Generalised Latent Variable Models for Location, Scale, and Shape parameters

Camilo Alberto Cárdenas Hurtado

A thesis presented for the degree of

Doctor of Philosophy

Department of Statistics

London School of Economics and Political Science

July 2023

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 37,518 words.

I confirm that Chapters 2 and 3 were jointly co-authored with Professor Irini Moustaki, Dr. Yunxiao Chen, and Professor Giampiero Marra, and I contributed 80% of this work.

Camilo Alberto Cárdenas Hurtado

July 2023

ɔ ƨ ɘ o ƀ o T

ɐ ɔ i l q ɯ o

ƨ ɘ ɟ ɹ ɐ o ƀ

ɘ ƨ ɿ ɘ ƨ ɘ ƀ

. o l l i ɔ ɹ

To those who believed in me

Thank you

Acknowledgements

I would like to begin by expressing gratitude to my supervisors, Professor Irini Moustaki and Dr. Yunxiao Chen. Throughout this journey, you both have been exceptional mentors.

Irini, you have been the kindest, most thoughtful, and caring supervisor I could ever wish for. Life during the pandemic was not easy, let alone navigating the perils of a PhD programme. You made my life easier when I most needed it, but also challenged me to push myself harder when I had to. I am thankful for the creative freedom you gave me in pursuing research projects that excited me, but I also appreciated your practicality and the different ways you brought me down to earth when I was off-track. Thank you for holding space and making time for our meetings. Thank you for talking to me and offering perspective when I was (secretly) feeling overwhelmed. Thank you for looking after me. Thank you for doing your best to make sure I do my best.

Yunxiao, I am grateful for the intellectual challenges you presented in every conversation we had. Our interactions not only made me a better Statistician, but also sharpened my critical thinking skills. Thank you for your diligence and patience. Thank you for being so sharp and competent, and for sharing your clear understanding of Statistics with me.

I also want to express my gratitude to Professor Giampiero Marra, who was always open to discussing my research ideas and providing technical assistance whenever I needed it. Thank you for your patience, Giampiero. Additionally, I am thankful to all the faculty members in the Department of Statistics, particularly Professor Wicher Bergsma, Professor Jouni Kuha, and Dr. Kostas Kalogeropoulos. My deepest gratitude also go to the examiners in my viva, Professor Jouni Kuha and Dr. Yang Liu. Their thought-provoking questions and insightful comments greatly contributed to the improvement of my dissertation.

I am extremely grateful to all the staff members in the Department of Statistics, with a special mention for Penny Montague. She is truly a silent hero who keeps the PhD programme running smoothly. I cannot express enough admiration and appreciation for her. Thank you, Penny, for taking care of all of us. Your dedication and tireless efforts make PhD students lives so much

easier.

To me, the PhD experience was more of an emotional and mental journey than just an academic one. While you may think you are being trained as a researcher, in reality, you are learning to pull long nights, consistently work hard, navigate frustration, deal with rejection, and silence that inner critic that questions almost everything you do. Completing a PhD requires a strong support circle. That includes your supervisors, but also your family, friends, and loved ones. With their support, and sometimes through them, you come to appreciate all the joy and growth that comes with the PhD journey. You need people to build memories, people who celebrate your accomplishments, and people to share your self-improvement and self-knowledge with. The list of individuals who have had a meaningful impact on my PhD journey is long, and I wish I could personally thank each one of you. I could not have done it without your support. Forgive me if you can't find your name below; I did my best.

I learned from others before me. I thank Dr. José Manuel Pedraza Ramírez, Dr. Tianlin Xu, Dr. Alice Pignatelli di Cerchiara, and Dr. Despoina Markou for showing me how it's done. I am grateful for the time at LSE when you were there; you gave me perspective and reassurance that things would always be fine. Special thanks to José, because representation matters. You are the face of Latino excellence everywhere you go, y eso inspira muchísimo.

I would like to thank my friends and colleagues: Shakeel Gavioli-Akilagun, Eduardo Ferioli-Gomes, Dr. Alexandros Pavlis, Sahoko Ishida, Dr. Gabriel Wallin, Dr. Anica Kostic, Dr. Patrick Aschermayr, Jinghan Tee, Xinyi Liu, Motonori Oka, and everyone on the fifth floor in Columbia House. You made my time in London very enjoyable, fun, and worth remembering for the rest of my life. I must also acknowledge my favourite non-LSE Italian visitors, Dr. Lucia Guastadisegni, and Dr. Giuseppe Alfonzetti. You were both awesome desk partners. Thank you for every coffee we drank together.

To my family in London: María José Herrera, Oriol Bosch, Eduardo Mercadante, Gabriel Granda, Parker Foe, and Roy White. You make the UK feel like home. Thanks for the dinners, trips, drinks, parties, and all the wonderful memories. I couldn't have found better friends to share this experience with than all of you. Importantly, a big shout-out to all the members at British Barbell. Although you'll probably never read this, power-lifting (a.k.a. "abusing my body by lifting heavy stuff, several times a week, for almost four years now, and counting") played a key role in keeping my sanity during the PhD. Training with all of you has been, and will always be, a privilege. Stay `#NaturallyMassive`.

To my family: my parents, Alberto and Diana, and my brother, Andrés. Thank you for

supporting my decisions and understanding during this period of my life. You have taught me everything I needed to become who I am today, and there are no words of appreciation that can fully describe how much I love you. You have always cheered for me, but you are the real heroes.

My gratitude goes to my friends back home, who have supported me and blessed me with their patience. Special thanks to Ricardo Salas, María Paula Gandur, Raúl Arce, José Fernando Moreno, Óscar Unás, Andrés Bacca, Juan Diago, Iván Lozano, Juan Sebastián Ramírez, and Manuel Preciado. You have witnessed me build this dream, step by step, since day one. You are a part of my identity. Thank you for being there when I needed it, thank you for your patience (x2) and love, thank you for your consistency, for your understanding. Thank you for being my outlet and for being part of my much-deserved resting time. To all of you: my absences were not intentional, and believe me when I say that I did my best to be present whenever we interacted. Please do.

Last, but not least, I would like to thank my partner, Nancy Breton. You have been my best support, companion, friend, partner, lover, roommate, travel buddy, soulmate, and more that I could ever ask for. This dissertation would not have seen the light without your continuous support. Life has been generous enough to cross our paths. I will never get tired of telling you how much of a gift you are to me. I consider myself the luckiest person for having met you, especially during this journey. Thank you for your countless motivating words and for being patient during all those late nights when I was working like a madman. Thank you for reminding me to take care of myself. Thank you for making me understand that I needed a break when I even didn't know. Thank you for believing in me. Thank you for being proud of me. And, of course, thank you for being late to the PhD academy orientation talk; it turned out to be a serendipitous event that changed my whole world for the better.

To all of you: many, many thanks.

Much love, Camilo

Abstract

Latent Variable Models (LVM) are widely used in social, behavioural, and educational sciences to uncover underlying associations in multivariate data using a smaller number of latent variables. However, the classical LVM framework has certain assumptions that can be restrictive in empirical applications. In particular, the distribution of the observed variables being from the exponential family and the latent variables influencing only the conditional mean of the observed variables. This thesis addresses these limitations and contributes to the current literature in two ways.

First, we propose a novel class of models called Generalised Latent Variable Models for Location, Scale, and Shape parameters (GLVM-LSS). These models use linear functions of latent factors to model location, scale, and shape parameters of the items' conditional distributions. By doing so, we model higher order moments such as variance, skewness, and kurtosis in terms of the latent variables, providing a more flexible framework compared to classical factor models. The model parameters are estimated using maximum likelihood estimation.

Second, we address the challenge of interpreting the GLVM-LSS, which can be complex due to its increased number of parameters. We propose a penalised maximum likelihood estimation approach with automatic selection of tuning parameters. This extends previous work on penalised estimation in the LVM literature to cases without closed-form solutions.

Our findings suggest that modelling the entire distribution of items, not just the conditional mean, leads to improved model fit and deeper insights into how the items reflect the latent constructs they are intended to measure. To assess the performance of the proposed methods, we conduct extensive simulation studies and apply it to real-world data from educational testing and public opinion research. The results highlight the efficacy of the GLVM-LSS framework in capturing complex relationships between observed variables and latent factors, providing valuable insights for researchers in various fields.

Contents

List of Figures

List of Tables

Chapter 1

Background: A short introduction to Latent Variable Models

In this chapter, we provide a comprehensive introduction to the Generalised Linear Latent Variable Model (GLLVM) framework. This Chapter provides an overview of the basic concepts that will be explored in this dissertation. In Section 1.2 we present a thorough discussion of model identification, and in Section 1.3 we discuss maximum likelihood estimation of the model parameters. Finally, in Section 1.4, we provide a brief introduction to factor scoring, which involves assigning values of the latent variables to observations in our sample. In Section 1.5 we conclude and provide an outline of this dissertation.

1.1. Introduction

Research questions in the social and behavioural sciences often involve analysing large and complex datasets that have measures on binary, categorical (nominal or ordinal), and/or metric (discrete or continuous) scales. Multivariate data analysis aims to identify common patterns and simplify complex structures in either an exploratory (i.e., data-driven) or confirmatory (i.e., hypothesis-driven) manner. Latent variable models (LVM) are a general class of statistical models that are commonly used to reduce the dimensionality of observed variables by explaining their associations through a set of lower-dimensional latent variables.

