribution of the PMLE coincides with that of the MLE as the sample size increases, which is desirable, as the MLE is the most efficient estimator.

Theorem 3.2 (Asymptotic distribution of the PMLE (II))**.** *If* $\max|N\boldsymbol{\mathcal{S}}_\eta(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0| = o\left(N^{\frac{3}{2}}\right)$*, and* $\max|N\boldsymbol{\mathcal{S}}_\eta(\boldsymbol{\theta}_0)| = o\left(N^{\frac{3}{2}}\right)$*, then*

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{N}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\right\}^{-1}\right).$$

Proof. See Appendix E.4. ■

Theorem 3.3 (Consistency)**.** *Suppose that $\eta \in [0, \infty)$ is fixed. Then, under the assumption of a convex unpenalized log-likelihood, the PMLE $\hat{\boldsymbol{\theta}}$ that minimizes $-\ell_p(\boldsymbol{\theta})$ is consistent, that is,*

$$\lim_{N\to\infty} \mathbb{P}\left(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2^2 > \bar{\varepsilon}\right) = 0 \quad \forall\ \bar{\varepsilon} > 0.$$

Proof. See Appendix E.5. ■

3.4.1 Intervals

At convergence, the covariance matrix of $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{\mathcal{J}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{J}}(\hat{\boldsymbol{\theta}})\boldsymbol{\mathcal{J}}_p(\hat{\boldsymbol{\theta}})^{-1}$. However, for practical purposes it is more convenient to employ the alternative Bayesian result $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{\mathcal{J}}_p(\hat{\boldsymbol{\theta}})^{-1}$. (For an unpenalized model $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$ and $\boldsymbol{V}_{\boldsymbol{\theta}}$ are equivalent as there is no penalty involved in the covariance matrices.) In fact, at finite sample sizes, $\boldsymbol{V}_{\boldsymbol{\theta}}$ can produce intervals with close to nominal "across-the-function" frequentist coverage probabilities (Marra & Wood, 2012) because the Bayesian covariance matrix includes both a bias and variance component in a frequentist sense, a feature not shared by $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$. This result can be justified using the distribution of $\boldsymbol{K}$ given in Section 3.1, making the large sample assumption that $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$ can be treated as fixed, and making the prior Bayesian assumption of $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, (N\boldsymbol{\mathcal{S}}_\eta(\tilde{\boldsymbol{\theta}}))^{-1})$. 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

$$\boldsymbol{\theta}|\boldsymbol{x}_N, \eta \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \boldsymbol{V}_{\boldsymbol{\theta}}).$$

Additional details are covered in Appendix E.6.

3.4.2 Bayesian interpretation

Introducing penalties in the estimation process is fundamentally motivated by the belief that in the population, the factor structures are more likely to be sparse than dense. This prior belief can be formalized by specifying the exponential prior

$\exp\left\{-\frac{N}{2}\boldsymbol{\theta}^T\boldsymbol{\mathcal{S}}_\eta(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right\}$ on the penalty function. This is equivalent to assuming for the parameter vector a zero-mean improper Gaussian prior distribution with precision matrix proportional to $\boldsymbol{\mathcal{S}}_\eta(\tilde{\boldsymbol{\theta}})$, i.e., $\boldsymbol{\theta} \propto \mathcal{N}(\mathbf{0}, (N\boldsymbol{\mathcal{S}}_\eta(\tilde{\boldsymbol{\theta}}))^{-1})$, where $\boldsymbol{\mathcal{S}}_\eta(\tilde{\boldsymbol{\theta}})^{-1}$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{\mathcal{S}}_\eta(\tilde{\boldsymbol{\theta}})$ (Wood, 2017). The proposed penalized approach can thus be viewed as an "empirical Bayes" method that gives good frequentist properties.

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; B. 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.

CHAPTER 4

Numerical and empirical evaluation of the penalized factor model

This chapter evaluates the validity of the penalized technique proposed in Chapter 3 through numerical and empirical examples. First, we illustrate a simulation study conducted to evaluate the performances of the PMLE and compare them to the ones of competing methods existing in the literature (Section 4.1). We investigate and assess 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. Then, the proposed model and its competitors are tested in a classical psychometric application on students' mental abilities (Section 4.2).

4.1 Simulation Study

An extensive simulation study was conducted to evaluate the performances of the proposed PMLE under a broad range of scenarios. For EFA, several works (Choi et al., 2010; Huang et al., 2017; Hirose & Yamamoto, 2014b; Jin et al., 2018; Scharf & Nestler, 2019) already demonstrated that penalized techniques generally outperform their unpenalized and rotated counterparts and, under certain

conditions, perform similarly to the oracle MLE. For this reason, instead of contrasting our model (implemented in the `R` package `GJRM`) to unpenalized maximum likelihood, we compared it to the penalized maximum likelihood solutions produced by the methods developed by Jacobucci et al. (2016) and by Huang et al. (2017) implemented in the `R` packages `regsem` (version 1.3.2; Jacobucci et al., 2019) and `lslx` (version 0.6.8; Huang & Hu, 2019), respectively. Despite the fact that other techniques to conduct penalized factor analysis exist (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 one to specify which parameters are fixed, which are free and which are penalized, as well to directly estimate the structural model.

We first illustrate the design of the study and then present the results.

4.1.1 Design and procedure

The simulation study was partly inspired by the empirical application (Section 4.2), 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 Section 3.1);

- **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:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0.85 & \underline{0} & \underline{0} \\ 0.75 & 0 & 0 \\ 0.65 & 0.3 & 0 \\ \underline{0} & 0.85 & \underline{0} \\ 0 & 0.75 & 0 \\ 0 & 0.65 & 0.3 \\ \underline{0} & \underline{0} & 0.85 \\ 0 & 0 & 0.75 \\ 0.3 & 0 & 0.65 \end{bmatrix} \qquad \boldsymbol{\Phi} = \begin{bmatrix} \underline{1} & 0.3 & 0.3 \\ & \underline{1} & 0.3 \\ & & \underline{1} \end{bmatrix}$$

and $\boldsymbol{\Psi} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix, which implies that the observed variables have been standardized. 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 non-zero 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` (version 3.5.1; R Core Team, 2018) according to the population parameters. Each data set was column-wise centered since the normal linear factor analysis model illustrated in Section 2.1 implicitly assumes that the observed variables have zero-means. The resulting data matrix was then analyzed in `GJRM`, `regsem` and `lslx` 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 matched the one of the corresponding population parameters. Based on the availability of the respective software implementations, 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 solutions (see Section 3.1 for the definition of admissibility).

4.1.2 Results

For the sake of clarity, we report in the following a selection of the most relevant results for particular configurations of alasso, scad and mcp, leaving the lasso in Section 4.1.3. Specifically, for `GJRM-alasso` $a = 2$, for `GJRM-scad` and `GJRM-mcp`

$a = 3$, whereas for the automatic procedure $\gamma = 4.5$. These configurations were found to produce the best models in terms of a number of different performance criteria (details are given below). 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). Due to its generally higher numerical stability in comparison to the Hessian, only `GJRM` models estimated with the Fisher information matrix are presented in the following. We evaluated the performance of the methods according to the criteria illustrated in Huang et al. (2017), which are briefly mentioned here. The overall accuracy of each estimator was assessed using the estimated mean squared error (MSE):

$$\widehat{\text{MSE}}(\hat{\boldsymbol{\theta}}) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\boldsymbol{\theta}}^{(l)} - \boldsymbol{\theta}_0)^T(\hat{\boldsymbol{\theta}}^{(l)} - \boldsymbol{\theta}_0), \tag{4.1}$$

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):

$$\widehat{\text{SB}}(\hat{\boldsymbol{\theta}}) = (\bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0)^T(\bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0), \tag{4.2}$$

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 $\mathcal{F} = \{q \,|\, \theta_{0q} \neq 0 \,\&\, \hat{\theta}_q \text{ penalized}\}$ indicate the set of indices associated to the true non-zero parameters that have been penalized (i.e., the penalized non-zero 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 non-zero parameters was evaluated via the estimated true positive rate (TPR):

$$\widehat{\text{TPR}}(\hat{\boldsymbol{\theta}}) = \frac{1}{L}\sum_{l=1}^{L}\frac{\sum_{q\in\mathcal{F}}\mathbb{1}\left(\hat{\theta}_q^{(l)} \neq 0\right)}{|\mathcal{F}|}. \tag{4.3}$$

Denote as $\mathcal{F}^c = \{q \,|\, \theta_{0q} = 0 \,\&\, \hat{\theta}_q \text{ penalized}\}$ 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 non-zero:

$$\widehat{\text{FPR}}(\hat{\boldsymbol{\theta}}) = \frac{1}{L}\sum_{l=1}^{L}\frac{\sum_{q\in\mathcal{F}^c}\mathbb{1}\left(\hat{\theta}_q^{(l)}\neq 0\right)}{|\mathcal{F}^c|}. \tag{4.4}$$

Lastly, selection consistency was assessed via the proportion of times the true model - for which all the true zero and non-zero 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):

$$\widehat{\text{PCTM}}(\hat{\boldsymbol{\theta}}) = \frac{1}{L}\sum_{l=1}^{L}\frac{\sum_{q\in\mathcal{F}}\mathbb{1}\left(\hat{\theta}_q^{(l)}\neq 0\right)+\sum_{q\in\mathcal{F}^c}\mathbb{1}\left(\hat{\theta}_q^{(l)}= 0\right)}{|\mathcal{F}|+|\mathcal{F}^c|}, \tag{4.5}$$

where $|\mathcal{F}|+|\mathcal{F}^c| = q^\star$. For the computation of PCTM and FPR, the parameter estimates were rounded to one decimal digit for all models.

By looking at the results in Table 4.1, we draw the following conclusions:

1. 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.

2. The MSE of all methods are very similar to each other and improve as the sample size increased.

3. The results with the lower bias were associated with the use of non-convex penalties, although the bias of `GJRM-alasso` very quickly converged to zero when the sample size increased, and hence the impact of the penalty decreased.

4. The true positive rates were always equal to 1.0, which showed that the inspected methods never suppressed the non-zero penalized parameters (i.e., the primary loadings and the cross-loadings).

5. In terms of both false positive rates and selection consistency, `GJRM-alasso` with automatic tuning parameter selection presented by far the best performances for all the sample sizes.

6. The mean squared error and bias of `GJRM-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 may indicate that the presence of a sparsity-inducing quantity (influence factor) in the optimization criterion helped the model obtain a nicer tradeoff between goodness of fit and model complexity.

7. By comparing the quality measures of the three methods for the same penalty function (i.e., the mcp), we notice that `GJRM` outperformed `lslx` and was generally close to `regsem` for MSE and SB and superior for FPR and PCTM.

8. 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 `GJRM-alasso` model for extreme values of the influence factor (i.e., $\gamma = 4.5$ in Table 4.1 and $\gamma = 1$ in Table 4.4 in Section 4.1.3). 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.

9. 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 $a = 2$, which is why it is the value of all `GJRM-alasso` models reported in Tables 4.1 and 4.4.

Computational efficiency

The investigated methods were compared in terms of their computational efficiency. 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 4.2 reports the minimum, median and standard error of the elapsed time for estimating one penalized factor model under every sample size scenario. The distributions of the elapsed times are visualized through violin plots under every sample size scenario in Figure 4.1. As

	GJRM				lslx	regsem
	ALASSO		**SCAD**	**MCP**	**MCP**	**MCP**
	grid	auto	grid	grid	grid	grid
MSE						
$N = 300$	0.073	0.075	0.074	0.074	0.075	0.071
$N = 500$	0.041	0.041	0.042	0.042	0.042	0.041
$N = 1000$	0.020	0.020	0.020	0.020	0.020	0.020
SB						
$N = 300$	0.003	0.004	0.002	0.002	0.003	0.000
$N = 500$	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
FPR						
$N = 300$	0.022	0.008	0.016	0.019	0.036	0.018
$N = 500$	0.012	0.004	0.007	0.008	0.016	0.012
$N = 1000$	0.003	0.001	0.002	0.002	0.004	0.009
PCTM						
$N = 300$	0.820	0.932	0.871	0.843	0.743	0.848
$N = 500$	0.898	0.962	0.936	0.925	0.877	0.897
$N = 1000$	0.974	0.991	0.982	0.979	0.966	0.923

Note: The values of the additional tuning parameters are $a = 2$ for `GJRM-alasso`, $\gamma = 4.5$ for the automatic procedure, $a = 3$ for `GJRM-scad` and `GJRM-mcp`, and $a = 3.7$ for `regsem-mcp` as per default software implementations. For `lslx-mcp` the values of both a and η were determined on the basis of grid-searches.

Table 4.1: Performance measures of the examined models by varying the sample size. MSE stands for mean-squared error, SB for squared bias, FPR for false positive rate and PCTM for proportion choosing the true model.

the number of observations increased, the computational times shortened because the penalized models converged faster. Specifically, the models fitted through the automatic tuning parameter procedure exhibited the lowest computational times, with an average of nearly 0.3 seconds per model, as well as the least variability. The `GJRM` models with grid-search presented comparable computational times of about 20 seconds per replicate, which is nearly half of the time it took `regsem` to fit one model. The computational times of `lslx` are noticeably inferior to those of the other grid-search techniques. This is a consequence of its underlying optimizer being implemented in C++, which significantly boosted the computations with respect to base `R` routines.

Coverage probabilities

We computed 95% coverage probabilities for the parameters of all fitted models using point-wise confidence intervals (Table 4.3). For clarity of presentation, we only report the inferential results of the models considered in Table 4.1. The standard errors for `GJRM` are based on the Bayesian result illustrated in Section 3.4.1. On the contrary, for `lslx`, they are computed using the frequentist expression of the covariance matrix based on the Fisher information. No coverage probabilities could be computed for `regsem` as the package does not currently provide any measure of uncertainty.

Because of the rationale discussed in Section 3.1, `GJRM` provides a standard error for every single model parameter, contrarily to `lslx` which does not provide this information for the parameters shrunken to zero. However, since the main intent of penalization is to get rid of the uninfluential elements, the inferential results are presented for the parameters remaining in the model, which are the effective quantities of interest. The coverage probabilities were furtherly split and averaged between those corresponding to the penalized parameters (i.e., the non-zero factor loadings) and the freely estimated ones (i.e., the factor covariances and unique variances).

Overall, the values of both `GJRM` and `lslx` are close to their true nominal level,

Elapsed time (seconds)	GJRM				lslx	regsem
	ALASSO		SCAD	MCP	MCP	MCP
	grid	auto	grid	grid	grid	grid
N = 300						
Minimum	10.70	0.20	15.07	11.81	3.38	19.77
Median	18.55	0.45	22.73	21.33	6.50	43.58
Standard error	2.81	0.31	5.30	6.63	3.86	6.17
N = 500						
Minimum	12.15	0.12	13.43	12.51	3.31	18.36
Median	17.19	0.34	20.93	21.01	7.46	41.98
Standard error	2.65	0.34	3.58	4.28	3.79	6.59
N = 1000						
Minimum	9.56	0.10	13.88	11.07	3.25	15.88
Median	15.29	0.23	19.59	20.22	5.90	41.04
Standard error	2.19	0.40	2.45	1.82	2.78	6.57

Table 4.2: Minimum, median and standard error of the elapsed time (seconds) for `GJRM-alasso` with grid (1-dim. grid for η; $a = 2$) and automatic procedure ($a = 2; \gamma = 4.5$), `GJRM-scad` (1-dim. grid for η; $a = 3$), `GJRM-mcp` (1-dim. grid for η; $a = 3$), `lslx-mcp` (2-dim. grid for η and a) and `regsem-mcp` (1-dim grid for η, $a = 3.7$ as per default software implementations) under each sample size scenario.

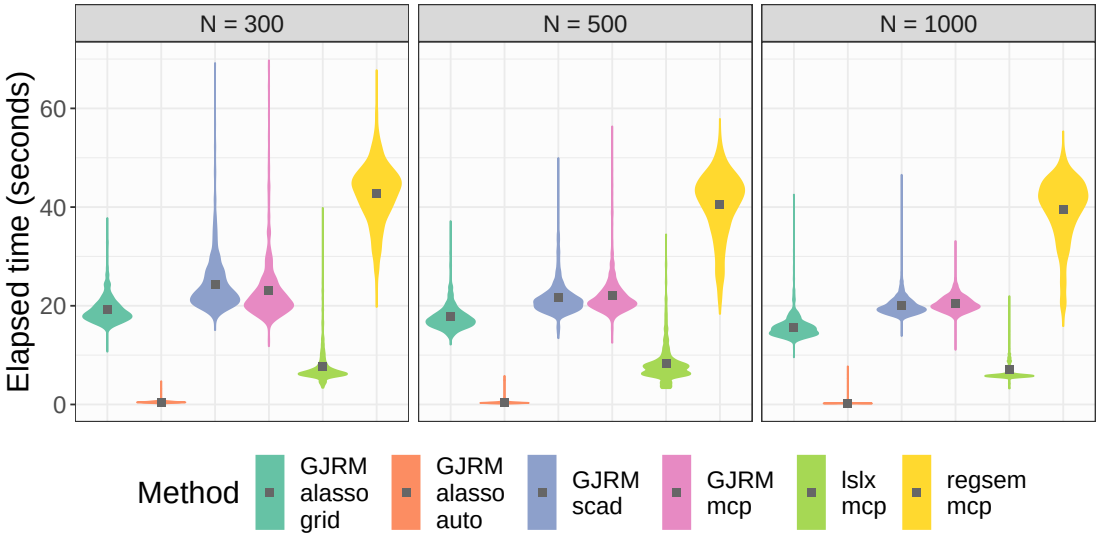


Figure 4.1: Distributions of the elapsed times of the investigated methods under each sample size scenario. The grey squares indicate the average times.

Sample size	GJRM								lslx	
	ALASSO				**SCAD**		**MCP**		**MCP**	
	grid		auto		grid		grid		grid	
	Pen.	Free	Pen.	Free	Pen.	Free	Pen.	Free	Pen.	Free
$N = 300$	0.922	0.942	0.900	0.942	0.916	0.942	0.918	0.942	0.924	0.942
$N = 500$	0.934	0.946	0.931	0.946	0.929	0.945	0.928	0.945	0.938	0.945
$N = 1000$	0.940	0.946	0.940	0.946	0.941	0.945	0.940	0.945	0.945	0.946

Note: Pen. indicates the penalized non-zero parameters and *free* the freely estimated parameters.

Table 4.3: Average coverage probabilities of the examined models by sample size and parameter type. For `GJRM-alasso` with grid $a = 2$, with the automatic procedure $a = 2$ and $\gamma = 4.5$, for `GJRM-scad` $a = 3$ and for `GJRM-mcp` $a = 3$.

the more so as the sample size increases, for all penalty functions, which proves that the selected models are also valid from an inferential point of view.

4.1.3 Additional models

In this section we report the performance measures of the `GJRM-alasso` model with the value of the influence factor $\gamma = 1$ (Table 4.4). As discussed in Section 4.1.2, the influence factor plays a decisive role in the final model fitting results. Specifically, the model with the larger γ (Table 4.1) resulted in visibly higher PCTM and lower FPR, at the expense of a slight increase in bias. This loss in bias, however, became negligible or nonexistent as the sample size grew. In this respect, it is interesting to look at the MSE, which encloses both the variance and the squared bias components of an estimator. Despite the model with $\gamma = 1$ always having a smaller bias, the one with $\gamma = 4.5$ produced such a decrease in the variability of the parameter estimates that its MSE ended up being always smaller than the one obtained with the inferior value of the influence factor. The TPR were equal to 1.0 for every sample size.

We complete the discussion of the simulation results by showing the performances of `GJRM-lasso` models when the tuning parameter was selected by grid-search or estimated with the automatic procedure (Table 4.4). The two models gave overall similar results, with the former having better FPR and PCTM and the

	ALASSO	**LASSO**	
	auto	grid	auto
	$\gamma = 1$		$\gamma = 4.5$
MSE			
$N = 300$	0.083	0.109	0.102
$N = 500$	0.049	0.066	0.061
$N = 1000$	0.024	0.034	0.031
SB			
$N = 300$	0.001	0.039	0.030
$N = 500$	0.000	0.024	0.017
$N = 1000$	0.000	0.013	0.008
FPR			
$N = 300$	0.154	0.094	0.113
$N = 500$	0.114	0.060	0.074
$N = 1000$	0.049	0.017	0.026
PCTM			
$N = 300$	0.256	0.409	0.321
$N = 500$	0.374	0.583	0.493
$N = 1000$	0.634	0.860	0.795

Table 4.4: Performance measures of `GJRM-alasso` and `GJRM-lasso` by sample size. The quantity γ denotes the influence factor. MSE stands for mean-squared error, SB for squared bias, FPR for false positive rate and PCTM for proportion choosing the true model.

latter lower MSE and bias. The TPR were equal to 1.0 in both cases and for every sample size. These results, however, are visibly less performing than the models where the alasso, scad and mcp were used. As a matter of fact, it is well known that the lasso tends to select an overfitted model, because it equally penalizes all model parameters. Therefore, we suggest opting for the other penalties, which have been specifically designed to improve the lasso.

4.2 Empirical application

The Holzinger & Swineford data set (Holzinger & Swineford, 1939) 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). Following Jacobucci et al. (2016) and Huang et al. (2017), we use a subset of $p = 9$ mental tests: visual perception (VISUAL), cubes (CUBES), flags (FLAGS), paragraph comprehension (PARAGRAP), sentence completion (SENTENCE), word meaning (WORDM), addition (ADDITION), counting groups of dots (COUNTING), straight and curved capitals (STRAIGHT). These tests are thought of as measuring $r = 3$ correlated abilities: spatial ability (VISUAL, CUBES, FLAGS), verbal intelligence (PARAGRAP, SENTENCE, WORDM), and speed (ADDITION, COUNTING, STRAIGHT).

The range of values of each variable is reported in the second and third column of Table 4.5. The data set was column-wise centered since the factor model in equation 2.1 implicitly assumes that the observed variables have zero-means. To mitigate the scaling effect, the data set was scaled as described in Yuan and Bentler (2006) to keep the marginal standard deviation of each variable between 1 and 2. After the centering and scaling, the ranges of the variables were as reported in the last two columns of Table 4.5.

The heat map of the covariance matrix of the scaled data set is presented in Figure 4.2; small, moderate and high covariances are represented in light blue, yellow and red, respectively. Besides the evident relationships of the tests designed to measure the same mental ability, there seem to be some connections between tests relative to distinct latent constructs. This may suggest that not all of the tests are pure measures, that is, they do not load only on the ability they were designed to measure. As a matter of fact, the CFA model assuming this simple structure (see the path diagram in Figure 4.3) presents a poor fit to the data (p-value of the

Observed variables	Original ranges		After centering and scaling	
	Minimum	Maximum	Minimum	Maximum
VISUAL	4	51	−4.27	3.56
CUBES	9	37	−3.84	3.16
FLAGS	2	36	−2.00	2.25
PARAGRAP	0	19	−3.06	3.27
SENTENCE	4	28	−3.34	2.66
WORDM	1	43	−2.04	3.96
ADDITION	30	171	−2.88	3.25
COUNTING	61	200	−2.48	4.47
STRAIGHT	100	333	−2.60	3.88

Table 4.5: Ranges of values of the observed variables of the Holzinger & Swineford data set before and after centering and scaling.

chi-square goodness of fit test < 0.001), which confirms the multi-dimensionality of some of the tests.

In these 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. 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 this thesis.

We fitted a large number of models involving all four penalties. For grid-search,

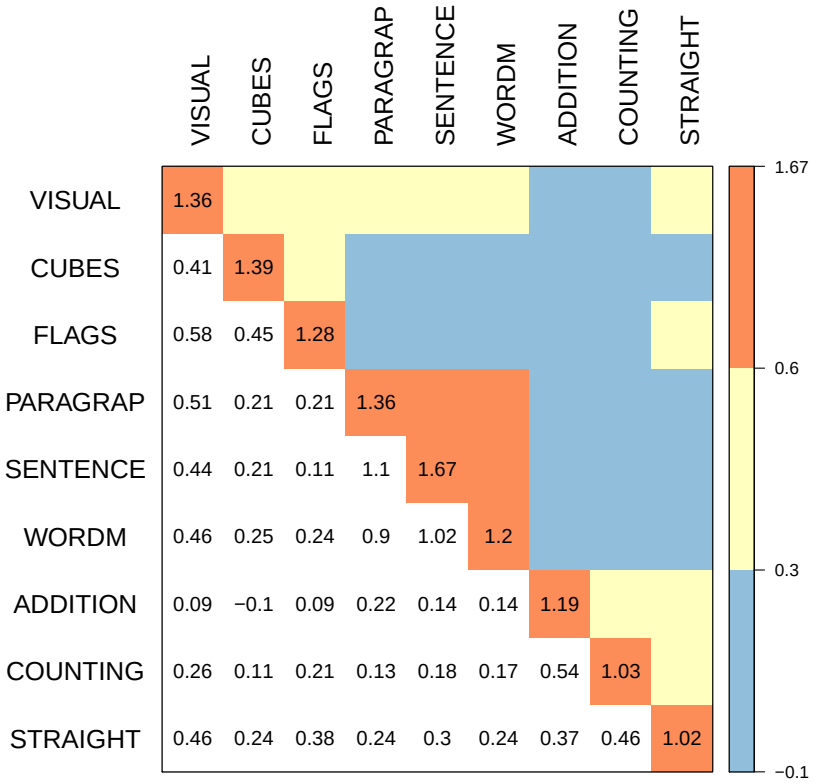


Figure 4.2: Covariance matrix of the Holzinger & Swineford data set.

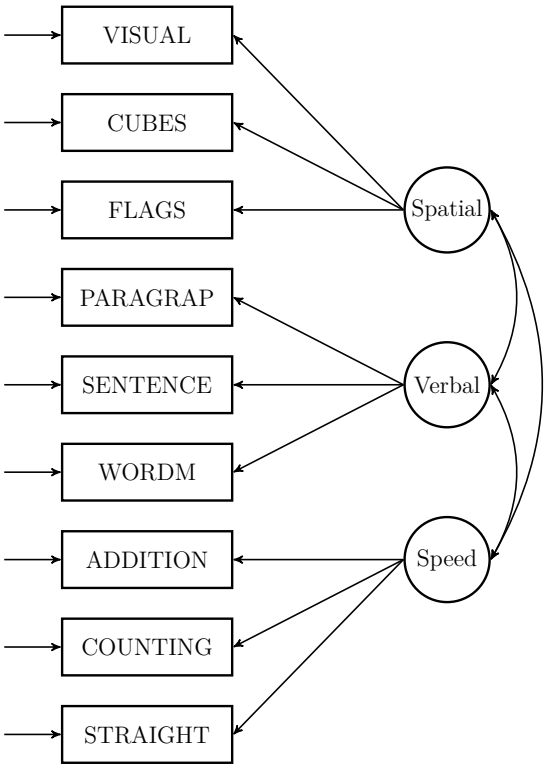


Figure 4.3: Path diagram of the CFA model assuming simple structure.

Method	Penalty	BIC
`GJRM`	ALASSO	7558.03
`GJRM`	MCP	7561.57
`GJRM`	SCAD	7561.68
`GJRM`	LASSO	7562.94
`regsem`	MCP	7565.21
`regsem`	SCAD	7565.21
`lslx`	MCP	7565.92
`regsem`	ALASSO	7571.39
`regsem`	LASSO	7584.91
`lslx`	LASSO	7585.07
CFA		7595.34
Unpenalized		7601.42

Table 4.6: BIC of the fitted models. For `GJRM-alasso` (automatic procedure) $a = 1$ and $\gamma = 4.5$, for `GJRM-scad` $a = 4.5$, for `GJRM-mcp` $a = 1.5$, and for `GJRM-lasso` (automatic procedure) $\gamma = 4.5$. For all `GJRM` models the Fisher information was used.

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 ($a = \{2.5, 3.7, 4.5\}$) and mcp ($a = \{1.5, 2, 2.5, 3, 3.5\}$). An effective way of "forcing" sparser solutions is increasing the value of the influence factor in the automatic procedure. Higher values are associated with sparser solutions, at the cost of a larger bias, which however tends to vanish as the sample size increases. We tested different values of the influence factor ($\gamma = \{1, 1.4, 2, 2.5, 3, 3.5, 4, 4.5\}$) for the automatic procedure. The data analysis was also conducted in `regsem` and `lslx` using the available penalties (i.e., lasso, alasso, scad, and mcp for the former, and lasso and mcp for the latter).

The BIC values were calculated for each of the fitted models and are ranked in Table 4.6 for some of the best instances of model configurations. The proposed method is placed at the top positions overall, showing the potential of the presented procedure. 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 `GJRM-lasso` with grid-search (7567.62) 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 (correspond-

ing to the factor analysis model in equation (2.1) 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 4.7 reports the parameter estimates of the unpenalized model and the best performing models for `GJRM`, `lslx` and `regsem`. A blank cell in the factor loading matrix indicates that the corresponding estimate was zero after one decimal rounding. The unpenalized model presented various cross-loadings, which resulted in a much more complex model. The factor structures of the three penalized models looked similar. Two penalized loadings were identified as non-zero ($\hat{\lambda}_{91}$, $\hat{\lambda}_{32}$) by all methods. Additionally, `GJRM` and `lslx` detected other secondary loadings, which were $\hat{\lambda}_{51}$ and $\hat{\lambda}_{81}$ for the former and $\hat{\lambda}_{51}$ for the latter.

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.

Measurement model	**Unpenalized model**				`GJRM-alasso`				`lslx-mcp`				`regsem-mcp`			
	Spatial	Verbal	Speed	$\boldsymbol{\Psi}$	Spatial	Verbal	Speed	$\boldsymbol{\Psi}$	Spatial	Verbal	Speed	$\boldsymbol{\Psi}$	Spatial	Verbal	Speed	$\boldsymbol{\Psi}$
VISUAL	0.81	*0*	*0*	0.70	0.83	*0*	*0*	0.63	0.85	*0*	*0*	0.62	0.84	*0*	*0*	0.66
CUBES	0.65	−0.12	−0.16	1.03	0.49			1.11	0.52			1.11	0.52			1.11
FLAGS	0.91	−0.33		0.69	0.76	−0.16		0.75	0.80	−0.17		0.73	0.86	−0.26		0.70
PARAGRAP	*0*	0.99	*0*	0.38	*0*	0.96	*0*	0.38	*0*	0.98	*0*	0.38	*0*	0.99	*0*	0.37
SENTENCE	−0.13	1.19		0.40	−0.06	1.11		0.42	−0.12	1.17		0.40		1.11		0.44
WORDM	0.07	0.87		0.37		0.89		0.36		0.91		0.36		0.91		0.36
ADDITION	*0*	*0*	0.77	0.59	*0*	*0*	0.70	0.67	*0*	*0*	0.66	0.74	*0*	*0*	0.66	0.75
COUNTING	0.30	−0.16	0.68	0.48	0.12		0.70	0.44			0.81	0.37			0.81	0.36
STRAIGHT	0.54	−0.14	0.43	0.55	0.41		0.42	0.56	0.37		0.45	0.57	0.38		0.44	0.57
Structural model	Spatial	Verbal	Speed		Spatial	Verbal	Speed		Spatial	Verbal	Speed		Spatial	Verbal	Speed	
Spatial	*1*	0.59	0.17		*1*	0.48	0.20		*1*	0.51	0.31		*1*	0.52	0.31	
Verbal	-	*1*	0.22		-	*1*	0.16		-	*1*	0.21		-	*1*	0.20	
Speed	-	-	*1*		-	-	*1*		-	-	*1*		-	-	*1*	

Table 4.7: Parameter estimates of the nine mental tests from the Holzinger & Swineford data set for the unpenalized model, `GJRM-alasso` (automatic procedure, $\hat{\eta} = 0.017, a = 1$ and $\gamma = 4.5$), `lslx-mcp` ($\hat{\eta} = 0.13, \hat{a} = 3.32$) and `regsem-mcp` ($\hat{\eta} = 1.28, a = 3.7$). Fixed parameters are italic and underlined. A blank cell in the factor loading matrix indicates that the corresponding estimate is zero.

CHAPTER 5

Sparsity and invariance in the multiple-group factor model

This chapter illustrates how the penalized likelihood-based approach described through Chapters 2-3 can be extended to multiple-group analyses, such as cross-national surveys. After an overview of the multiple-group factor analysis model (Section 5.1), we present a penalty that suitably combines sparsity in the loading matrices and invariance in the loadings and intercepts across groups (Section 5.2). This is easily achieved by aggregating multiple penalty terms, each of which is controlled by its own tuning parameter. The obtained penalty function is singular at the origin, so it is locally approximated. An example clarifying the formulation of the employed penalties is provided in Section 5.2.1. The estimation process and the procedure for the selection of the multiple tuning parameters substantially follow the rules delineated for the single-group factor analysis model and are briefly formulated in Section 5.3.

5.1 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

$$\boldsymbol{x}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{f}_g + \boldsymbol{\epsilon}_g \qquad \text{for } g = 1, \ldots, G, \tag{5.1}$$

where the subscript g denotes the group, and $\boldsymbol{\tau}_g$ the intercept terms. It is assumed that $\boldsymbol{f}_g \sim \mathcal{N}(\boldsymbol{\kappa}_g, \boldsymbol{\Phi}_g)$, $\boldsymbol{\epsilon}_g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$, with $\boldsymbol{\Psi}_g$ usually a diagonal matrix, and $\boldsymbol{f}_g$ is uncorrelated with $\boldsymbol{\epsilon}_g$. 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 $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g^T + \boldsymbol{\Psi}_g$.

For multiple-group analyses, one can apply the restrictions needed for scale setting and identification for a single group factor model (see Section 2.1) repeatedly within each of the G groups. The placement of these constraints is usually the same across groups. When the mean structure is present, the origin, as well as the scale of each latent factor, must be fixed. This implies that researchers need to specify in each group at least r constraints on the intercepts or the factor means, in addition to the r^2 constraints required to identify the covariance structure.

Two popular approaches exist to fix the metric of the common factors and the identification restrictions for a multiple-group factor model. The first method is known as "marker-variable" approach and relies on the selection of a representative variable (marker) for each factor in each group. Then, the intercepts of the markers are fixed to zero, the loadings of the markers on the factor they measure to 1.0, and the loadings of the markers on the remaining factors to zero. All of the other parameters are estimated. Because each common factor inherits the mean and the scale of the corresponding marker variable, the interpretation of the latent variable parameters is relative to the chosen marker. The choice of the markers is crucial and should be an accurate one (Millsap, 2001).

An alternative version of this approach proceeds as described, except that the fixed unit elements are placed on the factor variances, instead of appearing on the loading matrix, and the fixed zero elements are on the factor means, and not on the intercepts. The reader is referred to Millsap (2012) and Little et al. (2006) for an exposition of other approaches to metric setting and identification.

Given that no necessary and sufficient condition for global identification is available, except for special cases, researchers should make sure that the model is

locally identified, for instance, by examining whether the information matrix is positive definite or resorting to empirical tests of identification (see Bollen, 1989).

The free parameters of each group are collected in the m_g-dimensional vector $\boldsymbol{\theta}_g = (\text{vec}(\boldsymbol{\Lambda}_g)^T, \boldsymbol{\tau}_g^T, \text{diag}(\boldsymbol{\Psi}_g)^T, \text{vech}(\boldsymbol{\Phi}_g)^T, \boldsymbol{\kappa}_g^T)^T$, 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 common factors r is the same across groups, the fixed elements required for model 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 (see Appendix F.1):

$$\ell(\boldsymbol{\theta}) = -\sum_{g=1}^{G} \frac{N_g}{2} \{\log|\boldsymbol{\Sigma}_g| + \text{tr}(\boldsymbol{W}_g \boldsymbol{\Sigma}_g^{-1}) + p \log(2\pi)\}, \tag{5.2}$$

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. 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.

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.

5.2 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 $g < g'$. It has dimension $m_1\binom{G}{2} \times m$, where the binomial coefficient $\binom{G}{2}$ denotes the total number of pairwise group differences for a given factor loading. In its general form, $\boldsymbol{D}^{\boldsymbol{\Lambda}}_q$ is a matrix with zeros in every position, except the $((s-1)m_1+q, (g-1)m_1+q)$ entries, which contain a 1.0, and the entries $((s-1)m_1+q, (g'-1)m_1+q)$, which contain a -1.0, for $s = 1, \ldots, G$ and $g < g'$ (see Matrix $\boldsymbol{D}^{\boldsymbol{\Lambda}}_q$). For the other parameters (i.e., the intercepts, the unique variances and the structural 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.

$$
\boldsymbol{D}_q^{\boldsymbol{\Lambda}} = \begin{array}{c}
\begin{array}{ccccccccccccccccccccc}
 & & & & & m_1 & & & & & & & 2m_1 & (G-2)m_1 & & & & & (G-1)m_1 & & & & & Gm_1 \\
\end{array} \\
\left[\begin{array}{ccccc:ccccc:ccc:ccccc:ccccc}
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots & & & \vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots \\
0 & \dots & 1 & \dots & 0 & 0 & \dots & -1 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & & & \vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots \\
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\hdashline
\vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\
\vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\
\hdashline
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots & & & \vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots \\
0 & \dots & 1 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & -1 & \dots & 0 \\
\vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & & & \vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots \\
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\hdashline
\vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\
\vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\
\hdashline
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots & & & \vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots \\
0 & \dots & 0 & \dots & 0 & 0 & \dots & 1 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & -1 & \dots & 0 \\
\vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & & & \vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots \\
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\hdashline
\vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\
\vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\
\hdashline
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots & & & \vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots \\
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 1 & \dots & 0 & 0 & \dots & -1 & \dots & 0 \\
\vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & & & \vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots \\
0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0
\end{array}\right]
\end{array}
\begin{array}{l}
\left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} m_1 \\
\left.\begin{array}{c} \\ \\ \end{array}\right\} m_1(G-3) \\
\left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} m_1 \\
\left.\begin{array}{c} \\ \\ \end{array}\right\} m_1(G-3) \\
\left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} m_1 \\
\left.\begin{array}{c} \\ \\ \end{array}\right\} \frac{(G-3)(G-2)}{2} - 1 \\
\left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} m_1
\end{array}
$$

Matrix $\boldsymbol{D}_q^{\boldsymbol{\Lambda}}$: The general structure of the matrix $\boldsymbol{D}_q^{\boldsymbol{\Lambda}}$ computing the pairwise differences of the q^{th} factor loading across G groups.

If $G = 2$, the absolute difference of the q^{th} loading across the two groups is expressed as $||\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\boldsymbol{\theta}||_1 = |\theta_q - \theta_{m_1+q}|$, where $\boldsymbol{D}_q^{\boldsymbol{\Lambda}} = [\boldsymbol{R}_q \ -\boldsymbol{R}_q]$. 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.

The derivation of the expression of the penalty $\mathcal{P}_{\eta_2}^{\mathcal{T}}(\boldsymbol{\theta})$ shrinking the pairwise group differences of the factor loadings follows the same rationale described in Appendix B.1, with the only difference being that $\boldsymbol{R}_q\boldsymbol{\theta}$ is now replaced by $\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\boldsymbol{\theta}$. The forms of the lasso, alasso, scad, and mcp penalties for the differences are:

$$\mathcal{P}_{\eta_2}^{L}(\boldsymbol{\theta}) = \eta_2 \sum_{g<g'} \sum_{q=1}^{q^\star} |\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}|,$$

$$\mathcal{P}_{\eta_2}^{A}(\boldsymbol{\theta}) = \eta_2 \sum_{g<g'} \sum_{q=1}^{q^\star} \frac{|\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}|}{\left|\hat{\theta}_{(g-1)m_1+q} - \hat{\theta}_{(g'-1)m_1+q}\right|^a},$$

$$\begin{aligned}\mathcal{P}_{\eta_2}^{S}(\boldsymbol{\theta}) = \sum_{g<g'} \sum_{q=1}^{q^\star} \Bigg\{ & \eta_2 |\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}| \mathbb{1}(0 \leq |\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}| \leq \eta_2) \\ & - \left[\frac{(\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q})^2 + \eta_2^2 - 2\eta_2 a |\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}|}{2(a-1)}\right] \\ & \times \mathbb{1}(\eta_2 < |\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}| \leq a\eta_2) \\ & + \frac{\eta_2^2(a+1)}{2} \big(|\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}| > a\eta_2\big) \Bigg\},\end{aligned}$$

$$\begin{aligned}\mathcal{P}_{\eta_2}^{M}(\boldsymbol{\theta}) = \sum_{g<g'} \sum_{q=1}^{q^\star} \Bigg\{ & \left(\eta_2 |\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}| - \frac{(\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q})^2}{2a}\right) \\ & \times \mathbb{1}(0 \leq |\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}| \leq a\eta_2) \\ & + \frac{\eta_2^2 a}{2} \mathbb{1}(|\theta_{(g-1)m_1+q} - \theta_{(g'-1)m_1+q}| > a\eta_2) \Bigg\},\end{aligned}$$

where for the alasso $a > 0$, for the scad $a > 2$ and for the mcp $a > 1$.

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. Because of the presence of fixed elements in $\boldsymbol{\tau}_g$ for model identification, $k^\star$ is smaller than p. Let $\boldsymbol{D}_q^{\boldsymbol{\tau}}$, 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) $\boldsymbol{D}_q^\tau = \boldsymbol{O}_{m_1\binom{G}{2}\times m}$. The penalty inducing equal intercepts across groups is then written as

$$\mathcal{P}_{\eta_3}^{\mathcal{T}}(\boldsymbol{\theta}) = \sum_{q=1}^{m} \mathcal{P}_{\eta_3,q}^{\mathcal{T}}(||\boldsymbol{D}_q^\tau \boldsymbol{\theta}||_1),$$

where $\eta_3 \in [0, \infty)$ governs the amount of intercept invariance. The penalty inducing equal intercepts across groups has precisely the same structure of the penalty inducing equal loadings, the only difference being in the type of parameters among which the differences are computed.

Optionally, one can encourage the invariance of the unique variances. However, as argued by Little, Card, Slegers and Ledford (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 the penalties $\mathcal{P}_{\eta_2}^{\mathcal{T}}$ and $\mathcal{P}_{\eta_3}^{\mathcal{T}}$.