Formally, let $\mathbf{y} = (y_1, ..., y_p)^\intercal$ be a vector of p observed variables, also known as *items*, and $\mathbf{z} = (z_1, ..., z_q)^\intercal$ a vector of q latent variables, also known as latent *factors* or latent *traits*, with

$q \ll p$. The joint probability distribution of $(\mathbf{y}, \mathbf{z})$ can be expressed as

$$f(\mathbf{y}, \mathbf{z}) = f(\mathbf{y} \,|\, \mathbf{z})\ p(\mathbf{z}). \tag{1.1}$$

Here, the multivariate conditional probability function $f(\mathbf{y} \mid \mathbf{z})$, known as the measurement model, describes the relationship between the observed variables given the latent variables. In other words, it captures the associations between items in $\mathbf{y}$ resulting from the set of factors $\mathbf{z}$. Similarly, the multivariate density function $p(\mathbf{z})$ corresponds to the structural model and specifies the joint distribution of the latent variables.

It is commonly assumed that the data generating process $f(\mathbf{y}, \mathbf{z}; \Theta) = f(\mathbf{y} \mid \mathbf{z}; \Theta_y)\ p(\mathbf{z}; \Theta_z)$ follows a parametric probability function characterised by a vector of parameters $\Theta^\intercal = (\Theta_y^\intercal, \Theta_z^\intercal)$. Here, Θ_y and Θ_z are the vectors of parameters characterising the measurement and structural models, respectively. In some settings, it is convenient to assume that the observed variables are independent from each other conditional on the latent variables. This assumption, known as *local (or conditional) independence*, allows for expressing the measurement model as $f(\mathbf{y} \mid \mathbf{z}; \Theta_y) = \prod_{i=1}^{p} f_i(y_i \mid \mathbf{z}; \Theta_{y_i})$, where Θ_{y_i} are the parameters in the measurement model for item i. The conditional independence assumption is not a necessary assumption and can be relaxed in some settings (e.g., longitudinal data). Moreover, measurement models for individual items do not necessarily have to follow the same parametric form.

The LVM framework encompasses several statistical models such as the generalised random effects model (Laird and Ware, 1982; Zeger and Karim, 1991; Lee and Nelder, 1996, 2001, 2006), models for longitudinal data (Hedeker and Gibbons, 2006), latent class analysis models (LCA, Lazarsfeld and Henry, 1968), the factor analysis model (Lawley and Maxwell, 1962, 1971), the structural equation model (LISREL, Jöreskog, 1970a,b, 1973, and SEM, Bollen, 1989), and the class of item response theory models (IRT, Lord and Novick, 1968; Bartholomew, 1980; Bock and Aitkin, 1981; Bartholomew et al., 2011). These statistical models were originally created to answer different research questions, but they all serve similar purposes such as dimensionality reduction or measurement of latent constructs. They differ on the specific distributional assumptions on imposed on $f(\mathbf{y} \mid \mathbf{z}; \Theta_y)$ and $p(\mathbf{z}; \Theta_z)$. Most of the above models can be considered as particular cases of the generalised linear latent variable model (GLLVM, see, e.g., Bartholomew et al., 2011 or Skrondal and Rabe-Hesketh, 2004, 2007).

The GLLVM framework builds upon the generalised linear model framework (GLM, see, e.g., Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). Each of the items $i = 1, ..., p$ has a

measurement model $f_i(y_i \mid \mathbf{z})$ that follows a distribution from the exponential family:

$$f_i(y_i \mid \mathbf{z}; \zeta_i, \phi_i) = \exp\left\{ \frac{y_i \zeta_i(\mathbf{z}) - b_i(\zeta_i(\mathbf{z}))}{\phi_i} + c_i(y_i; \phi_i) \right\}, \tag{1.2}$$

where $\zeta_i(\mathbf{z})$ is the canonical parameter and the functional dependence on $\mathbf{z}$ denotes that is modelled in terms of the latent variables, ϕ_i is a dispersion parameter, and b_i and c_i are pre-specified distribution-specific functions. Moreover, for each item we assume a systematic component, denoted by η_i, which results from a linear combination of a set of covariates, in this case, the latent variables $\mathbf{z}$:

$$\eta_i = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} z_j \tag{1.3}$$

This linear equation is known as the measurement equation in the LVM literature. In the GLLVM, the relationship between items and factors is linear through the canonical parameter: $\zeta_i(\mathbf{z}) = \eta_i$.

The focus in the GLLVM is on modelling the conditional mean of the items as linear functions of the latent variables, that is $\mu_i(\mathbf{z}) := \mathbb{E}(y_i \mid \mathbf{z})$. Naturally, if f_i is from the exponential family, we have that $\mu_i(\mathbf{z}) = b_i'(\eta_i)$, where $b_i' = \mathrm{d}b_i/\mathrm{d}\zeta_i$. The mapping $\upsilon_i^{-1} := b_i'$ is known as the canonical link function, a monotonic differentiable function that connects the conditional mean of y_i with the latent variables, and thus

$$\upsilon_i(\mu_i(\mathbf{z})) = \eta_i$$

The link function υ_i depends on the distribution assumed for y_i. The conditional variance of an item is $\mathrm{Var}(y_i \mid \mathbf{z}) = \phi_i b_i''(\eta_i) = \phi_i \mu_i'(\mathbf{z})$. For some distributions in the exponential family, the scale parameter is $\phi_i = 1$, and it is only of interest in the continuous case. Higher order moments of the manifest variables, like the kurtosis and the skewness, are not modelled explicitly in terms of the latent variables $\mathbf{z}$ (or any set of covariates).

We use matrix notation to represent the measurement equations more concisely. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\intercal$ be a vector of conditional means. Also, let $\boldsymbol{\alpha}_0 = (\alpha_{10}, \ldots, \alpha_{p0})^\intercal$ be a vector of intercepts and A_μ be a matrix of model parameters (also known as the *factor loading matrix*) with p rows corresponding to the q-dimensional vectors $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iq})^\intercal$, for items $i = 1, \ldots, p$. Lastly, let $\upsilon(\cdot)$ be the vector function that applies the corresponding link function $\upsilon_i(\cdot)$ to each entry of $\boldsymbol{\mu}$. The measurement equations can be expressed in matrix notation as:

$$\upsilon(\boldsymbol{\mu}) = \boldsymbol{\alpha}_0 + \mathrm{A}_\mu \mathbf{z}, \tag{1.4}$$

Regarding the structural model for the latent variables, in this dissertation we focus on the

case of continuous latent variables. In particular, we assume $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \mathbf{\Phi})$, where $\mathbf{\Phi}$ is a covariance (or correlation) matrix.

Finally, let $K = \dim(\Theta)$ and define $\Xi \subseteq \mathbb{R}^K$ as the parameter space. Since the latent variables $\mathbf{z}$ are unobserved, we can only infer the model parameters $\Theta \in \Xi$ from the marginal (observed data) distribution, given by:

$$f(\mathbf{y}; \Theta) = \int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_i \,|\, \mathbf{z}; \Theta_{y_i}) \right] \, p(\mathbf{z}; \Theta_z) \, \mathrm{d}\mathbf{z} \tag{1.5}$$

In the following sections we discuss identification and estimation of the model parameters.

1.2. Model Identification

A parametric statistical model is identified if a point $\Theta \in \Xi$ generates a unique value for the marginal distribution $f(\mathbf{y}; \Theta)$ for the observed data. If the model is not identified, there could be multiple sets of parameters that could have produced the same observed data distribution, in that sense making the model parameters arbitrary and invalid for substantive scientific inference. The following definitions from Skrondal and Rabe-Hesketh (2004, Chapter 5) are useful:

Definition 1.2.1 (Observationally equivalent points)**.** Two distinct points Θ^* and $\Theta^\dagger$ in the parameter space Ξ are observationally equivalent if they generate the same marginal probability distribution for the observed variables, $f(\mathbf{y}; \Theta^*) = f(\mathbf{y}; \Theta^\dagger)$, for all $\mathbf{y} \in \mathbb{R}^p$.

Definition 1.2.2 (Globally identified parameter point)**.** The parameters of a statistical model are globally identified if for any given point $\Theta^* \in \Xi$ there is no other observationally equivalent point $\Theta^\dagger \in \Xi$.

Definition 1.2.3 (Locally identified parameter point)**.** A point $\Theta^* \in \Xi$ is locally identified if there exists an open neighbourhood around Θ^* containing no other point, say $\Theta^\dagger$, that is observationally equivalent to Θ^*.

Often conditions for global identification depend on the specifics of the model at hand and are difficult to verify (Shapiro, 1985). Thus, for general models, it is common practice to rely on the weaker notion of local identification. Local identification is usually achieved by imposing restrictions on the parameters in a mathematically structured (but sometimes arbitrary) manner. It is important to note that local identification throughout the parameter space Ξ is a necessary but not sufficient condition for global identification (Bentler and Weeks, 1980; McDonald, 1982).

Additionally, local identification at one point in Ξ does not guarantee local identification throughout Ξ, and in some cases, certain points in Ξ may not be locally identified even after imposing restrictions (McDonald, 1982).

1.2.1 Analytical identification

Under standard regularity assumptions for the marginal probability function $f(\mathbf{y};\Theta)$, a necessary and sufficient condition for local identification at a given point $\Theta^* \in \Xi$ is that the (theoretical) information matrix is non-singular when evaluated at that point (Rothenberg, 1971). That is, $\mathbb{E}[\nabla_\Theta \log f(\mathbf{y};\Theta)\nabla_\Theta \log f(\mathbf{y};\Theta)^\intercal]|_{\Theta=\Theta^*}$, must be non-singular. However, in practice, this approach is often unfeasible due to the analytical intractability of the information matrix in complex models.

An alternative approach for assessing local identifiability was proposed by Wald (1950), but it only applies to cases where the marginal distribution of the observed data is completely characterised by reduced-form parameters, typically related to the items' first- and second-order moments. In Wald's approach, we determine local identifiability of a point $\Theta^* \in \Xi$ by studying the functional relationship between the model parameters Θ and the reduced-form parameters.

Let $\mathbf{m}$ be a S-dimensional vector of reduced-form parameters in the parameter space $\mathcal{M} \subset \mathbb{R}^S$. Assume there exist S continuously differentiable known functions $m_s = h_s(\Theta)$ $(s = 1,\dots,S)$ mapping Ξ into $\mathcal{M}$ such that the vector of reduced-form parameters is $\mathbf{m} = (m_1,\dots,m_S)^\intercal$. For brevity, let $h(\cdot)$ denote the vector function that applies the corresponding mapping $h_s(\cdot)$ to each entry of Θ, i.e., $\mathbf{m} = h(\Theta)$. Moreover let $\tilde{f}(\mathbf{y};h(\Theta))$ be the (marginal) distribution function for the observed data $\mathbf{y}$ parameterised by the reduced-form parameters, such that $\tilde{f}(\mathbf{y};h(\Theta)) = \tilde{f}(\mathbf{y};\mathbf{m}) = f(\mathbf{y};\Theta)$.

As a clarifying example, consider the Normal linear factor model (NLFM) with one latent variable and p items, where the items have been mean-centred for simplicity. The NLFM can be expressed as $y_i = \alpha_{i1}z_1 + \epsilon_i$, where $z \sim \mathbb{N}(0,\psi_z)$, the error terms $\epsilon_i = \mathbb{N}(0,\sigma_i^2)$, and $\text{Cov}(\epsilon_i,\epsilon_{i'}) = 0$ for $i \neq i'$. On one hand, the model parameters are $\Theta = (\boldsymbol{\alpha}_1^\intercal, \text{vech}(\Sigma_\epsilon)^\intercal, \psi_z)^\intercal$, where $\boldsymbol{\alpha}_1 = (\alpha_{11},\dots,\alpha_{p1})^\intercal$ is a vector of factor loadings, $\Sigma_\epsilon = \text{diag}(\sigma_1^2,\dots,\sigma_p^2)$ is the diagonal covariance matrix of the error terms, and 'vech' denotes the half-vectorisation operator. Thus, $f(\mathbf{y};\Theta) = \mathbb{N}(\mathbf{0},\psi_z\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^\intercal + \Sigma_\epsilon)$. On the other hand, the reduced-form parameters $\mathbf{m}$ consist of the $p(p+1)/2$ non-redundant elements of the covariance matrix of $\mathbf{y}$, denoted by $\boldsymbol{\Sigma}_y$. That is, $\mathbf{m} = \text{vech}(\boldsymbol{\Sigma}_y)$. The marginal distribution of the observed data parameterised by the reduced-form parameters is $\tilde{f}(\mathbf{y};\mathbf{m}) = \mathbb{N}(\mathbf{0},\Sigma_y)$. Note how the reduced-form parameters can be expressed in terms of the model parameters: $\Sigma_y = \psi_z\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^\intercal + \Sigma_\epsilon$. Specifically, the diagonal elements of $\text{diag}(\Sigma_y)_{[i]} =$

$\text{Var}(y_i) = \alpha_{i1}^2\psi_z + \sigma_i^2$, where the sub-index $[i]$ denotes the i^{th} entry of the vector; and the off-diagonal elements of Σ_y being $\text{Cov}(y_i, y_{i'}) = \alpha_{i1}\alpha_{i'1}\psi_z$ for $i \neq i'$. Thus, for the NLFM we have $\tilde{f}(\mathbf{y}; \mathbf{m}) = f(\mathbf{y}; \Theta)$, for all $\mathbf{y} \in \mathbb{R}^p$. Other examples of 'simple' GLLVMs completely characterised by reduced-form parameters include the factor model with dichotomous or ordinal items (e.g., Muthén, 1984), and the Structural Equation Model (SEM, Bollen, 1989).

For a point $\Theta^* \in \Xi$ that generates the reduced-form parameters $\mathbf{m}^* = h(\Theta^*)$, Θ^* is identifiable if and only if Θ^* is the unique solution to the system of equations $\mathbf{m}^* = h(\Theta)$. Therefore, identification of the model parameters Θ depends on the characteristics of the mapping $h : \Xi \to \mathcal{M}$. A necessary, but not sufficient, condition for identification is that there are at least as many reduced-form parameters as there are model parameters, that is $K \leq S$. General parameter identification rules of this type have been long established in the LVM literature, such as the restrictions on the model parameters in Confirmatory Factor Analysis (CFA, Jöreskog, 1969) or the 't-rule' in SEM (Bollen, 1989). However, stronger identification results can be derived from the characteristics of the Jacobian matrix:

$$\text{J}(\Theta) = \left[\frac{\partial h_s}{\partial \Theta_{[k]}}, \quad 1 \leq s \leq S; 1 \leq k \leq K\right],$$

where the sub-index $[k]$ refers to the k^{th} entry in the parameter vector Θ. The following definition is required:

Definition 1.2.4 (Regular point, Skrondal and Rabe-Hesketh, 2004, Chapter 5)**.** A point $\Theta^* \in \Xi$ is a *regular point* if there is an open neighbourhood of Θ^* in which the Jacobian matrix has constant rank.

If Θ^* is a regular point, then the system of equations $\mathbf{m}^* = h(\Theta)$ has a unique solution Θ^* if and only if $\text{rank}(\text{J}(\Theta^*)) = K$. Therefore, local identification can be analytically assessed based on the following Lemma (Skrondal and Rabe-Hesketh, 2004, Chapter 5):

Lemma 1.2.1. Let $\Theta^* \in \Xi$ be a regular point of $\text{J}(\Theta)$. Then, Θ^* is locally identified if and only if $\text{rank}(\text{J}(\Theta^*)) = K$.

As discussed above, the approach proposed by Wald (1950) for assessing local identification is limited to 'simple' models in which i) there are reduced-form parameters that fully characterise the marginal distribution of the observed data, and ii) the functional forms of the mapping functions $h_s(\cdot)$ are known so that the analytical Jacobian matrix $\text{J}(\Theta)$ can be computed. Despite this limitation, the condition in Lemma 1.2.1 can be achieved in practice by imposing constrains on the model parameters Θ. These constraints are rank restrictions on $\text{J}(\Theta)$, which limit the number of free parameters in the model and ensure that the constrained point is a regular point of $\text{J}(\Theta)$. In the

next subsection, we discuss common restrictions on Θ used in the LVM literature. Throughout the rest of this dissertation, when we use the term 'identification' we are referring to local identification, unless otherwise stated.

1.2.2 Identification via restrictions on the model parameters

The model parameters $\Theta^\intercal = (\Theta_y^\intercal, \Theta_z^\intercal) \in \Xi$ in the GLLVM are not identified, partly due to the arbitrariness of the location and the scale of the latent variables. To achieve local identification, it is necessary (but not sufficient) to impose restrictions on Θ. Using simple examples, below we illustrate how the GLLVM parameters are not identifiable without additional constraints.

Scale and location indeterminacy: the unidimensional case ($q = 1$)

As a starting example, consider a GLLVM with only one latent variable and p items. In this case, the distributional assumption on the structural model is $z_1 \sim \mathbb{N}(0, \psi_z)$. The measurement equations are:

$$\upsilon_i(\mu_i(\mathbf{z})) = \alpha_{i0} + \alpha_{i1} z_1, \qquad i = 1, ..., p$$

The model parameters are $\Theta^\intercal = (\boldsymbol{\alpha}_0^\intercal, \boldsymbol{\alpha}_1^\intercal, \boldsymbol{\phi}^\intercal, 0, \psi_z)$, where $\boldsymbol{\alpha}_0 = (\alpha_{10}, \ldots, \alpha_{p0})^\intercal$ is the vector of intercepts, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{p1})^\intercal$ is the vector of factor loadings, and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\intercal$ is the vector of scale parameters. Note that some entries in $\boldsymbol{\phi}$ could be fixed to 1 for some items following certain distributions (e.g., Poisson, Binomial, etc.), and thus are not estimated. For pedagogical reasons we include the mean of the latent variable (0) in Θ, to show how the identification problem is related to the location and scale of the factor z_1. However, it's important to note that this is a fixed parameter and is not estimated.