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

$$\begin{aligned} \mathcal{P}_{\boldsymbol{\eta}}^{\mathcal{T}}(\boldsymbol{\theta}) &= \mathcal{P}_{\eta_1}^{\mathcal{T}}(\boldsymbol{\theta}) + \mathcal{P}_{\eta_2}^{\mathcal{T}}(\boldsymbol{\theta}) + \mathcal{P}_{\eta_3}^{\mathcal{T}}(\boldsymbol{\theta}) \\ &= \sum_{q=1}^{m} \left\{ \mathcal{P}_{\eta_1,q}^{\mathcal{T}}(||\boldsymbol{R}_q \boldsymbol{\theta}||_1) + \mathcal{P}_{\eta_2,q}^{\mathcal{T}}(||\boldsymbol{D}_q^{\boldsymbol{\Lambda}} \boldsymbol{\theta}||_1) + \mathcal{P}_{\eta_3,q}^{\mathcal{T}}(||\boldsymbol{D}_q^\tau \boldsymbol{\theta}||_1) \right\}, \end{aligned} \tag{5.3}$$

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 (5.3) can be any of the functions illustrated in Section 2.2, including lasso, alasso, scad and mcp, and different penalty functions can be

in principle combined.

By following the rationale described in Section 2.3, we replace each non-differentiable penalty in (5.3) with its differentiable local approximation:

$$\begin{aligned}
\mathcal{P}_{\eta_1}^{\mathcal{T}}(\boldsymbol{\theta}) &\approx \frac{1}{2}\boldsymbol{\theta}^T\left\{\sum_{q=1}^{m}\frac{\partial\mathcal{P}_{\eta_1,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{D}}_{\eta_1}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}, \\
\mathcal{P}_{\eta_2}^{\mathcal{T}}(\boldsymbol{\theta}) &\approx \frac{1}{2}\boldsymbol{\theta}^T\left\{\sum_{q=1}^{m}\frac{\partial\mathcal{P}_{\eta_2,q}^{\mathcal{T}}(||\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\theta}}||_1}\frac{1}{\sqrt{(\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\theta}})^T\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\theta}}+\bar{c}}}\boldsymbol{D}_q^{\boldsymbol{\Lambda}^T}\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\right\}\boldsymbol{\theta} \\
&= \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\mathcal{D}}_{\eta_2}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}, \\
\mathcal{P}_{\eta_3}^{\mathcal{T}}(\boldsymbol{\theta}) &\approx \frac{1}{2}\boldsymbol{\theta}^T\left\{\sum_{q=1}^{m}\frac{\partial\mathcal{P}_{\eta_3,q}^{\mathcal{T}}(||\boldsymbol{D}_q^{\boldsymbol{\tau}}\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{D}_q^{\boldsymbol{\tau}}\tilde{\boldsymbol{\theta}}||_1}\frac{1}{\sqrt{(\boldsymbol{D}_q^{\boldsymbol{\tau}}\tilde{\boldsymbol{\theta}})^T\boldsymbol{D}_q^{\boldsymbol{\tau}}\tilde{\boldsymbol{\theta}}+\bar{c}}}\boldsymbol{D}_q^{\boldsymbol{\tau}^T}\boldsymbol{D}_q^{\boldsymbol{\tau}}\right\}\boldsymbol{\theta} \\
&= \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\mathcal{D}}_{\eta_3}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}.
\end{aligned}$$

The matrix $\boldsymbol{\mathcal{D}}_{\eta_1}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ has the same form of the matrix $\boldsymbol{\mathcal{S}}_{\eta}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ described in equation (2.13), with the non-zero diagonal elements being the factor loadings in each of the groups. Let us now examine the elements that make up the matrix $\boldsymbol{\mathcal{D}}_{\eta_2}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$, namely,

$$d_q^{\mathcal{T}} = \frac{\partial\mathcal{P}_{\eta_2,q}^{\mathcal{T}}(||\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\theta}}||_1}\frac{1}{\sqrt{(\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\theta}})^T\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\theta}}+\bar{c}}}.$$

If $\boldsymbol{D}_q^{\boldsymbol{\Lambda}}$ for the parameter θ_q is non-null, the expressions of $d_q^{\mathcal{T}}$ for the lasso, alasso, scad and mcp penalties are:

$$d_q^L = \frac{\eta_2}{\sqrt{\sum_{g<g'}(\tilde{\theta}_{(g-1)m_1+q}-\tilde{\theta}_{(g'-1)m_1+q})^2+\bar{c}}},$$

$$d_q^A = \frac{\eta_2}{\left\{\sum_{g<g'}|\hat{\theta}_{(g-1)m_1+q}-\hat{\theta}_{(g'-1)m_1+q}|\right\}^a\sqrt{\sum_{g<g'}(\tilde{\theta}_{(g-1)m_1+q}-\tilde{\theta}_{(g'-1)m_1+q})^2+\bar{c}}},$$

$$d_q^S = \begin{cases} \dfrac{\eta_2}{\sqrt{\sum_{g<g'}(\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q})^2 + \bar{c}}} & \text{if } \sum_{g<g'}|\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q}| \leq \eta_2, \\ \dfrac{\dfrac{\max(a\eta_2 - \sum_{g<g'}|\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q}|, 0)}{a-1}}{\sqrt{\sum_{g<g'}(\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q})^2 + \bar{c}}} & \text{if } \sum_{g<g'}|\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q}| > \eta_2, \end{cases}$$

$$d_q^M = \begin{cases} \dfrac{\eta_2 - \dfrac{\sum_{g<g'}|\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q}|}{a}}{\sqrt{\sum_{g<g'}(\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q})^2 + \bar{c}}} & \text{if } \sum_{g<g'}|\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q}| \leq \eta_2 a, \\ 0 & \text{if } \sum_{g<g'}|\tilde{\theta}_{(g-1)m_1+q} - \tilde{\theta}_{(g'-1)m_1+q}| > \eta_2 a, \end{cases}$$

where for the alasso $a > 0$, for the scad $a > 2$ and for the mcp $a > 1$. The specification of the matrix $\boldsymbol{D}_q^{\boldsymbol{\tau}}$ computing the pairwise differences of the intercepts across groups and the corresponding expression of the approximated penalty matrix $\boldsymbol{\mathcal{D}}_{\eta_3}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ follows the same rationale just described for $\boldsymbol{D}_q^{\boldsymbol{\Lambda}}$ and $\boldsymbol{\mathcal{D}}_{\eta_2}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$.

These approximations lead to the following differentiable form of the combined penalty:

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

where $\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{D}}_{\eta_1}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{\mathcal{D}}_{\eta_2}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{\mathcal{D}}_{\eta_3}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ is the overall penalty matrix.

Partial invariance

The adequacy of an unpenalized multiple-group factor model is usually evaluated by testing the cross-group equality of any (set of) parameter(s) through likelihood ratio tests or local-fit measures. If factorial invariance is rejected, model modifications are conducted until one obtains a well-fitting model in which some, but not all, of the parameters are invariant ("partial invariance"; Steenkamp & Baumgartner, 1998).

The process of searching the non-invariant parameters in a multiple-group

analysis is the same as the one evaluating the plausibility of the fixed elements in a single-group analysis, but their determination is generally more difficult, error-prone, time-consuming in case of many observed variables and factors, and might change depending on the order of testing.

The proposed penalized approach can serve as an automatic tool for the detection of the optimal pattern of partial invariance, thus eluding invariance testing procedures.

Fused penalty

The first two penalties in (5.3) shrink the factor loadings within each group as well as their differences across groups. If $\mathcal{T} = L$, such penalty can be related to the generalized fused lasso proposed by Danaher, Wang and Witten (2014) in the context of multiple graphical models to penalize the off-diagonal elements of the precision matrices of different classes, as well as their differences across classes.

On a different note, that penalty can be viewed as an extension of the pairwise fused lasso illustrated by Petry (2011) to penalize the coefficients of a general linear model as well as their differences among any pair of regressors.

The next section provides an example clarifying the formulation of the presented penalty functions and matrices.

5.2.1 An example

For notational clarity, we illustrate the aforedescribed penalties in a simple example. Consider the following two-group factor model with $p = 6$ observed variables and $r = 2$ factors:

$$\boldsymbol{x}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{f}_g + \boldsymbol{\epsilon}_g \qquad \text{for } g = 1, 2,$$

where $\boldsymbol{f}_g \sim \mathcal{N}(\boldsymbol{\kappa}_g, \boldsymbol{\Phi}_g)$, $\boldsymbol{\epsilon}_g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$, with $\boldsymbol{\Psi}_g$ a diagonal matrix, and $\boldsymbol{f}_g$ is uncorrelated with $\boldsymbol{\epsilon}_g$. The parameter matrices are as follows, for $g = 1, 2$:

$$\boldsymbol{\Lambda}_g = \begin{bmatrix} \underline{1} & \underline{0} \\ \lambda_{21g} & \lambda_{22g} \\ \lambda_{31g} & \lambda_{32g} \\ \underline{0} & \underline{1} \\ \lambda_{51g} & \lambda_{52g} \\ \lambda_{61g} & \lambda_{62g} \end{bmatrix} \quad \boldsymbol{\tau}_g = \begin{bmatrix} \underline{0} \\ \tau_{2g} \\ \tau_{3g} \\ \underline{0} \\ \tau_{5g} \\ \tau_{6g} \end{bmatrix} \quad \boldsymbol{\Psi}_g = \begin{bmatrix} \psi_{11g} & 0 & 0 & 0 & 0 & 0 \\ & \psi_{22g} & 0 & 0 & 0 & 0 \\ & & \psi_{33g} & 0 & 0 & 0 \\ & & & \psi_{44g} & 0 & 0 \\ & & & & \psi_{55g} & 0 \\ & & & & & \psi_{66g} \end{bmatrix},$$

$$\boldsymbol{\Phi}_g = \begin{bmatrix} \phi_{11g} & \phi_{12g} \\ & \phi_{22g} \end{bmatrix} \qquad \boldsymbol{\kappa}_g = \begin{bmatrix} \kappa_{1g} \\ \kappa_{2g} \end{bmatrix}.$$

The factor loadings and intercepts of variables x_1 and x_4 have been fixed for metric setting and identification purposes, as illustrated in Section 5.1. The parameters of each group are collected in the m_g-dimensional vectors:

$$\begin{aligned} \boldsymbol{\theta}_1 &= (\text{vec}(\boldsymbol{\Lambda}_1)^T, \boldsymbol{\tau}_1^T, \text{diag}(\boldsymbol{\Psi}_1)^T, \text{vech}(\boldsymbol{\Phi}_1)^T, \boldsymbol{\kappa}_1^T)^T \\ &= (\lambda_{211}, \lambda_{311}, \lambda_{511}, \lambda_{611}, \lambda_{221}, \lambda_{321}, \lambda_{521}, \lambda_{621}, \tau_{21}, \tau_{31}, \tau_{51}, \tau_{61}, \psi_{111}, \psi_{221}, \psi_{331}, \\ &\quad \psi_{441}, \psi_{551}, \psi_{661}, \phi_{111}, \phi_{121}, \phi_{221}, \kappa_{11}, \kappa_{21})^T, \\ \boldsymbol{\theta}_2 &= (\text{vec}(\boldsymbol{\Lambda}_2)^T, \boldsymbol{\tau}_2^T, \text{diag}(\boldsymbol{\Psi}_2)^T, \text{vech}(\boldsymbol{\Phi}_2)^T, \boldsymbol{\kappa}_2^T)^T \\ &= (\lambda_{212}, \lambda_{312}, \lambda_{512}, \lambda_{612}, \lambda_{222}, \lambda_{322}, \lambda_{522}, \lambda_{622}, \tau_{22}, \tau_{32}, \tau_{52}, \tau_{62}, \psi_{112}, \psi_{222}, \psi_{332}, \\ &\quad \psi_{442}, \psi_{552}, \psi_{662}, \phi_{112}, \phi_{122}, \phi_{222}, \kappa_{12}, \kappa_{22})^T, \end{aligned}$$

where $m_1 = m_2 = 23$. The two group parameter vectors are combined into the m-dimensional vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, which can be conveniently expressed as

$$\begin{aligned} \boldsymbol{\theta} = (&\underbrace{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8}_{\substack{\textit{Factor loadings} \\ \textit{of Group 1}}}, \underbrace{\theta_9, \theta_{10}, \theta_{11}, \theta_{12}}_{\substack{\textit{Intercepts} \\ \textit{of Group 1}}}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{16}, \theta_{17}, \theta_{18}, \theta_{19}, \theta_{20}, \\ &\theta_{21}, \theta_{22}, \theta_{23}, \underbrace{\theta_{24}, \theta_{25}, \theta_{26}, \theta_{27}, \theta_{28}, \theta_{29}, \theta_{30}, \theta_{31}}_{\substack{\textit{Factor loadings} \\ \textit{of Group 2}}}, \underbrace{\theta_{32}, \theta_{33}, \theta_{34}, \theta_{35}}_{\substack{\textit{Intercepts} \\ \textit{of Group 2}}}, \theta_{36}, \theta_{37}, \theta_{38}, \\ &\theta_{39}, \theta_{40}, \theta_{41}, \theta_{42}, \theta_{43}, \theta_{44}, \theta_{45}, \theta_{46})^T, \end{aligned}$$

with $m = m_1 + m_2 = 2m_1 = 46$. Let $q^\star = 8$ be the number of factor loadings

in each group, and $k^\star = 4$ the number of intercepts in each group. Notice that the factor loadings in $\boldsymbol{\theta}$ are located in the positions determined by $q = (g-1)m_1 + 1, \ldots, (g-1)m_1 + q^\star$, for $g = 1, 2$, that is, $q = 1, \ldots, 8, 24, \ldots, 31$. Define the matrix $\boldsymbol{R}_q$:

$$\boldsymbol{R}_q = \begin{array}{c} \\ 1 \\ \\ q \\ \\ \\ 46 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & & q & & & 46 \end{array} \\ \begin{bmatrix} 0 & \ldots & 0 & \ldots & \ldots & 0 \\ \vdots & \ddots & \vdots & & & \vdots \\ 0 & \ldots & 1 & \ldots & \ldots & 0 \\ \vdots & & \vdots & \ddots & & \vdots \\ \vdots & & \vdots & & \ddots & \vdots \\ 0 & \ldots & 0 & \ldots & \ldots & 0 \end{bmatrix} \end{array} \quad \text{for } q = 1, \ldots, 8, 24, \ldots, 31,$$

and $\boldsymbol{R}_q = \boldsymbol{O}_{46\times 46}$ otherwise. Then, the penalty inducing sparsity on the factor loadings of each group is expressed as

$$\mathcal{P}_{\eta_1}(\boldsymbol{\theta}) = \sum_{q=1}^{46} \mathcal{P}_{\eta_1, q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1),$$

where $||\boldsymbol{R}_q\boldsymbol{\theta}||_1 = |\theta_q|$ for $q = 1, \ldots, 8, 24, \ldots, 31$, and 0 otherwise.

The pairwise differences of every loading across the two groups are $(\theta_q - \theta_{m_1+q})$, for $q = 1, \ldots, 8$, which consist of the set $\{(\theta_1 - \theta_{24}), (\theta_2 - \theta_{25}), (\theta_3 - \theta_{26}), (\theta_4 - \theta_{27}), (\theta_5 - \theta_{28}), (\theta_6 - \theta_{29}), (\theta_7 - \theta_{30}), (\theta_8 - \theta_{31})\}$. These differences can be specified through the matrix $\boldsymbol{D}_q^{\boldsymbol{\Lambda}}$, which, in case of two groups, for $q = 1, \ldots, 8$, is equal to:

$$\boldsymbol{D}_q = [\boldsymbol{R}_q \ -\boldsymbol{R}_q] = \begin{array}{c} \\ 1 \\ \\ q \\ \\ \\ 23 \end{array} \begin{array}{c} \begin{array}{ccccccccccc} 1 & & q & & & 23 & & 23+q & & & 46 \end{array} \\ \left[\begin{array}{ccccc|ccccc} 0 & \ldots & 0 & \ldots & \ldots & 0 & \ldots & 0 & \ldots & \ldots & 0 \\ \vdots & \ddots & \vdots & & & \vdots & & \vdots & & & \vdots \\ 0 & \ldots & 1 & \ldots & \ldots & 0 & \ldots & -1 & \ldots & \ldots & 0 \\ \vdots & & \vdots & \ddots & & \vdots & & \vdots & \ddots & & \vdots \\ \vdots & & \vdots & & \ddots & \vdots & & \vdots & & \ddots & \vdots \\ 0 & \ldots & 0 & \ldots & \ldots & 0 & \ldots & 0 & \ldots & \ldots & 0 \end{array}\right] \end{array}, \tag{5.4}$$

and $\boldsymbol{D}_q^{\boldsymbol{\Lambda}} = \boldsymbol{O}_{23\times 46}$ otherwise. Then, the penalty inducing equal loadings across groups can be written as

$$\mathcal{P}_{\eta_2}(\boldsymbol{\theta}) = \sum_{q=1}^{46} \mathcal{P}_{\eta_2,q}(||\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\boldsymbol{\theta}||_1),$$

where $||\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\boldsymbol{\theta}||_1 = |\theta_q - \theta_{m_1+q}|$ for $q = 1, \ldots, 8$, and 0 otherwise.

The pairwise differences of the intercepts across groups are computed similarly, the only difference being that the index q is now shifted by $q^\star$ units, that is, $q = (g-1)m_1 + q^\star + 1, \ldots, (g-1)m_1 + q^\star + k^\star = 9, \ldots, 12, 32, \ldots, 35$. Then, the penalty introducing equal intercepts across groups is written as

$$\mathcal{P}_{\eta_3}(\boldsymbol{\theta}) = \sum_{q=1}^{46} \mathcal{P}_{\eta_3,q}(||\boldsymbol{D}_q^{\boldsymbol{\tau}}\boldsymbol{\theta}||_1),$$

where $\boldsymbol{D}_q^{\boldsymbol{\tau}}$ is equal to the matrix in (5.4) for $q = 9, \ldots, 12$, and $\boldsymbol{D}_q^{\boldsymbol{\tau}} = \boldsymbol{O}_{23\times 46}$ otherwise, and $||\boldsymbol{D}_q^{\boldsymbol{\tau}}\boldsymbol{\theta}||_1 = |\theta_{q^\star+q} - \theta_{m_1+q^\star+q}|$ for $q = 9, \ldots, 12$, and 0 otherwise.

The penalty that simultaneously generates sparsity on the factor loading matrices and equivalent loadings and intercepts is:

$$\begin{aligned}\mathcal{P}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) &= \mathcal{P}_{\eta_1}(\boldsymbol{\theta}) + \mathcal{P}_{\eta_2}(\boldsymbol{\theta}) + \mathcal{P}_{\eta_3}(\boldsymbol{\theta}) \\ &= \sum_{q=1}^{46} \left\{\mathcal{P}_{\eta_1,q}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1) + \mathcal{P}_{\eta_2,q}(||\boldsymbol{D}_q^{\boldsymbol{\Lambda}}\boldsymbol{\theta}||_1) + \mathcal{P}_{\eta_3,q}(||\boldsymbol{D}_q^{\boldsymbol{\tau}}\boldsymbol{\theta}||_1)\right\}.\end{aligned}$$

5.3 Penalized maximum likelihood estimation

Similarly to what was done for the single-group factor model, we can express the penalized log-likelihood function employing the local approximations of the penalties described in equation (5.4) as

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - N\,\mathcal{P}_{\boldsymbol{\eta}}^{\mathcal{T}}(\boldsymbol{\theta}) = \left\{\ell(\boldsymbol{\theta}) - \frac{N}{2}\boldsymbol{\theta}^T\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right\}, \tag{5.5}$$

where the log-likelihood of the multiple-group factor model $\ell(\boldsymbol{\theta})$ is given in (5.2), and $\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{D}}_{\eta_1}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{\mathcal{D}}_{\eta_2}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{\mathcal{D}}_{\eta_3}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ is the sum of the three penalty matrices

introducing sparsity and loading and intercept invariance.

The estimation of the model parameters follows the same procedure described in Section 3.1, with the only difference being that the scalar tuning parameter η is now replaced with the tuning parameter vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)^T$. Due to the presence of parameters for the mean structure (i.e., the intercepts and the factor means) in addition to those for the covariance structure, we only considered the penalized Fisher information matrix $\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}) = \boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ as second-order derivative information in the trust-region algorithm. The expressions of the gradient vector and the Fisher information matrix for the multiple-group factor model are derived in Appendix F.2.

Although one may in principle conduct a grid-search combined with GBIC (as illustrated in Section 3.2) to determine the optimal values of the tuning parameters, this procedure inevitably becomes computationally intensive and inefficient due to the presence of three distinct tuning parameters, which requires fine grid-searches in three dimensions. The automatic tuning parameter procedure described in Section 3.3 really comes in handy here as it can be straightforwardly extended to estimate the multiple tuning parameters that compose the penalty $\mathcal{P}_{\boldsymbol{\eta}}^{\mathcal{T}}(\boldsymbol{\theta})$ in a fast, stable and efficient way.

The *edf* of the penalized multiple-group factor model are estimated as

$$edf = \text{tr}\left\{\boldsymbol{\mathcal{J}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{J}}(\hat{\boldsymbol{\theta}})\right\} = m - \text{tr}\left\{[\boldsymbol{\mathcal{J}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})]^{-1}N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\hat{\boldsymbol{\theta}})\right\}, \tag{5.6}$$

which shows that $edf \to m$ as $\boldsymbol{\eta} \to \mathbf{0}$, and $edf \to m - r^\star$ as $\boldsymbol{\eta} \to \boldsymbol{\infty}$, where $r^\star = G(q^\star + k^\star)$ is the total number of penalized elements; when $\mathbf{0} < \boldsymbol{\eta} < \boldsymbol{\infty}$, the $edf \in [m - r^\star, m]$.

All of the theoretical properties of the PMLE illustrated in Section 3.4 for the normal linear factor model (Theorems 3.1-3.3) continue to hold in the penalized multiple-group factor model.

CHAPTER 6

Numerical and empirical evaluation of the penalized multiple-group factor model

This chapter evaluates the validity of the penalized multiple-group factor model presented in Chapter 5 through numerical and empirical examples. First, we describe a simulation study examining progressive levels of non-invariance in the factor loadings and intercepts (Section 6.1). The performances of the PMLE are evaluated and compared to the ones of a competing method. We investigate the impact of several conditions, including the sample size, the size of the generated difference, the magnitude of the influence factor and the value of the additional tuning parameter. In addition, the proposed model and its competitor are tested on a well-known psychometric data set (Section 6.2).

6.1 Simulation study

A simulation study was conducted to evaluate the ability of the proposed PMLE technique in identifying the pattern of partial invariance in a multiple-group factor analysis model. We first describe the design of the study and then present the results. Since the current implementation of `regsem` does not allow for multiple-group analyses, our method is only compared with `lslx`.

6.1.1 Design and procedure

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 the simulation study in the single-group factor model (Section 4.1.2), we employ the alasso penalty for inducing sparsity and invariant loadings and intercepts, that is, $\mathcal{S}^A_{\boldsymbol{\eta}}(\tilde{\boldsymbol{\theta}}) = \mathcal{D}^A_{\eta_1}(\tilde{\boldsymbol{\theta}}) + \mathcal{D}^A_{\eta_2}(\tilde{\boldsymbol{\theta}}) + \mathcal{D}^A_{\eta_3}(\tilde{\boldsymbol{\theta}})$. 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 (see Section 6.2);
- **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 the simulation and empirical application for the single-group factor model (Chapter 4), 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

	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

Note: Under the null condition, the parameters of Group 2 coincide with those of Group 1.

Table 6.1: The factor loading matrices and intercepts of the two groups under each difference scenario. Elements fixed for origin and scale setting and identification purposes are italic and underlined.

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 6.1 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 6.1. 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 Section 5.2 (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 factor model produced admissible solutions, and analyzed them as described in the simulation for the single-group model (Section 4.1.1).

6.1.2 Results

The performances of the penalized models are evaluated through the criteria used in the simulation study for the single-group factor model reported in expressions (4.1)-(4.5), that is, mean-squared error (MSE), squared bias (SB), true positive rate (TPR), false positive rate (FPR) and proportion choosing the true model (PCTM). For the sake of conciseness, we report the results for the `GJRM-alasso` model ($a = 2$ and $\gamma = 4.5$) that produced the best solution in terms of these performance criteria.

By looking at the results in Table 6.2, we draw the following conclusions:

1. 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.

2. Higher difference sizes were associated with higher MSE and squared bias, with the lower values generally occurring for `GJRM-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 `GJRM-alasso`; the results are depicted in Figure 6.1 for MSE and Figure 6.2 for SB. 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.

3. The TPR were always equal to 1.0, which showed that the examined methods never suppressed the non-zero penalized parameters.

Difference scenario	Null		Small		Medium		Large	
	`GJRM`	`lslx`	`GJRM`	`lslx`	`GJRM`	`lslx`	`GJRM`	`lslx`
MSE								
$N = 300$	0.275	0.279	0.303	0.307	0.356	0.372	0.385	0.416
$N = 500$	0.165	0.164	0.189	0.189	0.220	0.239	0.221	0.235
$N = 1000$	0.083	0.082	0.102	0.104	0.105	0.115	0.103	0.101
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
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
FPR								
$N = 300$	0.006	0.010	0.005	0.012	0.005	0.019	0.004	0.035
$N = 500$	0.004	0.004	0.005	0.005	0.004	0.014	0.003	0.020
$N = 1000$	0.002	0.001	0.002	0.002	0.002	0.005	0.001	0.003

Table 6.2: Performance measures of `GJRM-alasso` and `lslx-mcp` models by sample size and difference scenario. MSE stands for mean-squared error, SB for squared bias, PCTM for proportion choosing the true model, FPR for false positive rate.

4. Whereas under the null and small scenarios the two methods produced similar measures, `GJRM-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 `GJRM-alasso`; even with the smallest sample size, `GJRM-alasso` identified the true heterogeneity pattern more than 90% of the times.

Computational efficiency

Thanks to the use of the automatic multiple tuning parameter procedure, under every sample size and difference scenario, the computational time to fit a `GJRM-alasso` model with three tuning parameters was much lower than the one necessary to fit a `lslx-mcp` model with a single shrinkage parameter η and the associated shape parameter a selected through a grid-search. Table 6.3 reports the

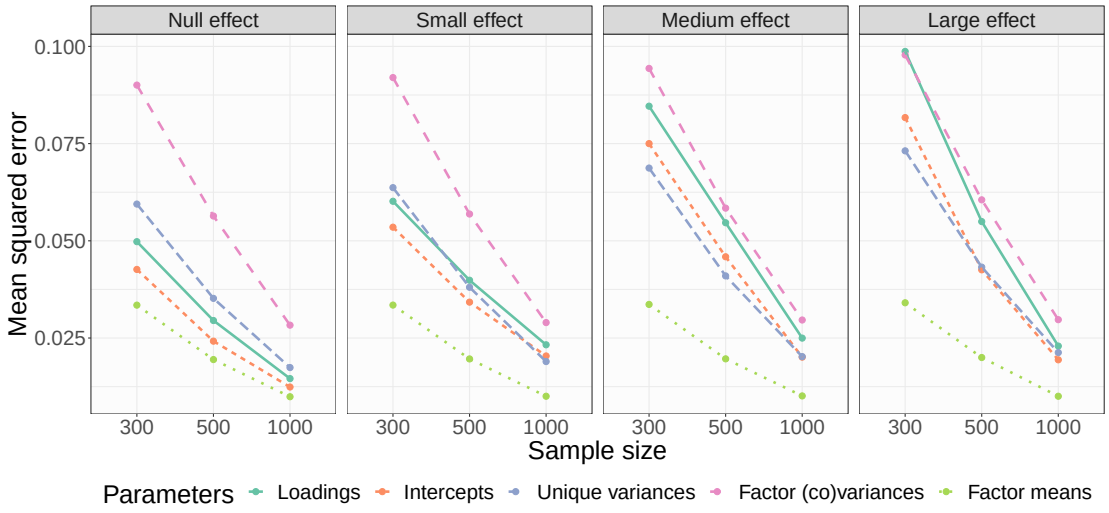


Figure 6.1: Average mean squared error of `GJRM-alasso` ($a = 2, \gamma = 4.5$) by difference scenario, sample size and parameter type.

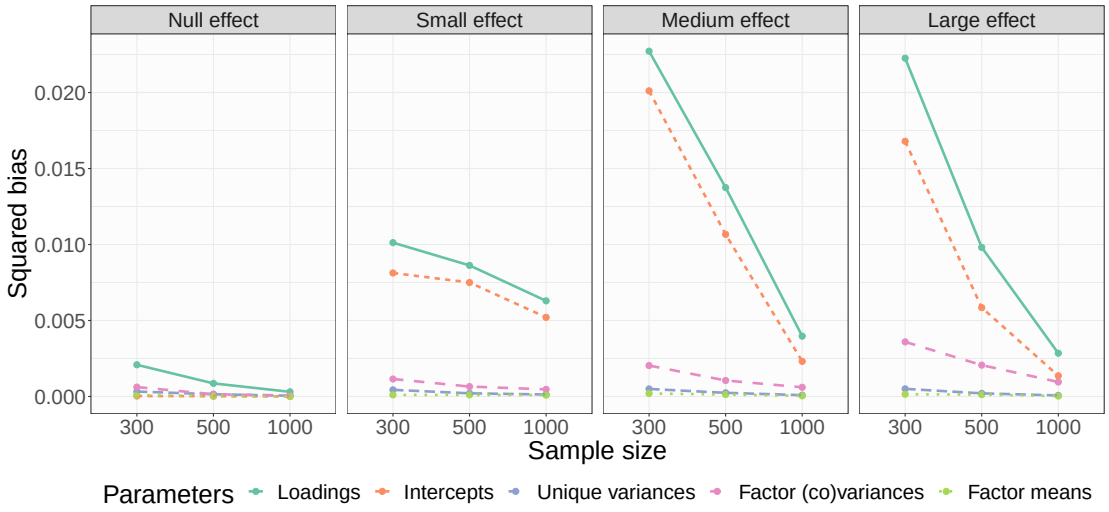


Figure 6.2: Average squared bias of `GJRM-alasso` ($a = 2, \gamma = 4.5$) by difference scenario, sample size and parameter type.

minimum, median and standard error of the elapsed time for estimating one penalized multiple-group factor analysis model under every sample size and difference scenario. The distributions of the elapsed times are visualized through violin plots in function of the sample size and the difference scenario in Figure 6.3. The times of `GJRM` models generally had higher standard errors, due to the larger variability in the number of iterations required by the automatic procedure, as opposed to the smaller variability of the times of `lslx` models, which were fitted through a grid-search and thus tended to be characterized by comparable computational times across replications. Nevertheless, `GJRM-alasso` was always by about 11 to 27 times faster than the competitor, depending on the condition. It is important to stress that the higher computational times of `lslx` are based on a unidimensional grid-search since the software uses a single tuning parameter for sparsity and loading and intercept invariance. If instead one were to consider three distinct tuning parameters, the method would require grid-searches in three dimensions. This procedure clearly becomes inefficient and prohibitive with the growth of the number of tuning parameters. The further problem with grid-searches is that they are essentially arbitrary due to the subjectivity in the choice of the grid size and the granularity (i.e., how much the elements are interspaced). On the contrary, the automatic tuning procedure can estimate tuning parameters that can take any positive value and scales well as the number of tuning parameters increases.

6.2 Empirical application

In Section 4.2, the Holzinger & Swineford data set was used to conduct an empirical analysis and demonstrate the proposed penalized technique for the normal linear factor analysis model. The data set on the mental ability tests also contains information about the school attended by the students. One school (Pasteur) includes students with parents who immigrated from Europe, whereas the other (Grant-White) is composed of students coming from middle-income American white families. Therefore, we can conduct a multiple-group analysis on these

Difference scenario	Null		Small		Medium		Large	
	GJRM	lslx	GJRM	lslx	GJRM	lslx	GJRM	lslx
N = 300								
Minimum	0.89	25.46	1.03	26.98	1.22	24.37	1.13	26.14
Median	1.84	45.57	2.19	48.61	4.70	47.09	4.64	49.77
Standard error	8.57	3.18	11.55	8.70	16.89	4.60	15.86	17.48
N = 500								
Minimum	0.86	23.73	0.81	24.59	1.06	33.03	1.00	25.70
Median	1.70	42.93	2.31	45.74	5.20	45.14	4.32	45.99
Standard error	9.06	2.21	11.56	100.03	15.32	3.93	10.37	3.01
N = 500								
Minimum	0.76	22.68	0.81	29.20	1.81	25.85	1.14	30.04
Median	1.56	40.65	4.18	46.55	4.19	42.70	3.36	43.94
Standard error	7.29	2.28	12.84	27.73	9.52	10.20	10.79	1.84

Table 6.3: Minimum, median and standard error of the elapsed time (seconds) for `GJRM-alasso` ($a = 2; \gamma = 4.5$) and `lslx-mcp` under each sample size and difference scenario.

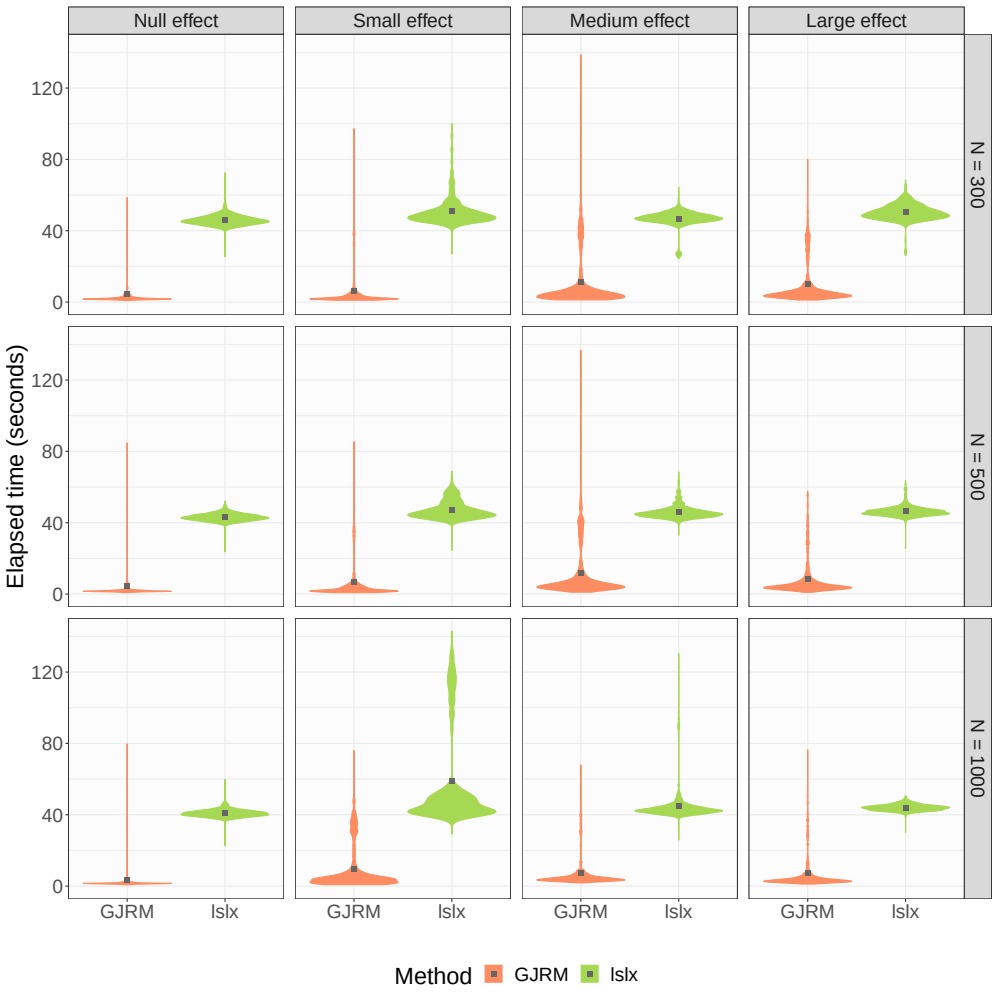


Figure 6.3: Distributions of the elapsed times of the investigated methods under each sample size and difference scenario. The grey squares indicate the average times.

Observed variables	Minimum	Maximum
VISUAL	4	51
CUBES	9	37
PAPER	6	25
FLAGS	2	36
GENERAL	8	84
PARAGRAP	0	19
SENTENCE	4	28
WORDC	10	43
WORDM	1	43
ADDITION	30	171
CODE	19	118
COUNTING	61	200
STRAIGHT	100	333
WORDR	121	198
NUMBERR	68	112
FIGURER	58	119
OBJECT	0	26
NUMBERF	0	20
FIGUREW	3	20

Table 6.4: Original ranges of values of the observed variables of the Holzinger & Swineford data set.

two sub-groups ($N_1 = 156, N_2 = 145$). Following Huang (2018), we consider the following $p = 19$ mental tests: visual perception (VISUAL), cubes (CUBES), paper from board (PAPER), flags (FLAGS), general information (GENERAL), paragraph comprehension (PARAGRAP), sentence completion (SENTENCE), word classification (WORDC), word meaning (WORDM), addition (ADDITION), code (CODE), counting groups of dots (COUNTING), straight and curved capitals (STRAIGHT), word recognition (WORDR), number recognition (NUMBERR), figure recognition (FIGURER), object-number (OBJECT), number-figure (NUMBERF), figure-word (FIGUREW). These tests are thought of as measuring $q = 4$ correlated abilities: spatial ability (VISUAL, CUBES, PAPER, FLAGS), verbal intelligence (GENERAL, PARAGRAP, SENTENCE, WORDC, WORDM), speed (ADDITION, CODE, COUNTING, STRAIGHT), and memory (WORDR, NUMBERR, FIGURER, OBJECT, NUMBERF, FIGUREW). The ranges of values of the observed variables are quite diverse (Table 6.4). As in Huang (2018), we standardized the data to handle the scaling effect.

The traditional approach to multiple-group analyses 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 a penalized technique employing penalties that combine sparsity and cross-group equivalence of loadings and intercepts, such as the one introduced in this thesis. 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 factor loadings and intercepts. The data analysis was also conducted in `lslx` with the mcp, 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 `GJRM-alasso` and `lslx-mcp` are reported in Tables 6.5 and 6.6, respectively. The better fit of `GJRM-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 parameter, 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. Both techniques produce sparse loading matrices with many zero-entries, but the presence of a couple of non-zero cross-loadings demonstrates that the structure hypothesized by a multiple-group CFA is too restrictive. Contrarily to `lslx-mcp`, which presents

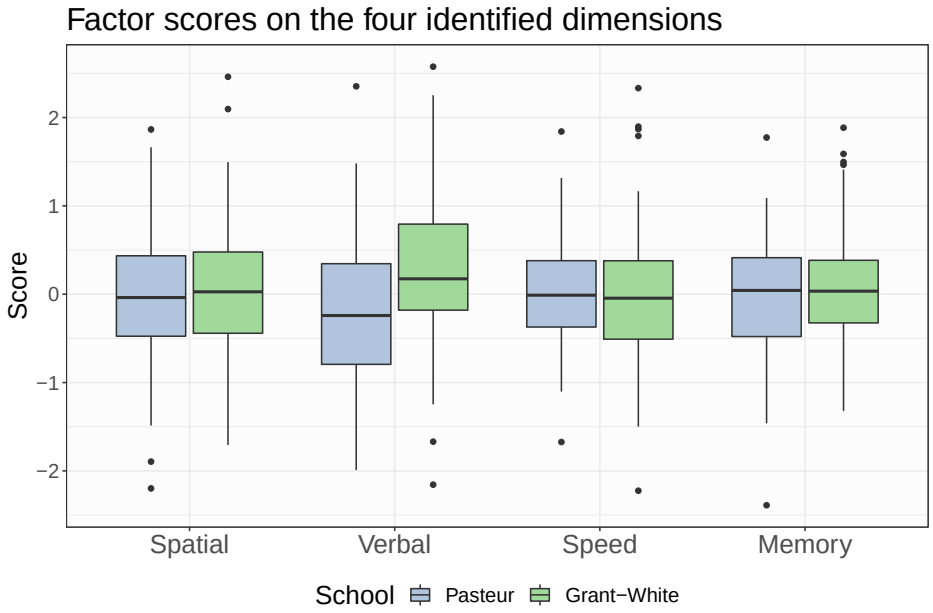


Figure 6.4: The distributions of the factor scores on the four identified dimensions and in the two schools for `GJRM-alasso` on the Holzinger & Swineford data set.

one non-invariant loading, the factor loading matrices of `GJRM-alasso` are fully equivalent, in agreement to the results of invariance testing. Conversely, the intercepts are not fully invariant, which is again in line with the findings from factorial invariance testing.

The students of the two schools can be scaled on every uncovered dimension through the calculation of the so-called factor scores, which are "estimates" or "predictions" of the values of the latent factors for each individual. Figure 6.4 shows the distributions of the factor scores on the four identified dimensions (i.e., spatial ability, verbal intelligence, speed and memory) and in the two groups for the `GJRM-alasso` model. From a visual inspection, the students from Grant-White school seem to score on average higher on the verbal construct, whereas the students' performances on the other factors appear comparable across schools. This result, however, should be interpreted with caution due to the lack of invariance detected in the intercepts as well as the indeterminacy problem that affects the factor scores (Grice, 2001).

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.