Consider a linear transformation of the latent variable, $\tilde{z}_1 = az_1 + c$, where a and c are arbitrary constants. This transformation results in a different structural model, as $\tilde{z}_1 \sim \mathbb{N}(c, a^2\psi_z)$. This change is compensated in the measurement part of the model:

$$\begin{aligned}\upsilon_i(\mu_i(\tilde{\mathbf{z}})) &= (\alpha_{i0} - \alpha_{i1}c/a) + (\alpha_{i1}/a)\tilde{z}_1 \\ &= \tilde{\alpha}_{i0} + \tilde{\alpha}_{i1}\tilde{z}_1\,,\end{aligned}$$

where $\tilde{\alpha}_{i0} = \alpha_{i0} - \alpha_{i1}c/a$ and $\tilde{\alpha}_{i1} = \alpha_{i1}/a$, for $i = 1, ..., p$. After this transformation, the model parameters can be expressed as $\tilde{\Theta}^\intercal = (\tilde{\boldsymbol{\alpha}}_0^\intercal, \tilde{\boldsymbol{\alpha}}_1^\intercal, \boldsymbol{\phi}^\intercal, c, a^2\psi_z)$, where the vectors $\tilde{\boldsymbol{\alpha}}_0$ and $\tilde{\boldsymbol{\alpha}}_1$ are defined similarly as above. Note how the conditional mean of the observed items are equivalent between

parametrisations: $\mu(\mathbf{z}) \equiv \mu(\tilde{\mathbf{z}})$. Moreover, $f(\mathbf{y}; \Theta) \equiv f(\mathbf{y}; \tilde{\Theta})$. Thus, the linear transformation of the latent variables leaves the distribution of the observed data unchanged.

Without further assumptions on the values for the model parameters, the unidimensional GLLVM is not identified, meaning that we cannot determine a *unique* point for the parameters that produces a *unique* value for the marginal distribution of the observed data. The latter is a good example of how the measurement model $f(\mathbf{y} \mid \mathbf{z})$ and the structural model $p(\mathbf{z})$ must be considered as a pair. Any change in one is balanced out by a compensating change in the other, and thus we are often required to 'fix' some parts of either the measurement or the structural model (or both) before estimating the model parameters. In other words, the parameters in the measurement model (intercepts $\boldsymbol{\alpha}_0$ and factor loadings A_μ) and the parameters in the structural model (factor covariance matrix $\boldsymbol{\Phi}$) are not independent of one another. In the unidimensional example above, we potentially rectify the indeterminacy of the location and the scale of the latent variable by fixing the values for the parameters in the structural model, say $z_1 \sim \mathbb{N}(0, 1)$. If there is a compelling reason to estimate the variance of the latent variable ψ_z, we can fix $\alpha_{i'1} = 1$ in the measurement equation for a chosen item i'.

Rotation indeterminacy: the multidimensional case ($q > 1$)

The issue of parameter identification for the multidimensional case ($q > 1$) has been long studied in the LVM literature. The seminal work of Anderson and Rubin (1956) is one such contribution that addresses this issue. In their work, the authors provide a comprehensive explanation of the indeterminacies that are inherent in the classical Normal linear factor model. In the context of multidimensional LVM, the indeterminacy pertaining to the latent variables' scale is commonly referred to as the rotational indeterminacy problem.

Let $\mathbf{z} \in \mathbb{R}^q$, where $q > 1$, and assume $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Phi}_z)$. For simplicity, assume the mean of the latent variables is fixed. Recall the measurement equations in matrix notation: $\upsilon(\mu) = \boldsymbol{\alpha}_0 + \mathrm{A}_\mu \mathbf{z}$. Now, consider an arbitrary rotation of the latent variable vector, denoted by $\tilde{\mathbf{z}} = \mathbb{M}\mathbf{z}$, where $\mathbb{M}$ is a non-singular $(q \times q)$ matrix. This rotation results in a different distribution of the latent variables in the structural model, $\tilde{\mathbf{z}} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Phi}_{\tilde{z}})$, where $\boldsymbol{\Phi}_{\tilde{z}} = \mathbb{M}\boldsymbol{\Phi}_z\mathbb{M}^\intercal$. Similar to the unidimensional case described above ($q = 1$), the compensating change in the measurement model is

$$\begin{aligned} \upsilon(\boldsymbol{\mu}) &= \boldsymbol{\alpha}_0 + \mathrm{A}_\mu \mathbb{M}^{-1}\mathbb{M}\mathbf{z} \\ &= \boldsymbol{\alpha}_0 + \tilde{\mathrm{A}}_\mu \tilde{\mathbf{z}} \end{aligned}$$

Following the rotation of the latent variables, the model that involves A_μ and $\mathbf{z}$ is empirically indistinguishable from the model that involves $\tilde{\mathrm{A}}_\mu$ and $\tilde{\mathbf{z}}$, where $\tilde{\mathrm{A}}_\mu = \mathrm{A}_\mu \mathbb{M}^{-1}$ and $\tilde{\mathbf{z}} = \mathbb{M}\mathbf{z}$. In contrast to the unidimensional case ($q = 1$), simply 'fixing' the scale of the latent variables as $\boldsymbol{\Phi}_z = \mathbb{I}_q$ (where $\mathbb{I}q$ represents the identity matrix of order q) does not solve the indeterminacy issue. Indeed, if the rotation matrix $\mathbb{M}$ is *orthogonal*, then $\mathbb{M}\mathbb{M}^\intercal = \mathbb{I}_q$ and thus we have two indistinguishable models sharing *the same* specification of the structural model, that is, $\boldsymbol{\Phi}_z = \boldsymbol{\Phi}_{\tilde{z}} = \mathbb{I}_q$, but with different factor loading matrices, as $\mathrm{A}_\mu \neq \tilde{\mathrm{A}}_\mu$. The GLLVM is still ill-identified if the rotation is *oblique* (i.e., $\mathbb{M}\mathbb{M}^\intercal = \boldsymbol{\Phi}_{\tilde{z}} \neq \mathbb{I}_q$), as we will have two models with different specifications for the structural model and different factor loadings that generate the same marginal distribution of the observed data. Thus, achieving identification requires not only anchoring the scale of the latent variables but also providing them with an associated 'direction'. This is accomplished by imposing additional constraints on the factor loading matrix A_μ, typically in the form of fixed zero-coefficients in selected positions and sign restrictions.

Before discussing common ways of imposing restrictions on model parameters, it is worth noting the benefits of assuming a multivariate Normal distribution for the latent variables. As Bartholomew et al. (2011) point out, this assumption is mainly for mathematical and computational convenience. By assuming a multivariate Normal distribution for the unrotated latent variables, we ensure that after rotation, the new latent variables follow a distribution $p(\tilde{\mathbf{z}})$ that belongs to the same class as the initial distribution $p(\mathbf{z})$. However, this is only true for the class of spherically symmetric distributions, which includes the (multivariate) Normal distribution (Ali, 1980). According to Maxwell's theorem (see, e.g., Feller, 1966, page 97), we can conclude that if $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \mathbb{I}_q)$, then the rotated factors $\tilde{z} = \mathbb{M}\mathbf{z}$ are also independent and Normally distributed for any orthogonal matrix $\mathbb{M}$. Normality and independence are interdependent, and therefore assuming a Normal distribution for the structural model protects against invariance under rotation.

Fixing rotational indeterminacy

We have demonstrated that the vector of parameters Θ that generates the marginal distribution of the observed data $f(\mathbf{y}; \Theta)$ is not unique due to the rotational indeterminacy described earlier. The orthogonal rotation matrix $\mathbb{M}$ has q^2 elements, and thus, a similar number of restrictions on the parameter vector should be imposed in a structured and consistent way to ensure the identifiability of the GLLVM. These restrictions yield necessary (but not sufficient) conditions for a unique solution for the parameter vector. As described by Anderson and Rubin (1956), two types of sets of (sometimes arbitrary) restrictions can be distinguished, depending on whether the GLLVM is used for exploratory or confirmatory purposes. Exploratory restrictions on A_μ and $\boldsymbol{\Phi}$

have no inherent interpretation, whereas confirmatory restrictions may have intrinsic meaning and are often informed by substantive knowledge or theoretical considerations of the researcher. The most common type of restrictions are factor loadings or latent variable covariances/correlations assumed to be zero, and factor variances assumed to be 1. Equality constraints among subsets of parameters can also be used.

In the LVM literature, common identification restrictions on the factor loading matrix A_μ and the factor covariance matrix $\mathbf{\Phi}$ include:

a) Let $\mathbf{\Phi} = \mathbb{I}_q$; and $\mathrm{A}_\mu^\intercal \mathrm{A}_\mu$ is a diagonal matrix where all its diagonal elements are positive, distinct, and arranged in decreasing order.

b) Let $\mathbf{\Phi} = \mathbb{I}_q$; and partition the factor loading matrix as $\mathrm{A}_\mu = \left[\mathrm{A}_{\mu,1}^\intercal, \mathrm{A}_{\mu,2}^\intercal\right]^\intercal$, where $\mathrm{A}_{\mu,1}$ is a lower triangular matrix with non-zero values in its diagonal.

c) Let $\mathbf{\Phi}$ be a covariance matrix; and partition the factor loading matrix as $\mathrm{A}_\mu = \left[\mathbb{I}_q, \mathrm{A}_{\mu,2}^\intercal\right]^\intercal$.

d) Let $\mathbf{\Phi}$ be a correlation matrix (i.e., $\mathrm{diag}(\mathbf{\Phi}) = \mathbb{I}_q$); and partition the factor loading matrix as $\mathrm{A}_\mu = \left[\mathrm{A}_{\mu,1}^\intercal, \mathrm{A}_{\mu,2}^\intercal\right]^\intercal$, where $\mathrm{A}_{\mu,1}$ is a diagonal with non-negative values in its diagonal.

e) Let A_μ have a '*simple structure*' (Thurstone, 1947, page 335. See also, e.g., Browne, 2001); and impose q normalisation restrictions on either the latent variables (i.e., $\mathbf{\Phi}$ is a correlation matrix) or on the columns of the factor loadings matrices, as well as an ordering of the columns of A_μ.

f) Let $\mathbf{\Phi}$ be a covariance or correlation matrix, with possibly some restricted parameters (correlations to zero, variances to 1); and fix at least $q - 1$ parameters in each column of A_μ to zero, based on substantive knowledge or theory.

Restrictions a)-d) are mathematically equivalent (Sagner, 2019), and they only differ on the interpretation of the model parameters. Restrictions a) and b) are commonly used in exploratory GLLVM, and they do not hold any inherent interpretation. Restriction a) is implicitly used in principal component analysis (PCA). Restriction b) is referred to as a '*recursive rotation*', due to its similarity with a triangular system of simultaneous equations. This is the default identification restriction used in this dissertation when the latent variables are assumed to be independent. Moreover, we implicitly assume independent latent variables. However, there are no reasons in advance to expect latent variables to be independent (Bartholomew et al., 2011, p.214). After imposing appropriate restrictions on A_μ, the assumption of independent latent variables is a matter of convenience to address rotational indeterminacy in the structural part of the model (as presented

in the previous subsection). Yet, in the Social Sciences latent constructs are often correlated. Moreover, letting the LVM to have a structure that allows for the estimation of $\mathbf{\Phi}$ can be useful when determining the existence of redundant factors in the structural model (Brown, 2015).

Restrictions c) and d) allow for better interpretation of the factor loadings and for correlated latent variables. In these cases, we combine elements from exploratory and confirmatory LVMs. Restriction c) is known in the LVM literature as the *error-in-variables* formulation, and it is inspired on the measurement error literature (see, e.g., Fuller, 1987). In restriction c), the first q items are interpreted as noisy measures of the latent variables, and the variances of the latent variables (i.e., the entries in $\text{diag}(\mathbf{\Phi})$) are defined by the metrics of their corresponding item (i.e., the factor loadings for the first q items are fixed to 1). Restriction d) is very similar, but instead of fixing the factor loadings, we assume $\mathbf{\Phi}$ to be a correlation matrix (i.e., $\text{diag}(\mathbf{\Phi}) = 1$). If we further restrict the factor loadings in the diagonal matrix $\mathrm{A}_{\mu,1}$ to be positive, we also address the column sign-flip indeterminacy. Restriction d) is the default restriction used in this dissertation when the latent variables are assumed to be correlated.

Lastly, and restrictions e) and f) are often used in confirmatory GLLVM (Jöreskog, 1969; Brown, 2015), as they usually reflect substantive theory, qualitative prior information, or hypotheses regarding the phenomenon of study. Here, we fix (at least) q^2 parameters either in the factor loading matrices and/or the factor covariance (correlation) matrix to solve the rotational indeterminacy. Restrictions of the type in e) are referred to '*simple structures*'. In this case we fix entries in A_μ to zero such that most items load on *at most* one factor. By doing so, we are effectively imposing a rank restriction on the factor loading matrix. Let $\mathrm{A}_\mu^{(q')}$ be the sub-matrix of A_μ obtained by deleting rows until $\mathrm{A}_\mu^{(q')}$ has only zero elements in the q'^{th} column, for $q' = 1, ..., q$. Formally, A_μ is identified if the sub-matrices $\mathrm{A}_\mu^{(q')}$ are of rank $q - 1$ (see, e.g., Reiersøl, 1950; Anderson and Rubin, 1956 for further details). Under a simple structure restriction, $\mathrm{A}_\mu^\intercal \mathrm{A}_\mu$ is a diagonal matrix. In this sense, if the normalisation is through the latent factors (i.e., $\mathbf{\Phi} = \mathbb{I}_q$), we end up satisfying the same conditions as in restriction a). On the other hand, if the normalisation is through the factor loading matrix, we have that $\mathrm{A}_\mu^\intercal \mathrm{A}_\mu = \mathbb{I}_q$ and also consistent with identification restrictions above. The main limitation of a *simple structure* restriction in the factor loading matrix is that, in practice, it might be unrealistic as items can reflect multiple latent constructs in many applied research areas. In restriction f), it is the researcher who fixes (at least) q^2 entries in A_μ and $\mathbf{\Phi}$ based on substantive theory or prior knowledge. Furthermore, we can address sign-flip if we require the first non-zero element in each column of A_μ to be non-negative.

Note, however, that the restrictions discussed above only identify the model parameters up to column permutation and column sign-flip. As such, imposing any restriction of the type a)-f) on

Θ only ensures local identifiability of the GLLVM.

1.2.3 Empirical identification

Analytical identification strategies are based on the unknown model parameters Θ. Empirical identification is an alternative that is based on the estimated parameters, $\hat{\Theta}$, which are introduced in the next Section. Empirical identification is verified by evaluating the estimated expected information matrix at the estimated solution, $\hat{\mathbb{E}}[\nabla_\Theta \log f(\mathbf{y};\Theta)\nabla_\Theta \log f(\mathbf{y};\Theta)^\intercal]|_{\Theta=\hat{\Theta}}$. A model is empirically (locally) identified for a sample if the estimated expected information matrix is non-singular at the maximum likelihood estimate $\hat{\Theta}$ (McDonald and Krane, 1977). It should be noted that this empirical condition is a counterpart of the identification condition based on the theoretical information matrix (Rothenberg, 1971) mentioned earlier.

According to Skrondal and Rabe-Hesketh (2004, Chapter 5), empirical identification offers several practical advantages over analytical identification. Firstly, the estimated information matrix is a byproduct of maximum likelihood estimation. Secondly, empirical identification is more general as it does not require reduced-form parameters to characterise the marginal distribution of the observed data. Thirdly, empirical identification assesses identification at a specific point of important interest, namely the maximum likelihood estimate. Finally, empirical identification addresses problems that may be specific to the sample used for estimation and inference. For complex models, we recommend verifying empirical identification at the estimated parameters, even after imposing restrictions on the model parameters.

Parameter identification is a fundamental issue in statistical modelling because of its close relationship with the existence of a consistent estimator $\hat{\Theta}$ of the true model parameters $\Theta^* \in \Xi$ generating the marginal distribution of the observed data. Estimation is discussed in the next Section. The following Theorem shows the connection between identification and consistency:

Theorem 1.2.1 (Theorem 2.2.1 in Bekker et al., 1994)**.** *Let $f(\mathbf{y};\Theta)$ be a continuous function of $\Theta \in \Xi$ for all $\mathbf{y} \in \mathbb{R}^p$. The true parameter generating the marginal distribution of the observed data $\Theta^* \in \Xi$ is locally identified if and only if there exists an open neighbourhood $\mathcal{O}_{\Theta^*}$ of Θ^* such that any sequence Θ^e, $e = 1, 2, ...,$ in $\Xi \cap \mathcal{O}_{\Theta^*}$ for which $f(\mathbf{y};\Theta^e) \to f(\mathbf{y};\Theta^*)$, for all $\mathbf{y} \in \mathbb{R}^p$, also satisfies $\Theta^e_{[k]} \to \Theta^*_{[k]}$, for $k = 1, \ldots, K$.*

Proof: See Bekker et al. (1994, page 18). □

1.3. Estimation

In line with standard statistical modelling practice, we draw a random sample of n independent units to estimate the model parameters $\Theta^\intercal = (\Theta_y^\intercal, \Theta_z^\intercal)$ that parameterise the GLLVM, as described by (1.5). Let $\mathbf{y}_m = (y_{1m}, \ldots, y_{pm})^\intercal$ denote the vector of items and $\mathbf{z}_m = (z_{1m}, \ldots, z_{qm})^\intercal$ the vector of latent variables for the m^{th} observation in the sample, $m = 1, \ldots, n$. In the GLLVM framework, the objective function of the estimation problem is the *marginal* likelihood:

$$\mathcal{L}(\Theta; \mathbf{y}) = \prod_{m=1}^{n} \int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \Theta_{y_i}) \right] \, p(\mathbf{z}; \Theta_z) \, \mathrm{d}\mathbf{z}$$

The model parameters characterising the GLLVM include the intercepts and factor loadings in $\Theta_y^\intercal = (\boldsymbol{\alpha}_0^\intercal, \mathrm{vec}(\mathrm{A}_\mu)^\intercal)$, with item specific parameters $\Theta_{y_i} = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{iq})^\intercal$; and the correlations among latent variables in $\Theta_z = \mathrm{vech}(\boldsymbol{\Phi})$, where 'vec' and 'vech' denote the vectorisation and half-vectorisation operators, respectively. The estimation procedure involves finding a vector $\hat{\Theta} \in \Xi$ that yields the largest possible likelihood for a given observed dataset. In practice, the objective function of the estimation problem is the marginal log-likelihood

$$\ell(\Theta; \mathbf{y}) := \log \mathcal{L}(\Theta; \mathbf{y}) = \sum_{m=1}^{n} \log \left[\int_{\mathbb{R}^q} \prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \Theta_{y_i}) \; p(\mathbf{z}; \Theta_z) \; \mathrm{d}\mathbf{z} \right] \tag{1.6}$$