GJRM - alasso

Measurement model	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
Structural model	$\boldsymbol{\kappa}_1$	Spatial	Verbal	Speed	Memory		$\boldsymbol{\kappa}_2$	Spatial	Verbal	Speed	Memory	
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	

Table 6.5: Parameter estimates of the 19 mental tests from the Holzinger & Swineford data set for `GJRM-alasso` (automatic procedure, $\hat{\boldsymbol{\eta}} = (0.006, 16221.852, 0.013)^T, a = 1, \gamma = 4$). 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 (*).

lslx - mcp

Measurement model	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*	0.48	*0*	*1*	*0*	*0*	*0*	0.45
CUBES	0.01	0.64				0.87	0.01	0.64				0.67
PAPER	0.00	0.66				0.81	0.00	0.66				0.71
FLAGS	0.27⋆	0.86				0.62	−0.28⋆	0.86				0.47
GENERAL	−0.03⋆		1.03		−0.15	0.26	0.02⋆		1.03		−0.15	0.31
PARAGRAP	0.00⋆		0.98			0.35	−0.01⋆		0.98			0.31
SENTENCE	0.00	−0.09	1.10		−0.10	0.25	0.00	−0.09	1.10		−0.10	0.21
WORDC	−0.09⋆	0.04	0.84			0.40	0.09⋆	0.04	0.84			0.44
WORDM	*0*	*0*	*1*	*0*	*0*	0.23	*0*	*0*	*1*	*0*	*0*	0.35
ADDITION	−0.02	−0.47		1.31		0.48	−0.02	−0.47		1.31		0.30
CODE	−0.01	0	0.15	0.91	0.10	0.43	−0.01		0.15	0.91	0.10	0.63
COUNTING	*0*	*0*	*0*	*1*	*0*	0.61	*0*	*0*	*0*	*1*	*0*	0.52
STRAIGHT	0.00	0.37		0.76		0.63	0.00	0.37		0.76		0.46
WORDR	*0*	*0*	*0*	*0*	*1*	0.61	*0*	*0*	*0*	*0*	*1*	0.57
NUMBERR	0.00		−0.09		0.88	0.69	0.00		−0.09		0.88	0.68
FIGURER	0.01	0.40			0.69	0.72	0.01	0.40			0.69	0.46
OBJECT	0.12⋆	−0.32		0.46	0.90	0.60	−0.15⋆	−0.32		0.46	0.90	0.45
NUMBERF	0.03				0.68	0.77	0.03			0.49⋆	0.68	0.62
FIGUREW	−0.28⋆			0.17	0.58	0.83	0.29⋆			0.17	0.58	0.60
Structural model	$\boldsymbol{\kappa}_1$	Spatial	Verbal	Speed	Memory		$\boldsymbol{\kappa}_2$	Spatial	Verbal	Speed	Memory	
Spatial	−0.07	0.54	0.25	0.17	0.16		0.07	0.55	0.32	0.29	0.21	
Verbal	−0.25		0.64	0.21	0.11		0.28		0.62	0.26	0.24	
Speed	0.12			0.36	0.12		−0.11			0.50	0.16	
Memory	−0.04				0.48		0.04				0.37	

Table 6.6: Parameter estimates of the 19 mental tests from the Holzinger & Swineford data set for `lslx-mcp` ($\hat{\eta} = 0.14, \hat{a} = 3$). 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 (⋆).

CHAPTER 7

Software implementation

The proposed methodology and estimation approach are implemented in the `R` package `GJRM` (Marra & Radice, 2019b) to enhance reproducible research and transparent dissemination of results. In this chapter, we describe the main functions for fitting single and multiple-group factor analysis models according to the penalized likelihood-based estimation framework proposed in this thesis (see Sections 7.1 and 7.2, respectively). To this end, we demonstrate how potential users can carry out the empirical analyses presented in Sections 4.2 and 6.2 through the package `GJRM`.

Get started

The subsequent analyses require the `R` package `GJRM`, so we install and load this package, and then progress with the analysis.

```
install.packages("GJRM", dependencies = TRUE)
library(GJRM)
```

7.1 Penalized estimation of a factor model

The empirical analysis presented in Section 4.2 employs the Holzinger & Swineford data set (Holzinger & Swineford, 1939), a classical psychometric application on students' mental abilities. The data set, already scaled as described in Yuan and Bentler (2006), is contained in the `R` package `lavaan` (Rosseel, 2012; Rosseel et al., 2019). Let us load and inspect the data.

```
data <- lavaan::HolzingerSwineford1939
summary(data)


##        id              sex            ageyr         agemo
##  Min.   :  1.0   Min.   :1.000   Min.   :11   Min.   : 0.000
##  1st Qu.: 82.0   1st Qu.:1.000   1st Qu.:12   1st Qu.: 2.000
##  Median :163.0   Median :2.000   Median :13   Median : 5.000
##  Mean   :176.6   Mean   :1.515   Mean   :13   Mean   : 5.375
##  3rd Qu.:272.0   3rd Qu.:2.000   3rd Qu.:14   3rd Qu.: 8.000
##  Max.   :351.0   Max.   :2.000   Max.   :16   Max.   :11.000
##
##          school        grade             x1               x2
##  Grant-White:145   Min.   :7.000   Min.   :0.6667   Min.   :2.250
##  Pasteur    :156   1st Qu.:7.000   1st Qu.:4.1667   1st Qu.:5.250
##                    Median :7.000   Median :5.0000   Median :6.000
##                    Mean   :7.477   Mean   :4.9358   Mean   :6.088
##                    3rd Qu.:8.000   3rd Qu.:5.6667   3rd Qu.:6.750
##                    Max.   :8.000   Max.   :8.5000   Max.   :9.250
##                    NA's   :1
##        x3              x4              x5              x6
##  Min.   :0.250   Min.   :0.000   Min.   :1.000   Min.   :0.1429
##  1st Qu.:1.375   1st Qu.:2.333   1st Qu.:3.500   1st Qu.:1.4286
##  Median :2.125   Median :3.000   Median :4.500   Median :2.0000
##  Mean   :2.250   Mean   :3.061   Mean   :4.341   Mean   :2.1856
##  3rd Qu.:3.125   3rd Qu.:3.667   3rd Qu.:5.250   3rd Qu.:2.7143
##  Max.   :4.500   Max.   :6.333   Max.   :7.000   Max.   :6.1429
##
##        x7              x8              x9
##  Min.   :1.304   Min.   : 3.050   Min.   :2.778
##  1st Qu.:3.478   1st Qu.: 4.850   1st Qu.:4.750
##  Median :4.087   Median : 5.500   Median :5.417
##  Mean   :4.186   Mean   : 5.527   Mean   :5.374
##  3rd Qu.:4.913   3rd Qu.: 6.100   3rd Qu.:6.083
##  Max.   :7.435   Max.   :10.000   Max.   :9.250
##
```

The data set contains information on the test scores (items x_1 to x_9) of $N = 301$ seventh-grade and eighth-grade students on $p = 9$ mental tests. Additional information is available, such as the age of the students and the attended school (i.e., Pasteur or Grant-White). Let us select and center the data subset constituted by the nine tests.

```
data <- scale(data[,7:15], center = TRUE, scale = FALSE)
```

The following sections describe how to specify and estimate a penalized factor analysis model using the adaptive lasso penalization to encourage a sparse factor loading matrix and the automatic tuning parameter procedure to select the optimal amount of sparsity. This combination of penalty and tuning selection strategy produced the model with the superior fit in the empirical analysis (see Table 4.6 with the BIC ranking).

7.1.1 Model specification

Before fitting the model, users should write a "model syntax" which describes the model to be estimated and specifies the relationships between the observed variables and the latent variables (i.e., the common factors). To facilitate its formulation, the rules for the syntax specification follow the ones required by the package `lavaan`, and are briefly reviewed below. Let us have a look at the following syntax, which is enclosed in single quotes.

```
syntax <-' # Measurement model
           spatial =~ x1 + x2 + x3 + 0*x4 + x5 + x6 + 0*x7 + x8 + x9
           verbal  =~ 0*x1 + x2 + x3 + x4 + x5 + x6 + 0*x7 + x8 + x9
           speed   =~ 0*x1 + x2 + x3 + 0*x4 + x5 + x6 + x7 + x8 + x9

            # Unit variances for common factors
           spatial ~~ 1*spatial
           verbal  ~~ 1*verbal
           speed   ~~ 1*speed '
```

The three common factors are referred to as `spatial`, `verbal` and `speed`, whereas the observed variables names range from `x1` to `x9`. The factors appear on the left-hand side, whereas the observed variables on the right-hand side. The special operator "`=~`" is read as "is measured by", and is used to list the observed variables loading on each factor. The factor variances and covariances are specified using the double tilde operator "`~~`". In order to fix a parameter to a given value, we pre-multiply (through the symbol "`*`") the corresponding variable in the formula by the specific numerical value.

The above syntax specifies a factor model with $r = 3$ common factors, where each observed variable loads on each of the factors, apart from the ones whose loadings are fixed to zero for identification purposes. The scales of the factors are specified by fixing their variances to 1.0. By default, the unique variances are automatically added to the model, and the common factors are allowed to correlate. These specifications can be easily modified by altering the syntax according to one's own preferences.

7.1.2 Model fitting

We now show how to estimate the factor analysis model specified in the syntax according to the penalized likelihood-based approach presented in this thesis. The estimation process is demonstrated for the alasso penalty and the automatic tuning procedure, but the rationale is similar for other choices of penalty functions. The alasso employs a set of adaptive weights correcting the bias issue of the lasso. A common choice for the weights is given by the maximum likelihood estimates from the unpenalized factor model. The unpenalized model can be estimated through the function `penfa` - a short form for *PENalized Factor Analysis* - as follows:

```
fit.mle <- penfa(model = syntax, data = data, information ="fisher",
                 shrink = "none")
```

The function `penfa` takes as first argument the user-specified model syntax, and as second argument the data set with the observed variables. The `information` argument allows users to choose between the penalized expected Fisher information ("fisher") or the penalized Hessian matrix ("hessian") as second-order derivatives to be used in the trust-region algorithm (the matrix $\boldsymbol{B}$ in expression (3.3)). In the `shrink` argument, users can specify the penalty function of interest; when it is set equal to "none", no penalization is applied, and the model is estimated by ordinary maximum likelihood. We can get an overview of the data set and the optimization process by printing the `fit.mle` object.

```
fit.mle

## GJRM reached convergence
##
##   Number of observations                                  301
##
##   Estimator                                               MLE
##   Optimization method                             trust-region
##   Information                                         expected
##   Strategy                                                grid
##   Number of iterations                                      15
##   Effective degrees of freedom                          33.000
##
```

The trust-region algorithm required a small number of iterations to converge. Since no penalization is imposed, the effective degrees of freedom coincide with the number of model parameters, that is, $edf = m = 33$. The parameter estimates can be extracted through the function `coef` together with their names. Each name is composed of three parts and reflects the part of the formula in which a given parameter was involved. The variable name appears on the left-hand side of the formula, the operator is placed in the middle, and the variable corresponding to the parameter on the right-hand side.

```
weights <- coef(fit.mle)
weights

##     spatial=~x1     spatial=~x2     spatial=~x3     spatial=~x5
##           0.814           0.652           0.909          -0.134
##     spatial=~x6     spatial=~x8     spatial=~x9      verbal=~x2
##           0.067           0.296           0.540          -0.118
##      verbal=~x3      verbal=~x4      verbal=~x5      verbal=~x6
##          -0.330           0.987           1.193           0.875
##      verbal=~x8      verbal=~x9       speed=~x2       speed=~x3
##          -0.158          -0.141          -0.161          -0.012
##       speed=~x5       speed=~x6       speed=~x7       speed=~x8
##           0.008          -0.020           0.767           0.680
##       speed=~x9          x1~~x1          x2~~x2          x3~~x3
##           0.433           0.696           1.035           0.692
##          x4~~x4          x5~~x5          x6~~x6          x7~~x7
##           0.377           0.403           0.365           0.594
##          x8~~x8          x9~~x9 spatial~~verbal  spatial~~speed
##           0.479           0.551           0.585           0.173
##   verbal~~speed
##           0.220
```

The estimation of the penalized factor model is again carried out through the function `penfa`, but with some new and different arguments. The alasso penalty function is specified in the `shrink` argument, whereas the adaptive weights are given in the `weights` argument. The value of the additional tuning parameter a of the alasso can be assigned through the `a.alasso` argument, whereas the `eta` argument allows users to provide a starting value for the shrinkage parameter η. The name given to the starting value - "lambda" in this case - reflects the parameter matrix or vector to be penalized. By default, all of its elements are penalized, which means here that the penalization is applied to all of the factor loadings. If "strategy" is specified equal to "grid", then a penalized model with the value of η given in `eta` is estimated, whereas the automatic tuning parameter procedure is carried out when `strategy` is set equal to "auto". Lastly, users can choose a specific value of the influence factor γ through the `gamma` argument.

```
fit <- penfa(model = syntax, data = data, information = "fisher",
             shrink = "alasso", weights = weights, a.alasso = 1,
             eta = list("shrink" = c("lambda" = 0.01)),
             strategy = "auto", gamma = 4.5)
fit

## GJRM reached convergence
##
##   Number of observations                                  301
##
##   Estimator                                              PMLE
##   Optimization method                           trust-region
##   Information                                       expected
##   Strategy                                              auto
##   Number of iterations (total)                            32
##   Number of two-steps (automatic)                          1
##   Effective degrees of freedom                        22.843
##
##   Penalty function:
##     Sparsity                                           alasso
##
```

Printing the fitted object gives an overview of the optimization and penalization processes, including the employed optimizer and penalty function, the total number of iterations and the number of outer iterations of the automatic procedure. The

automatic procedure is very fast, as it required a single outer iteration to reach convergence. The number of effective degrees of freedom of the penalized model is $edf = 22.843$, which is a fractional number, as opposed to the integer number that existing penalized factor analytic techniques report for the degrees of freedom.

The `summary` function provides detailed information on the model characteristics, the optimization and the penalization procedures, as well as the parameter estimates with associated standard errors and confidence intervals. The optimal value of the tuning parameter is $\hat{\eta} = 0.017$. The data set well supported the introduction of sparsity, as is demonstrated by the reduction in the Generalized Bayesian Information Criterion (GBIC) when moving from the unpenalized model `fit.mle` (7601.416) to its penalized counterpart `fit` (7558.026). The *Type* column distinguishes between the *fixed* parameters that have been set to specific values for identification purposes, the *free* parameters that have been estimated through ordinary maximum likelihood, and the penalized parameters (denoted as *pen*). The standard errors are computed as the square root of the inverse of the penalized Fisher information matrix (or alternatively, of the penalized Hessian if `information` = "hessian"). The last columns report 95% confidence intervals for the model parameters. The standard errors and the confidence intervals of the penalized parameters that were shrunken to zero are not reported. A different significance level can be specified through the `level` argument in the `summary` call.

```
summary(fit)

## GJRM reached convergence
##
##   Number of observations                              301
##   Number of groups                                      1
##   Number of observed variables                          9
##   Number of latent factors                              3
##
##   Estimator                                          PMLE
##   Optimization method                       trust-region
##   Information                                   expected
##   Strategy                                          auto
##   Number of iterations (total)                        32
##   Number of two-steps (automatic)                      1
##   Influence factor                                   4.5
```

```
##   Number of parameters:
##     Free                                                 12
##     Penalized                                            21
##   Effective degrees of freedom                       22.843
##   GIC                                              7473.346
##   GBIC                                             7558.026
##
##   Penalty function:
##     Sparsity                                         alasso
##
##   Additional tuning parameter
##     alasso                                                1
##
##   Optimal tuning parameter:
##     Sparsity
##      - Factor loadings                                0.017
##
## Parameter Estimates:
##
## Latent Variables:
##                  Type    Estimate    Std.Err       2.5%      97.5%
##   spatial =~
##     x1           pen        0.829      0.073      0.685      0.972
##     x2           pen        0.493      0.073      0.350      0.636
##     x3           pen        0.758      0.086      0.591      0.926
##     x4           fixed      0.000                 0.000      0.000
##     x5           pen       -0.060      0.034     -0.128      0.007
##     x6           pen        0.000
##     x7           fixed      0.000                 0.000      0.000
##     x8           pen        0.124      0.059      0.008      0.239
##     x9           pen        0.410      0.062      0.290      0.531
##   verbal =~
##     x1           fixed      0.000                 0.000      0.000
##     x2           pen       -0.000
##     x3           pen       -0.157      0.066     -0.286     -0.029
##     x4           pen        0.960      0.055      0.852      1.069
##     x5           pen        1.114      0.065      0.987      1.240
##     x6           pen        0.889      0.052      0.787      0.992
##     x7           fixed      0.000                 0.000      0.000
##     x8           pen       -0.000
##     x9           pen       -0.000
##   speed =~
##     x1           fixed      0.000                 0.000      0.000
##     x2           pen       -0.013
##     x3           pen        0.000
##     x4           fixed      0.000                 0.000      0.000
##     x5           pen        0.000
##     x6           pen        0.000
##     x7           pen        0.697      0.078      0.544      0.850
##     x8           pen        0.704      0.077      0.553      0.854
##     x9           pen        0.423      0.060      0.305      0.541
```

```
##
## Covariances:
##                  Type    Estimate    Std.Err       2.5%      97.5%
##   spatial ~~
##     verbal       free       0.481      0.065      0.354      0.609
##     speed        free       0.196      0.098      0.004      0.389
##   verbal ~~
##     speed        free       0.160      0.077      0.008      0.312
##
## Variances:
##                  Type    Estimate    Std.Err       2.5%      97.5%
##     spatial      fixed      1.000                 1.000      1.000
##     verbal       fixed      1.000                 1.000      1.000
##     speed        fixed      1.000                 1.000      1.000
##    .x1           free       0.623      0.095      0.438      0.809
##    .x2           free       1.110      0.099      0.917      1.304
##    .x3           free       0.748      0.092      0.567      0.930
##    .x4           free       0.380      0.048      0.287      0.473
##    .x5           free       0.418      0.059      0.303      0.533
##    .x6           free       0.363      0.043      0.279      0.447
##    .x7           free       0.669      0.097      0.479      0.859
##    .x8           free       0.444      0.087      0.273      0.616
##    .x9           free       0.560      0.059      0.444      0.676
```

The penalty matrix $\boldsymbol{\mathcal{S}}_{\hat{\eta}}(\hat{\boldsymbol{\theta}})$ at convergence is stored in the slot `@Penalize`. It is a diagonal matrix with the elements on the diagonal quantifying the extent to which each model parameter has been penalized.

```
round(diag(fit@Penalize@Sh.info$S.h), 2)
```

```
##     spatial=~x1     spatial=~x2     spatial=~x3     spatial=~x5
##            7.64           16.02            7.47          639.57
##     spatial=~x6     spatial=~x8     spatial=~x9      verbal=~x2
##       626389.20          140.69           23.27       427303.89
##      verbal=~x3      verbal=~x4      verbal=~x5      verbal=~x6
##           99.47            5.44            3.88            6.62
##      verbal=~x8      verbal=~x9       speed=~x2       speed=~x3
##       246589.16       347789.43         2446.04      4332622.32
##       speed=~x5       speed=~x6       speed=~x7       speed=~x8
##      6419433.77      2587290.17            9.63           10.76
##       speed=~x9          x1~~x1          x2~~x2          x3~~x3
##           28.16            0.00            0.00            0.00
##          x4~~x4          x5~~x5          x6~~x6          x7~~x7
##            0.00            0.00            0.00            0.00
##          x8~~x8          x9~~x9 spatial~~verbal  spatial~~speed
##            0.00            0.00            0.00            0.00
##   verbal~~speed
##            0.00
```

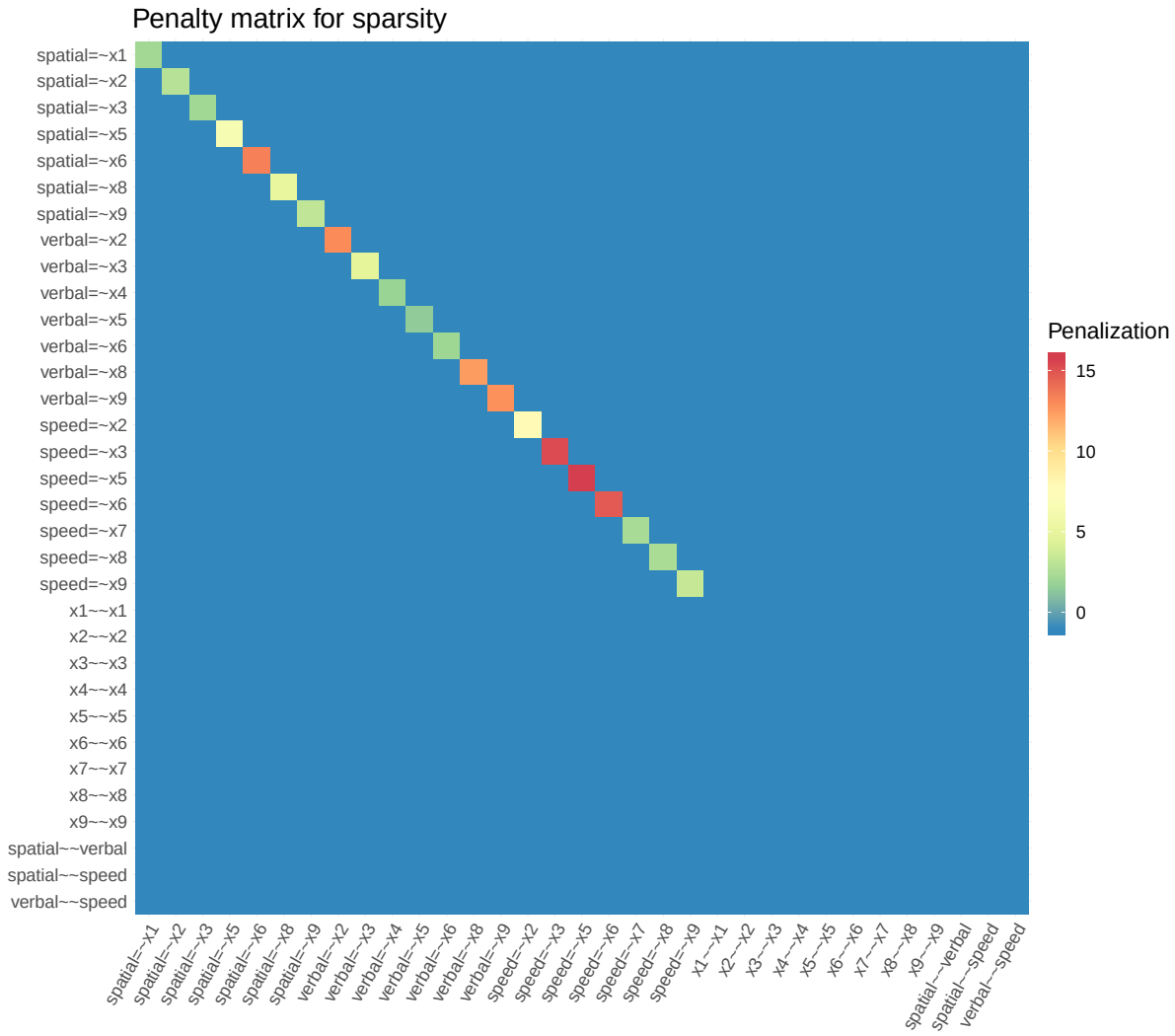


Figure 7.1: Heat map of the penalty matrix $\boldsymbol{\mathcal{S}}^A_{\hat{\eta}}(\hat{\boldsymbol{\theta}})$ on a log-scale for `GJRM-alasso` ($a = 1, \gamma = 4.5$) on the Holzinger & Swineford data set.

The values corresponding to the factor loadings are different from zero, as these are the parameters that have been penalized, whereas the values for the unique variances (`x1~~x1` to `x9~~x9`) and the factor covariances (`spatial~~verbal`, `spatial~~speed`, `verbal~~speed`) are zero, as these elements were not affected by the penalization. The magnitude of the penalization varied depending on the size of the factor loading to be penalized: small loadings received a considerable penalty, whereas large loadings a little one. Figure 7.1 shows the heat map of the penalty matrix $\boldsymbol{\mathcal{S}}^A_{\hat{\eta}}(\hat{\boldsymbol{\theta}})$ on a log-scale, given the wide range of its elements (from 0 to over 6×10^6).

7.2 Penalized estimation of a multiple-group factor model

As a followup, we consider the penalized estimation of a multiple-group factor model with the alasso penalty and the automatic multiple tuning procedure (Section 6.2). Interestingly, there are now multiple tuning parameters: one of them introduces sparsity in the factor loading matrices of each of the groups, whereas the other two encourage cross-group invariance of loadings and intercepts. For this example, we use the complete version of the Holzinger & Swineford data set in the `R` package `MBESS` (Kelley, 2019). An inspection at the data set structure reveals that `HS.data` contains the scores on 26 tests from $N = 301$ students attending the Pasteur and Grant-White schools. We analyze the subset consisting of the first $p = 19$ tests, which we standardized to handle the scaling effect. The variables were also renamed for convenience when formulating the syntax.

```
data <- HS.data[, 6:25]
summary(data)
```

```
##          school        visual          cubes           paper
##  Grant-White:145   Min.   : 4.00   Min.   : 9.00   Min.   : 6.00
##  Pasteur    :156   1st Qu.:25.00   1st Qu.:21.00   1st Qu.:12.00
##                    Median :30.00   Median :24.00   Median :14.00
##                    Mean   :29.61   Mean   :24.35   Mean   :14.23
##                    3rd Qu.:34.00   3rd Qu.:27.00   3rd Qu.:16.00
##                    Max.   :51.00   Max.   :37.00   Max.   :25.00
##      flags        general         paragrap        sentence
##  Min.   : 2   Min.   : 8.00   Min.   : 0.000   Min.   : 4.00
##  1st Qu.:11   1st Qu.:31.00   1st Qu.: 7.000   1st Qu.:14.00
##  Median :17   Median :41.00   Median : 9.000   Median :18.00
##  Mean   :18   Mean   :40.62   Mean   : 9.183   Mean   :17.36
##  3rd Qu.:25   3rd Qu.:49.00   3rd Qu.:11.000   3rd Qu.:21.00
##  Max.   :36   Max.   :84.00   Max.   :19.000   Max.   :28.00
##      wordc           wordm         addition         code
##  Min.   :10.00   Min.   : 1.0   Min.   : 30.00   Min.   : 19.00
##  1st Qu.:23.00   1st Qu.:10.0   1st Qu.: 80.00   1st Qu.: 60.00
##  Median :26.00   Median :14.0   Median : 94.00   Median : 68.00
##  Mean   :26.13   Mean   :15.3   Mean   : 96.24   Mean   : 69.16
##  3rd Qu.:30.00   3rd Qu.:19.0   3rd Qu.:113.00   3rd Qu.: 79.00
##  Max.   :43.00   Max.   :43.0   Max.   :171.00   Max.   :118.00
```

```
##     counting        straight         wordr          numberr
##  Min.   : 61.0   Min.   :100.0   Min.   :121.0   Min.   : 68
##  1st Qu.: 97.0   1st Qu.:171.0   1st Qu.:168.0   1st Qu.: 84
##  Median :110.0   Median :195.0   Median :176.0   Median : 90
##  Mean   :110.5   Mean   :193.4   Mean   :175.2   Mean   : 90
##  3rd Qu.:122.0   3rd Qu.:219.0   3rd Qu.:184.0   3rd Qu.: 96
##  Max.   :200.0   Max.   :333.0   Max.   :198.0   Max.   :112
##     figurer          object          numberf          figurew
##  Min.   : 58.0   Min.   : 0.000   Min.   : 0.000   Min.   : 3.00
##  1st Qu.: 98.0   1st Qu.: 5.000   1st Qu.: 6.000   1st Qu.:11.00
##  Median :103.0   Median : 8.000   Median : 9.000   Median :14.00
##  Mean   :102.5   Mean   : 8.216   Mean   : 9.395   Mean   :14.02
##  3rd Qu.:107.0   3rd Qu.:11.000   3rd Qu.:12.000   3rd Qu.:17.00
##  Max.   :119.0   Max.   :26.000   Max.   :20.000   Max.   :20.00

data[, 2:20]          <- scale(data[, 2:20])
colnames(data)[2:20] <- paste0("x", 1:19)
```

7.2.1 Model specification

The syntax becomes more elaborate, due to the additional specification of the mean structure.

```
syntax.mg <-'
# Measurement model
spatial =~ 1*x1  + x2 + x3 + x4 + x5 + x6 + x7 + x8 + 0*x9 + x10 +
             x11 + 0*x12 + x13 + 0*x14 + x15 + x16 + x17 + x18 + x19
verbal  =~ 0*x1  + x2 + x3 + x4 + x5 + x6 + x7 + x8 + 1*x9 + x10 +
             x11 + 0*x12 + x13 + 0*x14 + x15 + x16 + x17 + x18 + x19
speed   =~ 0*x1  + x2 + x3 + x4 + x5 + x6 + x7 + x8 + 0*x9 + x10 +
             x11 + 1*x12 + x13 + 0*x14 + x15 + x16 + x17 + x18 + x19
memory  =~ 0*x1  + x2 + x3 + x4 + x5 + x6 + x7 + x8 + 0*x9 + x10 +
             x11 + 0*x12 + x13 + 1*x14 + x15 + x16 + x17 + x18 + x19

# Estimate intercepts
x2 + x3 +  x4 + x5 + x6 + x7 + x8 + x10 + x11 +
                           x13 + x15 + x16 + x17 + x18 + x19 ~ 1
# Fixed intercepts
x1 + x9 + x12 + x14 ~ 0*1

# Structural model
spatial ~~ NA*spatial
verbal  ~~ NA*verbal
speed   ~~ NA*speed
memory  ~~ NA*memory
```

```
spatial ~ NA*1
verbal  ~ NA*1
speed   ~ NA*1
memory  ~ NA*1
'
```

The mean structure can be explicitly introduced by including "intercept formulas" in the model syntax. These expressions are constituted by the name of the variable, followed by the tilde operator "`~`", and the number 1. If the variable appearing in the formula is an observed variable, then the formula specifies the intercept term for that item; if the variable is latent (i.e., a common factor), then the formula specifies a factor mean. To avoid clutter, if users desire to introduce intercepts for multiple variables, they can specify on the left-hand side all the variables of interest, followed by plus ("+") signs. By default, the factor means are fixed to zero. Provided that identification restrictions are applied, users can force the estimation of any model parameter by pre-multiplying the variable name on the right-hand side by `NA`. This is done in the syntax for the means and the variances of the common factors.

The syntax above specifies a factor model with $r = 4$ factors and $p = 19$ observed variables. The metric of the factors is accommodated through the "marker-variable" approach, with the markers being x_1, x_9, x_{12}, x_{14}. The structural model is freely estimated. The fact that the syntax should prompt a multiple-group analysis will be communicated to the fitting function `penfa` through proper arguments (see below for details). By default, the model in the syntax is fitted to all groups.

Before carrying out the penalized estimation, we fit the unpenalized model to obtain the maximum likelihood estimates to be used as weights for the alasso. To facilitate the estimation process, we can provide informative starting values to (some of) the parameters. This can be done through the pre-multiplication mechanism employed to fix some parameter values, but the numeric constant becomes the argument of the function `start`. To fix parameters or provide starting values in case of multiple groups, we use the same pre-multiplication mechanism, but the numeric argument is a vector of arguments, one for each group. When

users provide a single value instead of a vector of values, that element is applied for all groups. The syntax below provides a starting value equal to 0.8 to the primary loadings of all factors.

```
syntax.mle.mg <- '
# Measurement model + starting values
spatial =~  1*x1 + start(0.8)*x2 + start(0.8)*x3 + start(0.8)*x4 +
            x5 + x6 + x7 + x8 + 0*x9 + x10 + x11 + 0*x12 + x13 +
            0*x14 + x15 + x16 + x17 + x18 + x19

verbal  =~  0*x1 + x2 + x3 + x4 + start(0.8)*x5 + start(0.8)*x6 +
            start(0.8)*x7 + start(0.8)*x8 + 1*x9 + x10 + x11 +
            0*x12 + x13 + 0*x14 + x15 + x16 + x17 + x18 + x19

speed   =~  0*x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + 0*x9 +
            start(0.8)*x10 + start(0.8)*x11 + 1*x12 +
            start(0.8)*x13 + 0*x14 + x15 + x16 + x17 + x18 + x19

memory  =~  0*x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + 0*x9 + x10 +
            x11 + 0*x12 + x13 + 1*x14 + start(0.8)*x15 +
            start(0.8)*x16 + start(0.8)*x17 + start(0.8)*x18 +
            start(0.8)*x19

# Estimate intercepts
x2 + x3 +  x4 + x5 + x6 + x7 + x8 + x10 + x11 + x13 + x15 + x16 +
                                                  x17 + x18 + x19 ~ 1
# Fix intercepts
x1 + x9 + x12 + x14 ~ 0*1

# Structural model
spatial ~~ NA*spatial
verbal  ~~ NA*verbal
speed   ~~ NA*speed
memory  ~~ NA*memory

spatial ~ NA*1
verbal  ~ NA*1
speed   ~ NA*1
memory  ~ NA*1 '
```

As for the single-group analysis, the fit of the unpenalized multiple-group factor model is carried out through the `penfa` function, with the specification of two new arguments: `meanstructure` and `group`. The argument `meanstructure` is set to `TRUE` to obtain the estimates of the means of the observed and the latent variables.

In the `group` argument, we indicate the name of the group variable in the data set, which is the "school" attended by the students.

```
fit.mle.mg <- penfa(model = syntax.mle.mg, data = data,
                    information = "fisher", meanstructure = TRUE,
                    group = "school", shrink = "none")
weights.mg <- coef(fit.mle.mg)
fit.mle.mg

## GJRM reached convergence
##
##   Number of observations per group:
##     Pasteur                                                156
##     Grant-White                                            145
##
##   Estimator                                                MLE
##   Optimization method                                trust-region
##   Information                                            expected
##   Strategy                                                  grid
##   Number of iterations                                        21
##   Effective degrees of freedom                           216.000
##
```

7.2.2 Model fitting

We can now proceed with the estimation of the penalized multiple-group factor model with the alasso penalization and the automatic tuning procedure to find the optimal value of the tuning parameter vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)^T$. The penalty function employed to shrink the pairwise group differences of the factor loadings and the intercepts can be specified through the `diff` argument. The argument `eta` is now a list that determines the starting values for each of the tuning parameters on the specified parameter matrices and vectors.

```
fit.mg <- penfa(model = syntax.mg, data = data,
                information = "fisher", meanstructure = TRUE,
                group = "school", shrink = "alasso", diff ="alasso",
                weights = weights.mg, a.alasso = 1,
                eta = list("shrink"=c("lambda" = 0.01),
                           "diff"  =c("lambda" = 0.1, "tau" =0.01)),
                strategy = "auto", gamma = 4)
```

From the summary of the fitted object, we can notice that the automatic tuning procedure required just a couple of iterations to converge. The optimal tuning parameters are $\hat{\eta}_1 = 0.006, \hat{\eta}_2 = 16221.852$ and $\hat{\eta}_3 = 0.013$. The analysis benefited from the encouragement of sparsity and loading and intercept invariance, as it is evident from the reduction in the GBIC after the penalization (from 15123.43 for the unpenalized model to 14658 for the penalized model).

```
summary(fit.mg)

## GJRM reached convergence
##
##   Number of observations per group:
##     Pasteur                                          156
##     Grant-White                                      145
##   Number of groups                                     2
##   Number of observed variables                        19
##   Number of latent factors                             4
##
##   Estimator                                         PMLE
##   Optimization method                        trust-region
##   Information                                    expected
##   Strategy                                           auto
##   Number of iterations (total)                        347
##   Number of two-steps (automatic)                       5
##   Influence factor                                      4
##   Number of parameters:
##     Free                                               66
##     Penalized                                         150
##   Effective degrees of freedom                    109.242
##   GIC                                           14253.027
##   GBIC                                          14657.998
##
##   Penalty functions:
##     Sparsity                                        alasso
##     Invariance                                      alasso
##
##   Additional tuning parameter
##     alasso                                              1
##
##   Optimal tuning parameters:
##     Sparsity
##      - Factor loadings                              0.006
##     Invariance
##      - Factor loadings                          16221.852
##      - Intercepts                                   0.013
##
##
```

```
## Parameter Estimates:
##
## Group 1 [Pasteur]:
##
## Latent Variables:
##                  Type    Estimate    Std.Err       2.5%      97.5%
##   spatial =~
##     x1           fixed      1.000                  1.000      1.000
##     x2           pen        0.583      0.082       0.423      0.744
##     x3           pen        0.618      0.082       0.457      0.779
##     x4           pen        0.863      0.094       0.678      1.047
##     x5           pen       -0.000
##     x6           pen        0.000
##     x7           pen       -0.121      0.045      -0.210     -0.032
##     x8           pen        0.000
##     x9           fixed      0.000                  0.000      0.000
##     x10          pen       -0.401      0.095      -0.588     -0.215
##     x11          pen        0.000
##     x12          fixed      0.000                  0.000      0.000
##     x13          pen        0.397      0.078       0.245      0.550
##     x14          fixed      0.000                  0.000      0.000
##     x15          pen        0.018
##     x16          pen        0.367      0.080       0.211      0.523
##     x17          pen       -0.231      0.077      -0.382     -0.080
##     x18          pen        0.001
##     x19          pen        0.059      0.042       0.024      0.142
##   verbal =~
##     x1           fixed      0.000                  0.000      0.000
##     x2           pen       -0.000
##     x3           pen        0.000
##     x4           pen       -0.087      0.051      -0.187      0.013
##     x5           pen        1.020      0.056       0.910      1.130
##     x6           pen        0.957      0.055       0.849      1.064
##     x7           pen        1.075      0.059       0.960      1.191
##     x8           pen        0.839      0.058       0.725      0.952
##     x9           fixed      1.000                  1.000      1.000
##     x10          pen        0.141      0.064       0.015      0.267
##     x11          pen        0.168      0.052       0.066      0.270
##     x12          fixed      0.000                  0.000      0.000
##     x13          pen       -0.000
##     x14          fixed      0.000                  0.000      0.000
##     x15          pen       -0.143      0.055      -0.250     -0.036
##     x16          pen       -0.000
##     x17          pen        0.000
##     x18          pen        0.000
##     x19          pen        0.000
##   speed =~
##     x1           fixed      0.000                  0.000      0.000
##     x2           pen       -0.000
##     x3           pen        0.000
##     x4           pen       -0.000
```

```
##     x5           pen        0.000                                    
##     x6           pen       -0.000                                    
##     x7           pen       -0.000                                    
##     x8           pen        0.000                                    
##     x9           fixed      0.000                0.000       0.000   
##     x10          pen        0.988       0.113    0.765       1.210   
##     x11          pen        0.744       0.089    0.570       0.918   
##     x12          fixed      1.000                1.000       1.000   
##     x13          pen        0.677       0.087    0.506       0.848   
##     x14          fixed      0.000                0.000       0.000   
##     x15          pen        0.000                                    
##     x16          pen        0.000                                    
##     x17          pen        0.321       0.078    0.168       0.475   
##     x18          pen        0.245       0.070    0.108       0.382   
##     x19          pen        0.093       0.045    0.005       0.181   
##   memory =~                                                          
##     x1           fixed      0.000                0.000       0.000   
##     x2           pen       -0.000                                    
##     x3           pen       -0.000                                    
##     x4           pen        0.000                                    
##     x5           pen       -0.109       0.045   -0.198      -0.020   
##     x6           pen        0.009                                    
##     x7           pen       -0.000                                    
##     x8           pen        0.028                                    
##     x9           fixed      0.000                0.000       0.000   
##     x10          pen        0.145       0.073    0.002       0.288   
##     x11          pen        0.267       0.079    0.113       0.422   
##     x12          fixed      0.000                0.000       0.000   
##     x13          pen       -0.000                                    
##     x14          fixed      1.000                1.000       1.000   
##     x15          pen        0.838       0.110    0.624       1.053   
##     x16          pen        0.632       0.100    0.435       0.828   
##     x17          pen        0.875       0.115    0.649       1.100   
##     x18          pen        0.647       0.098    0.455       0.840   
##     x19          pen        0.533       0.093    0.351       0.714   
##                                                                       
## Covariances:                                                          
##                  Type    Estimate    Std.Err       2.5%      97.5%   
##   spatial ~~                                                          
##     verbal       free       0.281       0.067    0.150       0.411   
##     speed        free       0.158       0.062    0.037       0.278   
##     memory       free       0.174       0.064    0.049       0.300   
##   verbal ~~                                                           
##     speed        free       0.185       0.059    0.071       0.300   
##     memory       free       0.104       0.059   -0.012       0.220   
##   speed ~~                                                            
##     memory       free       0.075       0.057   -0.038       0.187   
##                                                                       
##                  Type    Estimate    Std.Err       2.5%      97.5%   
##    .x2           pen        0.009       0.056   -0.100       0.119   
##    .x3           pen        0.001       0.056   -0.108       0.110   
```

```
##    .x4           pen        0.137      0.070      0.001      0.273
##    .x5           pen       -0.012      0.044     -0.099      0.074
##    .x6           pen       -0.007      0.044     -0.094      0.079
##    .x7           pen       -0.006      0.043     -0.091      0.079
##    .x8           pen       -0.081      0.055     -0.188      0.026
##    .x10          pen        0.145      0.078     -0.008      0.298
##    .x11          pen        0.000      0.053     -0.104      0.104
##    .x13          pen       -0.002      0.052     -0.104      0.099
##    .x15          pen        0.000      0.060     -0.117      0.118
##    .x16          pen        0.016      0.054     -0.090      0.121
##    .x17          pen        0.164      0.077      0.012      0.316
##    .x18          pen       -0.002      0.057     -0.114      0.110
##    .x19          pen       -0.204      0.075     -0.352     -0.057
##    .x1           fixed      0.000                 0.000      0.000
##    .x9           fixed      0.000                 0.000      0.000
##    .x12          fixed      0.000                 0.000      0.000
##    .x14          fixed      0.000                 0.000      0.000
##     spatial      free      -0.021      0.077     -0.173      0.130
##     verbal       free      -0.259      0.073     -0.402     -0.116
##     speed        free       0.089      0.074     -0.055      0.234
##     memory       free      -0.046      0.077     -0.198      0.105
##
## Variances:
##                 Type    Estimate    Std.Err       2.5%      97.5%
##     spatial      free       0.591      0.106      0.384      0.798
##     verbal       free       0.656      0.091      0.477      0.834
##     speed        free       0.441      0.087      0.271      0.612
##     memory       free       0.519      0.104      0.315      0.722
##    .x1           free       0.437      0.079      0.283      0.591
##    .x2           free       0.886      0.107      0.677      1.095
##    .x3           free       0.814      0.099      0.619      1.008
##    .x4           free       0.612      0.087      0.442      0.781
##    .x5           free       0.257      0.038      0.183      0.331
##    .x6           free       0.348      0.046      0.258      0.439
##    .x7           free       0.254      0.039      0.179      0.330
##    .x8           free       0.407      0.051      0.307      0.506
##    .x9           free       0.230      0.035      0.162      0.298
##    .x10          free       0.523      0.085      0.356      0.689
##    .x11          free       0.441      0.061      0.321      0.561
##    .x12          free       0.543      0.084      0.378      0.707
##    .x13          free       0.617      0.082      0.456      0.778
##    .x14          free       0.580      0.091      0.402      0.758
##    .x15          free       0.676      0.092      0.495      0.857
##    .x16          free       0.735      0.094      0.550      0.919
##    .x17          free       0.625      0.089      0.450      0.800
##    .x18          free       0.778      0.096      0.589      0.966
##    .x19          free       0.846      0.101      0.648      1.044
##
##
##
```

```
## Group 2 [Grant-White]:
##
## Latent Variables:
##                  Type    Estimate    Std.Err       2.5%      97.5%
##   spatial =~
##     x1           fixed      1.000                 1.000      1.000
##     x2           pen        0.583      0.082      0.423      0.744
##     x3           pen        0.618      0.082      0.457      0.779
##     x4           pen        0.863      0.094      0.678      1.047
##     x5           pen       -0.000
##     x6           pen        0.000
##     x7           pen       -0.121      0.045     -0.210     -0.032
##     x8           pen        0.000
##     x9           fixed      0.000                 0.000      0.000
##     x10          pen       -0.401      0.095     -0.588     -0.215
##     x11          pen        0.000
##     x12          fixed      0.000                 0.000      0.000
##     x13          pen        0.397      0.078      0.245      0.550
##     x14          fixed      0.000                 0.000      0.000
##     x15          pen        0.018
##     x16          pen        0.367      0.080      0.211      0.523
##     x17          pen       -0.231      0.077     -0.382     -0.080
##     x18          pen        0.001
##     x19          pen        0.059      0.042     -0.024      0.142
##   verbal =~
##     x1           fixed      0.000                 0.000      0.000
##     x2           pen       -0.000
##     x3           pen        0.000
##     x4           pen       -0.087      0.051     -0.187      0.013
##     x5           pen        1.020      0.056      0.910      1.130
##     x6           pen        0.957      0.055      0.849      1.064
##     x7           pen        1.075      0.059      0.960      1.191
##     x8           pen        0.839      0.058      0.725      0.952
##     x9           fixed      1.000                 1.000      1.000
##     x10          pen        0.141      0.064      0.015      0.267
##     x11          pen        0.168      0.052      0.066      0.270
##     x12          fixed      0.000                 0.000      0.000
##     x13          pen       -0.000
##     x14          fixed      0.000                 0.000      0.000
##     x15          pen       -0.143      0.055     -0.250     -0.036
##     x16          pen       -0.000
##     x17          pen        0.000
##     x18          pen        0.000
##     x19          pen        0.000
##   speed =~
##     x1           fixed      0.000                 0.000      0.000
##     x2           pen       -0.000
##     x3           pen        0.000
##     x4           pen       -0.000
##     x5           pen        0.000
##     x6           pen       -0.000
```

```
##     x7           pen       -0.000
##     x8           pen        0.000
##     x9           fixed      0.000                0.000      0.000
##     x10          pen        0.988       0.113    0.765      1.210
##     x11          pen        0.744       0.089    0.570      0.918
##     x12          fixed      1.000                1.000      1.000
##     x13          pen        0.677       0.087    0.506      0.848
##     x14          fixed      0.000                0.000      0.000
##     x15          pen        0.000
##     x16          pen        0.000
##     x17          pen        0.321       0.078    0.168      0.475
##     x18          pen        0.245       0.070    0.108      0.382
##     x19          pen        0.093       0.045    0.005      0.181
##   memory =~
##     x1           fixed      0.000                0.000      0.000
##     x2           pen       -0.000
##     x3           pen       -0.000
##     x4           pen        0.000
##     x5           pen       -0.109       0.045   -0.198     -0.020
##     x6           pen        0.009
##     x7           pen       -0.000
##     x8           pen        0.028
##     x9           fixed      0.000                0.000      0.000
##     x10          pen        0.145       0.073    0.002      0.288
##     x11          pen        0.267       0.079    0.113      0.422
##     x12          fixed      0.000                0.000      0.000
##     x13          pen       -0.000
##     x14          fixed      1.000                1.000      1.000
##     x15          pen        0.838       0.110    0.624      1.053
##     x16          pen        0.632       0.100    0.435      0.828
##     x17          pen        0.875       0.115    0.649      1.100
##     x18          pen        0.647       0.098    0.455      0.840
##     x19          pen        0.533       0.093    0.351      0.714
##
## Covariances:
##                  Type    Estimate    Std.Err       2.5%      97.5%
##   spatial ~~
##     verbal       free       0.363       0.071    0.223      0.503
##     speed        free       0.289       0.074    0.143      0.434
##     memory       free       0.242       0.064    0.117      0.367
##   verbal ~~
##     speed        free       0.231       0.067    0.100      0.362
##     memory       free       0.257       0.061    0.138      0.375
##   speed ~~
##     memory       free       0.158       0.062    0.037      0.279
##
## Intercepts:
##                  Type    Estimate    Std.Err       2.5%      97.5%
##    .x2           pen        0.011       0.056   -0.098      0.121
##    .x3           pen        0.001       0.056   -0.108      0.110
##    .x4           pen       -0.163       0.067   -0.294     -0.032
```

```
##    .x5           pen       -0.008      0.044     -0.095      0.079
##    .x6           pen       -0.007      0.044     -0.094      0.079
##    .x7           pen       -0.006      0.043     -0.091      0.079
##    .x8           pen        0.074      0.059     -0.041      0.189
##    .x10          pen       -0.179      0.072     -0.319     -0.038
##    .x11          pen       -0.000      0.053     -0.104      0.104
##    .x13          pen       -0.002      0.052     -0.104      0.099
##    .x15          pen       -0.000      0.060     -0.118      0.117
##    .x16          pen        0.016      0.054     -0.089      0.121
##    .x17          pen       -0.191      0.073     -0.335     -0.048
##    .x18          pen       -0.002      0.057     -0.114      0.110
##    .x19          pen        0.235      0.068      0.102      0.369
##    .x1           fixed      0.000                 0.000      0.000
##    .x9           fixed      0.000                 0.000      0.000
##    .x12          fixed      0.000                 0.000      0.000
##    .x14          fixed      0.000                 0.000      0.000
##     spatial      free       0.023      0.080     -0.134      0.180
##     verbal       free       0.289      0.075      0.141      0.436
##     speed        free      -0.085      0.082     -0.246      0.075
##     memory       free       0.052      0.075     -0.095      0.199
##
## Variances:
##                  Type    Estimate    Std.Err       2.5%      97.5%
##     spatial      free       0.597      0.108      0.385      0.808
##     verbal       free       0.625      0.092      0.445      0.805
##     speed        free       0.627      0.115      0.402      0.851
##     memory       free       0.420      0.088      0.248      0.591
##    .x1           free       0.435      0.074      0.290      0.579
##    .x2           free       0.683      0.086      0.515      0.851
##    .x3           free       0.712      0.090      0.536      0.888
##    .x4           free       0.472      0.070      0.334      0.610
##    .x5           free       0.311      0.045      0.222      0.400
##    .x6           free       0.313      0.044      0.226      0.400
##    .x7           free       0.219      0.037      0.147      0.292
##    .x8           free       0.446      0.058      0.333      0.560
##    .x9           free       0.346      0.049      0.250      0.442
##    .x10          free       0.335      0.067      0.203      0.467
##    .x11          free       0.615      0.082      0.454      0.775
##    .x12          free       0.442      0.075      0.295      0.590
##    .x13          free       0.443      0.065      0.315      0.570
##    .x14          free       0.556      0.084      0.392      0.719
##    .x15          free       0.674      0.091      0.496      0.851
##    .x16          free       0.471      0.065      0.344      0.598
##    .x17          free       0.464      0.069      0.328      0.601
##    .x18          free       0.649      0.083      0.487      0.812
##    .x19          free       0.600      0.075      0.453      0.746
```

The diagonal elements of the penalty matrix $\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}^{A}(\hat{\boldsymbol{\theta}})$ are roughly in the range $[-3 \times 10^{12}, 3 \times 10^{12}]$. In Figure 7.2a, we find the heat map of the penalty matrix

$\boldsymbol{\mathcal{D}}^{A}_{\hat{\eta}_1}(\hat{\boldsymbol{\theta}})$, which shrinks the small factor loadings of each group to zero. Because the range of the diagonal elements of the penalty matrix is very wide, we employed the log-scale. The non-zero diagonal elements correspond to the factor loadings of the two groups. All of the remaining entries of the penalty matrix are equal to zero. Figure 7.2b represents the heat map of the penalty matrix $\boldsymbol{\mathcal{D}}^{A}_{\hat{\eta}_2}(\hat{\boldsymbol{\theta}})$, which shrinks the pairwise group differences of the factor loadings towards zero. Similarly, the heat map of the penalty matrix $\boldsymbol{\mathcal{D}}^{A}_{\hat{\eta}_3}(\hat{\boldsymbol{\theta}})$ shrinks the pairwise group differences of the intercepts, and is depicted in Figure 7.2c.

Further details and options can be found in the documentation of the `R` package `GJRM` (https://cran.r-project.org/web/packages/GJRM/GJRM.pdf).

(a) *Heat map of* $\mathcal{D}^A_{\hat{\eta}_1}(\hat{\boldsymbol{\theta}})$ *on a log-scale.*

(b) *Heat map of* $\mathcal{D}^A_{\hat{\eta}_2}(\hat{\boldsymbol{\theta}})$ *on a log-scale.*

(c) *Heat map of* $\mathcal{D}^A_{\hat{\eta}_3}(\hat{\boldsymbol{\theta}})$ *on a log-scale.*

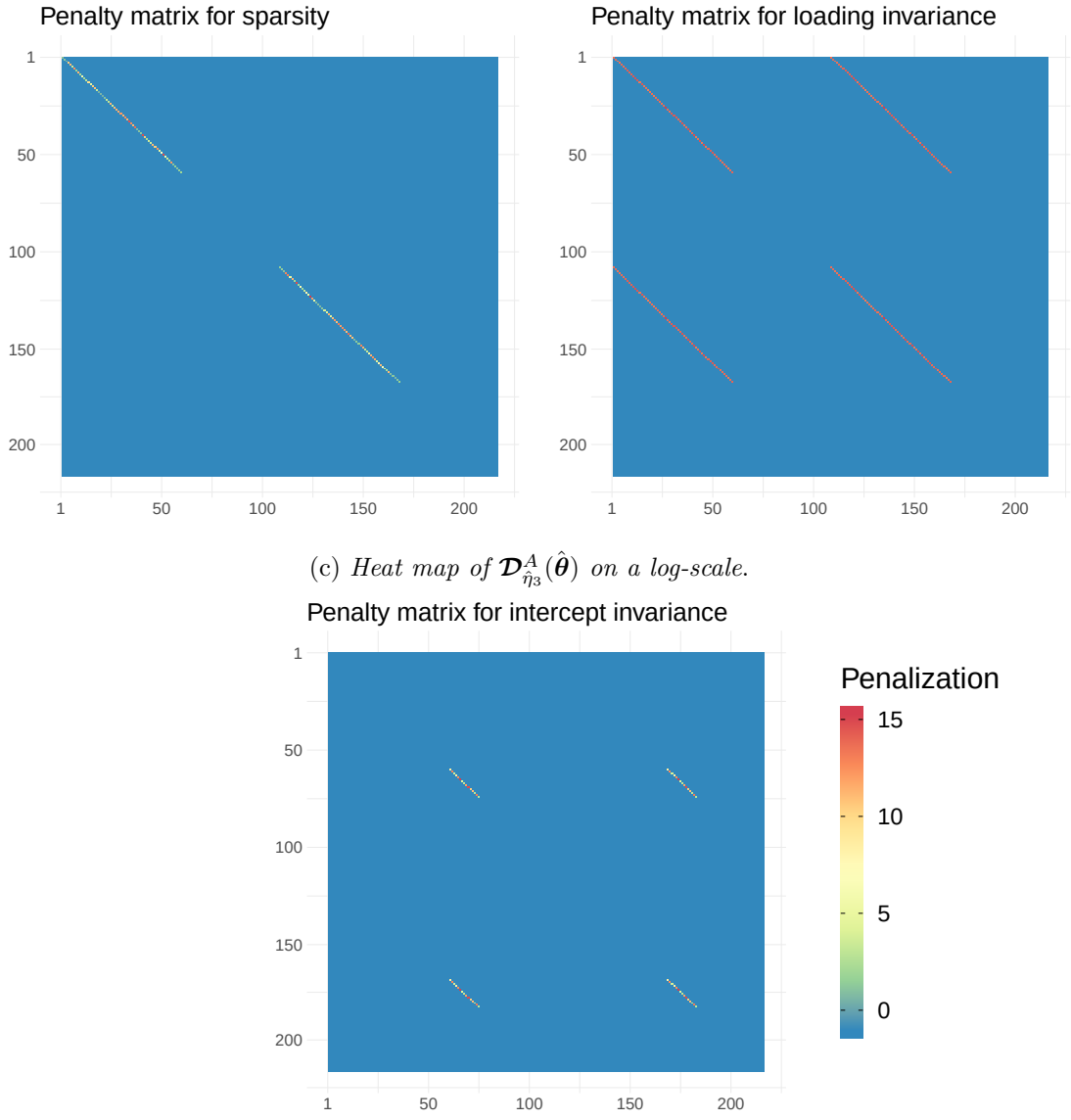


Figure 7.2: Representation of the penalty matrices for sparsity of the factor loadings and loading and intercept invariance on a log-scale for `GJRM-alasso` ($a = 1, \gamma = 4$) on the Holzinger & Swineford data set.

CHAPTER 8

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 singular at the origin, and thus non-differentiable. In this thesis, we adopted suitable local approximations of them. In this way, in the optimization process it was possible to employ 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 non-zero 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 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, but if more sparsity is desired, researchers 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 results described in this work were derived under the $N > p$ scenario with p a moderate number of indicators, as it is the case for many applications from the social and behavioral sciences. We tested the methodology in the frameworks common to confirmatory analyses, with the advantage of letting the zero loadings - as well as group-invariant measurement model parameters in the multiple-group case - freely emerge as a result of the penalization, as opposed to fixing their values

or constraining them to equivalence. Therefore, researchers are requested to have already an idea about the number of underlying factors and the observed variables serving as proper indicators of such latent constructs. However, penalized techniques can also be extremely useful in presence of many observed variables or in the high-dimensional case. Under the latter scenario, the sample covariance matrix of the observed variables is not positive-definite, which makes maximum likelihood estimation infeasible. Consequently, weights other than the maximum likelihood estimates should be used for the computation of the alasso penalty. It would be interesting to review and adapt the presented methodology in this demanding set-up.

Secondly, 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.

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).

APPENDIX

Details on the normal linear factor analysis model

A.1 Log-likelihood

For a random sample of deviation scores $\boldsymbol{x}_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of size N from a multivariate normal distribution, the likelihood function is:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) = \prod_{\alpha=1}^{N} f(\boldsymbol{x}_\alpha|\boldsymbol{\theta}) &= \prod_{\alpha=1}^{N} \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\boldsymbol{x}_\alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_\alpha\right\} \\ &= (2\pi)^{-\frac{N}{2}p}|\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{N}\boldsymbol{x}_\alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_\alpha\right\}, \end{aligned}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$. The log-likelihood, which is defined as the logarithm of $\mathcal{L}(\boldsymbol{\theta})$, takes the following form:

$$\begin{aligned} \ell(\boldsymbol{\theta}) \coloneqq \log\mathcal{L}(\boldsymbol{\theta}) &= -\frac{N}{2}p\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{\alpha=1}^{N}\boldsymbol{x}_\alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_\alpha \\ &= -\frac{N}{2}p\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{\alpha=1}^{N}\mathrm{tr}\left\{\boldsymbol{x}_\alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_\alpha\right\} \\ &= -\frac{N}{2}p\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\left\{\sum_{\alpha=1}^{N}\boldsymbol{x}_\alpha\boldsymbol{x}_\alpha^T\boldsymbol{\Sigma}^{-1}\right\} \\ &= -\frac{N}{2}p\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\right\} \end{aligned}$$

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

where $\boldsymbol{S} = \frac{1}{N}\sum_{\alpha=1}^{N} \boldsymbol{x}_\alpha \boldsymbol{x}_\alpha^T$ is the sample covariance matrix which could be estimated by maximum likelihood. Since $\boldsymbol{S}$ is a sufficient statistic for $\boldsymbol{\theta}$, it suffices the sample covariance matrix, and not the individual $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, to estimate the parameter vector and its covariance matrix.

The log-likelihood is made up by the determinant and the trace, which summarize important information about the matrices $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$. The determinant is a single number that reflects a generalized measure of variance for the entire set of variables contained in the matrix, whereas the trace of a matrix is the sum of the values on the diagonal. The objective of maximum likelihood is to minimize these matrix summaries.

The derivation of the log-likelihood function was established under the multivariate normality assumption of the observed variables. An alternative formulation, especially employed in the early days of factor analysis, starts with the assumption of a Wishart distribution for the unbiased sample covariance matrix.

A.2 Gradient, Hessian and Fisher information

Propositions 2.1-2.3 in Section 2.1 state the general expressions of the gradient of the log-likelihood $\boldsymbol{g}(\boldsymbol{\theta}) \coloneqq \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, the Hessian matrix of the second-order derivatives $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) \coloneqq \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$, and the expected Fisher information $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) \coloneqq \mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})^T] = -\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})\right]$ for the normal linear factor model. We now enunciate the specific forms of these derivatives with respect to each parameter matrix.

Proposition A.1 (First-order derivatives of the normal linear factor model with respect to the parameter matrices)**.** *The matrix expressions of the first-order derivatives of the log-likelihood of the normal linear factor analysis model in equation* (2.1) *with respect to the parameter matrices are:*

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Lambda}} = -N\, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}, \tag{A.2}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Phi}} = \begin{cases} -N\, \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} & \textit{non-diagonal elements}, \\ -\dfrac{N}{2}\, \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} & \textit{diagonal elements}, \end{cases} \tag{A.3}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Psi}} = -\frac{N}{2} \mathit{diag}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}). \tag{A.4}$$

Proof. See Appendix A.2.1.2. ■

Define the following matrices:

$$\boldsymbol{\omega} = \boldsymbol{\Sigma}^{-1}, \qquad \boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, \qquad \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi},$$

$$\boldsymbol{\gamma} = \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, \qquad \boldsymbol{\delta} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, \qquad \boldsymbol{\zeta} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi},$$

$$\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}, \qquad \boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{M}.$$

Proposition A.2 (Second-order derivatives of the normal linear factor model with respect to the parameter matrices)**.** *The Hessian of the normal linear factor analysis model in equation (2.1) is a symmetric block matrix of the form:*

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) = \left[\begin{array}{c:c:c} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} & \boldsymbol{H}_{13} \\ \hdashline \boldsymbol{H}_{12}^T & \boldsymbol{H}_{22} & \boldsymbol{H}_{23} \\ \hdashline \boldsymbol{H}_{13}^T & \boldsymbol{H}_{23}^T & \boldsymbol{H}_{33} \end{array}\right], \tag{A.5}$$

where, for $i, t = 1, \ldots, p$ and $g, h, j, l, q, s = 1, \ldots, r$, the sub-matrices are:

$$[\boldsymbol{H}_{11}]_{(ij,ts)} = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \lambda_{ts}} = -N \Big\{ [\boldsymbol{\Omega}]_{ti} [\boldsymbol{\Phi} - \boldsymbol{\zeta}]_{sj} + \omega_{ti} [\boldsymbol{\Phi} \boldsymbol{\Lambda}^T \boldsymbol{M} \boldsymbol{\Lambda} \boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Omega} \boldsymbol{\Lambda} \boldsymbol{\Phi}]_{is} \beta_{tj} + \beta_{is} [\boldsymbol{M} \boldsymbol{\Lambda} \boldsymbol{\Phi}]_{tj} \Big\}, \tag{A.6}$$

$$[\boldsymbol{H}_{12}]_{(ij,tt)} = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \psi_{tt}} = -N \left\{ [\boldsymbol{M}]_{it} \beta_{tj} - \omega_{it} [\boldsymbol{\Omega} \boldsymbol{\Lambda} \boldsymbol{\Phi}]_{tj} \right\}, \tag{A.7}$$

$$\begin{aligned} [\boldsymbol{H}_{13}]_{(ij,gh)} &= \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \phi_{gh}} \\ &= -\frac{N}{2} \Big\{ (2 - [\boldsymbol{I}]_{gh}) \Big([\boldsymbol{\Omega} \boldsymbol{\Lambda}]_{ig} [\boldsymbol{I} - \boldsymbol{\delta}^T]_{hj} + [\boldsymbol{\Omega} \boldsymbol{\Lambda}]_{ih} [\boldsymbol{I} - \boldsymbol{\delta}^T]_{gj} \\ &\qquad + \alpha_{ig} [\boldsymbol{\Lambda}^T \boldsymbol{M} \boldsymbol{\Lambda} \boldsymbol{\Phi}]_{hj} + \alpha_{ih} [\boldsymbol{\Lambda}^T \boldsymbol{M} \boldsymbol{\Lambda} \boldsymbol{\Phi}]_{gj} \Big) \Big\}, \end{aligned} \tag{A.8}$$

$$[\boldsymbol{H}_{22}]_{(ii,tt)} = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{ii} \partial \psi_{tt}} = -\frac{N}{2} \left\{ \omega_{it} [2\boldsymbol{M} - \boldsymbol{\omega}]_{it} \right\}, \tag{A.9}$$

$$[\boldsymbol{H}_{23}]_{(tt,gh)} = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{tt} \partial \phi_{gh}} = -\frac{N}{2} \Big\{ (2 - [\boldsymbol{I}]_{gh}) \Big(\alpha_{th} [\boldsymbol{M} \boldsymbol{\Lambda}]_{tg} - \alpha_{tg} [\boldsymbol{\Omega} \boldsymbol{\Lambda}]_{th} \Big) \Big\}, \tag{A.10}$$

$$\begin{aligned} [\boldsymbol{H}_{33}]_{(gh,lq)} &= \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \phi_{gh} \partial \phi_{lq}} \\ &= -\frac{N}{2} \Big\{ (2 - [\boldsymbol{I}]_{lq} - [\boldsymbol{I}]_{gh} + [\boldsymbol{I}]_{lq} [\boldsymbol{I}]_{gh}) ([\boldsymbol{\alpha}^T \boldsymbol{S} \boldsymbol{\alpha}]_{hl} \gamma_{qg} - \gamma_{hl} [\boldsymbol{\Lambda}^T \boldsymbol{\Omega} \boldsymbol{\Lambda}]_{qg}) \\ &\qquad + (2 - [\boldsymbol{I}]_{lq} - [\boldsymbol{I}]_{gh}) ([\boldsymbol{\Lambda}^T \boldsymbol{M} \boldsymbol{\Lambda}]_{gl} \gamma_{qh} - \gamma_{gl} [\boldsymbol{\Lambda}^T \boldsymbol{\Omega} \boldsymbol{\Lambda}]_{qh}) \Big\}. \end{aligned} \tag{A.11}$$

Proof. See Appendix A.2.2.2. ■

The above exact expressions for the second-order derivatives have a complicated form, and a considerable amount of computation is required to evaluate them all at each iteration of the optimization algorithm. Despite the complexity in getting their exact expressions, it is easy to find good approximations of them by employing the expected Fisher information matrix. We shall henceforth assume that N is reasonably large.

Proposition A.3 (Elements of the expected Fisher information of the normal linear factor model with respect to the parameter matrices)**.** *The expected Fisher information matrix of the normal linear factor analysis model in equation (2.1) is a block matrix of the form:*

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) = \left[\begin{array}{c:c:c} \boldsymbol{J}_{11} & \boldsymbol{J}_{12} & \boldsymbol{J}_{13} \\ \hdashline \boldsymbol{J}_{12}^T & \boldsymbol{J}_{22} & \boldsymbol{J}_{23} \\ \hdashline \boldsymbol{J}_{13}^T & \boldsymbol{J}_{23}^T & \boldsymbol{J}_{33} \end{array}\right], \tag{A.12}$$

where, for $i, t = 1, \ldots, p$ and $g, h, j, l, q, s = 1, \ldots, r$, the sub-matrices are:

$$[\boldsymbol{J}_{11}]_{(ij,ts)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \lambda_{ts}}\right] = N(\beta_{is}\beta_{tj} + \omega_{it}\zeta_{js}), \tag{A.13}$$

$$[\boldsymbol{J}_{12}]_{(ij,tt)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \psi_{tt}}\right] = N\omega_{it}\beta_{tj}, \tag{A.14}$$

$$[\boldsymbol{J}_{13}]_{(ij,gh)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \phi_{gh}}\right] = \frac{N}{2}(2 - [\boldsymbol{I}]_{gh})(\alpha_{ig}\delta_{jh} + \alpha_{ih}\delta_{jg}), \tag{A.15}$$

$$[\boldsymbol{J}_{22}]_{(ii,tt)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{ii} \partial \psi_{tt}}\right] = \frac{N}{2}\omega_{it}^2, \tag{A.16}$$

$$[\boldsymbol{J}_{23}]_{(tt,gh)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{tt} \partial \phi_{gh}}\right] = \frac{N}{2}(2 - [\boldsymbol{I}]_{gh})\alpha_{tg}\alpha_{th}, \tag{A.17}$$

$$[\boldsymbol{J}_{33}]_{(gh,lq)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \phi_{gh} \partial \phi_{lq}}\right] \tag{A.18}$$

$$= \frac{N}{4}(2 - [\boldsymbol{I}]_{gh})(2 - [\boldsymbol{I}]_{lq})(\gamma_{gl}\gamma_{hq} + \gamma_{gq}\gamma_{hl}). \tag{A.19}$$

Proof. See Appendix A.2.3.1. ■

Alternatively, the Fisher information matrix can be formulated more compactly as follows. Let $\text{vec}(\boldsymbol{B})$ be the vector stacking the columns of a $p \times p$ matrix $\boldsymbol{B}$, and $\text{vech}(\boldsymbol{B})$ the vector that contains only the $p^\star = \frac{p(p+1)}{2}$ non duplicated elements of $\boldsymbol{B}$ by leaving out the elements above the diagonal. Let $\boldsymbol{D}$ be the $p^2 \times p^\star$ duplication matrix (Magnus & Neudecker, 2019) such that $\text{vec}(\boldsymbol{B}) = \boldsymbol{D}\text{vech}(\boldsymbol{B})$. Denote $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{\theta}) = \text{vech}(\boldsymbol{\Sigma})$, $\boldsymbol{s} = \text{vech}(\boldsymbol{S})$, $\boldsymbol{E} = \frac{1}{2}\boldsymbol{D}^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{D}$, where $\otimes$ is the Kronecker product, and $\boldsymbol{\Delta} = \dfrac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$ the $p^\star \times m$ Jacobian matrix of the partial derivatives of the model with respect to the parameters. Then, the expected Fisher information can be written as (Yuan & Bentler, 2006):

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) = N\boldsymbol{\Delta}^T\boldsymbol{E}\boldsymbol{\Delta}. \tag{A.20}$$

The propositions on the form of the gradient, the Hessian matrix and the expected Fisher information are proved in Appendices A.2.1-A.2.3, respectively.

A.2.1 Gradient vector

A.2.1.1 Proof of proposition 2.1

Proof. Consider the first-order partial derivative of the log-likelihood function in equation (2.2) with respect to an arbitrary scalar variable θ_q:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_q} &= -\frac{N}{2}\frac{\partial}{\partial \theta_q}\left[\log|\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + p\log(2\pi)\right] \\
&= -\frac{N}{2}\left\{\frac{\partial}{\partial \theta_q}\log|\boldsymbol{\Sigma}| + \frac{\partial}{\partial \theta_q}\text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + \cancel{\frac{\partial}{\partial \theta_q} p\log(2\pi)}\right\} \\
&= -\frac{N}{2}\left\{\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right)\right] - \text{tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right)\right]\right\} \\
&= -\frac{N}{2}\text{tr}\left\{(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1})\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right)\right\} \\
&= -\frac{N}{2}\text{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right\} = -\frac{N}{2}\text{tr}\left\{\boldsymbol{\Omega}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right\},
\end{aligned} \tag{A.21}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}$. ■

A.2.1.2 Proof of Proposition A.1

Proof. To find the expressions in equations (A.2)-(A.4), we need the partial derivatives of the matrix $\boldsymbol{\Sigma}$ with respect to the model parameter matrices $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. These quantities are:

$$\begin{aligned}\frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ij}} &= \frac{\partial(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})}{\partial \lambda_{ij}} = \frac{\partial \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T}{\partial \lambda_{ij}} + \cancel{\frac{\partial \boldsymbol{\Psi}}{\partial \lambda_{ij}}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\frac{\partial \boldsymbol{\Lambda}^T}{\partial \lambda_{ij}} + \frac{\partial \boldsymbol{\Lambda}\boldsymbol{\Phi}}{\partial \lambda_{ij}}\boldsymbol{\Lambda}^T \\ &= \boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \left[\cancel{\boldsymbol{\Lambda}\frac{\partial \boldsymbol{\Phi}}{\partial \lambda_{ij}}} + \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{ij}}\boldsymbol{\Phi}\right]\boldsymbol{\Lambda}^T = \boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T, \end{aligned} \tag{A.22}$$

$$\begin{aligned}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}} &= \frac{\partial(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})}{\partial \phi_{gh}} = \frac{\partial \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T}{\partial \phi_{gh}} + \cancel{\frac{\partial \boldsymbol{\Psi}}{\partial \phi_{gh}}} = \cancel{\boldsymbol{\Lambda}\boldsymbol{\Phi}\frac{\partial \boldsymbol{\Lambda}^T}{\partial \phi_{gh}}} + \frac{\partial \boldsymbol{\Lambda}\boldsymbol{\Phi}}{\partial \phi_{gh}}\boldsymbol{\Lambda}^T \\ &= \left[\boldsymbol{\Lambda}\frac{\partial \boldsymbol{\Phi}}{\partial \phi_{gh}} + \cancel{\frac{\partial \boldsymbol{\Lambda}}{\partial \phi_{gh}}\boldsymbol{\Phi}}\right]\boldsymbol{\Lambda}^T = \boldsymbol{\Lambda}\frac{\partial \boldsymbol{\Phi}}{\partial \phi_{gh}}\boldsymbol{\Lambda}^T \\ &= \boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T, \end{aligned} \tag{A.23}$$

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{ii}} = \frac{\partial(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})}{\partial \psi_{ii}} = \cancel{\frac{\partial \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T}{\partial \psi_{ii}}} + \frac{\partial \boldsymbol{\Psi}}{\partial \psi_{ii}} = \mathbf{1}_{ii}, \tag{A.24}$$

where $\mathbf{1}_{ab}$ is a matrix with zeros in every position, except the entry (a, b), which contains a 1.0. By substituting expressions (A.22), (A.23) and (A.24) in equation (A.21), we get the following set of first-order derivatives of $\ell(\boldsymbol{\theta})$ with respect to the factor loadings, the factor variances and covariances, and the unique variances, respectively (Mulaik, 1971):

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta})}{\partial \lambda_{ij}} &= -\frac{N}{2}\text{tr}\left\{\boldsymbol{\Omega}\frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ij}}\right\} = -\frac{N}{2}\text{tr}\left\{\boldsymbol{\Omega}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T)\right\} \\ &= -\frac{N}{2}\text{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \boldsymbol{\Omega}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\ &= -\frac{N}{2}\left\{\text{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] + \text{tr}\left[\boldsymbol{\Omega}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right]\right\} \\ &= -\frac{N}{2}\left\{\text{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] + \text{tr}\left[\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\right]\right\} \\ &= -\frac{N}{2}\left\{\text{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] + \text{tr}\left[(\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega})^T\right]\right\}\end{aligned}$$

$$
\begin{aligned}
&= -\frac{N}{2}\left\{\operatorname{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] + \operatorname{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right]\right\} \\
&= -N\operatorname{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] = -N\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\right]_{ij} \\
&= -N\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\right]_{ij}, \qquad \text{(A.25)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \phi_{gh}} &= -\frac{N}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}}\right\} = -\frac{N}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh}+\mathbf{1}_{hg}-\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\right\} \\
&= -\frac{N}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh}+\mathbf{1}_{hg}-\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\right\} \\
&= -\frac{N}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T+\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T-\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right\} \\
&= -\frac{N}{2}\left\{\operatorname{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right]+\operatorname{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\right]-\operatorname{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right]\right\} \\
&= -\frac{N}{2}\left\{\operatorname{tr}\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right]+\operatorname{tr}\left[\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\right]-\operatorname{tr}\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\right\} \\
&= -\frac{N}{2}\left\{2\operatorname{tr}\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right]-\operatorname{tr}\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\right\} \\
&= -N\left\{\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\right]_{gh}-\frac{1}{2}\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\right]_{hg}\left[\boldsymbol{I}\right]_{hg}\right\} \\
&= -N\left(1-\frac{1}{2}\left[\boldsymbol{I}\right]_{gh}\right)\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\right]_{gh} \\
&= -N\left(1-\frac{1}{2}\left[\boldsymbol{I}\right]_{gh}\right)\left[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right]_{gh}, \qquad \text{(A.26)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \psi_{ii}} &= -\frac{N}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{ii}}\right\} \\
&= -\frac{N}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}\mathbf{1}_{ii}\right\} = -\frac{N}{2}[\boldsymbol{\Omega}]_{ii} = -\frac{N}{2}[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}]_{ii}. \qquad \text{(A.27)}
\end{aligned}
$$

The analytical first-order derivatives in matrix expression are then:

$$
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Lambda}} = -N\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}, \qquad \text{(A.28)}
$$

$$
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Phi}} = \begin{cases} -N\,\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} & \text{non-diagonal elements,} \\ -\dfrac{N}{2}\,\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} & \text{diagonal elements,} \end{cases} \qquad \text{(A.29)}
$$

$$
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Psi}} = -\frac{N}{2}\operatorname{diag}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}), \qquad \text{(A.30)}
$$

with the understanding that the elements of the three matrices on the left corresponding to the positions of fixed elements of $\boldsymbol{\Lambda}, \boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are taken to be zero. For instance, if the factors were chosen to have unit variance, the diagonal elements of $\dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Phi}}$ would be zero. ■

A.2.2 Hessian matrix

A.2.2.1 Proof of proposition 2.2

Proof. The second partial derivative of $\ell(\boldsymbol{\theta})$ with respect to two arbitrary scalar variables θ_q and $\theta_{q'}$ is:

$$
\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_q \partial \theta_{q'}} &= \frac{\partial}{\partial \theta_q}\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{q'}}\right) = \frac{\partial}{\partial \theta_q}\left\{-\frac{N}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right]\right\} \\
&= -\frac{N}{2}\mathrm{tr}\left\{\frac{\partial[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}]}{\partial \theta_q}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}} + \boldsymbol{\Omega}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_q \partial \theta_{q'}}\right\} \\
&= -\frac{N}{2}\left\{\mathrm{tr}\left[\frac{\partial[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}]}{\partial \theta_q}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] + \mathrm{tr}\left[\boldsymbol{\Omega}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_q \partial \theta_{q'}}\right]\right\} \\
&= -\frac{N}{2}\left\{\mathrm{tr}\left[\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \theta_q}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] - \mathrm{tr}\left[\frac{\partial[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}]}{\partial \theta_q}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] + \mathrm{tr}\left[\boldsymbol{\Omega}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_q \partial \theta_{q'}}\right]\right\} \\
&= -\frac{N}{2}\Bigg\{ - \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] - \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \theta_q}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] \\
&\qquad - \mathrm{tr}\left[\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \theta_q}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] + \mathrm{tr}\left[\boldsymbol{\Omega}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_q \partial \theta_{q'}}\right]\Bigg\} \\
&= -\frac{N}{2}\Bigg\{ - \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] + \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\Sigma^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] \\
&\qquad + \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] + \mathrm{tr}\left[\boldsymbol{\Omega}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_q \partial \theta_{q'}}\right]\Bigg\} \\
&= -\frac{N}{2}\Bigg\{\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] - 2\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] \\
&\qquad + \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] \\
&\qquad + \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right] + \mathrm{tr}\left[\boldsymbol{\Omega}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_q \partial \theta_{q'}}\right]\Bigg\}
\end{aligned}
$$

$$= -\frac{N}{2}\left\{\text{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\right] + \text{tr}\left[\boldsymbol{\Omega}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\theta_q\partial\theta_{q'}}\right] - 2\text{tr}\left[\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\right]\right\}$$

$$= -\frac{N}{2}\left\{\text{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\right) + \text{tr}\left[\boldsymbol{\Omega}\left(\frac{\partial^2\boldsymbol{\Sigma}}{\partial\theta_q\partial\theta_{q'}} - 2\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\right)\right]\right\}, \tag{A.31}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}$. ■

A.2.2.2 Proof of proposition A.2

Proof. To find expressions (A.6)-(A.11), we need the second partial derivatives of $\boldsymbol{\Sigma}$ with respect to the model parameters, which are as follows, for $i, t = 1, \ldots, p$ and $g, h, j, l, q, s = 1, \ldots, r$:

$$\begin{aligned}
\frac{\partial^2\boldsymbol{\Sigma}}{\partial\lambda_{ij}\partial\lambda_{ts}} &= \frac{\partial}{\partial\lambda_{ij}}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ts}}\right) = \frac{\partial}{\partial\lambda_{ij}}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st} + \mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right) \\
&= \frac{\partial}{\partial\lambda_{ij}}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\right) + \frac{\partial}{\partial\lambda_{ij}}\left(\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right) = \frac{\partial\boldsymbol{\Lambda}}{\partial\lambda_{ij}}\left(\boldsymbol{\Phi}\mathbf{1}_{st}\right) + \left(\mathbf{1}_{ts}\boldsymbol{\Phi}\right)\frac{\partial\boldsymbol{\Lambda}^T}{\partial\lambda_{ij}} \\
&= \mathbf{1}_{ij}\boldsymbol{\Phi}\mathbf{1}_{st} + \mathbf{1}_{ts}\boldsymbol{\Phi}\mathbf{1}_{ji} = \mathbf{1}_{ti}\phi_{sj} + \mathbf{1}_{it}\phi_{js}, \qquad \text{(A.32)} \\
\frac{\partial^2\boldsymbol{\Sigma}}{\partial\lambda_{ij}\partial\phi_{gh}} &= \frac{\partial}{\partial\lambda_{ij}}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{gh}}\right) = \frac{\partial}{\partial\lambda_{ij}}\left(\boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\right) \\
&= \boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\frac{\partial\boldsymbol{\Lambda}^T}{\partial\lambda_{ij}} + \frac{\partial(\boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right])}{\partial\lambda_{ij}}\boldsymbol{\Lambda}^T \\
&= \boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\mathbf{1}_{ji} + \frac{\partial\boldsymbol{\Lambda}}{\partial\lambda_{ij}}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T \\
&= \boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\mathbf{1}_{ji} + \mathbf{1}_{ij}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T, \qquad \text{(A.33)} \\
\frac{\partial^2\boldsymbol{\Sigma}}{\partial\lambda_{ij}\partial\psi_{tt}} &= \frac{\partial}{\partial\lambda_{ij}}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}^2}\right) = \frac{\partial}{\partial\lambda_{ij}}\mathbf{1}_{tt} = 0, \qquad \text{(A.34)} \\
\frac{\partial^2\boldsymbol{\Sigma}}{\partial\phi_{gh}\partial\phi_{lq}} &= \frac{\partial}{\partial\phi_{gh}}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{lq}}\right) = \frac{\partial}{\partial\phi_{gh}}\left(\boldsymbol{\Lambda}\left[\mathbf{1}_{lq} + \mathbf{1}_{ql} - \mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq}\right]\boldsymbol{\Lambda}^T\right) = 0, \qquad \text{(A.35)} \\
\frac{\partial^2\boldsymbol{\Sigma}}{\partial\phi_{gh}\partial\psi_{tt}} &= \frac{\partial}{\partial\phi_{gh}}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}^2}\right) = \frac{\partial\mathbf{1}_{tt}}{\partial\phi_{gh}} = 0, \qquad \text{(A.36)} \\
\frac{\partial^2\boldsymbol{\Sigma}}{\partial\psi_{ii}\partial\psi_{tt}} &= \frac{\partial}{\partial\psi_{ii}}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}^2}\right) = \frac{\partial\mathbf{1}_{tt}}{\partial\psi_{ii}} = 0. \qquad \text{(A.37)}
\end{aligned}$$

We now have the necessary quantities to obtain the second derivatives of the log-likelihood function. For simplicity, we compute the second-order derivatives of the function $F = -\frac{2}{N}\ell(\boldsymbol{\theta})$, that is,

$$F = \log|\boldsymbol{\Sigma}| + \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) + p\log(2\pi);$$

the second-order derivatives of $\ell(\boldsymbol{\theta})$ are then easily found by multiplying the resulting expressions by the factor $-\frac{N}{2}$. Based on the result in (2.5) and after some computations, we have that

$$\frac{\partial^2 F}{\partial\theta_q\partial\theta_{q'}} = \operatorname{tr}\left[\boldsymbol{\Omega}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\theta_q\partial\theta_{q'}}\right] - \operatorname{tr}\left[\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right] + \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\boldsymbol{M}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right]. \quad \text{(A.38)}$$

The derivation of each second-order derivative is carried out by substituting the respective matrix expressions of the first and second derivatives of $\boldsymbol{\Sigma}$ into (A.38) and simplifying the resulting expressions. The traces of the resulting matrix expressions are obtained by application of the properties of the trace and, in particular, its invariance under cyclic permutations. After taking the traces of these expressions and simplifying the result, we obtain the following set of second partial derivatives of F with respect to the model parameters, for $i, t = 1, \ldots, p$ and $g, h, j, l, q, s = 1, \ldots, r$ (Mulaik, 1971).