The maximum likelihood estimate (MLE), denoted as $\hat{\Theta}$, is the value of the parameter vector that maximises the *marginal* log-likelihood, i.e.,:

$$\hat{\Theta} = \underset{\Theta \,\in\, \Xi}{\arg\max}\, \ell(\Theta; \mathbf{y})$$

The theory and the associated computational methodologies for the marginal maximum likelihood (MML) estimation of latent factor models with continuous and discrete data are well established in the LVM literature. There are two main MML estimation methods for LVMs: full-information (FI-) and limited-information (LI-) procedures. In FI-MML methods, the model parameters are estimated using all the items and units contributions to the log-likelihood function in (1.6). This approach can be computationally expensive, as it usually involves numerical evaluation of high-dimensional integrals. In LI-MML methods (e.g., Jöreskog and Moustaki, 2001; Katsikatsou et al., 2012), the parameters are estimated only using partial information from the lower-order margins of the observed data (e.g., univariate and/or bivariate distributions of the manifest variables) and therefore, the integrals involved in the estimation process are of lower dimensionality. This

results in lower computational complexity. Both FI-MML and LI-MML estimators are consistent and asymptotically normal. However, FI-MML estimators will be more efficient than LI-MML. The decision regarding the estimation strategy should be made on a case-by-case basis, depending on the characteristics of the problem at hand. We do not intend to present an exhaustive list of references on this aspect, and refer the readers to Chen and Zhang (2021); Chen et al. (in press, 2023) for a comprehensive overview. In this dissertation we focus on the FI-MML estimation, but it is worth mentioning that alternative estimation approaches have been proposed in the LVM literature, such as least-squared-based estimators (e.g., Jöreskog and Golberger, 1972; Browne, 1984), Bayesian estimation methods (e.g., Lee, 2007a), and joint maximum likelihood estimation methods (e.g., Chen et al., 2019).

Computation

Finding the MLE via FI-MML requires solving the score equations $\nabla_\Theta \ell = \mathbf{0}$. However, in most cases, the solution for $\hat{\Theta}$ is not available in a closed form and iterative optimisation algorithms are required to solve the estimation problem. Let $\boldsymbol{\alpha}_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{iq})^\intercal$. The score vector for the intercepts and factor loadings in the measurement equation for item i is given by

$$\begin{aligned} \frac{\partial \ell(\Theta; \mathbf{y})}{\partial \boldsymbol{\alpha}_i} &= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_i} \right] \, p(\mathbf{z} \,|\, \mathbf{y}_m) \, \mathrm{d}\mathbf{z} \\ &= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \frac{1}{\phi_i} \left[y_{im} \frac{\partial \eta_i}{\partial \boldsymbol{\alpha}_i} - \frac{\partial b_i(\eta_i)}{\partial \boldsymbol{\alpha}_i} \right] \, p(\mathbf{z} \,|\, \mathbf{y}_m) \, \mathrm{d}\mathbf{z} \,, \end{aligned} \tag{1.7}$$

which depends on the functional form of b_i in f_i for item i. Let the covariance (correlation) between latent variables j and j' $(j < j')$ be denoted by $\psi_{j,j'} = \boldsymbol{\Phi}_{[j,j']}$, where the sub-script refers to the entry in the $[j, j']^{\text{th}}$ position of $\boldsymbol{\Phi}$. Individual entries in the score vector for $\boldsymbol{\Phi}$ are

$$\begin{aligned} \frac{\partial \ell(\Theta; \mathbf{y})}{\partial \psi_{j,j'}} &= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial \log p(\mathbf{z})}{\partial \psi_{j,j'}} \right] \, p(\mathbf{z} \,|\, \mathbf{y}_m) \, \mathrm{d}\mathbf{z} \\ &= -\frac{n}{2} \mathrm{tr} \left(\boldsymbol{\Phi}^{-1} \mathrm{D}_{jj'} \right) + \frac{1}{2} \sum_{m=1}^{n} \left[\mathrm{tr} \left(\mathrm{G}_{jj'} \mathbb{V}_m \right) + \breve{\mathbf{z}}_m \mathcal{G}_{jj'} \breve{\mathbf{z}}_m \right] \end{aligned} \tag{1.8}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, $\mathrm{D}_{jj'} = \partial \boldsymbol{\Phi} / \partial \psi_{j,j'}$ is a matrix of zeroes except in the $[j, j']^{\text{th}}$ entry, where it takes the value of 1, and $\mathrm{G}_{jj'} = \boldsymbol{\Phi}^{-1} \mathrm{D}_{jj'} \boldsymbol{\Phi}^{-1}$. The conditional mean, $\breve{\mathbf{z}}_m = \mathbb{E}(\mathbf{z} | \mathbf{y}_m)$, and the conditional variance $\mathbb{V}_m = \mathbb{E}\left[(\mathbf{z} - \breve{\mathbf{z}}_m)(\mathbf{z} - \breve{\mathbf{z}}_m)^\intercal \,|\, \mathbf{y}_m \right]$ of the latent variables for unit m in our sample, are computed using the properties of the trace operator and the linearity of the conditional expectation.

We resort to iterative optimisation algorithms to compute the MLE. Popular alternatives for the direct maximisation of the log-likelihood are the gradient descent (GD) or the family of (quasi-)Newton algorithms. These algorithms iteratively search for a new value of the parameter vector in the direction determined by the score vectors, and then update this vector by a magnitude defined by a step size (learning rate), until convergence. In the GD, the step size is a positive number (which can be adaptive), while the (quasi-)Newton algorithms use the observed or expected information (i.e., the matrix of (expected values of) second-order derivatives of the marginal log-likelihood), or an approximation using the score vectors, to determine an optimal step size at each iteration. Some advantages of these algorithms are their fast (super-linear) convergence rate (Broyden et al., 1973) and estimates of the information matrix as a byproduct, which can be later used for the computation of standard errors for the parameter estimates. Possible limitations include sensitivity to starting values and heavy computational operations involving the inversion and computation of the information matrices.

Alternatively, we can treat the latent variables as missing data and then estimate the model parameters using the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977). MML estimation using the EM algorithm was introduced for the case of LVM with binary items in Bock and Aitkin (1981), and later formulated within a FI-MML framework independently by Moustaki (1996); Moustaki and Knott (2000) and Sammel et al. (1997) for items following a distribution from the exponential family. The EM algorithm is an iterative procedure that alternates between the *E-step*, where the approximation to the marginal log-likelihood is computed, and the *M-step*, where the parameters are updated, until convergence.

Let $(\mathbf{y}, \mathbf{z})$ denote the complete data, where the items $\mathbf{y}$ are the observed data and the latent variables $\mathbf{z}$ the missing data. The *complete-data* log-likelihood function is:

$$\begin{aligned} \ell_c(\Theta; \mathbf{y}, \mathbf{z}) &= \sum_{m=1}^{n} \log f(\mathbf{y}_m, \mathbf{z}_m; \Theta) \\ &= \sum_{m=1}^{n} \left[\log f(\mathbf{y}_m \,|\, \mathbf{z}_m; \Theta_y) + \log p(\mathbf{z}_m; \Theta_z)\right] \end{aligned} \tag{1.9}$$

The *E-step* involves computing and approximation function $\mathcal{Q}(\Theta; \Theta^{[t]})$, which is the expected value of the complete-data log-likelihood in (1.9) over the distribution of $\mathbf{z}$ conditional on $\mathbf{y}$, evaluated at the (current) estimates for the model parameters at the t^{th} iteration, $\Theta^{[t]}$:

$$\mathcal{Q}(\Theta; \Theta^{[t]}) = \mathbb{E}_{\mathbf{z} \,|\, \mathbf{y}, \Theta^{[t]}} \left[\, \ell_c(\Theta; \mathbf{y}, \mathbf{z}) \,\right]$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^q} \sum_{m=1}^{n} \left[\log f(\mathbf{y}_m \,|\, \mathbf{z}_m; \Theta_y) + \log p(\mathbf{z}_m; \Theta_z) \right] \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \sum_{i=1}^{p} \log f_i(y_{im} \,|\, \mathbf{z}; \Theta_{y,i}) \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&\quad + \sum_{m=1}^{n} \int_{\mathbb{R}^q} \log p(\mathbf{z}_m; \Theta_z) \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z}
\end{aligned} \tag{1.10}
$$

In the *M-step*, we update the parameter vector as $\Theta^{[t+1]} = \arg\max \mathcal{Q}(\Theta; \Theta^{[t]})$. In practice, it suffices to find a solution that results in an increase of the approximation function, i.e., $\mathcal{Q}(\Theta^{[t+1]}; \Theta^{[t]}) \geq \mathcal{Q}(\Theta^{[t]}; \Theta^{[t]})$. The optimisation problem involved in the M-step can be solved using any of the gradient-based algorithms described above. However, the observed and expected information matrices are easier to compute for the approximation function $\mathcal{Q}(\Theta; \Theta^{[t]})$ than for the marginal log-likelihood $\ell(\Theta; \mathbf{y})$, and thus a Newton-Raphson update scheme is usually preferred. The *E-step* and *M-step* are repeated until convergence. The score vectors for the approximation function in the EM-algorithm and the marginal log-likelihood are equivalent (Louis, 1982), and thus the solutions obtained through the EM-algorithm or the direct optimisation are the same (provided we use the same starting point). The main advantage of the EM-algorithm is its relatively simple implementation. However, due to its (sub-)linear convergence rate (McLachlan and Krishnan, 2008), the EM-algorithm can be slow to reach the mode.