Factor loadings

$$\begin{aligned}
\frac{\partial^2 F}{\partial\lambda_{ij}\partial\lambda_{ts}} = \;& \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\lambda_{ij}\partial\lambda_{ts}}\right\} - \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ts}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ij}}\right\} + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ts}}\boldsymbol{M}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ij}}\right\} \\
= \;& \operatorname{tr}(\boldsymbol{\Omega}[\mathbf{1}_{ti}\phi_{sj} + \mathbf{1}_{it}\phi_{js}]) - \operatorname{tr}\{\boldsymbol{\Omega}[\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st} + \mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T]\boldsymbol{\Sigma}^{-1}[\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T]\} \\
& - \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}[\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st} + \mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T]\boldsymbol{M}[\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T]\} \\
= \;& \operatorname{tr}(\boldsymbol{\Omega}\mathbf{1}_{ti})\phi_{sj} + \operatorname{tr}(\mathbf{1}_{it}\boldsymbol{\Omega})\phi_{js} - \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} - \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\} \\
& - \operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} - \operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\} \\
& + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\} \\
& + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\}
\end{aligned}$$

$$
\begin{aligned}
&= 2\mathrm{tr}(\boldsymbol{\Omega}\mathbf{1}_{ti})\phi_{sj} - \mathrm{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} - \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\} \\
&\quad - \mathrm{tr}\{\boldsymbol{\Omega}\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} - \mathrm{tr}\{\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\mathbf{1}_{ts}\} \\
&\quad + \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} + \mathrm{tr}\{\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\} \\
&\quad + \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} + \mathrm{tr}\{\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ts}\} \\
&= 2[\boldsymbol{\Omega}]_{it}[\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - [\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{js} \\
&\quad - [\boldsymbol{\Omega}]_{it}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{jt} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} \\
&\quad + [\boldsymbol{M}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{js} + [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} + [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt} \\
&= 2[\boldsymbol{\Omega}]_{it}[\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - [\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{js} \\
&\quad + [\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{js} - [\boldsymbol{\Omega}]_{it}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{jt} \\
&\quad + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} + [\boldsymbol{M}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{js} \\
&\quad + [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} + [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt} \\
&= 2[\boldsymbol{\Omega}]_{it}[\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - 2[\boldsymbol{\Omega}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} \\
&\quad + 2[\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{jt} \\
&\quad + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} + [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt} \\
&= 2[\boldsymbol{\Omega}]_{it}[\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - 2[\boldsymbol{\Omega}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} \\
&\quad + 2[\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt} + [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}]_{jt} \\
&\quad + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} + [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt} \\
&= 2[\boldsymbol{\Omega}]_{it}[\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - 2[\boldsymbol{\Omega}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} \\
&\quad + 2[\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} - [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T(\boldsymbol{\Sigma}^{-1} - \boldsymbol{M})]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt} \\
&\quad + [\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{si}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}]_{jt} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} \\
&= 2[\boldsymbol{\Omega}]_{it}[\boldsymbol{\Phi}]_{sj} - 2[\boldsymbol{\Omega}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} + 2[\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} \\
&\quad - 2[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} + 2[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} \\
&= 2[\boldsymbol{\Omega}]_{ti}[\boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} + 2[\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} \\
&\quad - 2[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} + 2[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj}. \qquad \text{(A.39)}
\end{aligned}
$$

Factor loading and factor covariance

$$
\begin{aligned}
\frac{\partial^2 F}{\partial\lambda_{ij}\partial\phi_{gh}} = \; & \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\lambda_{ij}\partial\phi_{gh}}\right\} - \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{gh}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ij}}\right\} + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{gh}}\boldsymbol{M}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ij}}\right\} \\
= \; & \operatorname{tr}\Big\{\boldsymbol{\Omega}\Big[\boldsymbol{\Lambda}\left[\mathbf{1}_{gh}+\mathbf{1}_{hg}-\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\mathbf{1}_{ji}+\mathbf{1}_{ij}\left[\mathbf{1}_{gh}+\mathbf{1}_{hg}-\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\Big]\Big\} \\
& - \operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh}+\mathbf{1}_{hg}-\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}[\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}+\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T]\right\} \\
& + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh}+\mathbf{1}_{hg}-\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\boldsymbol{M}[\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}+\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T]\right\} \\
= \; & \operatorname{tr}\Big\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh}+\mathbf{1}_{hg}-\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\mathbf{1}_{ji}\Big\} + \operatorname{tr}\Big\{\boldsymbol{\Omega}\mathbf{1}_{ij}\left[\mathbf{1}_{gh}+\mathbf{1}_{hg}-\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\Big\} \\
& - \operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right\} - \operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\
& - \operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right\} - \operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\
& + \operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right\} + \operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\
& + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right\} + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\
& + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right\} + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\
& - \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right\} - \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\
= \; & \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}(\mathbf{1}_{gh}+\mathbf{1}_{hg})\mathbf{1}_{ji}\} + \operatorname{tr}\{\mathbf{1}_{ij}(\mathbf{1}_{gh}+\mathbf{1}_{hg})\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\} - \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\mathbf{1}_{ji}\} \\
& - \operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{ij}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} - [\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{hj} \\
& - \operatorname{tr}\left\{\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right\} - [\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{gj} \\
& - \operatorname{tr}\left\{\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{hg}\right\} + \operatorname{tr}\left\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right\} \\
& + \operatorname{tr}\left\{\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right\} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{hj} \\
& + \operatorname{tr}\left\{\boldsymbol{\Lambda}^T\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right\} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{gj} \\
& + \operatorname{tr}\left\{\boldsymbol{\Lambda}^T\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\right\} - \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right\} \\
& - \operatorname{tr}\left\{\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right\}
\end{aligned}
$$

$$
\begin{aligned}
=\ & 2\mathrm{tr}\{\boldsymbol{\Omega\Lambda}(\mathbf{1}_{gh}+\mathbf{1}_{hg})\mathbf{1}_{ji}\} - \mathrm{tr}\{\boldsymbol{\Omega\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\mathbf{1}_{ji}\} - \mathrm{tr}\{\boldsymbol{\Omega}\mathbf{1}_{ij}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
& - [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Omega\Lambda}]_{jg} - [\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} \\
& - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Omega\Lambda}]_{jh} + [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} \\
& + \mathrm{tr}\left\{\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right\} - \mathrm{tr}\left\{\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}\mathbf{1}_{gh}\right\} \\
& + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} + [\boldsymbol{\Lambda}^T\boldsymbol{M}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg} \\
& + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} + [\boldsymbol{\Lambda}^T\boldsymbol{M}]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jh} \\
& - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} - [\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{M}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg}
\end{aligned}
$$

$$
\begin{aligned}
=\ & 2\mathrm{tr}\{\boldsymbol{\Omega\Lambda}\mathbf{1}_{gh}\mathbf{1}_{ji}\} + 2\mathrm{tr}\{\boldsymbol{\Omega\Lambda}\mathbf{1}_{hg}\mathbf{1}_{ji}\} - \mathrm{tr}\{\boldsymbol{\Omega\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\mathbf{1}_{ji}\} - \mathrm{tr}\{\boldsymbol{\Omega}\mathbf{1}_{ij}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
& - [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg} + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jg} \\
& - [\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jh} + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jh} \\
& + [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} + [\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg} \\
& - [\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jg} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{M}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} + [\boldsymbol{\Lambda}^T\boldsymbol{M}]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jh} \\
& - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} - [\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{M}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg}
\end{aligned}
$$

$$
\begin{aligned}
=\ & 2[\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hj} + 2[\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I}]_{gj} - [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{I}]_{hj} - [\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I}]_{gh}[\boldsymbol{I}]_{gj} \\
& - [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} - [\boldsymbol{\Lambda}^T(\boldsymbol{\Sigma}^{-1}-\boldsymbol{M})]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jg} - [\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} \\
& - [\boldsymbol{\Lambda}^T(\boldsymbol{\Sigma}^{-1}-\boldsymbol{M})]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jh} + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jh} \\
& + [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} + [\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T(\boldsymbol{\Sigma}^{-1}-\boldsymbol{M})]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg} \\
& - [\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jg} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} \\
& + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj}
\end{aligned}
$$

$$
\begin{aligned}
= \; & 2[\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hj} + 2[\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I}]_{gj} - [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{I}]_{hj} - [\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I}]_{gh}[\boldsymbol{I}]_{gj} \\
& - [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} - [\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg} + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jg} \\
& - [\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} - [\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jh} + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jh} \\
& + [\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} + [\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg} \\
& - [\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi}[\boldsymbol{\Phi\Lambda}^T\boldsymbol{M\Lambda}]_{jg} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} \\
& + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} \\
= \; & 2[\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hj} - 2[\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} + 2[\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I}]_{gj} - 2[\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} \\
& + 2[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} + 2[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} \\
& - [\boldsymbol{I}]_{gh}[\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I}]_{hj} + [\boldsymbol{I}]_{gh}[\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} - [\boldsymbol{I}]_{gh}[\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I}]_{gj} \\
& + [\boldsymbol{I}]_{gh}[\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} - [\boldsymbol{I}]_{gh}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} \\
& - [\boldsymbol{I}]_{gh}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} \\
= \; & 2[\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I} - \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} + 2[\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I} - \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} \\
& + 2[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} + 2[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} \\
& - [\boldsymbol{I}]_{gh}[\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I} - \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} - [\boldsymbol{I}]_{gh}[\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I} - \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} \\
& - [\boldsymbol{I}]_{gh}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} - [\boldsymbol{I}]_{gh}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} \\
= \; & (2 - [\boldsymbol{I}]_{gh})\Big([\boldsymbol{\Omega\Lambda}]_{ig}[\boldsymbol{I} - \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{hj} + [\boldsymbol{\Omega\Lambda}]_{ih}[\boldsymbol{I} - \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}]_{gj} \\
& + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{hj} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{M\Lambda\Phi}]_{gj} \Big). \qquad \text{(A.40)}
\end{aligned}
$$

Factor loading and unique variance

$$
\begin{aligned}
\frac{\partial^2 F}{\partial\lambda_{ij}\partial\psi_{tt}} = \; & \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\lambda_{ij}\partial\psi_{tt}}\right\} - \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ij}}\right\} + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}}\boldsymbol{M}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ij}}\right\} \\
= \; & \cancel{\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{0})} - \operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi\Lambda}^T)\} \\
& + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M}(\boldsymbol{\Lambda\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi\Lambda}^T)\} \\
= \; & -\operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}\mathbf{1}_{ji} + \boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi\Lambda}^T\} + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M\Lambda\Phi}\mathbf{1}_{ji} \\
& + \boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi\Lambda}^T\}
\end{aligned}
$$

$$
\begin{aligned}
&= -\operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} - \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\mathbf{1}_{tt}\} + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\} \\
&\quad + \operatorname{tr}\{\boldsymbol{M}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\} \\
&= -[\boldsymbol{\Omega}]_{it}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - [\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{jt} + [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} \\
&\quad + [\boldsymbol{M}]_{it}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt} \\
&= -[\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} + [\boldsymbol{M}]_{it}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{jt} \\
&\quad + [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} + [\boldsymbol{M}]_{it}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} \\
&= 2[\boldsymbol{M}]_{it}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi} - \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} \\
&= 2[\boldsymbol{M}]_{it}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - 2[\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} \\
&= 2\Big([\boldsymbol{M}]_{it}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj} - [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj}\Big). \qquad \text{(A.41)}
\end{aligned}
$$

Factor covariances

$$
\frac{\partial^2 F}{\partial\phi_{gh}\partial\phi_{lq}} = \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\phi_{gh}\partial\phi_{lq}}\right\} - \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{lq}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{gh}}\right\} + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{lq}}\boldsymbol{M}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{gh}}\right\}
$$

$$
\begin{aligned}
&= \cancel{\operatorname{tr}\{\boldsymbol{\Omega}\mathbf{0}\}} - \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}(\mathbf{1}_{lq} + \mathbf{1}_{ql} - \mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq})\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh})\boldsymbol{\Lambda}^T\} \\
&\quad + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\mathbf{1}_{lq} + \mathbf{1}_{ql} - \mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq})\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}(\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh})\boldsymbol{\Lambda}^T\}
\end{aligned}
$$

$$
\begin{aligned}
= &-\operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} - \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\} \\
&+ \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} - \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{ql}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
&- \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{ql}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\} + \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{ql}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
&+ \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} + \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\} \\
&- \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
&+ \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\} - \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
&+ \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{ql}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{ql}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\} \\
&- \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{ql}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} - \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
&- \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\} + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\}
\end{aligned}
$$

$$
\begin{aligned}
= & -\operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} - \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} \\
& + \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} - \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{ql}\} \\
& - \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{ql}\} + \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{ql}\} \\
& + \operatorname{tr}\{\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} + \operatorname{tr}\{\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} \\
& - \operatorname{tr}\{\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} + \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} \\
& + \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} - \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} \\
& + \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{ql}\} + \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{ql}\} \\
& - \operatorname{tr}\{\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{ql}\} - \operatorname{tr}\{\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} \\
& - \operatorname{tr}\{\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\} + \operatorname{tr}\{\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\}
\end{aligned}
$$

$$
\begin{aligned}
= & -[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hl} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{gl} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hl} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{lg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hq} \\
& - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{lh}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{gq} + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{lg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hq} \\
& + [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hl} + [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{gl} \\
& - [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hl} + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl} - [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{lg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hq} + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{lh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gq} \\
& - [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{lg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hq} - [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
& - [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl} + [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}
\end{aligned}
$$

$$
\begin{aligned}
= & -[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl} \\
& - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl} + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl} \\
& - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{lg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hq} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{lh}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{gq} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{lg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hq} + [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
& - [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl} + [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl} \\
& - [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl} - [\boldsymbol{I}]_{ql}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
& + [\boldsymbol{I}]_{ql}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl} + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl} - [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{lg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hq} + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{lh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gq} \\
& - [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{lg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hq} - [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
& - [\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl} + [\boldsymbol{I}]_{ql}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{qg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl} \\
= & -[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T(\boldsymbol{\Sigma}^{-1} - \boldsymbol{M})\boldsymbol{\Lambda}]_{qg} + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} \\
& - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T(\boldsymbol{\Sigma}^{-1} - \boldsymbol{M})\boldsymbol{\Lambda}]_{qh} + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} \\
& + [\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T(\boldsymbol{\Sigma}^{-1} - \boldsymbol{M})\boldsymbol{\Lambda}]_{qg} - [\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} \\
& - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qh} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qg} \\
& + [\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qh} + [\boldsymbol{I}]_{lq}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qg} \\
& - [\boldsymbol{I}]_{lq}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} + [\boldsymbol{I}]_{lq}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qh} \\
& - [\boldsymbol{I}]_{lq}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} - [\boldsymbol{I}]_{lq}[\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qg} \\
& + [\boldsymbol{I}]_{lq}[\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} \\
& + [\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} - [\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh}
\end{aligned}
$$

$$
\begin{aligned}
= \; & 2[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} + 2[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} - 2[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qh} \\
& - 2[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qg} + [\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qg} \\
& - [\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} + [\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qh} \\
& + [\boldsymbol{I}]_{lq}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qg} - [\boldsymbol{I}]_{lq}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} \\
& + [\boldsymbol{I}]_{lq}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qh} - [\boldsymbol{I}]_{lq}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} \\
& - [\boldsymbol{I}]_{lq}[\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qg} + [\boldsymbol{I}]_{lq}[\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} \\
& - [\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} \\
= \; & (2 - [\boldsymbol{I}]_{lq} - [\boldsymbol{I}]_{gh} + [\boldsymbol{I}]_{lq}[\boldsymbol{I}]_{gh})([\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qg} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qg}) \\
& + (2 - [\boldsymbol{I}]_{lq} - [\boldsymbol{I}]_{gh})([\boldsymbol{\Lambda}^T\boldsymbol{M}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} - [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{qh}).
\end{aligned}
\tag{A.42}
$$

Factor covariance and unique variance

$$
\begin{aligned}
\frac{\partial^2 F}{\partial\phi_{gh}\partial\psi_{tt}} = \; & \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\phi_{gh}\partial\psi_{tt}}\right\} - \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{gh}}\right\} + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}}\boldsymbol{M}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{gh}}\right\} \\
= \; & \cancel{\operatorname{tr}(\boldsymbol{\Omega}\mathbf{0})} - \operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\} \\
& + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\} \\
= \; & - \operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} - \operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\} \\
& + \operatorname{tr}\{\boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
& + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\} - \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\} \\
= \; & - \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\mathbf{1}_{tt}\} - \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\mathbf{1}_{tt}\} \\
& + \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\mathbf{1}_{tt}\} + \operatorname{tr}\{\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\} \\
& + \operatorname{tr}\{\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\} - \operatorname{tr}\{\boldsymbol{M}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\} \\
= \; & - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{gt} \\
& + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} + [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} \\
& + [\boldsymbol{M}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gt} - [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht}
\end{aligned}
$$

$$
\begin{aligned}
&= -[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T(\boldsymbol{\Sigma}^{-1}-\boldsymbol{M})]_{gt} \\
&\quad + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} + [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} \\
&\quad + [\boldsymbol{M}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gt} - [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} \\
&= -[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gt} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{M}]_{gt} \\
&\quad + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} + [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} \\
&\quad + [\boldsymbol{M}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gt} - [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} \\
&= -[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} - [(\boldsymbol{\Sigma}^{-1}-\boldsymbol{M})\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gt} \\
&\quad + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{M}]_{gt} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} \\
&\quad + [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} - [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} \\
&= -[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} - [\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gt} + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{M}]_{gt} \\
&\quad + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}]_{ht} + [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} \\
&\quad - [\boldsymbol{M}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} \\
&= 2[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{M}\boldsymbol{\Lambda}]_{tg} - 2[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Omega}\Lambda]_{th} - [\boldsymbol{I}]_{gh}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{M}\boldsymbol{\Lambda}]_{tg} \\
&\quad + [\boldsymbol{I}]_{gh}[\boldsymbol{\Sigma}^{-1}\Lambda]_{tg}[\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{th} \\
&= (2-[\boldsymbol{I}]_{gh})\Big([\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{M}\boldsymbol{\Lambda}]_{tg} - [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Omega}\Lambda]_{th}\Big). \qquad \text{(A.43)}
\end{aligned}
$$

Unique variances

$$
\begin{aligned}
\frac{\partial^2 F}{\partial\psi_{ii}\partial\psi_{tt}} &= \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\psi_{ii}\partial\psi_{tt}}\right\} - \operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{ii}}\right\} + \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{tt}}\boldsymbol{M}\frac{\partial\boldsymbol{\Sigma}}{\partial\psi_{ii}}\right\} \\
&= \cancel{\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\theta})} - \operatorname{tr}(\boldsymbol{\Omega}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ii}) + \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{M}\mathbf{1}_{ii}) \\
&= -[\boldsymbol{\Omega}]_{it}[\boldsymbol{\Sigma}^{-1}]_{ti} + [\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{M}]_{ti} = [\boldsymbol{\Sigma}^{-1}]_{it}\left(-[\boldsymbol{\Omega}]_{it} + [\boldsymbol{M}]_{ti}\right) \\
&= [\boldsymbol{\Sigma}^{-1}]_{it}\left(-[\boldsymbol{\Sigma}^{-1}]_{it} + [\boldsymbol{M}]_{it} + [\boldsymbol{M}]_{ti}\right) = [\boldsymbol{\Sigma}^{-1}]_{it}[2\boldsymbol{M} - \boldsymbol{\Sigma}^{-1}]_{it}. \qquad \text{(A.44)}
\end{aligned}
$$

The expressions in (A.6)-(A.11) are obtained by using the fact that $\dfrac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_q\partial\theta_{q'}} = -\dfrac{N}{2}\dfrac{\partial^2 F}{\partial\theta_q\partial\theta_{q'}}$. ■

A.2.3 Fisher information matrix

A.2.3.1 Proof of Proposition A.3

Proof. We now compute the approximate second-order derivatives which coincide with the elements of the expected Fisher information matrix, for $i, t = 1, \ldots, p$ and $g, h, j, l, q, s = 1, \ldots, r$ (Jöreskog, 1969).

Factor loadings

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \lambda_{ts}}\right] &= \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ij}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ts}}\right\} \\
&= \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st} + \mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T)\right\} \\
&= \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\right\} + \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\
&\quad + \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{st}\right\} \\
&\quad + \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ts}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right\} \\
&= \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is} + \frac{N}{2}[\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} \\
&\quad + \frac{N}{2}[\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{js} + \frac{N}{2}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{si} \\
&= N[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{is} + N[\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{sj} \\
&= N(\beta_{tj}\beta_{is} + \omega_{it}\zeta_{js}).
\end{aligned}
$$

Factor loading and unique variance

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \psi_{tt}}\right] &= \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ij}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{tt}}\right\} \\
&= \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T)\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\right\} \\
&= \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\right\} + \frac{N}{2} \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\right\} \\
&= \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{tj}[\boldsymbol{\Sigma}^{-1}]_{it} + \frac{N}{2}[\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{jt} = N\omega_{it}\beta_{tj}.
\end{aligned}
$$

Factor loading and factor covariance

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \phi_{gh}}\right] &= \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ij}}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}}\right\} \\
&= \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\right\} \\
&= \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right\} + \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\right\} \\
&\quad - \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right\} \\
&\quad + \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right\} \\
&\quad + \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\right\} \\
&\quad - \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right\} \\
&= \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{hj} + \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{gj} \\
&\quad - \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{hj} + \frac{N}{2}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi} \\
&\quad + \frac{N}{2}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gi} - \frac{N}{2}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi} \\
&= N[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{hj} + N[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{gj} \\
&\quad - \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{hj} - \frac{N}{2}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg}[\boldsymbol{I}]_{hg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{hi} \\
&= \frac{N}{2}\Big\{(2 - [\boldsymbol{I}]_{gh})([\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jh} \\
&\quad + [\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ih}[\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{jg})\Big\} = \frac{N}{2}(2 - [\boldsymbol{I}]_{gh})(\alpha_{ig}\delta_{jh} + \alpha_{ih}\delta_{jg}).
\end{aligned}
$$

Unique variances

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{ii} \partial \psi_{tt}}\right] &= \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{ii}}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{tt}}\right\} = \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ii}\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\right\} \\
&= \frac{N}{2}[\boldsymbol{\Sigma}^{-1}]_{ti}[\boldsymbol{\Sigma}^{-1}]_{it} = \frac{N}{2}\omega_{it}^2.
\end{aligned}
$$

Unique variance and factor covariance

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{tt} \partial \phi_{gh}}\right] &= \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{tt}}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}}\right\} \\
&= \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right\} + \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\right\} \\
&\quad - \frac{N}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{tt}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right\} \\
&= \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{tg}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{ht} + \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gt} \\
&\quad - \frac{N}{2}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{th}[\boldsymbol{I}]_{gh}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}]_{gt} \\
&= N\alpha_{tg}\alpha_{th} - \frac{N}{2}[\boldsymbol{I}]_{gh}\alpha_{tg}\alpha_{th} = \frac{N}{2}(2-[\boldsymbol{I}]_{gh})\alpha_{tg}\alpha_{th}.
\end{aligned}
$$

Factor covariances

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\phi_{gh}\partial\phi_{lq}}\right] &= -\mathbb{E}\left[\frac{\partial}{\partial\phi_{lq}}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\phi_{gh}}\right] \\
&= -\mathbb{E}\left[\frac{\partial}{\partial\phi_{lq}}\left\{-\frac{N}{2}(2-[\boldsymbol{I}]_{gh})[\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gh}\right\}\right] \\
&= \frac{N}{2}(2-[\boldsymbol{I}]_{gh})\left[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\left(\frac{\boldsymbol{\Sigma}}{\partial\phi_{lq}}\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right]_{gh} \\
&= \frac{N}{2}(2-[\boldsymbol{I}]_{gh})\left[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\mathbf{1}_{lq}+\mathbf{1}_{ql}-\mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq})\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right]_{gh} \\
&= \frac{N}{2}(2-[\boldsymbol{I}]_{gh})\ \mathrm{tr}(\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}) \\
&\quad + \frac{N}{2}(2-[\boldsymbol{I}]_{gh})\ \mathrm{tr}(\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{ql}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}) \\
&\quad - \frac{N}{2}(2-[\boldsymbol{I}]_{gh})\ \mathrm{tr}(\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{lq}\boldsymbol{I}\mathbf{1}_{lq}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{hg}) \\
&= \frac{N}{2}(2-[\boldsymbol{I}]_{gh})[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} \\
&\quad + \frac{N}{2}(2-[\boldsymbol{I}]_{gh})[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gq}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{lh} \\
&\quad - \frac{N}{2}(2-[\boldsymbol{I}]_{gh})[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{I}]_{ql}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{qh} \\
&= \frac{N}{2}\frac{1}{2}(2-[\boldsymbol{I}]_{gh})(2-[\boldsymbol{I}]_{lq})([\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gl}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hq} \\
&\quad + [\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{gq}[\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{hl}) \\
&= \frac{N}{4}(2-[\boldsymbol{I}]_{gh})(2-[\boldsymbol{I}]_{lq})(\gamma_{gl}\gamma_{hq}+\gamma_{gq}\gamma_{hl}).
\end{aligned}
$$

■

APPENDIX B

Locally approximated penalties

Appendix B.1 contains the derivations of the expressions of the penalty functions examined in this work (i.e., lasso, alasso, scad and mcp), whereas Appendix B.2 reports the associated penalty matrices resulting from the local approximations of the non-differentiable penalties.

B.1 The penalty functions

We consider the case where the interest lies in the shrinkage of the factor loadings, although other model parameters could be in principle penalized. Let us write the parameter vector as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{q^\star}, \theta_{q^\star+1}, \ldots, \theta_m)^T$, where the sub-vector $(\theta_1, \ldots, \theta_{q^\star})^T$ collects the penalized parameters (i.e., the factor loadings), whereas $(\theta_{q^\star+1}, \ldots, \theta_m)^T$ the unpenalized parameters (i.e., the free elements in $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$). Define the diagonal matrix $\boldsymbol{R}_q = \text{diag}(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ for $q = 1, \ldots, q^\star$ where the 1 on the $(q, q)^{\text{th}}$ entry of the matrix corresponds to the q^{th} parameter in $\boldsymbol{\theta}$, and $\boldsymbol{R}_q = \boldsymbol{O}_{m \times m}$ for $q = q^\star + 1, \ldots, m$. Let $\boldsymbol{e}_q = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ be the canonical vector with a 1 in the q^{th} position for $q = 1, \ldots, q^\star$, and the null vector otherwise.

The overall penalty $\mathcal{T}$ is given by the sum of the penalty terms for each parameter, that is,

$$\mathcal{P}_\eta^{\mathcal{T}}(\boldsymbol{\theta}) = \sum_{q=1}^{m} \mathcal{P}_{\eta,q}^{\mathcal{T}}(||\boldsymbol{R}_q \boldsymbol{\theta}||_1),$$

where $\mathcal{T} = \{L, A, S, M\}$ stands for lasso, alasso, scad, and mcp, respectively. The

term $||\boldsymbol{R}_q\boldsymbol{\theta}||_1 = |\boldsymbol{e}_q^T\boldsymbol{\theta}| = |\theta_q|$ for $q = 1, \ldots, q^\star$, and is equal to zero otherwise. Let us detail the expression of the penalty term for each of these penalties.

Lasso

$$\begin{aligned}
\mathcal{P}_\eta^L(\boldsymbol{\theta}) &= \sum_{q=1}^{m} \mathcal{P}_{\eta,q}^L(||\boldsymbol{R}_q\boldsymbol{\theta}||_1) = \sum_{q=1}^{m} \eta||\boldsymbol{R}_q\boldsymbol{\theta}||_1 \\
&= \eta \sum_{q=1}^{m} \left\{(\boldsymbol{R}_q\boldsymbol{\theta})^T(\boldsymbol{R}_q\boldsymbol{\theta})\right\}^{\frac{1}{2}} = \eta \sum_{q=1}^{m} \left\{(\boldsymbol{e}_q^T\boldsymbol{\theta})^2\right\}^{\frac{1}{2}} \\
&= \eta \sum_{q=1}^{m} |\boldsymbol{e}_q^T\boldsymbol{\theta}| = \eta \sum_{q=1}^{q^\star} |\theta_q|.
\end{aligned}$$

Alasso

$$\begin{aligned}
\mathcal{P}_\eta^A(\boldsymbol{\theta}) &= \sum_{q=1}^{m} \mathcal{P}_{\eta,q}^A(||\boldsymbol{R}_q\boldsymbol{\theta}||_1) = \eta \sum_{q=1}^{m} \frac{||\boldsymbol{R}_q\boldsymbol{\theta}||_1}{||\boldsymbol{R}_q\hat{\boldsymbol{\theta}}||_1^a} \\
&= \eta \sum_{q=1}^{m} \frac{\left\{(\boldsymbol{R}_q\boldsymbol{\theta})^T(\boldsymbol{R}_q\boldsymbol{\theta})\right\}^{\frac{1}{2}}}{\left\{\left(\boldsymbol{R}_q\hat{\boldsymbol{\theta}}\right)^T\left(\boldsymbol{R}_q\hat{\boldsymbol{\theta}}\right)\right\}^{\frac{a}{2}}} = \eta \sum_{q=1}^{m} \frac{\left\{(\boldsymbol{e}_q^T\boldsymbol{\theta})^2\right\}^{\frac{1}{2}}}{\left\{\left(\boldsymbol{e}_q^T\hat{\boldsymbol{\theta}}\right)^2\right\}^{\frac{a}{2}}} \\
&= \eta \sum_{q=1}^{m} \frac{|\boldsymbol{e}_q^T\boldsymbol{\theta}|}{\left|\boldsymbol{e}_q^T\hat{\boldsymbol{\theta}}\right|^a} = \eta \sum_{q=1}^{q^\star} \frac{|\theta_q|}{|\hat{\theta}_q|^a},
\end{aligned}$$

where $\hat{\boldsymbol{\theta}}$ is generally the maximum likelihood estimator $\hat{\boldsymbol{\theta}}^{\text{MLE}}$ and $a > 0$ an additional tuning parameter.

Scad

$$\begin{aligned}
\mathcal{P}_\eta^S(\boldsymbol{\theta}) &= \sum_{q=1}^{m} \mathcal{P}_{\eta,q}^S(||\boldsymbol{R}_q\boldsymbol{\theta}||_1) \\
&= \sum_{q=1}^{m} \Bigg\{ \eta||\boldsymbol{R}_q\boldsymbol{\theta}||_1 \mathbb{1}\left(0 \leq ||\boldsymbol{R}_q\boldsymbol{\theta}||_1 \leq \eta\right) \\
&\qquad - \left[\frac{(\boldsymbol{R}_q\boldsymbol{\theta})^T(\boldsymbol{R}_q\boldsymbol{\theta}) + \eta^2 - 2\eta a||\boldsymbol{R}_q\boldsymbol{\theta}||_1}{2(a-1)}\right] \\
&\qquad \times \mathbb{1}\left(\eta < ||\boldsymbol{R}_q\boldsymbol{\theta}||_1 \leq a\eta\right) + \frac{\eta^2(a+1)}{2}\mathbb{1}\left(||\boldsymbol{R}_q\boldsymbol{\theta}||_1 > a\eta\right) \Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{q=1}^{m} \Bigg\{ \eta \left[(\boldsymbol{R}_q\boldsymbol{\theta})^T (\boldsymbol{R}_q\boldsymbol{\theta}) \right]^{\frac{1}{2}} \mathbb{1}\left(0 \leq \left[(\boldsymbol{R}_q\boldsymbol{\theta})^T (\boldsymbol{R}_q\boldsymbol{\theta}) \right]^{\frac{1}{2}} \leq \eta \right) \\
&\qquad - \left[\frac{(\boldsymbol{R}_q\boldsymbol{\theta})^T (\boldsymbol{R}_q\boldsymbol{\theta}) + \eta^2 - 2\eta a \left[(\boldsymbol{R}_q\boldsymbol{\theta})^T (\boldsymbol{R}_q\boldsymbol{\theta}) \right]^{\frac{1}{2}}}{2(a-1)} \right] \\
&\qquad \times \mathbb{1}\left(\eta < \left[(\boldsymbol{R}_q\boldsymbol{\theta})^T (\boldsymbol{R}_q\boldsymbol{\theta}) \right]^{\frac{1}{2}} \leq a\eta \right) \\
&\qquad + \frac{\eta^2(a+1)}{2} \mathbb{1}\left(\left[(\boldsymbol{R}_q\boldsymbol{\theta})^T (\boldsymbol{R}_q\boldsymbol{\theta}) \right]^{\frac{1}{2}} > a\eta \right) \Bigg\} \\
&= \sum_{q=1}^{m} \Bigg\{ \eta \left[(\boldsymbol{e}_q^T\boldsymbol{\theta})^2 \right]^{\frac{1}{2}} \mathbb{1}\left(0 \leq \left[(\boldsymbol{e}_q^T\boldsymbol{\theta})^2 \right]^{\frac{1}{2}} \leq \eta \right) \\
&\qquad - \left[\frac{(\boldsymbol{e}_q^T\boldsymbol{\theta})^2 + \eta^2 - 2\eta a \left[(\boldsymbol{e}_q^T\boldsymbol{\theta})^2 \right]^{\frac{1}{2}}}{2(a-1)} \right] \mathbb{1}\left(\eta < \left[(\boldsymbol{e}_q^T\boldsymbol{\theta})^2 \right]^{\frac{1}{2}} \leq a\eta \right) \\
&\qquad + \frac{\eta^2(a+1)}{2} \mathbb{1}\left(\left[(\boldsymbol{e}_q^T\boldsymbol{\theta})^2 \right]^{\frac{1}{2}} > a\eta \right) \Bigg\} \\
&= \sum_{q=1}^{m} \Bigg\{ \eta |\boldsymbol{e}_q^T\boldsymbol{\theta}| \mathbb{1}\left(0 \leq |\boldsymbol{e}_q^T\boldsymbol{\theta}| \leq \eta \right) \\
&\qquad - \left[\frac{(\boldsymbol{e}_q^T\boldsymbol{\theta})^2 + \eta^2 - 2\eta a |\boldsymbol{e}_q^T\boldsymbol{\theta}|}{2(a-1)} \right] \mathbb{1}\left(\eta < |\boldsymbol{e}_q^T\boldsymbol{\theta}| \leq a\eta \right) \\
&\qquad + \frac{\eta^2(a+1)}{2} \mathbb{1}\left(|\boldsymbol{e}_q^T\boldsymbol{\theta}| > a\eta \right) \Bigg\} \\
&= \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) \\
&\qquad\qquad + \frac{\eta^2(a+1)}{2} \big(|\theta_q| > a\eta \big) \Bigg\},
\end{aligned}
$$

where $a > 2$ is an additional tuning parameter.

Mcp

$$
\begin{aligned}
\mathcal{P}_{\eta}^{M}(\boldsymbol{\theta}) &= \sum_{q=1}^{m} \mathcal{P}_{\eta,q}^{M}(||\boldsymbol{R}_q\boldsymbol{\theta}||_1) \\
&= \sum_{q=1}^{q^\star} \Bigg\{ \left(\eta||\boldsymbol{R}_q\boldsymbol{\theta}||_1 - \frac{(\boldsymbol{R}_q\boldsymbol{\theta})^T(\boldsymbol{R}_q\boldsymbol{\theta})}{2a}\right) \mathbb{1}\left(0 \leq ||\boldsymbol{R}_q\boldsymbol{\theta}||_1 \leq a\eta\right) \\
&\qquad + \frac{\eta^2 a}{2}\mathbb{1}\left(||\boldsymbol{R}_q\boldsymbol{\theta}||_1 > a\eta\right) \Bigg\} \\
&= \sum_{q=1}^{m} \Bigg\{ \left(\eta\left[(\boldsymbol{R}_q\boldsymbol{\theta})^T(\boldsymbol{R}_q\boldsymbol{\theta})\right]^{\frac{1}{2}} - \frac{(\boldsymbol{R}_q\boldsymbol{\theta})^T(\boldsymbol{R}_q\boldsymbol{\theta})}{2a}\right) \mathbb{1}\left(0 \leq \left[(\boldsymbol{R}_q\boldsymbol{\theta})^T(\boldsymbol{R}_q\boldsymbol{\theta})\right]^{\frac{1}{2}} \leq a\eta\right) \\
&\qquad + \frac{\eta^2 a}{2}\mathbb{1}\left(\left[(\boldsymbol{R}_q\boldsymbol{\theta})^T(\boldsymbol{R}_q\boldsymbol{\theta})\right]^{\frac{1}{2}} > a\eta\right) \Bigg\} \\
&= \sum_{q=1}^{m} \Bigg\{ \left(\eta\left[(\boldsymbol{e}_q^T\boldsymbol{\theta})^2\right]^{\frac{1}{2}} - \frac{(\boldsymbol{e}_q^T\boldsymbol{\theta})^2}{2a}\right) \mathbb{1}\left(0 \leq \left[(\boldsymbol{e}_q^T\boldsymbol{\theta})^2\right]^{\frac{1}{2}} \leq a\eta\right) \\
&\qquad + \frac{\eta^2 a}{2}\mathbb{1}\left(\left[(\boldsymbol{e}_q^T\boldsymbol{\theta})^2\right]^{\frac{1}{2}} > a\eta\right) \Bigg\} \\
&= \sum_{q=1}^{m} \left\{ \left(\eta|\boldsymbol{e}_q^T\boldsymbol{\theta}| - \frac{(\boldsymbol{e}_q^T\boldsymbol{\theta})^2}{2a}\right) \mathbb{1}\left(0 \leq |\boldsymbol{e}_q^T\boldsymbol{\theta}| \leq a\eta\right) + \frac{\eta^2 a}{2}\mathbb{1}\left(|\boldsymbol{e}_q^T\boldsymbol{\theta}| > a\eta\right) \right\} \\
&= \sum_{q=1}^{q^\star} \left\{ \left(\eta|\theta_q| - \frac{\theta_q^2}{2a}\right) \mathbb{1}(0 \leq |\theta_q| \leq a\eta) + \frac{\eta^2 a}{2}\mathbb{1}(|\theta_q| > a\eta) \right\},
\end{aligned}
$$

where $a > 1$ is an additional tuning parameter.

B.2 The penalty matrices

Based on the approximation derived in Section 2.3, the penalty matrix $\boldsymbol{\mathcal{S}}_{\eta}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ is defined as

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

for $\mathcal{T} = \{L, A, S, M\}$. Recall that $\boldsymbol{R}_q = \text{diag}(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ for $q = 1, \ldots, q^\star$ where the 1 on the $(q,q)^{\text{th}}$ entry of the matrix corresponds to the q^{th} parameter in $\boldsymbol{\theta}$, and $\boldsymbol{R}_q = \boldsymbol{O}_{m\times m}$ for $q = q^\star + 1, \ldots, m$. Therefore, the penalty

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

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

The first block is composed by the $q^\star \times q^\star$ diagonal matrix $\boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ and corresponds to the penalized parameters (i.e., the $q^\star$ factor loadings), whereas the second block is an $(m - q^\star)$-dimensional null matrix relative to the unpenalized parameters (i.e., the factor variances and covariances and the unique variances). The matrix $\boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ has the following structure

$$\boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} m_1^{\mathcal{T}} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & \dots & m_q^{\mathcal{T}} & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & 0 & \dots & m_{q^\star}^{\mathcal{T}} \end{bmatrix},$$

where the diagonal entries

$$m_q^{\mathcal{T}} = \frac{\partial \mathcal{P}_{\eta,q}^{\mathcal{T}}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T \boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \qquad \text{for } q = 1, \dots, q^\star \tag{B.1}$$

determine the amount of shrinkage on $\tilde{\theta}_q$ controlled by the tuning η and required by penalty $\mathcal{T}$. We now derive their expressions for the lasso, alasso, scad and mcp.

Lasso

The derivative of the lasso penalty with respect to the L_1 norm of its argument is simply the tuning parameter, that is,

$$\frac{\partial \mathcal{P}_{\eta,q}^{L}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} = \frac{\partial \left(\eta ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1\right)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} = \eta.$$

Therefore,

$$\left[\boldsymbol{\mathcal{M}}_{\eta}^{L}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^L = \frac{\partial \mathcal{P}_{\eta,q}^{L}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+\bar{c}}}$$
$$= \frac{\eta}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+\bar{c}}} = \frac{\eta}{\sqrt{\tilde{\theta}_q^2+\bar{c}}}.$$

Alasso

Similarly, the derivative of the alasso penalty with respect to the L_1 norm of its argument is the tuning parameter multiplied by the adaptive weight, that is,

$$\frac{\partial \mathcal{P}_{\eta,q}^{A}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} = \frac{\partial}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\left(\eta\frac{||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}{\left|\left|\boldsymbol{R}_q\hat{\boldsymbol{\theta}}\right|\right|_1^a}\right)$$
$$= \eta\frac{1}{\left|\left|\boldsymbol{R}_q\hat{\boldsymbol{\theta}}\right|\right|_1^a} = \eta\frac{1}{|\hat{\theta}_q|^a} = \eta\, w_q.$$

Therefore,

$$\left[\boldsymbol{\mathcal{M}}_{\eta}^{A}(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^A = \frac{\partial \mathcal{P}_{\eta,q}^{A}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}\frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+\bar{c}}}$$
$$= \eta\, w_q\frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}+\bar{c}}}$$
$$= \frac{\eta}{|\hat{\theta}_q|^a\sqrt{\tilde{\theta}_q^2+\bar{c}}},$$

where $\hat{\boldsymbol{\theta}}$ is generally the maximum likelihood estimator $\hat{\boldsymbol{\theta}}^{\text{MLE}}$.

Scad

The derivative of the scad penalty with respect to the L_1 norm of its argument has the form:

$$\frac{\partial \mathcal{P}_{\eta,q}^{S}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} = \eta\left\{\mathbb{1}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1\leq\eta) + \frac{\max(a\eta-||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1,0)}{(a-1)\eta}\mathbb{1}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1>\eta)\right\}$$

$$= \begin{cases} \eta & \text{if } |\tilde{\theta}_q| \leq \eta, \\ \dfrac{\max(a\eta - |\tilde{\theta}_q|, 0)}{a-1} & \text{if } |\tilde{\theta}_q| > \eta, \end{cases}$$

which leads to the following expression

$$\begin{aligned} \left[\boldsymbol{\mathcal{M}}_\eta^S(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^S &= \frac{\partial \mathcal{P}_{\eta,q}^S(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \\ &= \eta \left\{ \mathbb{1}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1 \leq \eta) + \frac{\max(a\eta - ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1, 0)}{(a-1)\eta} \mathbb{1}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1 > \eta) \right\} \\ &\quad \times \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \\ &= \frac{\eta \left[\mathbb{1}(|\tilde{\theta}_q| \leq \eta) + \dfrac{\max(a\eta - |\tilde{\theta}_q|, 0)}{(a-1)\eta} \mathbb{1}(|\tilde{\theta}_q| > \eta) \right]}{\sqrt{\tilde{\theta}_q^2 + \bar{c}}}. \end{aligned}$$

Mcp

The derivative of the mcp penalty with respect to the L_1 norm of its argument is

$$\begin{aligned} \frac{\partial \mathcal{P}_{\eta,q}^M(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} &= \left(\eta - \frac{||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}{a}\right) \mathbb{1}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1 < \eta a) \\ &= \begin{cases} \eta - \dfrac{|\tilde{\theta}_q|}{a} & \text{if } |\tilde{\theta}_q| \leq \eta a, \\ 0 & \text{if } |\tilde{\theta}_q| > \eta a, \end{cases} \end{aligned}$$

which implies that

$$\begin{aligned} \left[\boldsymbol{\mathcal{M}}_\eta^M(\tilde{\boldsymbol{\theta}})\right]_{qq} = m_q^M &= \frac{\partial \mathcal{P}_{\eta,q}^M(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1)}{\partial ||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1} \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \\ &= \left(\eta - \frac{||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1}{a}\right) \mathbb{1}(||\boldsymbol{R}_q\tilde{\boldsymbol{\theta}}||_1 < \eta a) \frac{1}{\sqrt{(\boldsymbol{R}_q\tilde{\boldsymbol{\theta}})^T\boldsymbol{R}_q\tilde{\boldsymbol{\theta}} + \bar{c}}} \\ &= \frac{\left(\eta - \dfrac{|\tilde{\theta}_q|}{a}\right) \mathbb{1}(|\tilde{\theta}_q| < \eta a)}{\sqrt{\tilde{\theta}_q^2 + \bar{c}}}. \end{aligned}$$

APPENDIX

Generalized Information Criterion

This appendix illustrates how the degrees of freedom of the penalized model can be found by deriving the bias term of the Generalized Information Criterion (GIC; Konishi & Kitagawa, 1996), an extension of the Akaike Information Criterion (AIC; Akaike, 1974) to the case where the estimation is not conducted through ordinary maximum likelihood. We follow the exposition in Konishi and Kitagawa (2008) and adapt it to the current context.

Suppose that N observations $\boldsymbol{x}_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\alpha, \ldots, \boldsymbol{x}_N\}$ generated from the unknown true distribution function $G(\boldsymbol{x})$ having density function $g(\boldsymbol{x})$ are realizations of the random vector $\boldsymbol{\mathcal{X}}_N = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\alpha, \ldots, \boldsymbol{X}_N)^T$. In order to capture the structure of the given phenomena, we assume a parametric model that consists of a family of parametric distributions $\{f(\boldsymbol{x}|\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m\}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$ is the m-dimensional vector of unknown parameters and $\boldsymbol{\Theta}$ an open subset of $\mathbb{R}^m$. We assume that the distribution $g(\boldsymbol{x})$ that generated the data is included in the class of parametric models, that is, there exists a parameter vector $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $g(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}_0)$. A statistical model $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ is then obtained by replacing the parameter vector $\boldsymbol{\theta}$ with the penalized maximum likelihood estimator (PMLE) $\hat{\boldsymbol{\theta}}$.

For convenience, we assume that each parameter θ_q in $\boldsymbol{\theta}$ can be expressed in the form of a real-valued function of the distribution of G, that is, the functional $T_q(G)$, where $T_q(G)$ is a function defined on the set of all distributions on the

sample space and does not depend on the sample size N. Then, given data $\boldsymbol{x}_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\alpha, \ldots, \boldsymbol{x}_N\}$, the estimator $\hat{\theta}_q$ for the q^{th} parameter θ_q is given by

$$\hat{\theta}_q = \hat{\theta}_q(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\alpha, \ldots, \boldsymbol{x}_N) = T_q(\hat{G}) \qquad \text{for} \quad q = 1, \ldots, m,$$

in which the unknown probability distribution G has been replaced with the empirical distribution function $\hat{G}$ based on the data. The empirical distribution function is the distribution function for the probability function $\hat{g}(\boldsymbol{x}_\alpha) = \dfrac{1}{N}$ $(\alpha = 1, \ldots, N)$ that gives the equal probability $\dfrac{1}{N}$ for each of the N observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\alpha, \ldots, \boldsymbol{x}_N\}$. Because the estimator $\hat{\theta}_q = T_q(\hat{G})$ depends on the data only through the empirical distribution function $\hat{G}$, the functional is referred to as *statistical functional*.

Let us write the m-dimensional functional vector with $T_q(G)$ as the q^{th} element as

$$\boldsymbol{T}(G) = (T_1(G), \ldots, T_q(G), \ldots, T_m(G))^T,$$

where $\boldsymbol{T}(G)$ is defined as the solution of the implicit equations

$$\int \boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{T}(G)) dG(\boldsymbol{x}) = \mathbf{0}. \tag{C.1}$$

The function $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)^T$ collects the real-valued functions $\psi_q(\boldsymbol{x}, \boldsymbol{T}(G))$ defined on the product space of the sample space and the parameter space $\boldsymbol{\Theta}$. The $\boldsymbol{\psi}$-function $\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{T}(G))$ of the PMLE defined in Section 3.1 is

$$\begin{aligned} \boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{T}(G)) &= \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(\boldsymbol{x}|\boldsymbol{\theta}) - \mathcal{P}_{\boldsymbol{\eta}}^{\mathcal{T}}(\boldsymbol{\theta}) \right\} \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\theta} \right\} \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)}, \end{aligned}$$

where the penalty term $\mathcal{P}_{\boldsymbol{\eta}}^{\mathcal{T}}(\boldsymbol{\theta}) = \dfrac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\theta}$ is a twice-continuously differentiable function, $\mathcal{T} = \{L, A, S, M\}$ and $\tilde{\boldsymbol{\theta}}$ is an initial value close to the true value of $\boldsymbol{\theta}$. In case of the normal linear factor model (Section 2.1), the log-likelihood of the sample is as in equation (2.2), the vector of the tuning parameters $\boldsymbol{\eta}$

reduces to the scalar η, and $\boldsymbol{\mathcal{S}}_{\eta}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ is as in equation (2.13). In case of the multiple-group factor model (Section 5.1), the log-likelihood of the sample is as in (5.2), the vector of tuning parameters $\boldsymbol{\eta}$ is equal to the triplet $(\eta_1, \eta_2, \eta_3)^T$, and $\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{D}}_{\eta_1}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{\mathcal{D}}_{\eta_2}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{\mathcal{D}}_{\eta_3}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$. Then, the m-dimensional PMLE $\hat{\boldsymbol{\theta}}$ can be expressed as

$$\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G}) = (T_1(\hat{G}), \ldots, T_q(\hat{G}), \ldots, T_m(\hat{G}))^T,$$

where $\boldsymbol{T}(\hat{G})$ is defined as the solution of the system of penalized likelihood equations

$$\sum_{\alpha=1}^{N} \boldsymbol{\psi}(\boldsymbol{x}_\alpha, \boldsymbol{T}(\hat{G})) = \sum_{\alpha=1}^{N} \boldsymbol{\psi}(\boldsymbol{x}_\alpha, \hat{\boldsymbol{\theta}}) = \mathbf{0},$$

with

$$\boldsymbol{\psi}(\boldsymbol{x}_\alpha, \hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(\boldsymbol{x}_\alpha|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\boldsymbol{\theta})\boldsymbol{\theta} \right\} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Once the model has been constructed, the interest usually lies in its evaluation from the standpoint of making a prediction. The idea is thus to evaluate the expected goodness of the estimated model $f(\boldsymbol{z}|\hat{\boldsymbol{\theta}})$ when it is used to predict the independent future data $\boldsymbol{Z} = \boldsymbol{z}$ generated from the unknown true distribution $g(\boldsymbol{z})$. Specifically, the goodness of the statistical model $f(\boldsymbol{z}|\hat{\boldsymbol{\theta}})$ can be assessed by evaluating its closeness to the true distribution $g(\boldsymbol{z})$ in terms of the Kullback-Leibler (K-L) information

$$\begin{aligned} I(g(\boldsymbol{z}); f(\boldsymbol{z}|\hat{\boldsymbol{\theta}})) &:= \mathbb{E}_{G(\boldsymbol{z})}\left[\log\left\{\frac{g(\boldsymbol{Z})}{f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}})}\right\}\right] = \int \log\left\{\frac{g(\boldsymbol{z})}{f(\boldsymbol{z}|\hat{\boldsymbol{\theta}})}\right\} g(\boldsymbol{z})d\boldsymbol{z} \\ &= \int g(\boldsymbol{z}) \log g(\boldsymbol{z})d\boldsymbol{z} - \int g(\boldsymbol{z}) \log f\left(\boldsymbol{z}|\hat{\boldsymbol{\theta}}\right) d\boldsymbol{z}, \end{aligned} \tag{C.2}$$

where the expectation is taken with respect to the unknown true probability distribution function $G(\boldsymbol{z})$. Because the first term on the right-hand side of equation (C.2) is a constant that depends solely on the true model g, in order to compare different models it is sufficient to consider only the second term on the right-hand side, called the expected log-likelihood:

$$\varphi(\boldsymbol{\mathcal{X}}_N; G) \coloneqq \mathbb{E}_{G(\boldsymbol{z})}[\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}}(\boldsymbol{\mathcal{X}}_N))] = \int g(\boldsymbol{z}) \log f(\boldsymbol{z}|\hat{\boldsymbol{\theta}}) d\boldsymbol{z} = \int \log f(\boldsymbol{z}|\hat{\boldsymbol{\theta}}) dG(\boldsymbol{z}). \tag{C.3}$$

The larger this value is for a model, the smaller its K-L information and the closer the model is to the true one. The expected log-likelihood still depends on the true distribution g and is an unknown quantity that eludes explicit computation. A good estimate of the expected log-likelihood can be obtained from the data by replacing G with $\hat{G}$, that is,

$$\varphi(\boldsymbol{\mathcal{X}}_N; \hat{G}) = \mathbb{E}_{\hat{G}}[\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}})] = \int \log f(\boldsymbol{z}|\hat{\boldsymbol{\theta}}) d\hat{G}(\boldsymbol{z}) = \sum_{\alpha=1}^{N} \hat{g}(\boldsymbol{x}_\alpha) \log f(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{\alpha=1}^{N} \log f(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}). \tag{C.4}$$

According to the law of large numbers, when the number of observations N tends to infinity, the mean of the random variables $\boldsymbol{Y}_\alpha = \log f(\boldsymbol{X}_\alpha)$ $(\alpha = 1, \ldots, N)$ converges in probability to its expectation, that is,

$$\varphi(\boldsymbol{\mathcal{X}}_N; \hat{G}) = \frac{1}{N} \sum_{\alpha=1}^{N} \log f(\boldsymbol{X}_\alpha) = \frac{1}{N} \log f(\boldsymbol{\mathcal{X}}_N|\hat{\boldsymbol{\theta}}(\boldsymbol{\mathcal{X}}_N)) \xrightarrow{N\to\infty} \mathbb{E}_G[\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}})] = \varphi(\boldsymbol{\mathcal{X}}_N; G).$$

Therefore, the estimate based on the empirical distribution function is a natural estimate of the expected log-likelihood. The estimate of the expected log-likelihood multiplied by N is the log-likelihood of the statistical model $f(\boldsymbol{z}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_N))$

$$N \int \log f(\boldsymbol{z}|\hat{\boldsymbol{\theta}}) d\hat{G}(\boldsymbol{z}) = \sum_{\alpha=1}^{N} \log f(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_N)) = \log f(\boldsymbol{x}_N|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_N)) = \ell(\hat{\boldsymbol{\theta}}).$$

It is worth noting that the estimator of the expected log-likelihood $\mathbb{E}_G[\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}})]$ is $\frac{1}{N}\ell(\hat{\boldsymbol{\theta}})$ and that the log-likelihood $\ell(\hat{\boldsymbol{\theta}})$ is an estimator of $N\,\mathbb{E}_G[\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}})]$.

In this procedure, the log-likelihood in (C.4) was obtained by estimating the

expected log-likelihood $\mathbb{E}_G[\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}})]$ by reusing the data $\boldsymbol{x}_N$ that were initially used to estimate the model $f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}})$ in place of the future data. The use of the same data twice for estimating the parameters and the evaluation measure (expected log-likelihood) of the goodness of the estimated model gives rise to bias. Specifically, the bias of the log-likelihood as an estimator of the expected log-likelihood given in (C.3) is defined as

$$\begin{aligned} b(G) &:= \mathbb{E}_G\{\varphi(\boldsymbol{\mathcal{X}}_N;\hat{G}) - \varphi(\boldsymbol{\mathcal{X}}_N;G)\} \\ &= \mathbb{E}_{G(\boldsymbol{x}_N)}\left[\frac{1}{N}\log f(\boldsymbol{\mathcal{X}}_N|\hat{\boldsymbol{\theta}}(\boldsymbol{\mathcal{X}}_N)) - \mathbb{E}_{G(\boldsymbol{z})}[\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}}(\boldsymbol{\mathcal{X}}_N))]\right], \end{aligned}$$

where the expectation $\mathbb{E}_{G(\boldsymbol{x}_N)}$ is taken with respect to the joint distribution $G(\boldsymbol{x}_N) = \prod_{\alpha=1}^{N} G(\boldsymbol{x}_\alpha)$ of the sample $\boldsymbol{\mathcal{X}}_N$. The prerequisite for a fair comparison of models is thus the evaluation of and the correction for this bias term. The general form of the Generalized Information Criterion, which is defined as a bias-corrected log-likelihood, can be constructed by evaluating the bias and correcting for it as follows:

$$\begin{aligned} GIC(\boldsymbol{\mathcal{X}}_N;\hat{G}) &= -2N\left(\frac{1}{N}\sum_{\alpha=1}^{N}\log f(\boldsymbol{X}_\alpha|\hat{\boldsymbol{\theta}}) - b(\hat{G})\right) \\ &= -2\sum_{\alpha=1}^{N}\log f(\boldsymbol{X}_\alpha|\hat{\boldsymbol{\theta}}) + 2N\, b(\hat{G}). \end{aligned} \tag{C.5}$$

The GIC represents an extension of the AIC (see Konishi & Kitagawa, 2008 for a full exposition on the topic). In the same spirit, we can formulate a Generalized Bayesian Information Criterion (GBIC) as an extension of the Bayesian Information Criterion (BIC; Schwarz, 1978)

$$GBIC(\boldsymbol{\mathcal{X}}_N;\hat{G}) = -2\sum_{\alpha=1}^{N}\log f(\boldsymbol{X}_\alpha|\hat{\boldsymbol{\theta}}) + \log(N)N\, b(\hat{G}), \tag{C.6}$$

by changing the weight given to the bias term $b(\hat{G})$ from 2 to $\log(N)$ used in the BIC.

Konishi and Kitagawa (1996) showed that the asymptotic bias of the log-

likelihood in the estimation of the expected log-likelihood can be represented as the integral of the product of the influence function of the employed estimator and the score function of the probability model, i.e.,

$$\begin{aligned}\mathbb{E}_G[\varphi(\boldsymbol{\mathcal{X}}_N;\hat{G})-\varphi(\boldsymbol{\mathcal{X}}_N;G)] &= \left[\frac{1}{N}\sum_{\alpha=1}^{N}\log f(\boldsymbol{X}_\alpha|\hat{\boldsymbol{\theta}})-\int\log f(\boldsymbol{z}|\hat{\boldsymbol{\theta}})dG(\boldsymbol{z})\right]\\ &= \frac{1}{N}b_1(G)+o\left(\frac{1}{N}\right),\end{aligned}$$

where

$$b_1(G)=\mathrm{tr}\left\{\int\boldsymbol{T}^{(1)}(\boldsymbol{z};G)\frac{\partial\log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)}dG(\boldsymbol{z})\right\}. \tag{C.7}$$

The quantity $\boldsymbol{T}^{(1)}(\boldsymbol{z};G)$ is the influence function of the m-dimensional functional $\boldsymbol{T}(G)$ at the true distribution G. The influence function $\boldsymbol{T}^{(1)}(\boldsymbol{z};G)=(T_1^{(1)}(\boldsymbol{z};G),\ldots,T_q^{(1)}(\boldsymbol{z};G),\ldots,T_m^{(1)}(\boldsymbol{z};G))^T$ describes the effect of an infinitesimal contamination at $\boldsymbol{z}$. Its components $T_q^{(1)}(\boldsymbol{z},G)\,(q=1,\ldots,m)$ are defined in terms of the directional derivative of the functional $T_q(G)$ with respect to G, that is,

$$\begin{aligned}\lim_{\epsilon\to0}\frac{T_q((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}})-T_q(G)}{\epsilon} &= \frac{\partial}{\partial\epsilon}\{T_q((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}})\}\Big|_{\epsilon=0}\\ &= \int T_q^{(1)}(\boldsymbol{z};G)d\delta_{\boldsymbol{z}}:=T_q^{(1)}(\boldsymbol{z};G),\end{aligned}$$

where $\delta_{\boldsymbol{z}}$ is a point mass at $\boldsymbol{z}$.

The expression of the influence function of the PMLE can be found by calculating the derivative of the corresponding functional. Firstly, substitute $(1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}}$ for G in equation (C.1):

$$\int\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{T}((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}}))d\{(1-\epsilon)G(\boldsymbol{x})+\epsilon\delta_{\boldsymbol{z}}(\boldsymbol{x})\} =$$

$$\int\frac{\partial}{\partial\boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta})-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}})}d\{(1-\epsilon)G(\boldsymbol{x})+\epsilon\delta_{\boldsymbol{z}}(\boldsymbol{x})\} = \mathbf{0}.$$

Secondly, differentiate both sides of the equation with respect to ϵ:

$$\int \frac{\partial}{\partial \epsilon}\left\{\frac{\partial}{\partial \boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}})} d\{(1-\epsilon)G(\boldsymbol{x}) + \epsilon\delta_{\boldsymbol{z}}(\boldsymbol{x})\}\right\} = \mathbf{0}$$

$$\int \frac{\partial}{\partial \boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}})} \frac{\partial}{\partial \epsilon} d\{(1-\epsilon)G(\boldsymbol{x}) + \epsilon\delta_{\boldsymbol{z}}(\boldsymbol{x})\}$$

$$+ \int \frac{\partial}{\partial \epsilon}\left\{\frac{\partial}{\partial \boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}})}\right\} d\{(1-\epsilon)G(\boldsymbol{x}) + \epsilon\delta_{\boldsymbol{z}}(\boldsymbol{x})\} = \mathbf{0}$$

$$\int \frac{\partial}{\partial \boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}})} d\{-G(\boldsymbol{x}) + \delta_{\boldsymbol{z}}(\boldsymbol{x})\}$$

$$+ \int \frac{\partial}{\partial \boldsymbol{\theta}^T}\frac{\partial}{\partial \boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}((1-\epsilon)G+\epsilon\delta_{\boldsymbol{z}})}$$
$$\times \frac{\partial}{\partial \epsilon}\{\boldsymbol{T}((1-\epsilon)G + \epsilon\delta_{\boldsymbol{z}})\} d\{(1-\epsilon)G(\boldsymbol{x}) + \epsilon\delta_{\boldsymbol{z}}(\boldsymbol{x})\} = \mathbf{0}.$$

Then set $\epsilon = 0$:

$$\int \frac{\partial}{\partial \boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} d\{\delta_{\boldsymbol{z}}(\boldsymbol{x}) - G(\boldsymbol{x})\}$$

$$+ \int \frac{\partial}{\partial \boldsymbol{\theta}}\frac{\partial}{\partial \boldsymbol{\theta}^T}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \frac{\partial}{\partial \epsilon}\{\boldsymbol{T}((1-\epsilon)G + \epsilon\delta_{\boldsymbol{z}})\}\Big|_{\epsilon=0} dG(\boldsymbol{x}) = \mathbf{0}$$

$$\int \frac{\partial}{\partial \boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} d\delta_{\boldsymbol{z}}(\boldsymbol{x})$$
$$\underbrace{- \int \frac{\partial}{\partial \boldsymbol{\theta}}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{x})}_{=0 \text{ by eq. (C.1)}}$$
$$+ \int \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{x})\frac{\partial}{\partial \epsilon}\left\{\boldsymbol{T}((1-\epsilon)G + \epsilon\delta_{\boldsymbol{z}})\right\}\Big|_{\epsilon=0} = \mathbf{0}$$

$$\int \frac{\partial}{\partial \boldsymbol{\theta}} \left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} d\delta_{\boldsymbol{z}}(\boldsymbol{x})$$

$$+ \int \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{x}) \frac{\partial}{\partial \epsilon} \left\{ \boldsymbol{T}((1-\epsilon)G + \epsilon\delta_{\boldsymbol{z}}) \right\} \Big|_{\epsilon=0} = \mathbf{0}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\log f(\boldsymbol{z}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)}$$

$$+ \int \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \left[\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{x}) \underbrace{\frac{\partial}{\partial \epsilon} \left\{ \boldsymbol{T}((1-\epsilon)G + \epsilon\delta_{\boldsymbol{z}}) \right\} \Big|_{\epsilon=0}}_{=\boldsymbol{T}^{(1)}(\boldsymbol{z};G)} = \mathbf{0}.$$

Consequently, the influence function $\boldsymbol{T}^{(1)}(\boldsymbol{z}; G)$ that defines the PMLE is given by

$$\begin{aligned}
\boldsymbol{T}^{(1)}(\boldsymbol{z}; G) &:= \frac{\partial}{\partial \epsilon} \left\{ \boldsymbol{T}((1-\epsilon)G + \epsilon\delta_{\boldsymbol{z}}) \right\} \Big|_{\epsilon=0} \\
&= -\left\{ \int \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \left[\log f(\boldsymbol{z}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \right\}^{-1} \\
&\qquad \times \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \left[\log f(\boldsymbol{z}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \right\} \\
&= \boldsymbol{R}(\boldsymbol{\psi}, G)^{-1} \boldsymbol{\psi}(\boldsymbol{z}; \boldsymbol{T}(G)),
\end{aligned} \tag{C.8}$$

where $\boldsymbol{R}(\boldsymbol{\psi}, G)$ is an $m \times m$ matrix defined as

$$\begin{aligned}
\boldsymbol{R}(\boldsymbol{\psi}, G) &= -\int \frac{\partial \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\
&= -\int \frac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) + \int \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \left(\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}).
\end{aligned}$$

More specifically, for the normal linear factor model, if we denote $\boldsymbol{\theta} = (\boldsymbol{\theta}^\star, \check{\boldsymbol{\theta}})^T$, where $\boldsymbol{\theta}^\star$ collects the penalized parameters and $\check{\boldsymbol{\theta}}$ the unpenalized parameters, we have that:

$$\frac{\partial \boldsymbol{\psi}(\boldsymbol{z},\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \dfrac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\star \partial \boldsymbol{\theta}^{\star T}} - \boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) & \dfrac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\star \partial \check{\boldsymbol{\theta}}^T} \\ \dfrac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \check{\boldsymbol{\theta}} \partial \boldsymbol{\theta}^{\star T}} & \dfrac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \check{\boldsymbol{\theta}} \partial \check{\boldsymbol{\theta}}^T} \end{bmatrix},$$

where $\boldsymbol{\mathcal{M}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ is the sub-matrix of $\boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}})$ corresponding to the penalized parameters defined in Section 2.3, and the tuning parameter vector $\boldsymbol{\eta}$ reduces to the scalar η.

By substituting the expression of the influence function of the PMLE into equation (C.7), we get the following expression of the bias:

$$\begin{aligned} b_1(G) &= \operatorname{tr}\left\{ \int \boldsymbol{T}^{(1)}(\boldsymbol{z};G) \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \right\} \\ &= \operatorname{tr}\left\{ \int \boldsymbol{R}(\boldsymbol{\psi},G)^{-1} \boldsymbol{\psi}(\boldsymbol{z},\boldsymbol{T}(G)) \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \right\} \\ &= \operatorname{tr}\left\{ \boldsymbol{R}(\boldsymbol{\psi},G)^{-1} \int \boldsymbol{\psi}(\boldsymbol{z};\boldsymbol{T}(G)) \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \right\} \\ &= \operatorname{tr}\left\{ \boldsymbol{R}(\boldsymbol{\psi},G)^{-1} \boldsymbol{Q}(\boldsymbol{\psi},G) \right\}, \end{aligned}$$

where $\boldsymbol{Q}(\boldsymbol{\psi},G)$ is an $m \times m$ matrix defined as

$$\begin{aligned} \boldsymbol{Q}(\boldsymbol{\psi},G) &= \int \boldsymbol{\psi}(\boldsymbol{z};\boldsymbol{T}(G)) \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(\boldsymbol{z}|\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\theta} \right\}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\ &= \int \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\ &\qquad - \int \boldsymbol{\mathcal{S}}_\eta^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{T}(G) \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\ &= \int \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) \\ &= -\int \frac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) = \boldsymbol{Q}(G). \end{aligned}$$

The fourth line follows from the fact that as $N \to \infty$ (see the conditions in Appendix E.1)

$$\mathbf{0} = \int \frac{\partial}{\partial \boldsymbol{\theta}} \left[\log f(\boldsymbol{z}|\boldsymbol{\theta}) - \left(\frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\theta} \right) \right] \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}) = \int \frac{\partial \log f(\boldsymbol{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} dG(\boldsymbol{z}).$$

Let $b_1(\hat{G})$ be a bias estimate obtained by replacing the unknown distribution G with the empirical distribution $\hat{G}$:

$$\begin{aligned} b_1(\hat{G}) &= \operatorname{tr} \left\{ \frac{1}{N} \sum_{\alpha=1}^{N} \boldsymbol{T}^{(1)}(\boldsymbol{x}_\alpha, \hat{G}) \frac{\partial \log f(\boldsymbol{x}_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(\hat{G})} \right\} \\ &= \operatorname{tr} \left\{ \boldsymbol{R}(\boldsymbol{\psi}, \hat{G})^{-1} \boldsymbol{Q}(\hat{G}) \right\}. \end{aligned} \tag{C.9}$$

The quantity $\boldsymbol{T}^{(1)}(\boldsymbol{x}_\alpha, \hat{G})$ represents the vector of empirical influence functions, whose components $T_q^{(1)}(\boldsymbol{x}_\alpha, \hat{G})$ are defined as the derivative of $T_q(\hat{G})$ with respect to the probability measure $\delta_{\boldsymbol{x}_\alpha}$ being the point mass at $\boldsymbol{x}_\alpha$, that is,

$$T_q^{(1)}(\boldsymbol{x}_\alpha, \hat{G}) = \lim_{\epsilon \to 0} \frac{T_q((1-\epsilon)\hat{G} + \epsilon \delta_{\boldsymbol{x}_\alpha}) - T_q(\hat{G})}{\epsilon}.$$

The matrices $\boldsymbol{R}(\boldsymbol{\psi}, \hat{G})$ and $\boldsymbol{Q}(\hat{G})$ are as follows:

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

$$\boldsymbol{Q}(\hat{G}) = -\frac{1}{N} \sum_{\alpha=1}^{N} \frac{\partial^2 \log f(\boldsymbol{x}_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(\hat{G})} = -\frac{1}{N} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\boldsymbol{T}(\hat{G})} = -\frac{1}{N} \boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}).$$

The estimated bias $b_1(\hat{G})$ is an estimate of the effective degrees of freedom (*edf*) of the penalized model, that is,

$$
\begin{aligned}
edf = b_1(\hat{G}) &= \operatorname{tr}\left\{\left[-\frac{1}{N}\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})\right]^{-1}\left[-\frac{1}{N}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right]\right\} \\
&= \operatorname{tr}\left\{\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\}. \qquad \text{(C.10)}
\end{aligned}
$$

By substituting the asymptotic bias estimate in equation (C.10) into the expressions of the GIC (eq. C.5) and the GBIC (eq. C.6), the following generalized information criteria are obtained:

$$
\begin{aligned}
GIC(\boldsymbol{\mathcal{X}}_N; \hat{G}) &= -2N\left\{\frac{1}{N}\sum_{\alpha=1}^{N}\log f(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) - \frac{1}{N}b_1(\hat{G})\right\} \\
&= -2\sum_{\alpha=1}^{N}\log f(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + 2\operatorname{tr}\{\boldsymbol{R}(\boldsymbol{\psi}, \hat{G})^{-1}\boldsymbol{Q}(\hat{G})\} \\
&= -2\,\ell(\hat{\boldsymbol{\theta}}) + 2\operatorname{tr}\left\{\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\}, \\
GBIC(\boldsymbol{\mathcal{X}}_N; \hat{G}) &= -2\,\ell(\hat{\boldsymbol{\theta}}) + \log(N)\operatorname{tr}\left\{\boldsymbol{\mathcal{H}}_p(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right\}.
\end{aligned}
$$

The vector of tuning parameters $\boldsymbol{\eta}$ enters through the penalty matrix, which is included in $\boldsymbol{\mathcal{H}}_p$. The determination of the tuning parameter(s) can be viewed as a model selection and evaluation problem. Therefore, information criteria evaluating a penalized model can be used as tuning parameter selectors. By evaluating statistical models determined according to grid(s) of values of $\boldsymbol{\eta}$, we take the optimal vector of the tuning parameter $\hat{\boldsymbol{\eta}}$ to be the one minimizing the value of the GBIC (since the BIC generally selects more sparse models than does the AIC), that is,

$$
\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} GBIC(\boldsymbol{\mathcal{X}}_N; \hat{G}).
$$

APPENDIX D

Details on the penalized estimation framework

This appendix covers the theoretical derivations necessary for the development of the penalized likelihood-based estimation framework proposed in Chapter 3. We maintain a general viewpoint and assume that the vector $\boldsymbol{\eta}$ collects multiple tuning parameters. This tuning vector reduces to the scalar η in the case of the normal linear factor model (Section 2.1), and the triplet $(\eta_1, \eta_2, \eta_3)^T$ in the multiple-group extension (Section 5.1).

D.1 A general expression for the PMLE

To avoid notational clutter, we omit the superscript $\mathcal{T} = \{L, A, S, M\}$ in the expression of the penalty matrix. By using a first-order Taylor expansion of $\boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}+1]})$ at $\boldsymbol{\theta}^{[\mathrm{t}]}$ it follows that

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

where $\boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}]}) = \boldsymbol{g}(\boldsymbol{\theta}^{[\mathrm{t}]}) - N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[\mathrm{t}]})\boldsymbol{\theta}^{[\mathrm{t}]}$ and $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}^{[\mathrm{t}]}) = \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}^{[\mathrm{t}]}) - N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[\mathrm{t}]})$. Define $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]}) = -\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}^{[\mathrm{t}]})$, then

$$\mathbf{0} = \boldsymbol{g}_p(\boldsymbol{\theta}^{[\mathrm{t}]}) + \left[-\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[\mathrm{t}]}) - N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[\mathrm{t}]})\right](\boldsymbol{\theta}^{[\mathrm{t}+1]} - \boldsymbol{\theta}^{[\mathrm{t}]}).$$

By rearranging the above equation, we get:

$$\begin{aligned}
\boldsymbol{g}_p(\boldsymbol{\theta}^{[t]}) &= \left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[t]})\right](\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^{[t]}) \\
\boldsymbol{g}(\boldsymbol{\theta}^{[t]}) - N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[t]})\boldsymbol{\theta}^{[t]} &= \left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[t]})\right]\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]})\boldsymbol{\theta}^{[t]} - N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[t]})\boldsymbol{\theta}^{[t]} \\
\boldsymbol{\theta}^{[t+1]}[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[t]})] &= \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]})\boldsymbol{\theta}^{[t]} + \boldsymbol{g}(\boldsymbol{\theta}^{[t]}) \\
\boldsymbol{\theta}^{[t+1]}[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}}^{[t]})] &= \sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]})}\left[\sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]})}\boldsymbol{\theta}^{[t]} + \sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]})}^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{[t]})\right].
\end{aligned}$$

Therefore, the vector parameter estimator can be expressed as

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

where $\boldsymbol{K}^{[t]} = \boldsymbol{\mu}_{\boldsymbol{K}}^{[t]} + \boldsymbol{\vartheta}^{[t]}$ with $\boldsymbol{\mu}_{\boldsymbol{K}}^{[t]} = \sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]})}\boldsymbol{\theta}^{[t]}$ and $\boldsymbol{\vartheta}^{[t]} = \sqrt{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]})}^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{[t]})$. The square root of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{[t]})$ and its inverse are obtained via eigenvalue decomposition (see Appendix D.2).

D.2 Correction for positive-definiteness

An eigenvalue decomposition is a technique that allows one to express an $m \times m$ symmetric matrix $\boldsymbol{B}$ as

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T,$$

where $\boldsymbol{U}$ is an orthogonal matrix with the eigenvectors in its columns, and $\boldsymbol{D}$ is a diagonal matrix with the corresponding eigenvalues $d_{11}, \ldots, d_{qq}, \ldots, d_{mm}$ in the main diagonal, sorted in descending order. If all the eigenvalues are strictly positive, the matrix is said to be positive-definite, and its inverse is found as $\boldsymbol{B}^{-1} = \boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^T$.

However, if at least one of its eigenvalues is null or negative, the matrix is non-positive definite, and it must be corrected before its inversion takes place. An effective procedure that adjusts the problematic eigenvalues of a non-positive definite matrix, and eventually makes the matrix positive-definite, is the following.

Without loss of generality, assume that all the eigenvalues of $\boldsymbol{B}$ are strictly positive except for the last one, i.e., $d_{qq} > 0$ for $q = 1, \ldots, m-1$ and $d_{mm} \leq 0$. Define $l = \sum_{q=2}^{m} d_{qq}$ and $t = 100l^2 + 1$. The non-positive eigenvalue d_{mm} is then substituted with the positive quantity

$$\tilde{d}_{mm} = d_{m-1,m-1}\frac{(l - d_{mm})^2}{t},$$

where $d_{m-1,m-1}$ is the smallest positive eigenvalue of $\boldsymbol{B}$. By defining $\tilde{\boldsymbol{D}} = \text{diag}(d_{11}, \ldots, d_{qq}, \ldots, \tilde{d}_{mm})$, the corrected positive-definite matrix $\tilde{\boldsymbol{B}}$ can be found as

$$\tilde{\boldsymbol{B}} = \boldsymbol{U}\tilde{\boldsymbol{D}}\boldsymbol{U}^T,$$

and its inverse as

$$\tilde{\boldsymbol{B}}^{-1} = \boldsymbol{U}\tilde{\boldsymbol{D}}^{-1}\boldsymbol{U}^T.$$

We employed this procedure to compute and, if necessary, to correct the square root of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ and its inverse.

D.3 Derivation of the UBRE criterion

Let $\boldsymbol{A}_{\boldsymbol{\eta}} = \sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}\left[\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}})\right]^{-1}\sqrt{\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})}$, where $\boldsymbol{A}_{\boldsymbol{\eta}}$ is used as a shortcut for $\boldsymbol{A}_{\boldsymbol{\eta}}^{\mathcal{T}}$ for $\mathcal{T} = \{L, A, S, M\}$. Based on the derivation in Appendix D.1, we can work out the expression of the UBRE criterion, i.e., the expectation of the average squared distance of $\hat{\boldsymbol{\mu}}_{\boldsymbol{K}} = \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{K}$ from its expected value $\boldsymbol{\mu}_{\boldsymbol{K}}$:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}||\boldsymbol{\mu}_{\boldsymbol{K}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}||_2^2\right] &= \mathbb{E}\left[\frac{1}{N}||(\boldsymbol{K} - \boldsymbol{\vartheta}) - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{K}||_2^2\right] \\
&= \mathbb{E}\left[\frac{1}{N}||(\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{K}) - \boldsymbol{\vartheta}||_2^2\right] \\
&= \frac{1}{N}\mathbb{E}\left[||\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{K}||_2^2 + \boldsymbol{\vartheta}^T\boldsymbol{\vartheta} - 2\boldsymbol{\vartheta}^T(\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{K})\right] \\
&= \frac{1}{N}\mathbb{E}\left[||\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{K}||_2^2\right] + \frac{1}{N}\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{\vartheta}\right] \\
&\quad - \frac{2}{N}\mathbb{E}\left[\boldsymbol{\vartheta}^T[\boldsymbol{\mu}_{\boldsymbol{K}} + \boldsymbol{\vartheta} - \boldsymbol{A}_{\boldsymbol{\eta}}(\boldsymbol{\mu}_{\boldsymbol{K}} + \boldsymbol{\vartheta})]\right]
\end{aligned}$$

$$= \frac{1}{N}\mathbb{E}\left[||\boldsymbol{K} - \boldsymbol{A}_\eta \boldsymbol{K}||_2^2\right] - \frac{1}{N}\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{\vartheta}\right] - \frac{2}{N}\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{\mu}_{\boldsymbol{K}}\right] + \frac{2}{N}\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{A}_\eta\boldsymbol{\mu}_{\boldsymbol{K}}\right] + \frac{2}{N}\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{A}_\eta\boldsymbol{\vartheta}\right].$$

We now use the following results (Wood, 2017, Section 1.8.6)

$$\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{\vartheta}\right] = \mathbb{E}\left[\sum_{\alpha=1}^{N}\boldsymbol{\vartheta}_i^2\right] = N,$$

$$\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{\mu}_{\boldsymbol{K}}\right] = \mathbb{E}\left[\boldsymbol{\vartheta}^T\right]\boldsymbol{\mu}_{\boldsymbol{K}} = \mathbf{0},$$

$$\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{A}_\eta\boldsymbol{\mu}_{\boldsymbol{K}}\right] = \mathbb{E}\left[\boldsymbol{\vartheta}^T\right]\boldsymbol{A}_\eta\boldsymbol{\mu}_{\boldsymbol{K}} = \mathbf{0},$$

$$\mathbb{E}\left[\boldsymbol{\vartheta}^T\boldsymbol{A}_\eta\boldsymbol{\vartheta}\right] = \mathbb{E}\left[\mathrm{tr}\{\boldsymbol{\vartheta}^T\boldsymbol{A}_\eta\boldsymbol{\vartheta}\}\right] = \mathbb{E}\left[\mathrm{tr}\{\boldsymbol{A}_\eta\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T\}\right] = \mathrm{tr}\{\mathbb{E}\left[\boldsymbol{A}_\eta\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T\right]\}$$
$$= \mathrm{tr}\left\{\boldsymbol{A}_\eta\mathbb{E}\left[\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T\right]\right\} = \mathrm{tr}\{\boldsymbol{A}_\eta\boldsymbol{I}\} = \mathrm{tr}(\boldsymbol{A}_\eta).$$

Then the expression of the UBRE criterion is:

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

D.4 Equivalence to the AIC

This section shows that $\mathcal{V}(\boldsymbol{\eta})$ is approximately proportional to the Akaike information criterion (AIC). The AIC of a model is defined as

$$\mathrm{AIC} \coloneqq -2\ell(\boldsymbol{\theta}) + 2m,$$

where m is the number of estimated parameters in the model. Consider the following Taylor expansion of $-2\ell(\hat{\boldsymbol{\theta}})$ about $-2\ell(\boldsymbol{\theta})$:

$$\begin{aligned} -2\ell(\hat{\boldsymbol{\theta}}) &\approx -2\ell(\boldsymbol{\theta}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\nabla_{\boldsymbol{\theta}}[-2\ell(\boldsymbol{\theta})] + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}^T}[-2\ell(\boldsymbol{\theta})](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ &\approx -2\ell(\boldsymbol{\theta}) - 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\boldsymbol{g} - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \end{aligned} \tag{D.1}$$

where we wrote $\boldsymbol{g} \coloneqq \boldsymbol{g}(\boldsymbol{\theta})$ and $\boldsymbol{\mathcal{H}} \coloneqq \boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$ for simplicity of notation. By denoting

$\boldsymbol{\mathcal{I}} = -\boldsymbol{\mathcal{H}}$ and recalling that $\boldsymbol{K} = \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\theta} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}$, we have that

$$\begin{aligned}
(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\boldsymbol{g} &= (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\sqrt{\boldsymbol{\mathcal{I}}}\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} = \left[\sqrt{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right]^T\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= \left[\sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}} - \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\theta}\right]^T\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} = \left[\sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}} - \boldsymbol{K} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right]^T\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= -\left[\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}\right]^T\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} + \boldsymbol{g}^T\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= -\left[\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}\right]^T\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} + ||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 \\
&= -\left\langle \boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle + ||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2, \qquad \text{(D.2)}
\end{aligned}$$

$$\begin{aligned}
-(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = ||\sqrt{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})||_2^2 \\
&= ||\sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}} - \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\theta}||_2^2 = ||\sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}} - \boldsymbol{K} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 \\
&= ||\left(\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}\right) - \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 \\
&= ||\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}||_2^2 + ||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 - 2\left\langle \boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle, \qquad \text{(D.3)}
\end{aligned}$$

where we used the fact that $||\boldsymbol{a}||_2^2 = ||-\boldsymbol{a}||_2^2$ for any vector $\boldsymbol{a}$, and $\langle\cdot,\cdot\rangle$ represents the inner product. By substituting equations (D.2) and (D.3) into expression (D.1), we obtain:

$$\begin{aligned}
-2\ell(\hat{\boldsymbol{\theta}}) &\approx -2\ell(\boldsymbol{\theta}) + 2\left\langle \boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle - 2||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 \\
&\quad + ||\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}||_2^2 + ||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 - 2\left\langle \boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle \\
&= -2\ell(\boldsymbol{\theta}) - ||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 + ||\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}||_2^2.
\end{aligned}$$

It then follows that

$$\begin{aligned}
\text{AIC} &= -2\ell(\boldsymbol{\theta}) + 2m \approx -2\ell(\boldsymbol{\theta}) - ||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 + ||\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}||_2^2 + 2m \\
&\approx -2\ell(\boldsymbol{\theta}) - ||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2 + ||\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}||_2^2 + 2\text{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}), \qquad \text{(D.4)}
\end{aligned}$$

where $\text{tr}(\boldsymbol{A}_{\boldsymbol{\eta}})$ denotes the number of estimated parameters in the model, and thus,

$m = \text{tr}(\boldsymbol{A}_{\boldsymbol{\eta}})$. Since we want to optimize the criterion with respect to the tuning parameter vector $\boldsymbol{\eta}$, we ignore any terms that are not affected by it, like $-2\ell(\boldsymbol{\theta})$ and $||\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}||_2^2$. After dropping these constants, expression (D.4) becomes proportional to the AIC, that is,

$$\text{AIC} = ||\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}||_2^2 + 2\text{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}) \propto \mathcal{V}(\boldsymbol{\eta}),$$

where $||\boldsymbol{K} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\theta}}||_2^2$ is a quadratic approximation of $-2\ell(\hat{\boldsymbol{\theta}})$ and $\text{tr}(\boldsymbol{A}_{\boldsymbol{\eta}})$ represents the effective degrees of freedom of the model.