Lastly, it is common practice to perform rotations on the estimated factor loading matrix to obtain more interpretable and/or sparse solutions (see, e.g., Browne, 2001; Jennrich, 2001, 2002, 2004, 2006, 2007; Liu et al., 2023, for a relevant and thorough explanation on rotation techniques).

1.4. Factor Scores

Factor scoring, also known as scoring, involves assigning values to the latent variables for individuals in a sample. In many applied contexts, such as educational and psychological testing, scoring is the primary objective of latent variable modelling. In this dissertation, we assume that the latent variables are continuous and normally distributed, $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Phi})$, and therefore we expect the factor scores to have similar characteristics.

In factor scoring, the parameter estimates $\hat{\Theta}$ are treated as known. As discussed in Skrondal and Rabe-Hesketh (2004, Chapter 7) and Bartholomew et al. (2011, Chapters 2 and 4), all the information about the latent variables is contained in the posterior distribution of the latent variables, which is the distribution of the latent variables conditional on the observed data and

evaluated at the maximum likelihood estimate, $p(\mathbf{z}\,|\,\mathbf{y};\hat{\Theta})$. Following Bayes' theorem, this posterior distribution is

$$p(\mathbf{z}\,|\,\mathbf{y};\hat{\Theta}) = \frac{f(\mathbf{y}\,|\,\mathbf{z};\hat{\Theta}_y)\; p(\mathbf{z};\hat{\Theta}_z)}{\int_{\mathbb{R}^q} f(\mathbf{y}\,|\,\mathbf{z}';\hat{\Theta}_y)\; p(\mathbf{z}';\hat{\Theta}_z)\; \mathrm{d}\mathbf{z}'} \tag{1.11}$$

From (1.11) above, we can obtain information about the factors $\mathbf{z}_m$ for individuals $m = 1, \ldots, n$ in our sample. A popular scoring method is the *empirical Bayes* (EB) approach, which allows us to compute factor scores for unit $m = 1, ..., n$ as follows:

$$\tilde{\mathbf{z}}_m^{\mathrm{EB}} = \mathbb{E}(\mathbf{z}\,|\,\mathbf{y}_m;\hat{\Theta}) = \int_{\mathbb{R}^q} \mathbf{z} \cdot p(\mathbf{z}\,|\,\mathbf{y}_m;\hat{\Theta})\; \mathrm{d}\mathbf{z}$$

For the classical Normal linear factor model, the EB scores $\tilde{\mathbf{z}}_m^{\mathrm{EB}}$ can be obtained analytically by using a linear combination of the manifest item values in $\mathbf{y}_m$ (see Skrondal and Rabe-Hesketh (2004, Section 7.3) for a detailed explanation). This approach is often referred to as the 'regression method' for factor scoring. However, when dealing with non-normal items, numerical integration methods are necessary to compute the conditional expected value.

An alternative to using the posterior mean as the factor scores, as in EB, is to use the (log-) posterior mode instead. The *maximum a-posteriori* (MAP) factor scores for unit $m = 1, ..., n$ are given by:

$$\tilde{\mathbf{z}}_m^{\mathrm{MAP}} = \underset{\mathbf{z}\,\in\,\mathbb{R}^q}{\arg\max}\; \log p(\mathbf{z}\,|\,\mathbf{y}_m;\hat{\Theta})$$

Generally, closed form solutions for the MAP scoring problem are not available, and iterative methods must be used to solve the optimisation problem. The MAP scores, denoted by $\tilde{\mathbf{z}}_m^{\mathrm{MAP}}$, are obtained by solving the following set of equations for $m = 1, \ldots, n$:

$$\frac{\partial}{\partial \mathbf{z}} \log p(\mathbf{z}\,|\,\mathbf{y}_m;\hat{\Theta}) = \frac{\partial}{\partial \mathbf{z}} \log p(\mathbf{z};\hat{\Theta}_z) + \sum_{i=1}^{p} \frac{\partial}{\partial \mathbf{z}} \log f_i(y_{im}\,|\,\mathbf{z};\hat{\Theta}_{y,i}) = \mathbf{0}.$$

Compared to the EB scoring method, the MAP scoring method does not require numerical integration.

Some scoring methods rely on simple aggregates of the responses from the test and are considered more heuristic in nature. An example is the *sum score* (also known as raw score) method, which assumes that all manifest variables measure the same latent construct similarly (with some measurement error), or, in other words, that all factor loadings are equal, as in congeneric tests (see, e.g., Jöreskog, 1971). The sum score for factor $j = 1, \ldots, q$ in unit $m = 1, \ldots, n$ can be calculated as $\tilde{z}_{jm}^{\mathrm{SS1}} = \sum_{i=1}^{p} y_{im}$, where y_{im} is the response of unit m to item i. Another version of the sum

score uses the estimated factor loadings, denoted by $\hat{\alpha}_{ij}$, and is defined as $\tilde{z}_{jm}^{\text{SS2}} = \sum_{i=1}^{p} \hat{\alpha}_{ij} y_{im}$ for binary items and $\tilde{z}_{jm}^{\text{SS2}} = \sum_{i=1}^{p} (\hat{\alpha}_{ij} y_{im})/\hat{\sigma}_i$ for continuous items, where $\hat{\sigma}_i$ is the estimated standard deviation of item i. For distributions in the exponential family, $\tilde{z}_{jm}^{\text{CS2}}$ corresponds to a minimal sufficient statistic of the latent variable (in a Bayesian sense), assuming a linear relationship between the latent variables and the observed variables in the measurement equations (see, e.g., Bartholomew, 1984; Bartholomew et al., 2011, Sections 2.5 and 2.15). Sum scores are attractive because they are easy to compute and account for the relative contribution of each item to the common factor.

In the unidimensional case, the scoring methods discussed above generally produce similar results and will usually rank individuals with the same response patterns in the same order. Knott and Albanese (1993) provide a general framework for scoring binary items and show that, under certain conditions, the methods above are equivalent. Moustaki and Knott (2000) extended this framework to the GLLVM with items in the exponential family. For more information on this equivalence, we refer readers to the aforementioned references and to Bartholomew et al. (2011, Section 2.15). Regardless of the scoring method used, the measurement model $f(\mathbf{y} \mid \mathbf{z}; \hat{\Theta}_y)$ plays a central role in the scoring process.

1.5. Summary and Outline of this dissertation

In this chapter, we have provided a comprehensive overview of the Generalised Linear Latent Variable Model (GLLVM) framework proposed by Bartholomew et al. (2011) and Skrondal and Rabe-Hesketh (2004). We began by introducing the model structure and the fundamental concepts of latent variable models in Section 1.1. The GLLVM assumes that observed variables follow distributions from the exponential family, given the latent variables. The relationship between the observed and latent variables is linear, up to a conditional mean, where the mean of the items results from a linear combination of the latent variables through a possibly nonlinear transformation. To simplify the model, we assume that the latent variables follow a Normal distribution. However, this assumption can be relaxed to allow for discrete latent variables (as in LCA), or for an explicit modelling of the relationships between latent factors (as in SEM).

In Section 1.2, we provide a comprehensive discussion of model identification. In general, a model is considered identified if two distinct parameter values produce unique marginal probability distributions for any given observed data. For LVMs, identifying a model often requires imposing restrictions on the model parameters, such as fixing factor loadings and/or factor covariances/-correlations. We outline several strategies for imposing these restrictions in a structured manner.

Achieving model identification is crucial for ensuring the consistency of the maximum likelihood estimator, which we introduce in Section 1.3. Here, we discuss common computational techniques used in the LVM literature, including the EM algorithm, which is an iterative approach treating the latent variables as missing data, and direct optimization of the marginal log-likelihood using (quasi-)Newton solvers. While both methods yield an equivalent maximum likelihood solution, their implementation and convergence speeds differ. Finally, in Section 1.4, we delve into factor scoring, the process of assigning values to the latent variables for the observations in our sample.

In Chapter 2, we address the limitations of the GLLVM framework, specifically the assumption of items following a distribution from the exponential family, which can be restrictive in many applications. To overcome this, we propose a new class of Latent Variable Models for Location, Scale, and Shape parameters (GLVM-LSS). This model employs linear functions of latent variables to model the location, scale, and shape parameters of observed items' conditional distributions, enabling a more effective way to model the mean, variance, skewness, and kurtosis of items in terms of the latent variables. However, the increased flexibility of this model also results in more complexity, making it difficult to interpret. In Chapter 3, we introduce a penalised maximum likelihood estimation of the GLVM-LSS to obtain factor loadings matrices that are sparse and easier to interpret. We also discuss an automatic, data-driven method for selecting the tuning parameters that determine the amount of penalisation, which simplifies the estimation process and avoids computationally intensive techniques such as grid-search or cross-validation. Lastly, in Chapter 4, we conclude by suggesting future research directions.