D.5 Automatic multiple tuning parameter estimation

This section describes how the approach by Wood (2004) for multiple smoothing parameter estimation of generalized additive models and the like can be adapted to the current context. We are interested in estimating the tuning parameters in $\boldsymbol{\eta}$ controlling the amount of penalization. The vector $\boldsymbol{\eta}$ reduces to the scalar η for the normal linear factor model, and the triplet $(\eta_1, \eta_2, \eta_3)^T$ for the multiple-group extension. This procedure implements a Newton's method that evaluates in a stable and computationally efficient way the components in $\mathcal{V}(\boldsymbol{\eta})$ (see equation (3.17)) and their first and second derivatives with respect to the tuning parameters. This numerical strategy for estimating the tuning parameters is called "performance iteration" (Gu, 2013) and consists of the minimization of the UBRE score and the selection of the tuning parameters of the penalized model in each iteration. The technique uses a series of pivoted QR and singular value decompositions (SVD) which make the evaluations of the quantities involving $\boldsymbol{A}_{\boldsymbol{\eta}}^{[\text{t+1}]}$, for a new trial value of $\boldsymbol{\eta}$, cheap and derivative calculations efficient and stable.

In the following, we follow the exposition in Wood (2017, Section 6.5.1) and refer interested readers to it for additional details. Given a tuning parameter vector

value for $\boldsymbol{\eta}$, we can rewrite the iterative equation

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

in a penalized iteratively re-weighted least squares form, that is,

$$||\boldsymbol{K} - \boldsymbol{A}_{\hat{\boldsymbol{\eta}}}\boldsymbol{K}||_2^2 + \frac{N}{2}\boldsymbol{\theta}^T\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta}.$$

The superscript $[\mathfrak{t}]$ has been suppressed from the quantities above and is omitted to avoid clutter; $\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}$ is a shortcut for $\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}$ for $\mathcal{T} = L, A$. The presented approach approximates the UBRE/AIC criterion $\mathcal{V}(\boldsymbol{\eta}) = \frac{1}{N}||\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{K}||_2^2 + \frac{2}{N}\gamma\,\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}) - 1$ in the vicinity of the current best estimate of the tuning parameters with the quadratic function

$$\mathcal{V}(\boldsymbol{\eta}) \approx \mathcal{V}(\boldsymbol{\eta}^{[\mathfrak{t}]}) + (\boldsymbol{\eta} - \boldsymbol{\eta}^{[\mathfrak{t}]})^T\boldsymbol{z} + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^{[\mathfrak{t}]})^T\boldsymbol{Z}(\boldsymbol{\eta} - \boldsymbol{\eta}^{[\mathfrak{t}]}),$$

where $\boldsymbol{z} = \dfrac{\partial\mathcal{V}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}$ and $\boldsymbol{Z} = \dfrac{\partial^2\mathcal{V}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}$ are the first derivative vector and second derivative matrix of $\mathcal{V}$ with respect to the tuning parameters. It can be shown that the minimum of the approximating quadratic function is at

$$\boldsymbol{\eta}^{[\mathfrak{t}+1]} = \arg\min_{\boldsymbol{\eta}} \mathcal{V}(\boldsymbol{\eta}) = \boldsymbol{\eta}^{[\mathfrak{t}]} - \boldsymbol{Z}^{-1}\boldsymbol{z},$$

which can be used as the next estimate of the tuning parameters. A new approximating quadratic is then found by expansion about $\boldsymbol{\eta}^{[\mathfrak{t}+1]}$, and this is minimized to find $\boldsymbol{\eta}^{[\mathfrak{t}+2]}$, with the process being repeated until convergence. This procedure may occasionally fail to converge. Consider the case where, at some iteration, a set of tuning parameter estimates and coefficient estimates, $\{\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}\}$ is obtained; this set in turn implies a certain model and an UBRE score which yield the new set of estimates $\{\check{\boldsymbol{\eta}}, \check{\boldsymbol{\theta}}\}$; this new set of estimates itself yields a new model and UBRE score, which yield a new set of estimates, but these turn out to be $\{\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}\}$. If this happens cyclically, convergence never occurs. Similar problems may involve cycling

through a larger number of sets of estimates. This might happen in the presence of multicollinearity, so users should carefully specify all the dependencies present in the model.

If $\boldsymbol{Z}$ is not positive definite, the quadratic approximation has no unique minimum. In this case, it is advised to search in the steepest descent direction, $-\boldsymbol{z}$, for parameter values that will reduce the score. Also, if the quadratic approximation is poor, stepping to its minimum actually increases the real $\mathcal{V}$. In this case, it is worth trying to successively half the length of the step until a step is found that decreases $\mathcal{V}$; if this fails, then steepest descent can be tried.

Since the expensive part of evaluating the UBRE/AIC criterion is the evaluation of the trace of the influence matrix $\boldsymbol{A}_{\boldsymbol{\eta}} = \sqrt{\hat{\boldsymbol{\mathcal{I}}}} \left[\hat{\boldsymbol{\mathcal{I}}} + N\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}}\right]^{-1} \sqrt{\hat{\boldsymbol{\mathcal{I}}}}$, where $\hat{\boldsymbol{\mathcal{I}}} = \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}} = \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}})$, it is this influence matrix that must be considered first. The first step consists of a QR decomposition of $\sqrt{\hat{\boldsymbol{\mathcal{I}}}}$, i.e., $\sqrt{\hat{\boldsymbol{\mathcal{I}}}} = \boldsymbol{QR}$, where the columns of $\boldsymbol{Q}$ are columns of an orthogonal matrix and $\boldsymbol{R}$ is upper triangular. Wood (2017) suggests the use of a pivoted QR decomposition for maximum stability.

Define $\boldsymbol{B}$ any matrix square root of $N\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}}$, such that $\boldsymbol{B}^T\boldsymbol{B} = N\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}}$. The matrix $\boldsymbol{B}$ can be obtained efficiently by pivoted Choleski decomposition or eigendecomposition of the symmetric matrix $\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}}$. Augmenting $\boldsymbol{R}$ with $\boldsymbol{B}$, a singular value decomposition is then obtained as

$$\begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{B} \end{bmatrix} = \boldsymbol{UDV}^T.$$

The columns of $\boldsymbol{U}$ are columns of an orthogonal matrix, whereas $\boldsymbol{V}$ is an orthogonal matrix. $\boldsymbol{D}$ is the diagonal matrix of singular values: the examination of these is the most reliable way of detecting numerical rank deficiency of the fitting problem (Golub & Van Loan, 2012). Rank deficiency of the fitting problem is dealt with at this stage by removing from $\boldsymbol{D}$ the rows and columns containing the singular values that are "too small", along with the corresponding columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. This has the effect of recasting the original fitting problem into a reduced space in which the model parameters are identifiable. "Too small" is judged with reference

to the largest singular value: for example, singular values less than the largest singular value multiplied by the square root of the machine precision might be deleted.

Now let $\boldsymbol{U}_1$ be the sub-matrix of $\boldsymbol{U}$ such that $\boldsymbol{R} = \boldsymbol{U}_1\boldsymbol{D}\boldsymbol{V}^T$. This implies that $\sqrt{\hat{\boldsymbol{\mathcal{I}}}} = \boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{D}\boldsymbol{V}^T$, while $\hat{\boldsymbol{\mathcal{I}}} + N\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}_1^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{D}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$, and

$$\begin{aligned}\boldsymbol{A}_{\boldsymbol{\eta}} &= \sqrt{\hat{\boldsymbol{\mathcal{I}}}}\left[\hat{\boldsymbol{\mathcal{I}}} + N\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}}\right]^{-1}\sqrt{\hat{\boldsymbol{\mathcal{I}}}} = \boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}_1^T\boldsymbol{Q}^T \\ &= \boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{U}_1^T\boldsymbol{Q}^T.\end{aligned}$$

Hence the trace of the influence matrix is efficiently computed as

$$\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\eta}}) = \mathrm{tr}\{\boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{U}_1^T\boldsymbol{Q}^T\} = \mathrm{tr}\{\boldsymbol{U}_1\boldsymbol{U}_1^T\boldsymbol{Q}^T\boldsymbol{Q}\} = \mathrm{tr}(\boldsymbol{U}_1\boldsymbol{U}_1^T).$$

Notice that the main computational cost is the QR decomposition, but thereafter the evaluation of $\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\eta}})$ is relatively cheap for new trial values of $\boldsymbol{\eta}$.

For efficient minimization of the tuning selection criterion, we also need the expressions of the derivatives of the criterion with respect to the tuning parameters. To this end, it is helpful to write the influence matrix as $\boldsymbol{A}_{\boldsymbol{\eta}} = \sqrt{\hat{\boldsymbol{\mathcal{I}}}}\boldsymbol{G}^{-1}\sqrt{\hat{\boldsymbol{\mathcal{I}}}}$ where $\boldsymbol{G} = \hat{\boldsymbol{\mathcal{I}}} + N\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$ and hence $\boldsymbol{G}^{-1} = \left[\hat{\boldsymbol{\mathcal{I}}} + N\hat{\boldsymbol{\mathcal{S}}}_{\boldsymbol{\eta}}\right]^{-1} = \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T$. Since the tuning parameters must be positive, we can avoid the algorithm to step to negative values by using $\rho_i = \log\eta_i$ as the optimization parameters. We then have that

$$\frac{\partial\boldsymbol{G}^{-1}}{\partial\rho_i} = -\boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\rho_i}\boldsymbol{G}^{-1} = -\eta_i\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T\hat{\boldsymbol{\mathcal{S}}}_{\eta_i}\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T,$$

and so,

$$\begin{aligned}\frac{\partial\boldsymbol{A}_{\boldsymbol{\eta}}}{\partial\rho_i} &= \sqrt{\hat{\boldsymbol{\mathcal{I}}}}\frac{\partial\boldsymbol{G}^{-1}}{\partial\rho_i}\sqrt{\hat{\boldsymbol{\mathcal{I}}}} = -\eta_i\,\boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T\hat{\boldsymbol{\mathcal{S}}}_{\eta_i}\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}_1^T\boldsymbol{Q} \\ &= -\eta_i\,\boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{D}^{-1}\boldsymbol{V}^T\hat{\boldsymbol{\mathcal{S}}}_{\eta_i}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}_1^T\boldsymbol{Q}.\end{aligned}$$

Turning to the second derivatives, we have:

$$\frac{\partial^2\boldsymbol{G}^{-1}}{\partial\rho_i\partial\rho_j} = \boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\rho_j}\boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\rho_i}\boldsymbol{G}^{-1}\boldsymbol{G}^{-1} - \boldsymbol{G}^{-1}\frac{\partial^2\boldsymbol{G}}{\partial\rho_i\partial\rho_j}\boldsymbol{G}^{-1} + \boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\rho_i}\boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\rho_j}\boldsymbol{G}^{-1}\boldsymbol{G}^{-1},$$

and then,

$$\begin{aligned}\frac{\partial^2 \boldsymbol{A}_{\boldsymbol{\eta}}}{\partial\rho_i\partial\rho_j} &= \sqrt{\hat{\boldsymbol{\mathcal{I}}}}\frac{\partial^2 \boldsymbol{G}^{-1}}{\partial\rho_i\partial\rho_j}\sqrt{\hat{\boldsymbol{\mathcal{I}}}} \\ &= \eta_i\eta_j\,\boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{D}^{-1}\boldsymbol{V}^T[\hat{\boldsymbol{\mathcal{S}}}_{\eta_j}\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T\hat{\boldsymbol{\mathcal{S}}}_{\eta_i}]^{\ddagger}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}_1^T\boldsymbol{Q}^T + \delta_j^i\frac{\partial \boldsymbol{A}_{\boldsymbol{\eta}}}{\partial\rho_i},\end{aligned}$$

where $\boldsymbol{B}^{\ddagger} \equiv \boldsymbol{B} + \boldsymbol{B}^T$ and $\delta_j^i = 1$ if $i = j$ and zero otherwise. Writing $\alpha = ||\boldsymbol{K} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{K}||_2^2$, we can now find convenient expressions for the component derivatives needed to find the derivatives of the UBRE score. Define $\boldsymbol{y}_1 = \boldsymbol{U}_1^T\boldsymbol{Q}^T\boldsymbol{K}$, $\boldsymbol{Z}_i = \boldsymbol{D}^{-1}\boldsymbol{V}^T\hat{\boldsymbol{\mathcal{S}}}_{\eta_i}\boldsymbol{V}\boldsymbol{D}^{-1}$, $\boldsymbol{C}_i = \boldsymbol{Z}_i\boldsymbol{U}_1^T\boldsymbol{U}_1$. Some manipulation then shows that:

$$\begin{aligned}\operatorname{tr}\left(\frac{\partial \boldsymbol{A}_{\boldsymbol{\eta}}}{\partial\rho_i}\right) &= -\eta_i\operatorname{tr}(\boldsymbol{C}_i), \\ \operatorname{tr}\left(\frac{\partial^2 \boldsymbol{A}_{\boldsymbol{\eta}}}{\partial\rho_i\partial\rho_j}\right) &= 2\eta_i\eta_j\operatorname{tr}(\boldsymbol{Z}_j\boldsymbol{C}_i) - \delta_j^i\eta_i\operatorname{tr}(\boldsymbol{C}_i), \\ \frac{\partial\alpha}{\partial\rho_i} &= 2\eta_i[\boldsymbol{y}_1^T\boldsymbol{Z}_i\boldsymbol{y}_1 - \boldsymbol{y}_1^T\boldsymbol{C}_i\boldsymbol{y}_1], \\ \frac{\partial^2\alpha}{\partial\rho_i\partial\rho_j} &= 2\eta_i\eta_j\boldsymbol{y}_1^T[\boldsymbol{Z}_i\boldsymbol{C}_j + \boldsymbol{Z}_j\boldsymbol{C}_i - \boldsymbol{Z}_i\boldsymbol{Z}_j - \boldsymbol{Z}_j\boldsymbol{Z}_i + \boldsymbol{K}_i\boldsymbol{Z}_j]\boldsymbol{y}_1 + \delta_j^i\frac{\partial\alpha}{\partial\rho_i}.\end{aligned}$$

These derivatives are used to find the derivatives of $\mathcal{V}(\boldsymbol{\eta})$ with respect to ρ_i. Define $\mathcal{W} = N - \gamma\operatorname{tr}(\boldsymbol{A}_{\boldsymbol{\eta}})$, so that $\mathcal{V}(\boldsymbol{\eta}) = \frac{1}{N}\alpha - \frac{2}{N}\mathcal{W} + 1$, then

$$\begin{aligned}[\boldsymbol{z}]_i &= \frac{\partial\mathcal{V}(\boldsymbol{\eta})}{\partial\rho_i} = \frac{1}{N}\frac{\partial\alpha}{\partial\rho_i} - \frac{2}{N}\frac{\partial\mathcal{W}}{\partial\rho_i}, \\ [\boldsymbol{Z}]_{ij} &= \frac{\partial^2\mathcal{V}(\boldsymbol{\eta})}{\partial\rho_i\partial\rho_j} = \frac{1}{N}\frac{\partial^2\alpha}{\partial\rho_i\partial\rho_j} - \frac{2}{N}\frac{\partial^2\mathcal{W}}{\partial\rho_i\partial\rho_j}.\end{aligned}$$

For each trial $\boldsymbol{\eta}$, these derivatives are obtained at a reasonable computational cost, so that Newton's method backed up with steepest descent is used to find the optimum $\boldsymbol{\eta}$ fairly efficiently. Given the estimated $\hat{\boldsymbol{\eta}}$, the best fit vector $\boldsymbol{\theta}$ is simply $\hat{\boldsymbol{\theta}} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{y}_1 = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}_1^T\boldsymbol{Q}^T\boldsymbol{K}$.

APPENDIX

Theoretical aspects

In this appendix, we discuss and derive some asymptotic properties of the PMLE. For notational convenience, let $\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}$ be the shorthand for $\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}^{\mathcal{T}}$, for $\mathcal{T} = \{L, A, S, M\}$, and $\boldsymbol{\theta}_0$ the true parameter vector. We maintain a general viewpoint and assume that the vector $\boldsymbol{\eta}$ collects multiple tuning parameters. This tuning vector reduces to the scalar η in the case of the normal linear factor model (Section 2.1), and the triplet $(\eta_1, \eta_2, \eta_3)^T$ in the multiple-group extension (Section 5.1).

E.1 Regularity conditions

In all of the theorems derived in this work, we consider the following assumptions:

(A1) $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ which is a compact subset of $\mathbb{R}^m$.

(A2) $\boldsymbol{\beta}(\boldsymbol{\theta}) = \boldsymbol{\beta}(\boldsymbol{\theta}_0)$ only when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\beta}(\boldsymbol{\theta}) = (\boldsymbol{\mu}^T, \boldsymbol{\sigma}^T)^T$ and $\boldsymbol{\sigma} = \text{vech}(\boldsymbol{\Sigma})$. For the normal linear factor model, $\boldsymbol{\beta}(\boldsymbol{\theta})$ reduces to $\boldsymbol{\sigma}$ due to the absence of a mean-structure (see equation 2.1).

(A3) $\boldsymbol{\beta}(\boldsymbol{\theta})$ is twice continuously differentiable.

(A4) $\dfrac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}^T}$ is of full rank.

(A5) $(\boldsymbol{x}_\alpha^T, \text{vech}^T\{(\boldsymbol{x}_\alpha - \boldsymbol{\mu}_0)(\boldsymbol{x}_\alpha - \boldsymbol{\mu}_0)^T\})^T$ has a covariance matrix that is of full rank.

(A6) Let $\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0)$ denote the normalized score defined as $\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0) = \frac{1}{N}\boldsymbol{g}(\boldsymbol{\theta}_0) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_0)] = \frac{1}{N}\boldsymbol{g}(\boldsymbol{\theta}_0)$ for $\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_0)] \approx \mathbf{0}$. Assume that $\boldsymbol{g}(\boldsymbol{\theta}_0) \equiv \sqrt{N}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0) = \mathcal{O}_P\left(N^{\frac{1}{2}}\right)$, where $\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0) = \mathcal{O}_P(1)$.

(A7) $\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)\right] = -\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) = \mathcal{O}(N)$. For independent and identically distributed random variables, $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) \equiv N\boldsymbol{\mathcal{J}}_\alpha(\boldsymbol{\theta}_0)$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0) \equiv N\boldsymbol{\mathcal{H}}_\alpha(\boldsymbol{\theta}_0)$ $(\alpha = 1, \ldots, N)$ where $\boldsymbol{\mathcal{J}}_\alpha(\boldsymbol{\theta}_0)$ and $\boldsymbol{\mathcal{H}}_\alpha(\boldsymbol{\theta}_0)$ denote the expected and observed Fisher information for a single observation, respectively. It then follows that $\boldsymbol{\mathcal{J}}_\alpha(\boldsymbol{\theta}_0) = \mathcal{O}(1)$.

(A8) $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0) - \mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)\right] = \mathcal{O}_P\left(N^{\frac{1}{2}}\right)$. This results by decomposing $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)$ in its mean and stochastic part, that is, $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0) = \mathbb{E}[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)] + \boldsymbol{\varepsilon}$, where we assume that $\boldsymbol{\varepsilon} = \mathcal{O}_P\left(N^{\frac{1}{2}}\right)$ (Kauermann, 2005).

(A9) $\boldsymbol{\eta} \to \mathbf{0}$ and $\sqrt{N}\boldsymbol{\eta} \to \boldsymbol{\infty}$ as $N \to \infty$, or equivalently, $N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0) = o\left(N^{\frac{3}{2}}\right)$.

Assumption (A1) and (A3) are the standard regularity conditions and are generally satisfied in practice. Assumption (A2) implies that the model structure is identified. If the model is properly parameterized, assumption (A4) is satisfied. Conditions (A1) and (A2) are for consistency of parameter estimates, whereas (A3) and (A4) are needed to establish asymptotic normality. Assumption (A5) is needed in order for the parameter estimates to have proper asymptotic distributions, and is satisfied when $\boldsymbol{x}_\alpha \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ is full rank (Yuan & Bentler, 2006). Furthermore, assumptions (A6)-(A8) are the classical conditions for the consistency of the MLE (Barndorff-Nielsen & Cox, 1994, Ch. 3, pp. 82–83), while assumption (A9) ensures that, as the sample size increases, the tuning parameter vector gets larger and the penalty function vanishes. In Appendices E.2, E.3, E.4, E.5 we derive three usual theorems on the PMLE by adapting to the current context the results exposed in Fan and Li (2001), Oelker and Tutz (2013) and Filippou et al. (2017).

E.2 Asymptotic distribution of the PMLE (I)

Theorem 1 (Asymptotic distribution of the PMLE (I)). *Under certain regularity conditions, the PMLE has the following asymptotic distribution:*

$$\sqrt{N}\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)\left\{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0+\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right\}\xrightarrow{d}\mathcal{N}(\mathbf{0},N\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)),$$

and thus the asymptotic bias of $\hat{\boldsymbol{\theta}}$ *is equal to* $-\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0$*, and the asymptotic covariance* $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}=\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}$*, where* $\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)=\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)+N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)$*.*

Proof. The proof involves a Taylor expansion of the score in the neighbourhood of $\boldsymbol{\theta}_0$. For simplicity of notation, we omit all terms of order higher than 1 and assume that higher-order derivatives of the log-likelihood behave in a similar manner as those defined in the regularity conditions in Appendix E.1. The first-order Taylor expansion of $\boldsymbol{g}_p(\cdot)$ around $\boldsymbol{\theta}_0$ implies

$$\begin{aligned}\boldsymbol{g}_p(\hat{\boldsymbol{\theta}}) &= \boldsymbol{g}_p(\boldsymbol{\theta}_0)+\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+\text{ higher order terms}\\ &\approx \boldsymbol{g}_p(\boldsymbol{\theta}_0)+\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).\end{aligned}\tag{E.1}$$

By using the fact that $\boldsymbol{g}_p(\hat{\boldsymbol{\theta}})=\mathbf{0}$, and by multiplying all terms by $\sqrt{N}$, we have that

$$\sqrt{N}\boldsymbol{g}_p(\boldsymbol{\theta}_0)+\sqrt{N}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=\mathbf{0}.$$

Inverting the above series results in

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\left[\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)\right]^{-1}\sqrt{N}\boldsymbol{g}_p(\boldsymbol{\theta}_0).$$

We now divide both $\boldsymbol{g}_p(\boldsymbol{\theta}_0)$ and $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)$ by N, that is,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\left[\frac{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1}\sqrt{N}\frac{\boldsymbol{g}_p(\boldsymbol{\theta}_0)}{N}.\tag{E.2}$$

Let us now consider the set of random variables $\{\boldsymbol{g}_{p,1}(\boldsymbol{\theta}_0),\ldots,\boldsymbol{g}_{p,N}(\boldsymbol{\theta}_0)\}$, such that

$\boldsymbol{g}_p(\boldsymbol{\theta}_0) = \sum_{\alpha=1}^{N} \boldsymbol{g}_{p,\alpha}(\boldsymbol{\theta}_0)$. They are independent and identically distributed random variables, with common expectation and variance given by

$$\begin{aligned}
\mathbb{E}[\boldsymbol{g}_{p,\alpha}(\boldsymbol{\theta}_0)] &= \mathbb{E}[\boldsymbol{g}_\alpha(\boldsymbol{\theta}_0) - \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0] = \mathbb{E}[\boldsymbol{g}_\alpha(\boldsymbol{\theta}_0)] - \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 = -\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0, \\
Cov(\boldsymbol{g}_{p,\alpha}(\boldsymbol{\theta}_0)) &= Var(\boldsymbol{g}_\alpha(\boldsymbol{\theta}_0) - \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0) = Var(\boldsymbol{g}_\alpha(\boldsymbol{\theta}_0)) = \mathbb{E}[\boldsymbol{g}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{g}_\alpha(\boldsymbol{\theta}_0)^T] \\
&= -\mathbb{E}[\boldsymbol{\mathcal{H}}_\alpha(\boldsymbol{\theta}_0)] = \boldsymbol{\mathcal{J}}_\alpha(\boldsymbol{\theta}_0).
\end{aligned}$$

Since the quantity $\dfrac{\boldsymbol{g}_p(\boldsymbol{\theta}_0)}{N}$ in expression (E.2) can be seen as the mean of the random sample $\{\boldsymbol{g}_{p,1}(\boldsymbol{\theta}_0), \ldots, \boldsymbol{g}_{p,N}(\boldsymbol{\theta}_0)\}$, we can apply the central limit theorem and conclude that:

$$\sqrt{N}\left\{\frac{\boldsymbol{g}_p(\boldsymbol{\theta}_0)}{N} + \boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right\} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{1}{N}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\right),$$

and thus,

$$\sqrt{N}\,\frac{\boldsymbol{g}_p(\boldsymbol{\theta}_0)}{N} \xrightarrow{d} \mathcal{N}\left(-\sqrt{N}\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0, \frac{1}{N}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\right).$$

By the law of large numbers, the penalized observed information $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)$ converges to the penalized expected Fisher information $N\mathbb{E}[\boldsymbol{\mathcal{H}}_{p,\alpha}(\boldsymbol{\theta}_0)] = \mathbb{E}[\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)] = -\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)$ as the sample size increases, and thus,

$$-\left[\frac{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1} \xrightarrow{d} \left[\frac{\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1}.$$

Therefore, we have that:

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\left[\frac{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1} \sqrt{N}\frac{\boldsymbol{g}_p(\boldsymbol{\theta}_0)}{N} \\
&\Big\downarrow \text{ d} \\
&\mathcal{N}\left(\left[\frac{\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1}\left[-\sqrt{N}\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right], \left[\frac{\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1}\frac{\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)}{N}\left[\frac{\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1}\right). \quad \text{(E.3)}
\end{aligned}$$

From the above result, we can find an expression for the asymptotic bias and covariance matrix of the estimator $\hat{\boldsymbol{\theta}}$, that is,

$$\begin{aligned}
\text{BIAS}(\hat{\boldsymbol{\theta}}) &= \mathbb{E}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0] = \frac{1}{\sqrt{N}}\mathbb{E}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)] \approx \frac{1}{\sqrt{N}}\left[\frac{\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1}\left[-\sqrt{N}\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right] \\
&= -N\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 = -N\left\{-\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 \\
&= -N\left\{\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right\}^{-1}\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0,
\end{aligned}$$

$$\begin{aligned}
\text{Cov}(\hat{\boldsymbol{\theta}}) &= \frac{1}{N}\text{Cov}(\sqrt{N}\hat{\boldsymbol{\theta}}) \approx \frac{1}{N}\left[\frac{\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1}\frac{\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)}{N}\left[\frac{\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)}{N}\right]^{-1} \\
&= \boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1} \\
&= \left\{-\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\left\{-\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right]\right\}^{-1} \\
&= \left\{\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right\}^{-1}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\left\{\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right\}^{-1} \\
&= \boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1},
\end{aligned}$$

where $\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0) = \boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) = -\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)\right]$ is the expected Fisher information of the unpenalized model.

After some manipulation expression (E.3) becomes

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(-\sqrt{N}N\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0, N\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}\right)$$

$$\sqrt{N}\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(-\sqrt{N}N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0, N\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\right)$$

$$\sqrt{N}\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \sqrt{N}N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 \xrightarrow{d} \mathcal{N}(\mathbf{0}, N\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)).$$

Therefore, the final asymptotic distribution of the estimator is

$$\sqrt{N}\boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \boldsymbol{\mathcal{J}}_p(\boldsymbol{\theta}_0)^{-1}N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right] \xrightarrow{d} \mathcal{N}(\mathbf{0}, N\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)),$$

which completes the proof. ■

E.3 Asymptotic orders

Asymptotic order of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$

Under the assumptions of Appendix E.1 the asymptotic consistency of $\hat{\boldsymbol{\theta}}$ is of order $N^{-\frac{1}{2}}$, that is,

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \mathcal{O}_P\left(N^{-\frac{1}{2}}\right) \quad \text{as } N \to \infty.$$

Proof. By rearranging expression (E.1), noticing that $\boldsymbol{g}_p(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, and inverting the series, we have that:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 &= -[\boldsymbol{\mathcal{H}}_p(\boldsymbol{\theta}_0)]^{-1}\boldsymbol{g}_p(\boldsymbol{\theta}_0) + \dots \\
&= -[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)]^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0) + \dots \\
&= -[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0) - \mathbb{E}[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)] + \mathbb{E}[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)] - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)]^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_0) - N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0) + \dots \\
&= -\left\{\mathcal{O}_P\left(N^{\frac{1}{2}}\right) + \mathcal{O}(N) - o\left(N^{\frac{3}{2}}\right)\right\}^{-1}\left\{\mathcal{O}_P\left(N^{\frac{1}{2}}\right) - o\left(N^{\frac{3}{2}}\right)\right\} \\
&= [\mathcal{O}_P(N)]^{-1}\mathcal{O}_P\left(N^{\frac{1}{2}}\right) = \mathcal{O}_P(N)^{-1}\mathcal{O}_P\left(N^{\frac{1}{2}}\right) \\
&= \mathcal{O}_P\left(N^{-\frac{1}{2}}\right) \quad \text{as } N \to \infty.
\end{aligned}$$

■

Asymptotic order of Bias$(\hat{\boldsymbol{\theta}})$

Under the assumptions of Appendix E.1, the asymptotic bias of $\hat{\boldsymbol{\theta}}$ has order $N^{-\frac{1}{2}}$.

Proof.

$$\begin{aligned}
\text{BIAS}(\hat{\boldsymbol{\theta}}) &\approx -\left\{-\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)\right] + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right\}^{-1} N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 \\
&= -\left\{-\mathcal{O}(N) + o\left(N^{\frac{3}{2}}\right)\right\}^{-1} o\left(N^{\frac{3}{2}}\right) \\
&= \mathcal{O}(N^{-1})o\left(N^{\frac{3}{2}}\right) = o\left(N^{-\frac{1}{2}}\right).
\end{aligned}$$

■

Asymptotic order of $\text{Cov}(\hat{\boldsymbol{\theta}})$

Under the regularity conditions of Appendix E.1, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ has order N^{-1}.

Proof.

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\theta}}) &\approx \{-\mathbb{E}[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)] + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\}^{-1}\{-\mathbb{E}[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)]\}\{-\mathbb{E}[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_0)] + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\}^{-1} \\ &= \left\{\mathcal{O}(N) + o\left(N^{\frac{3}{2}}\right)\right\}^{-1}\mathcal{O}(N)\left\{\mathcal{O}(N) + o\left(N^{\frac{3}{2}}\right)\right\}^{-1} \\ &= \mathcal{O}(N^{-1})\mathcal{O}(N)\mathcal{O}(N^{-1}) = \mathcal{O}(N^{-1}).\end{aligned}$$

■

When $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)$ is near singular, $\text{Cov}(\hat{\boldsymbol{\theta}}^{\text{MLE}}) \to \infty$ and $\text{Cov}(\hat{\boldsymbol{\theta}}) \to \mathbf{0}$. This verifies that asymptotically the PMLE has smaller variance than the MLE, and thus may perform better.

E.4 Asymptotic distribution of the PMLE (II)

The next theorem shows that the asymptotic distribution of the PMLE coincides with the one of the MLE as the sample size increases, which is desirable, as the MLE is the most efficient estimator.

Theorem 2 (Asymptotic distribution of the PMLE (II)). *If* $\max|N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0| = o\left(N^{\frac{3}{2}}\right)$, *and* $\max|N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)| = o\left(N^{\frac{3}{2}}\right)$, *then*

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{N}\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0)\right\}^{-1}\right).$$

Proof. If $\max|N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0| = o\left(N^{\frac{3}{2}}\right)$ and $\max|N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)| = o\left(N^{\frac{3}{2}}\right)$, it follows that $\frac{1}{N\sqrt{N}}\max|N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0| \to \mathbf{0}$, and $\frac{1}{N\sqrt{N}}\max|N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)| \to \mathbf{0}$ as $N \to \infty$. Given these two conditions, we have that

$$\mathbb{E}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)] \approx \left\{\frac{\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N}\right\}^{-1}\left(-\sqrt{N}\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right)$$

$$
\begin{aligned}
&= \left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N}\right\}^{-1}\left(-\frac{1}{\sqrt{N}}N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right)\\
&= \left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N^2}\right\}^{-1}\left(-\frac{N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0}{N\sqrt{N}}\right)\\
&\to \left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N^2}\right\}^{-1}\cdot\mathbf{0} = \mathbf{0},
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}(\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)) &\approx \left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N}\right\}^{-1}\left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)}{N}\right\}\left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N}\right\}^{-1}\\
&= N\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)\right\}^{-1}\\
&= \frac{1}{N}\left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N\sqrt{N}}\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0) + N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N\sqrt{N}}\right\}^{-1}\frac{1}{N}\\
&= \frac{1}{N}\left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)}{N\sqrt{N}} + \frac{N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N\sqrt{N}}\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)}{N\sqrt{N}} + \frac{N\boldsymbol{\mathcal{S}}_{\boldsymbol{\eta}}(\boldsymbol{\theta}_0)}{N\sqrt{N}}\right\}^{-1}\frac{1}{N}\\
&\to \frac{1}{N}\left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)}{N\sqrt{N}} + \mathbf{0}\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\left\{\frac{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)}{N\sqrt{N}} + \mathbf{0}\right\}^{-1}\frac{1}{N}\\
&\to \frac{N\sqrt{N}}{N}\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)]\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\right\}^{-1}\frac{N\sqrt{N}}{N}\\
&\to N\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)^{-1} = \left[\frac{1}{N}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\right]^{-1}.
\end{aligned}
$$

Therefore,

$$
\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{N}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\right\}^{-1}\right).
$$

■

E.5 Consistency

Theorem 3 (Consistency). *Suppose that $\boldsymbol{\eta} \in [\mathbf{0}, \boldsymbol{\infty})$ is fixed. Then, under the assumption of a convex unpenalized log-likelihood, the PMLE $\hat{\boldsymbol{\theta}}$ that minimizes $-\ell_p(\boldsymbol{\theta})$ is consistent, that is,*

$$
\lim_{N\to\infty} \mathbb{P}\left(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2^2 > \bar{\varepsilon}\right) = 0 \quad \forall\ \bar{\varepsilon} > 0.
$$

Proof. If $\hat{\boldsymbol{\theta}}$ minimizes $-\ell_p(\boldsymbol{\theta})$, then it also minimizes $-\frac{\ell_p(\boldsymbol{\theta})}{N}$. Similarly, $\hat{\boldsymbol{\theta}}^{MLE}$ minimizes $-\ell(\boldsymbol{\theta})$, as well as $-\frac{\ell(\boldsymbol{\theta})}{N}$. Because $\boldsymbol{\eta}$ is fixed, $-\frac{\ell_p(\hat{\boldsymbol{\theta}})}{N} \to -\frac{\ell(\hat{\boldsymbol{\theta}}^{\text{MLE}})}{N}$, and $-\frac{\ell_p(\hat{\boldsymbol{\theta}})}{N} \to -\frac{\ell(\hat{\boldsymbol{\theta}})}{N}$; thus, $-\frac{\ell(\hat{\boldsymbol{\theta}})}{N} \to -\frac{\ell(\hat{\boldsymbol{\theta}}^{\text{MLE}})}{N}$ holds as well. Since $\hat{\boldsymbol{\theta}}^{\text{MLE}}$ is a unique minimizer of $-\frac{\ell(\boldsymbol{\theta})}{N}$, and $-\frac{\ell(\boldsymbol{\theta})}{N}$ is convex, it follows that $\hat{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}^{\text{MLE}}$. The consistency of $\hat{\boldsymbol{\theta}}$ follows from the consistency of $\hat{\boldsymbol{\theta}}^{\text{MLE}}$. ■

E.6 Confidence intervals

As illustrated in Section 3.4, point-wise confidence intervals for each model parameter can be obtained using $\boldsymbol{\theta} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \boldsymbol{V}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}))$, where $\boldsymbol{V}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{J}}_p(\hat{\boldsymbol{\theta}})^{-1} = (\boldsymbol{\mathcal{J}}(\hat{\boldsymbol{\theta}}) + N\boldsymbol{\mathcal{S}}_{\hat{\boldsymbol{\eta}}}(\hat{\boldsymbol{\theta}}))^{-1}$ is the covariance matrix of the PMLE based on the Bayesian result derived in Marra and Wood (2012). Confidence intervals for non-linear functions of the parameter vector $\boldsymbol{\theta}$ can be conveniently obtained by simulation from the posterior of $\boldsymbol{\theta}$ as follows. Let $T(\boldsymbol{\theta})$ be any function of the parameters, then

Step 1 Draw N_{sim} random vectors $\boldsymbol{\theta}_h^\star$ (for $h = 1, \ldots, N_{\text{sim}}$) from $\mathcal{N}(\hat{\boldsymbol{\theta}}, \boldsymbol{V}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}))$;

Step 2 Compute $T_h^\star := T(\boldsymbol{\theta}_h^\star)\, \forall\, h$, and define $T_\alpha^\star$ to be the $[N_{\text{sim}} \cdot \alpha]^{\text{th}}$ smallest value of the ordered sample $\{T_1^\star, \ldots, T_{N_{\text{sim}}}^\star\}$, with $[a]$ denoting the integer part of $a \in \mathbb{R}$;

Step 3 Obtain an approximate $(1-\alpha)\%$ confidence interval for $T(\hat{\boldsymbol{\theta}})$ using $\left[T_{\frac{\alpha}{2}}^\star, T_{1-\frac{\alpha}{2}}^\star\right]$.

Small values of N_{sim} are typically tolerable. The quantity α is usually set to 0.05.

APPENDIX F

Details on the multiple-group factor analysis model

F.1 Log-likelihood

Given random samples of sizes $N_1, \ldots, N_G$ from a multivariate normal distribution, with $N = \sum_{g=1}^{G} N_g$ the total sample size across groups, the likelihood of the multiple-group factor model is:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{g=1}^{G}\prod_{\alpha=1}^{N_g} f(\boldsymbol{x}_{\alpha g};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) \\
&= \prod_{g=1}^{G}\prod_{\alpha=1}^{N_g} \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_g|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)\right\} \\
&= \prod_{g=1}^{G}(2\pi)^{-\frac{N_g}{2}p}|\boldsymbol{\Sigma}_g|^{-\frac{N_g}{2}} \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{N_g}(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)\right\}.
\end{aligned}$$

The log-likelihood, which is defined as the logarithm of $\mathcal{L}(\boldsymbol{\theta})$, takes the form:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &\coloneqq \log\mathcal{L}(\boldsymbol{\theta}) \\
&= -\sum_{g=1}^{G}\left\{\frac{N_g}{2}p\log(2\pi)+\frac{N_g}{2}\log|\boldsymbol{\Sigma}_g|+\frac{1}{2}\sum_{\alpha=1}^{N_g}(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)\right\} \\
&= -\sum_{g=1}^{G}\left\{\frac{N_g}{2}p\log(2\pi)+\frac{N_g}{2}\log|\boldsymbol{\Sigma}_g|+\frac{1}{2}\sum_{\alpha=1}^{N_g}\operatorname{tr}\left[(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)\right]\right\}
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{g=1}^{G}\left\{\frac{N_g}{2}p\log(2\pi)+\frac{N_g}{2}\log|\boldsymbol{\Sigma}_g|+\frac{1}{2}\mathrm{tr}\left[\sum_{\alpha=1}^{N_g}(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)(\boldsymbol{x}_{\alpha g}-\boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_g^{-1}\right]\right\}\\
&= -\sum_{g=1}^{G}\left\{\frac{N_g}{2}p\log(2\pi)+\frac{N_g}{2}\log|\boldsymbol{\Sigma}_g|\right.\\
&\qquad\left.+\frac{1}{2}\mathrm{tr}\left[\sum_{\alpha=1}^{N_g}[(\boldsymbol{x}_{\alpha g}-\bar{\boldsymbol{x}}_g)+(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)][(\boldsymbol{x}_{\alpha g}-\bar{\boldsymbol{x}}_g)+(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)]^T\boldsymbol{\Sigma}_g^{-1}\right]\right\}\\
&= -\sum_{g=1}^{G}\left\{\frac{N_g}{2}p\log(2\pi)+\frac{N_g}{2}\log|\boldsymbol{\Sigma}_g|\right.\\
&\qquad\left.+\frac{1}{2}\mathrm{tr}\left[\left[N_g\boldsymbol{S}_g+N_g(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)^T\right]\boldsymbol{\Sigma}_g^{-1}\right]\right\}\\
&= -\sum_{g=1}^{G}\frac{N_g}{2}\left\{\log|\boldsymbol{\Sigma}_g|+\mathrm{tr}(\boldsymbol{S}_g\boldsymbol{\Sigma}_g^{-1})+(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_g^{-1}(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)+p\log(2\pi)\right\}\\
&= -\sum_{g=1}^{G}\frac{N_g}{2}\left\{\log|\boldsymbol{\Sigma}_g|+\mathrm{tr}(\boldsymbol{W}_g\boldsymbol{\Sigma}_g^{-1})+p\log(2\pi)\right\},
\end{aligned}
\tag{F.1}
$$

where $\boldsymbol{W}_g = \boldsymbol{S}_g + (\bar{\boldsymbol{x}}_g - \boldsymbol{\mu}_g)(\bar{\boldsymbol{x}}_g - \boldsymbol{\mu}_g)^T$, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_g(\boldsymbol{\theta}_g) = \boldsymbol{\Lambda}_g\boldsymbol{\Phi}_g\boldsymbol{\Lambda}_g^T + \boldsymbol{\Psi}_g$ and $\boldsymbol{\mu}_g = \boldsymbol{\mu}_g(\boldsymbol{\theta}_g) = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g\boldsymbol{\kappa}_g$.

F.2 Gradient and Fisher information

Before deriving the gradient and Fisher information matrix for a multiple-group factor model, it is more convenient to first examine these formulas for a (single-group) mean and covariance structure factor model (Appendix F.2.1). Note that the resulting expressions differ from those derived in Appendix A.2 as the model now involves a mean structure $\boldsymbol{\mu}(\boldsymbol{\theta})$ and a covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The gradient and Fisher information for the multiple-group factor model are then easily found by combining the obtained results across groups (Appendix F.2.2).

F.2.1 Mean and covariance structure factor model

Consider the mean and covariance structure factor model:

$$\boldsymbol{x} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\epsilon}. \tag{F.2}$$

It is assumed that $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{\kappa}, \boldsymbol{\Phi})$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi}$ usually a diagonal matrix, and $\boldsymbol{f}$ is uncorrelated with $\boldsymbol{\epsilon}$. The log-likelihood of the model in (F.2) is

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

where $\boldsymbol{W} = \boldsymbol{S} + (\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T$, $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$. The propositions below enunciate the expressions of the gradient $\boldsymbol{g}(\boldsymbol{\theta}) \coloneqq \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and the expected Fisher information matrix $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) \coloneqq \mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})^T] = -\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})\right]$ of the mean and covariance structure factor model in (F.2).

Proposition F.1 (Gradient of the mean and covariance structure factor model)**.** *The gradient of the log-likelihood of the mean and covariance structure factor model in equation* (F.2) *with respect to an arbitrary scalar variable* θ_q *takes the form:*

$$[\boldsymbol{g}(\boldsymbol{\theta})]_q = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_q} = -\frac{N}{2} tr\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{W})\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right\} + N(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \theta_q}. \tag{F.3}$$

Proof. See Appendix F.2.1.1. ■

Proposition F.2 (First-order derivatives of the mean and covariance structure factor model with respect to the parameter matrices)**.** *The matrix expressions of the first-order derivatives of the log-likelihood of the mean and covariance structure factor model in equation* (F.2) *with respect to the parameter matrices are:*

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Lambda}} = -N\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{W})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi} + N\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\boldsymbol{\kappa}^T, \tag{F.4}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Phi}} = \begin{cases} -N\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{W})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, & \textit{non-diagonal elements,} \\ -\dfrac{N}{2}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{W})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, & \textit{diagonal elements,} \end{cases} \tag{F.5}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Psi}} = -N\, diag(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{W})\boldsymbol{\Sigma}^{-1}), \tag{F.6}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\tau}} = N\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}), \tag{F.7}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\kappa}} = N\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}). \tag{F.8}$$

Proof. See Appendix F.2.1.2. ■

Define the following matrices: $\boldsymbol{\omega} = \boldsymbol{\Sigma}^{-1}, \boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}, \boldsymbol{\gamma} = \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$, $\boldsymbol{\delta} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, \boldsymbol{\zeta} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}, \boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{W})\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Omega}_{\boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})$.

Proposition F.3 (Expected Fisher information of the mean and covariance structure factor model)**.** *The expected Fisher information matrix of the mean and covariance structure factor model in equation* (F.2) *with respect to two arbitrary scalar variables* θ_q *and* $\theta_{q'}$ *takes the form:*

$$[\boldsymbol{\mathcal{J}}(\boldsymbol{\theta})]_{qq'} = \frac{N}{2} tr\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{q'}}\right) + N\frac{\partial \boldsymbol{\mu}^T}{\partial \theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \theta_{q'}}. \tag{F.9}$$

Proof. See Appendix F.2.1.3. ■

Proposition F.4 (Elements of the expected Fisher information of the mean and covariance structure factor model with respect to the parameter matrices)**.** *The expected Fisher information of the mean and covariance structure factor model matrix in equation* (F.2) *is a block matrix of the form:*

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) = \left[\begin{array}{c:c:c:c:c} \boldsymbol{J}_{11} & \boldsymbol{J}_{12} & \boldsymbol{J}_{13} & \boldsymbol{J}_{14} & \boldsymbol{J}_{15} \\ \hdashline \boldsymbol{J}_{12}^T & \boldsymbol{J}_{22} & \boldsymbol{J}_{23} & \boldsymbol{J}_{24} & \boldsymbol{J}_{25} \\ \hdashline \boldsymbol{J}_{13}^T & \boldsymbol{J}_{23}^T & \boldsymbol{J}_{33} & \boldsymbol{J}_{34} & \boldsymbol{J}_{35} \\ \hdashline \boldsymbol{J}_{14}^T & \boldsymbol{J}_{24}^T & \boldsymbol{J}_{34}^T & \boldsymbol{J}_{44} & \boldsymbol{J}_{45} \\ \hdashline \boldsymbol{J}_{15}^T & \boldsymbol{J}_{25}^T & \boldsymbol{J}_{35}^T & \boldsymbol{J}_{45}^T & \boldsymbol{J}_{55} \end{array}\right], \tag{F.10}$$

where, for $i, t = 1, \ldots, p$ *and* $g, h, j, l, q, s = 1, \ldots, r$*, the sub-matrices are:*

$$[\boldsymbol{J}_{11}]_{(ij,ts)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \lambda_{ts}}\right] = N(\beta_{is}\beta_{tj} + \omega_{it}[\boldsymbol{\zeta} + \boldsymbol{\kappa}\boldsymbol{\kappa}^T]_{js}]), \tag{F.11}$$

$$[\boldsymbol{J}_{12}]_{(ij,t)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \tau_{t}}\right] = N\omega_{it}\kappa_j, \tag{F.12}$$

$$[\boldsymbol{J}_{13}]_{(ij,tt)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \psi_{tt}}\right] = N\omega_{it}\beta_{tj}, \tag{F.13}$$

$$[\boldsymbol{J}_{14}]_{(ij,gh)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \phi_{gh}}\right] = \frac{N}{2}(2 - [\boldsymbol{I}]_{gh})(\alpha_{ig}\delta_{jh} + \alpha_{ih}\delta_{jg}), \tag{F.14}$$

$$[\boldsymbol{J}_{15}]_{(ij,g)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \kappa_{g}}\right] = N\alpha_{ig}\kappa_j, \tag{F.15}$$

$$[\boldsymbol{J}_{22}]_{(i,t)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_{i} \partial \tau_{t}}\right] = N\omega_{it}, \tag{F.16}$$

$$[\boldsymbol{J}_{23}]_{(i,tt)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_{i} \partial \psi_{tt}}\right] = 0, \tag{F.17}$$

$$[\boldsymbol{J}_{24}]_{(i,gh)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_{i} \partial \phi_{gh}}\right] = 0, \tag{F.18}$$

$$[\boldsymbol{J}_{25}]_{(i,g)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_{i} \partial \kappa_{g}}\right] = N\alpha_{ig}, \tag{F.19}$$

$$[\boldsymbol{J}_{33}]_{(ii,tt)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{ii} \partial \psi_{tt}}\right] = \frac{N}{2}\omega_{it}^2, \tag{F.20}$$

$$[\boldsymbol{J}_{34}]_{(tt,gh)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{tt} \partial \phi_{gh}}\right] = \frac{N}{2}(2 - [\boldsymbol{I}]_{gh})\alpha_{tg}\alpha_{th}, \tag{F.21}$$

$$[\boldsymbol{J}_{35}]_{(ii,g)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{ii} \partial \kappa_{g}}\right] = 0, \tag{F.22}$$

$$[\boldsymbol{J}_{44}]_{(gh,lq)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \phi_{gh} \partial \phi_{lq}}\right] \tag{F.23}$$

$$= \frac{N}{4}(2 - [\boldsymbol{I}]_{gh})(2 - [\boldsymbol{I}]_{lq})(\gamma_{gl}\gamma_{hq} + \gamma_{gq}\gamma_{hl}), \tag{F.24}$$

$$[\boldsymbol{J}_{45}]_{(gh,l)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \phi_{gh} \partial \kappa_{l}}\right] = 0, \tag{F.25}$$

$$[\boldsymbol{J}_{55}]_{(g,h)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \kappa_{g} \partial \kappa_{h}}\right] = N\gamma_{gh}. \tag{F.26}$$

Proof. See Appendix F.2.1.4. ■

Alternatively, the Fisher information matrix of the factor model with a mean structure $\boldsymbol{\mu}(\boldsymbol{\theta})$ and a covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ can be formulated more compactly as follows. Denote $\boldsymbol{\sigma} = \text{vech}(\boldsymbol{\Sigma})$, $\boldsymbol{s} = \text{vech}(\boldsymbol{S})$, $\boldsymbol{\beta} = (\boldsymbol{\mu}^T, \boldsymbol{\sigma}^T)^T$, $\hat{\boldsymbol{\beta}} = (\bar{\boldsymbol{x}}^T, \boldsymbol{s}^T)^T$, and $\boldsymbol{D}$ the $p^2 \times \frac{p(p+1)}{2}$ duplication matrix such that $\text{vec}(\boldsymbol{B}) = \boldsymbol{D}\text{vech}(\boldsymbol{B})$ for a $p \times p$ matrix $\boldsymbol{B}$. Define the block diagonal matrix $\boldsymbol{E} = \text{diag}(\boldsymbol{\Sigma}^{-1}, \boldsymbol{E}_c)$, where $\boldsymbol{E}_c = \frac{1}{2}\boldsymbol{D}^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{D}$. Let $\boldsymbol{\Delta} = \dfrac{\partial \boldsymbol{\beta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$ be the $\dfrac{p(p+3)}{2} \times m$ Jacobian matrix of the partial derivatives of the model with respect to the parameters. Then, the expected Fisher information matrix can be expressed as (Yuan & Bentler, 2006):

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) = N\boldsymbol{\Delta}^T\boldsymbol{E}\boldsymbol{\Delta}. \tag{F.27}$$

F.2.1.1 Proof of proposition F.1

Proof. Let us consider the function:

$$F = \frac{1}{2}\left\{\log|\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + \text{tr}\left[(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right] + p\log(2\pi)\right\}. \tag{F.28}$$

The first-order partial derivative of F with respect to an arbitrary scalar variable θ_q is:

$$\begin{aligned}
\frac{\partial F}{\partial \theta_q} &= \frac{1}{2}\frac{\partial}{\partial \theta_q}\left\{ \log|\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + \text{tr}\left[(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right] + p\log(2\pi) \right\} \\
&= \frac{1}{2}\left\{ \frac{\partial}{\partial \theta_q}\log|\boldsymbol{\Sigma}| + \frac{\partial}{\partial \theta_q}\text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + \frac{\partial}{\partial \theta_q}\text{tr}\left[(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right] \right. \\
&\qquad \left. + \frac{\partial}{\partial \theta_q}\cancel{p\log(2\pi)} \right\} \\
&= \frac{1}{2}\left\{ \text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right)\right] - \text{tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right)\right] \right. \\
&\qquad \left. + \text{tr}\left[\frac{\partial}{\partial \theta_q}\left[(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right]\right] \right\} \\
&= \frac{1}{2}\left\{ \text{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_q}\right)\right] \right.
\end{aligned}$$

$$
\begin{aligned}
&+ \operatorname{tr}\left[-(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\boldsymbol{\Sigma}^{-1} - 2(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right]\Bigg\} \\
&= \frac{1}{2}\Bigg\{\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right] \\
&\quad - \operatorname{tr}\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\right] - 2\operatorname{tr}\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right]\Bigg\} \\
&= \frac{1}{2}\Bigg\{\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right] \\
&\quad - \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right] - 2\operatorname{tr}\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right]\Bigg\} \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S}-(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T)\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right] - \operatorname{tr}\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{W})\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right] - (\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q} \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Omega}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right] - \boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\frac{\partial\boldsymbol{\mu}}{\partial\theta_q},
\end{aligned}
\tag{F.29}
$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{W})\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Omega}_{\boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})$. Given that $F = -\frac{1}{N}\ell(\boldsymbol{\theta})$, it follows that the gradient of the log-likelihood is:

$$
[\boldsymbol{g}(\boldsymbol{\theta})]_q = \frac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_q} = -N\frac{\partial F}{\partial\theta_q} = -\frac{N}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right\} + N\boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}. \tag{F.30}
$$

If the mean structure is absent, $\boldsymbol{W} = \boldsymbol{S}$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Omega}_{\boldsymbol{\mu}} = \mathbf{0}$, and we get expression (2.4). ■

F.2.1.2 Proof of proposition F.2

Proof. In order to compute equations (F.4)-(F.8), we need the derivatives of the matrix $\boldsymbol{\Sigma}$ and the vector $\boldsymbol{\mu}$ with respect to each model parameter. Let us find the partial derivatives of the model-implied moments taken with respect to the elements of $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$, respectively:

$$
\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ij}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T \qquad \text{by equation (A.22)}, \tag{F.31}
$$

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}} = \boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T \quad \text{by equation (A.23),} \tag{F.32}$$

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{ii}} = \mathbf{1}_{ii} \quad \text{by equation (A.24),} \tag{F.33}$$

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \tau_i} = \frac{\partial(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})}{\partial \tau_i} = 0, \tag{F.34}$$

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \kappa_j} = \frac{\partial(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})}{\partial \kappa_j} = 0, \tag{F.35}$$

$$\frac{\partial \boldsymbol{\mu}}{\partial \lambda_{ij}} = \frac{\partial(\boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa})}{\partial \lambda_{ij}} = \frac{\partial \boldsymbol{\Lambda}\boldsymbol{\kappa}}{\partial \lambda_{ij}} = \mathbf{1}_{ij}\boldsymbol{\kappa}, \tag{F.36}$$

$$\frac{\partial \boldsymbol{\mu}}{\partial \phi_{gh}} = \frac{\partial(\boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa})}{\partial \phi_{gh}} = 0, \tag{F.37}$$

$$\frac{\partial \boldsymbol{\mu}}{\partial \psi_{ii}} = \frac{\partial(\boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa})}{\partial \psi_{ii}} = 0, \tag{F.38}$$

$$\frac{\partial \boldsymbol{\mu}}{\partial \tau_i} = \frac{\partial(\boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa})}{\partial \tau_i} = \frac{\partial \boldsymbol{\tau}}{\partial \tau_i} = \mathbf{1}_{i1}, \tag{F.39}$$

$$\frac{\partial \boldsymbol{\mu}}{\partial \kappa_j} = \frac{\partial(\boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa})}{\partial \kappa_j} = \frac{\partial \Lambda \boldsymbol{\kappa}}{\partial \kappa_j} = \boldsymbol{\Lambda}\mathbf{1}_{j1}, \tag{F.40}$$

where $\mathbf{1}_{ab}$ is a matrix with zeros in every position, except the entry (a, b), which contains a 1.0, and $\mathbf{1}_{a1}$ is a column vector with zeros in every position, except the entry a, which contains a 1.0. By substituting expressions (F.31)-(F.40) into equation (F.29), we get the following set of first-order derivatives of F with respect to the model parameters:

$$\begin{aligned}
\frac{\partial F}{\partial \lambda_{ij}} &= \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ij}}\right)\right] - \boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\frac{\partial \boldsymbol{\mu}}{\partial \lambda_{ij}} = \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji} + \mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right)\right] - \boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\mathbf{1}_{ij}\boldsymbol{\kappa} \\
&= \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\mathbf{1}_{ij}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T\right] - \mathrm{tr}(\boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\mathbf{1}_{ij}\boldsymbol{\kappa}) \\
&= \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] - \mathrm{tr}(\boldsymbol{\kappa}^T\mathbf{1}_{ji}\boldsymbol{\Omega}_{\boldsymbol{\mu}}) \\
&= \mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}\mathbf{1}_{ji}\right] - \mathrm{tr}(\boldsymbol{\Omega}_{\boldsymbol{\mu}}\boldsymbol{\kappa}^T\mathbf{1}_{ji}) = [\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi}]_{ij}[\boldsymbol{\Omega}_{\boldsymbol{\mu}}\boldsymbol{\kappa}^T]_{ij}, \\
\frac{\partial F}{\partial \phi_{gh}} &= \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}}\right)\right] - \boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\frac{\partial \boldsymbol{\mu}}{\partial \phi_{gh}} = \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\left[\mathbf{1}_{gh} + \mathbf{1}_{hg} - \mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right]\boldsymbol{\Lambda}^T\right] = \\
&= \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right] + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\right] - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\boldsymbol{\Lambda}^T\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right] + \frac{1}{2}\mathrm{tr}\left[\mathbf{1}_{hg}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\right] - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\boldsymbol{I}\mathbf{1}_{gh}\right] \\
&= \mathrm{tr}\left[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}\mathbf{1}_{gh}\right] - \frac{1}{2}[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{hg}[\boldsymbol{I}]_{hg} \\
&= \left(1 - \frac{1}{2}[\boldsymbol{I}]_{gh}\right)[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}]_{gh},
\end{aligned}
$$

$$
\frac{\partial F}{\partial \psi_{ii}} = \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{ii}}\right)\right] - \boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\frac{\partial \boldsymbol{\mu}}{\partial \psi_{ii}} = \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\mathbf{1}_{ii}\right] = \frac{1}{2}[\boldsymbol{\Omega}]_{ii},
$$

$$
\frac{\partial F}{\partial \tau_i} = \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \tau_i}\right)\right] - \boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\frac{\partial \boldsymbol{\mu}}{\partial \tau_i} = -\boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\mathbf{1}_{i1} = -[\boldsymbol{\Omega}_{\boldsymbol{\mu}}^T]_i,
$$

$$
\begin{aligned}
\frac{\partial F}{\partial \kappa_j} &= \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \kappa_j}\right)\right] - \boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\frac{\partial \boldsymbol{\mu}}{\partial \kappa_j} = -\boldsymbol{\Omega}_{\boldsymbol{\mu}}^T\boldsymbol{\Lambda}\mathbf{1}_{j1} = -\mathrm{tr}(\mathbf{1}_{j1}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}_{\boldsymbol{\mu}}) \\
&= -\mathrm{tr}(\boldsymbol{\Lambda}^T\boldsymbol{\Omega}_{\boldsymbol{\mu}}\mathbf{1}_{j1}) = -[\boldsymbol{\Lambda}^T\boldsymbol{\Omega}_{\boldsymbol{\mu}}]_j.
\end{aligned}
$$

The analytical first-order derivatives of the log-likelihood $\ell(\boldsymbol{\theta}) = -N\,F$ in matrix expression are then:

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Lambda}} &= -N\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Phi} + N\boldsymbol{\Omega}_{\boldsymbol{\mu}}\boldsymbol{\kappa}^T, \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Phi}} &= \begin{cases} -N\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}, & \text{non-diagonal elements,} \\ -\dfrac{N}{2}\boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}, & \text{diagonal elements,} \end{cases} \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\Psi}} &= -N\mathrm{diag}(\boldsymbol{\Omega}), \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\tau}} &= N\boldsymbol{\Omega}_{\boldsymbol{\mu}}, \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\kappa}} &= N\boldsymbol{\Lambda}^T\boldsymbol{\Omega}_{\boldsymbol{\mu}},
\end{aligned}
$$

with the understanding that the elements of the parameter matrices on the left corresponding to the positions of fixed elements of $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$ are taken to be zero. ■

F.2.1.3 Proof of proposition F.3

Proof. The second partial derivative of F with respect to two arbitrary scalar variables θ_q and $\theta_{q'}$ is:

$$\begin{aligned}
\frac{\partial^2 F}{\partial\theta_q\partial\theta_{q'}} &= \frac{\partial}{\partial\theta_{q'}}\left\{\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right]\right. \\
&\qquad -\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right] \\
&\qquad \left.-\mathrm{tr}\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right]\right\} \\
&= \frac{\partial}{\partial\theta_{q'}}\left\{\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right]\right\} \\
&\qquad -\frac{\partial}{\partial\theta_{q'}}\left\{\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right]\right\} \\
&\qquad -\frac{\partial}{\partial\theta_{q'}}\mathrm{tr}\left\{(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right\} \\
&= \boldsymbol{T}_1+\boldsymbol{T}_2+\boldsymbol{T}_3. \qquad \text{(F.41)}
\end{aligned}$$

Due to the presence of the mean structure as well as the covariance structure, the derivation of the exact second-order derivatives is a lengthy and tedious process. We will thus employ approximate second-order derivatives by disregarding the terms involving the second-order derivatives of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. Then, given that $\mathbb{E}[\bar{\boldsymbol{x}}-\boldsymbol{\mu}]=\mathbf{0}$ and $\mathbb{E}[\boldsymbol{S}-\boldsymbol{\Sigma}]=\boldsymbol{O}$ as $N\to\infty$, the resulting quantities coincide with the expected Fisher information matrix. Let us examine the terms in (F.41):

$$\begin{aligned}
\boldsymbol{T}_1 &= \frac{\partial}{\partial\theta_{q'}}\left\{\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right]\right\} \\
&= \frac{1}{2}\left\{\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right]\right. \\
&\quad \left.+\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\left(\frac{\partial^2\boldsymbol{\Sigma}}{\partial\theta_{q'}\partial\theta_q}-2\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)\right]\right\} \quad \text{by equation (A.31)} \\
&\approx \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\right)=\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\right),
\end{aligned}$$

$$
\begin{aligned}
\boldsymbol{T}_3 &= -\frac{\partial}{\partial\theta_{q'}}\mathrm{tr}\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right] = -\mathrm{tr}\left\{\frac{\partial}{\partial\theta_{q'}}\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right]\right\} \\
&= -\mathrm{tr}\left\{\frac{\partial\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right]}{\partial\theta_{q'}}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q} + (\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial^2\boldsymbol{\mu}}{\partial\theta_{q'}\partial\theta_q}\right\} \\
&= -\mathrm{tr}\left\{(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial^2\boldsymbol{\mu}}{\partial\theta_{q'}\partial\theta_q}\right\} - \mathrm{tr}\left\{\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\theta_{q'}} + \frac{\partial(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\right]\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right\} \\
&= -\mathrm{tr}\left\{(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial^2\boldsymbol{\mu}}{\partial\theta_{q'}\partial\theta_q}\right\} + \mathrm{tr}\left\{(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right\} \\
&\quad + \mathrm{tr}\left\{\frac{\partial\boldsymbol{\mu}^T}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right\} \\
&= -\mathrm{tr}\left\{(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\left[\frac{\partial^2\boldsymbol{\mu}}{\partial\theta_{q'}\partial\theta_q} - \frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right]\right\} + \mathrm{tr}\left\{\frac{\partial\boldsymbol{\mu}^T}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right\} \\
&\approx \mathrm{tr}\left(\frac{\partial\boldsymbol{\mu}^T}{\partial\theta_{q'}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_q}\right) = \mathrm{tr}\left(\frac{\partial\boldsymbol{\mu}^T}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_{q'}}\right).
\end{aligned}
$$

The expected Fisher information matrix is then easily obtained as

$$
\begin{aligned}
\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) &= -\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_q\partial\theta_{q'}}\right] = -\mathbb{E}\left[-N\frac{\partial^2 F}{\partial\theta_q\partial\theta_{q'}}\right] \\
&= \frac{N}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_{q'}}\right) + N\mathrm{tr}\left(\frac{\partial\boldsymbol{\mu}^T}{\partial\theta_q}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_{q'}}\right).
\end{aligned}
$$

since the term involving $\boldsymbol{T}_2$ vanishes given that $\mathbb{E}[\bar{\boldsymbol{x}}-\boldsymbol{\mu}] = \mathbf{0}$. ■

F.2.1.4 Proof of proposition F.4

Proof. By using the results in Appendix A.2.3.1 and equations (F.31)-(F.40), we obtain the following expressions of the sub-matrices of the Fisher information matrix, for $i, t = 1, \ldots, p$ and $g, h, j, l, q, s = 1, \ldots, r$:

$$
\begin{aligned}
[\boldsymbol{J}_{11}]_{(ij,ts)} &= -\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\lambda_{ij}\partial\lambda_{ts}}\right] = \frac{N}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ij}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\lambda_{ts}}\right) + N\mathrm{tr}\left(\frac{\partial\boldsymbol{\mu}^T}{\partial\lambda_{ij}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\lambda_{ts}}\right) \\
&= N(\beta_{is}\beta_{tj} + \omega_{it}\zeta_{js}) + N\mathrm{tr}(\boldsymbol{\kappa}^T\mathbf{1}_{ji}\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ts}\boldsymbol{\kappa}) \\
&= N(\beta_{is}\beta_{tj} + \omega_{it}\zeta_{js}) + N\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{1}_{ts}\boldsymbol{\kappa}\boldsymbol{\kappa}^T\mathbf{1}_{ji}) \\
&= N(\beta_{is}\beta_{tj} + \omega_{it}\zeta_{js}) + N[\boldsymbol{\Sigma}^{-1}]_{it}[\boldsymbol{\kappa}\boldsymbol{\kappa}^T]_{sj} \\
&= N(\beta_{is}\beta_{tj} + \omega_{it}\zeta_{js}) + N\omega_{it}[\boldsymbol{\kappa}\boldsymbol{\kappa}^T]_{sj} = N(\beta_{is}\beta_{tj} + \omega_{it}[\boldsymbol{\zeta} + \boldsymbol{\kappa}\boldsymbol{\kappa}^T]_{js}]),
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{J}_{12}]_{(ij,t)} &= -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \tau_t}\right] = N\mathrm{tr}\left(\frac{\partial \boldsymbol{\mu}^T}{\partial \lambda_{ij}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \tau_t}\right) \\
&= N\mathrm{tr}\left\{\boldsymbol{\kappa}^T \mathbf{1}_{ji} \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t1}\right\} = N[\boldsymbol{\kappa}^T]_{1j}[\boldsymbol{\Sigma}^{-1}]_{it} = N\omega_{it}\kappa_j,
\end{aligned}
$$

$$
[\boldsymbol{J}_{13}]_{(ij,tt)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \psi_{tt}}\right] = \frac{N}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ij}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{tt}}\right) = N\omega_{it}\beta_{tj},
$$

$$
\begin{aligned}
[\boldsymbol{J}_{14}]_{(ij,gh)} &= -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \phi_{gh}}\right] = \frac{N}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{ij}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}}\right) \\
&= \frac{N}{2}(2 - [\boldsymbol{I}]_{gh})(\alpha_{ig}\delta_{jh} + \alpha_{ih}\delta_{jg}),
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{J}_{15}]_{(ij,g)} &= -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda_{ij} \partial \kappa_g}\right] = N\mathrm{tr}\left(\frac{\partial \boldsymbol{\mu}^T}{\partial \lambda_{ij}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \kappa_g}\right) \\
&= N\mathrm{tr}\left(\boldsymbol{\kappa}^t \mathbf{1}_{ji} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \mathbf{1}_{g1}\right) = [\boldsymbol{\kappa}^T]_{1j}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ig} = N\alpha_{ig}\kappa_j,
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{J}_{22}]_{(i,t)} &= -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_i \partial \tau_t}\right] = N\mathrm{tr}\left(\frac{\partial \boldsymbol{\mu}^T}{\partial \tau_i} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \tau_t}\right) \\
&= N\mathrm{tr}\left(\mathbf{1}_{1i} \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t1}\right) = N\mathrm{tr}(\boldsymbol{I}\mathbf{1}_{1i} \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t1}) = N\omega_{it},
\end{aligned}
$$

$$
[\boldsymbol{J}_{23}]_{(i,tt)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_i \partial \psi_{tt}}\right] = 0,
$$

$$
[\boldsymbol{J}_{24}]_{(i,gh)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_i \partial \phi_{gh}}\right] = 0,
$$

$$
\begin{aligned}
[\boldsymbol{J}_{25}]_{(i,g)} &= -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \tau_i \partial \kappa_g}\right] = N\mathrm{tr}\left(\frac{\partial \boldsymbol{\mu}^T}{\partial \tau_i} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \kappa_g}\right) \\
&= N\mathrm{tr}\left(\mathbf{1}_{1i} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \mathbf{1}_{g1}\right) = N\mathrm{tr}(\boldsymbol{I}\mathbf{1}_{1i} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \mathbf{1}_{g1}) = N\alpha_{ig},
\end{aligned}
$$

$$
[\boldsymbol{J}_{33}]_{(ii,tt)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{ii} \partial \psi_{tt}}\right] = \frac{N}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{ii}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{tt}}\right) = \frac{N}{2}\omega_{it}^2,
$$

$$
[\boldsymbol{J}_{34}]_{(tt,gh)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{tt} \partial \phi_{gh}}\right] = \frac{N}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \psi_{tt}} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}}\right) = \frac{N}{2}(2 - [\boldsymbol{I}]_{gh})\alpha_{tg}\alpha_{th},
$$

$$
[\boldsymbol{J}_{35}]_{(ii,g)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \psi_{ii} \partial \kappa_g}\right] = 0,
$$

$$
\begin{aligned}
[\boldsymbol{J}_{44}]_{(gh,lq)} &= -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \phi_{gh} \partial \phi_{lq}}\right] = \frac{N}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{gh}}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{lq}}\right) \\
&= \frac{N}{4}(2 - [\boldsymbol{I}]_{gh})(2 - [\boldsymbol{I}]_{lq})(\gamma_{gl}\gamma_{hq} + \gamma_{gq}\gamma_{hl}),
\end{aligned}
$$

$$
[\boldsymbol{J}_{45}]_{(gh,l)} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \phi_{gh} \partial \kappa_{l}}\right] = 0,
$$

$$
\begin{aligned}
[\boldsymbol{J}_{55}]_{(g,h)} &= -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \kappa_{g} \partial \kappa_{h}}\right] = N\mathrm{tr}\left(\frac{\partial \boldsymbol{\mu}^T}{\partial \kappa_{g}}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \kappa_{h}}\right) \\
&= N\mathrm{tr}\left(\mathbf{1}_{1g}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{h1}\right) = N\mathrm{tr}(\boldsymbol{I}\mathbf{1}_{1g}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{1}_{h1}) = N\gamma_{gh}.
\end{aligned}
$$

■

F.2.2 Multiple-group factor model

We now generalize to the case of multiple groups the results derived in Appendix F.2.1. Consider G independent groups, each of size N_g, with $N = \sum_{g=1}^{G} N_g$ the total sample size across groups. Let $\boldsymbol{\sigma}_g = \mathrm{vech}(\boldsymbol{\Sigma}_g)$ be the vector of non-duplicated elements of the implied covariance matrix in group g, that is, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_g(\boldsymbol{\theta}_g) = \boldsymbol{\Lambda}_g\boldsymbol{\Phi}_g\boldsymbol{\Lambda}_g^T + \boldsymbol{\Psi}_g$, where $\boldsymbol{\theta}_g$ is the corresponding parameter vector. The mean structure $\boldsymbol{\mu}_g = \boldsymbol{\mu}_g(\boldsymbol{\theta}_g) = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g\boldsymbol{\kappa}_g$ and the covariance structure $\boldsymbol{\sigma}_g$ are gathered in the vector $\boldsymbol{\beta}_g = \boldsymbol{\beta}_g(\boldsymbol{\theta}_g) = (\boldsymbol{\mu}_g^T, \boldsymbol{\sigma}_g^T)^T$. The non-duplicated elements of the sample covariance matrix $\boldsymbol{s}_g = \mathrm{vech}(\boldsymbol{S}_g)$ and the sample mean vector $\bar{\boldsymbol{x}}_g$ are collected in $\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\beta}}_g(\boldsymbol{\theta}_g) = (\bar{\boldsymbol{x}}_g^T, \boldsymbol{s}_g^T)^T$. We also define $\boldsymbol{\Omega}_g = \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\Sigma}_g - \boldsymbol{W}_g)\boldsymbol{\Sigma}_g^{-1}$, where $\boldsymbol{W}_g = \boldsymbol{S}_g + (\bar{\boldsymbol{x}}_g - \boldsymbol{\mu}_g)(\bar{\boldsymbol{x}}_g - \boldsymbol{\mu}_g)^T$, $\boldsymbol{\Omega}_{\boldsymbol{\mu}g} = \boldsymbol{\Sigma}_g^{-1}(\bar{\boldsymbol{x}}_g - \boldsymbol{\mu}_g)$, $\boldsymbol{\Delta}_g = \dfrac{\partial \boldsymbol{\beta}_g}{\partial \boldsymbol{\theta}_g^T}$, $\boldsymbol{E}_g = \mathrm{diag}(\boldsymbol{\Sigma}_g^{-1}, \frac{1}{2}\boldsymbol{D}^T(\boldsymbol{\Sigma}_g^{-1} \otimes \boldsymbol{\Sigma}_g^{-1})\boldsymbol{D})$, and $\boldsymbol{V}_g = N_g\boldsymbol{E}_g$.

The log-likelihood of group g is

$$
\ell_g(\boldsymbol{\theta}_g) = -\frac{N_g}{2}\left\{\log|\boldsymbol{\Sigma}_g| + \mathrm{tr}(\boldsymbol{W}_g\boldsymbol{\Sigma}_g^{-1}) + p\log(2\pi)\right\};
$$

the gradient $\boldsymbol{g}_g(\boldsymbol{\theta}_g)$ is obtained by concatenating the free elements in

$$
\frac{\partial \ell_g(\boldsymbol{\theta}_g)}{\partial \boldsymbol{\Lambda}_g} = -N_g\boldsymbol{\Omega}_g\boldsymbol{\Lambda}_g\boldsymbol{\Phi}_g + N_g\boldsymbol{\Omega}_{\boldsymbol{\mu}g}\boldsymbol{\kappa}_g^T,
$$

$$
\begin{aligned}
\frac{\partial \ell_g(\boldsymbol{\theta}_g)}{\partial \boldsymbol{\Phi}_g} &= \begin{cases} -N_g \boldsymbol{\Lambda}_g^T \boldsymbol{\Omega}_g \boldsymbol{\Lambda}_g, & \text{non-diagonal elements,} \\ -\dfrac{N_g}{2} \boldsymbol{\Lambda}_g^T \boldsymbol{\Omega}_g \boldsymbol{\Lambda}_g, & \text{diagonal elements,} \end{cases} \\
\frac{\partial \ell_g(\boldsymbol{\theta}_g)}{\partial \boldsymbol{\Psi}_g} &= -N_g \text{diag}(\boldsymbol{\Omega}_g), \\
\frac{\partial \ell_g(\boldsymbol{\theta}_g)}{\partial \boldsymbol{\tau}_g} &= N_g \boldsymbol{\Omega}_{\boldsymbol{\mu} g}, \\
\frac{\partial \ell_g(\boldsymbol{\theta}_g)}{\partial \boldsymbol{\kappa}_g} &= N_g \boldsymbol{\Lambda}_g^T \boldsymbol{\Omega}_{\boldsymbol{\mu} g};
\end{aligned}
$$

the expected Fisher information matrix is $\boldsymbol{\mathcal{J}}_g(\boldsymbol{\theta}_g) = \boldsymbol{\Delta}_g^T \boldsymbol{V}_g \boldsymbol{\Delta}_g$. We can define the following quantities by assembling the group-specific elements over groups:

$$
\begin{aligned}
\boldsymbol{\theta} &= \left(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T, \ldots, \boldsymbol{\theta}_G^T\right)^T, \\
\boldsymbol{\sigma} &= \left(\boldsymbol{\sigma}_1^T, \ldots, \boldsymbol{\sigma}_g^T, \ldots, \boldsymbol{\sigma}_G^T\right)^T, \\
\boldsymbol{\mu} &= \left(\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_g^T, \ldots, \boldsymbol{\mu}_G^T\right)^T, \\
\boldsymbol{\beta} &= \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}^T} = \left(\boldsymbol{\beta}_1^T(\boldsymbol{\theta}_1), \ldots, \boldsymbol{\beta}_g^T(\boldsymbol{\theta}_g), \ldots, \boldsymbol{\beta}_G^T(\boldsymbol{\theta}_G)\right)^T = \left(\boldsymbol{\mu}^T, \boldsymbol{\sigma}^T\right)^T, \\
\boldsymbol{s} &= \left(\boldsymbol{s}_1^T, \ldots, \boldsymbol{s}_g^T, \ldots, \boldsymbol{s}_G^T\right)^T, \\
\bar{\boldsymbol{x}} &= \left(\bar{\boldsymbol{x}}_1^T, \ldots, \bar{\boldsymbol{x}}_g^T, \ldots, \bar{\boldsymbol{x}}_G^T\right)^T, \\
\hat{\boldsymbol{\beta}} &= \left(\hat{\boldsymbol{\beta}}_1^T(\boldsymbol{\theta}_1), \ldots, \hat{\boldsymbol{\beta}}_g^T(\boldsymbol{\theta}_g), \ldots, \hat{\boldsymbol{\beta}}_G^T(\boldsymbol{\theta}_G)\right)^T = \left(\bar{\boldsymbol{x}}^T, \boldsymbol{s}^T\right)^T, \\
\boldsymbol{\Delta} &= \text{diag}\left(\frac{\partial \boldsymbol{\beta}_1}{\partial \boldsymbol{\theta}_1^T}, \ldots, \frac{\partial \boldsymbol{\beta}_g}{\partial \boldsymbol{\theta}_g^T}, \ldots, \frac{\partial \boldsymbol{\beta}_G}{\partial \boldsymbol{\theta}_G^T}\right) = \text{diag}(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_g, \ldots, \boldsymbol{\Delta}_G), \\
\boldsymbol{E} &= \text{diag}(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_g, \ldots, \boldsymbol{E}_G), \\
\boldsymbol{V} &= \text{diag}(N_1 \boldsymbol{E}_1, \ldots, N_g \boldsymbol{E}_g, \ldots, N_G \boldsymbol{E}_G),
\end{aligned}
$$

where $\text{diag}(\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_G)$ denotes a block diagonal matrix with blocks $\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_G$. Then, the log-likelihood of the multiple-group factor model in equation (5.1) is $\ell(\boldsymbol{\theta}) = \sum_{g=1}^G \ell_g(\boldsymbol{\theta}_g)$ (see equation (F.1)), the gradient is $\boldsymbol{g}(\boldsymbol{\theta}) = \left(\boldsymbol{g}_1(\boldsymbol{\theta}_1)^T, \ldots, \boldsymbol{g}_g(\boldsymbol{\theta}_g)^T, \ldots, \boldsymbol{g}_G(\boldsymbol{\theta}_G)^T\right)^T$, and the expected Fisher information is $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) = \boldsymbol{\Delta}^T \boldsymbol{V} \boldsymbol{\Delta} = \text{diag}\left(\boldsymbol{\mathcal{J}}_1(\boldsymbol{\theta}_1), \ldots, \boldsymbol{\mathcal{J}}_g(\boldsymbol{\theta}_g), \ldots, \boldsymbol{\mathcal{J}}_G(\boldsymbol{\theta}_G)\right)$.

Index

References

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

Anderson, T. W. (1989). Linear latent variable models and covariance structures. *Journal of Econometrics*, *41*(1), 91–119.

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.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438.

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

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and asymptotics* (Vol. 52). Chapman & Hall. Monographs on Statistics and Applied Probability.

Bartholomew, D. J., Knott, M. & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.

Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford, England: John Wiley & Sons.

Bollen, K. A. & Jöreskog, K. G. (1985). Uniqueness does not imply identification: A note on confirmatory factor analysis. *Sociological Methods & Research*, *14*(2), 155–163.

Bollen, K. A. & Long, J. S. (1993). *Testing structural equation models* (Vol. 154). Sage Publications.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Browne, M. W. (2001). An overview of 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, US: The Guilford Press.

Conn, A. R., Gould, N. I. & Toint, P. L. (2000). *Trust region methods* (Vol. 1). Siam. MPS/SIAM Series on Optimization.

Danaher, P., Wang, P. & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 373–397.

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.

Golub, G. H. & Van Loan, C. F. (2012). *Matrix computations*. JHU press.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450.

Gu, C. (2013). *Smoothing spline anova models*. Springer Science & Business Media.

Hägglund, G. (1982). Factor analysis by instrumental variables methods. *Psychometrika*, *47*(2), 209–222.

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*.

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. (in press). *lslx: Semi-confirmatory structural equation modeling via penalized likelihood.* (Journal of Statistical Software)

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

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)

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)

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

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. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133.

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.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, *49*(1), 169–186.

Kelley, K. (2019). MBESS: The MBESS R package [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=MBESS (R package version 4.4.3)

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.

Konishi, S. & Kitagawa, G. (2008). *Information criteria and statistical modeling.* Springer Science & Business Media.

Little, T. D., Card, N. A., Slegers, D. W. & Ledford, E. C. (2012). Representing con-

textual 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). Routledge New York, NY.

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.

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.

Magnus, J. R. & Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons.

Marra, G. & Radice, R. (2019a). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, 1–20.

Marra, G. & Radice, R. (2019b). GJRM: Generalised joint regression modelling [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=GJRM/ (R package version 0.2-2)

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

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 99–130). Lawrence Erlbaum Associates.

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*. Routledge.

Mulaik, S. A. (1971). A note on some equations of confirmatory factor analysis. *Psychometrika*, *36*(1), 63–70.

Mulaik, S. A. (2009). *Foundations of factor analysis*. 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.

Muthén, L. & Muthén, B. (2020). *Mplus. the comprehensive modelling program for applied researchers: user's guide 5.*

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

Oelker, M. & Tutz, G. (2013). *A general family of penalties for combining differing types of penalties in generalized structured models* (Tech. Rep.). Univerisity of Münich, Münich, Germany.

Petry, S. (2011). *Regularization approaches for generalized linear models and single index models* (Unpublished doctoral dissertation). Ludwig–Maximilians–Universität München.

R Core Team. (2018). R: A language 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.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*(2), 1–36.

Rosseel, Y. et al. (2019). lavaan: Latent Variable Analysis [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=lavaan/ (R package version 0.6-5)

Scharf, F. & Nestler, S. (2019). Should regularization replace simple structure

rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 576-590.

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

Steenkamp, J. E. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90.

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, US: 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.* Verlag Dr. Hut.

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.* 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.