Chapter 2

Generalised Latent Variable Models for Location, Scale, and Shape parameters

In this Chapter, we introduce a novel class of Generalised Latent Variable Models for Location, Scale, and Shape parameters (GLVM-LSS). These models use linear functions of latent factors to model the location, scale, and shape parameters of the items' conditional distributions. By doing so, we model higher order moments such as variance, skewness, and kurtosis in terms of the latent variables, providing a more flexible framework compared to classical factor models. The model parameters are estimated using maximum likelihood estimation. Our findings suggest that modelling the entire distribution of items, not just the conditional mean, leads to improved model fit and deeper insights into how the items reflect the latent constructs they are intended to measure. To assess the performance of the proposed methods, we conduct extensive simulation studies and apply it to real-world data from educational testing and public opinion research. The results highlight the efficacy of the GLVM-LSS framework in capturing complex relationships between observed variables and latent factors, providing valuable insights for researchers in various fields.

2.1. Introduction

Latent Variable Models (LVM) are widely used in social, behavioural, and educational sciences to measure unobserved constructs and reduce dimensionality. These models explain the associations between a set of observed variables (also known as *items*) through a much smaller set of

latent variables (also known as *factors*). Many probabilistic, parametric LVMs fall within the Generalised Linear Latent Variable Model framework (GLLVM, Skrondal and Rabe-Hesketh, 2004; Bartholomew et al., 2011), as discussed in Section 1.1. In the GLLVM, we assume that i) conditional on the latent variables, items follow a distribution from the exponential family, and ii) the relationship between an item and the latent variables is linear through the (conditional) mean, treating other distributional parameters as nuisance parameters. For further details on estimation and applications of GLLVMs, we refer the readers to Moustaki and Knott (2000); Skrondal and Rabe-Hesketh (2004); Bartholomew et al. (2011).

To some extent, the GLLVM can be seen as a case of independent and simultaneous generalised linear models (GLM, McCullagh and Nelder, 1989) where the conditional mean of the outcome variables is modelled in terms of latent predictors. However, in many cases, we are interested not only in the conditional mean, but also in how higher order moments of the outcome variable (such as variance, skewness, and kurtosis) relate to the predictors. In regression analysis, where all variables are observed, there has been a growing interest in modelling the entire conditional distribution of the response variable given one or more predictors, rather than just the conditional mean. This approach leads to a more flexible modelling framework and a more comprehensive understanding of the relationship between the response variable and the predictors.

A popular distributional regression framework is the Generalised Additive Model for Location, Scale and Shape (GAMLSS) Rigby and Stasinopoulos, 2005; Klein et al., 2015; Umlauf et al., 2018). In the GAMLSS framework, we model the conditional distribution of the outcome variable (given the covariates) by expressing the distributional parameters that characterise such conditional distribution as functions of the explanatory variables. GAMLSS offers several advantages over traditional GLMs, including greater flexibility and less restrictive assumptions on the response variable's distribution, better estimation of the relationships between dependent and independent variables, modelling of extreme events and outcomes, and improved forecasting. Moreover, in GAMLSS the likelihood is available due to the parametric distributional assumption, enabling likelihood-based inferences on the parameter estimates. The GAMLSS framework is especially useful when covariate effects on higher order moments are of substantive interest. For a comprehensive review of this topic, see Stasinopoulos et al. (2017); Kneib (2013); Kneib et al. (2023); Fahrmeir et al. (2021).

2.1.1 Motivation: A distributional approach to LVM

A distributional approach to latent variable modelling offers numerous potential advantages over the classic GLLVM. We present some examples where such an approach would be beneficial.

Going beyond the exponential family assumption: The most relevant application of a distributional approach to LVM is when real-world data do not meet the distributional assumptions required by the GLLVM framework. Various issues related to this have been addressed separately and independently in the LVM literature by many authors who have proposed methodological contributions with varying estimation and inference approaches. For example, continuous items often exhibit heteroscedasticity (Meijer and Mooijaart, 1996; Lewin-Koh and Amemiya, 2003; Hessen and Dolan, 2009), skewness (Montanari and Viroli, 2010; Lin et al., 2015), and/or kurtosis (Asparouhov and Muthén, 2016); discrete, count, and bounded continuous (e.g., in the unit interval) items often display zero/one/maximum value inflation and/or heaping due to respondents rounding their numerical answers to the nearest five or ten (Wang, 2010; Wall et al., 2015b; Magnus and Thissen, 2017; Molenaar et al., 2022), and data can be censored or truncated (Moustaki and Steele, 2005). Although limited information and robust estimation methods exist to control for deviations from the distributional assumptions under the GLLVM framework (e.g., Browne, 1984; Bollen, 1996; Moustaki and Victoria-Feser, 2006), these features of the items are crucial and ignoring them can result in underestimated standard errors or biased parameter estimates. Despite being an active line of research in the LVM literature, these contributions are not part of a general parametric-based distributional LVM framework, but rather independent models that address separate problems in applied research.

Substantive interest in higher order moments: A distributional approach to latent variable modelling may also be of interest when examining the effects of latent variables on higher order moments of items. For example, in studies involving clustered data or repeated measures obtained through experience sampling methods, researchers focus not only on the conditional mean of an item, but also on how its variance is related to the level of a latent factor. This is particularly common in psychopathology studies, where researchers are interested in the intra-individual variability or stability of emotional responses, in addition to deviations from an individual's baseline mood, to have a more nuanced understanding of the underlying phenomena. Examples of works in this area include Hedecker et al. (2006, 2008, 2012), and Wang et al. (2012).

Assessing item quality: Distributional LVMs could also be applied in the item quality control literature. In this context, discrimination refers to an item's ability to distinguish between individuals with different levels of the latent construct being measured, and it is crucial when assessing the psychometric properties of an item or a test. Items with low discrimination power may not be contributing to the test's accuracy and may need to be revised or removed. The slope coefficient in the equation describing the linear relationship between the latent factor and the item observed score is often used to assess item discrimination. However, poor discrimination can occur if the linear relationship does not hold, or if the item exhibits conditional heteroscedasticity or skewness. In the latter case, different values of the conditional variance or higher order moments along the latent scale can lower the item's discrimination power (Hessen and Dolan, 2009). Heteroscedastic items, for example, will have respondents with different levels of the latent factor responding similar values for the item score. It is therefore important to take into account such item characteristics when evaluating its quality.

While distributional modelling techniques have the potential to yield numerous advantages, they have not been widely explored within the LVM framework. Some authors have proposed *quantile-based* factor models (see e.g., Sagner, 2019; Chen et al., 2021). However, to the best of our knowledge, there is currently no unified *parametric-based* distributional framework for LVM analysis. The existing developments consist of independent, disconnected methods with no common estimation or inferential frameworks. In response, this paper introduces a flexible class of parametric distributional latent variable models to fill that gap. Specifically, we present a Generalised Latent Variable model for Location, Scale, and Shape parameters (GLVM-LSS), which adapts the GAMLSS distributional regression framework (Rigby and Stasinopoulos, 2005; Klein et al., 2015) to models with latent variables. In the GLVM-LSS, the location, scale, and shape parameters of the items' conditional distributions are assumed to be linear functions of latent variables. This allows the modelling of the entire conditional distribution of the item, including the mean and higher order moments, in terms of the latent variables. To an extent, the GLVM-LSS serves as an umbrella class of LVMs that includes previous works as particular cases. We present the GLVM-LSS model in detail in Section 2.2, followed by a discussion of the full-information marginal maximum likelihood estimation procedure in Section 2.3. We conduct simulation studies in Section 2.4 to demonstrate the properties of the proposed method under finite sample settings and then, in Section 2.5, apply it to real-world settings using educational data from the PISA 2018 mathematics exam and data on public opinion research from the American National Election Study 2020.

2.2. Latent Variable Models for Location, Scale and Shape parameters (GLVM-LSS)

Let $\mathbf{y} = (y_1, ..., y_p)^\intercal \in \mathbb{R}^p$ be a vector of items and $\mathbf{z} = (z_1, ..., z_q)^\intercal \in \mathbb{R}^q$ a vector of latent variables, with $q \ll p$. The density for the observed data is written as $f(\mathbf{y}) = \int_{\mathbb{R}_q} f(\mathbf{y} \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}$, where the conditional distribution $f(\mathbf{y} \,|\, \mathbf{z})$, also known as the measurement component of the LVM, describes the relationship between $\mathbf{y}$ and $\mathbf{z}$; while $p(\mathbf{z})$, the structural component, specifies the relationships among the latent variables. In the LVM framework, the correlations among the manifest variables are fully accounted for by the latent variables, and thus we assume the items are conditionally independent given $\mathbf{z}$. We further assume that the conditional distributions of the items follow a known parametric form indexed by a (vector of) distributional parameter(s) $\boldsymbol{\theta}_i$. For the structural model, we adopt the multivariate Normal distribution, $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Phi})$, which is a common choice in the LVM literature due to its mathematical and computational convenience (see, e.g., Bartholomew et al., 2011, Chapter 2). However, this assumption can be relaxed to allow for greater flexibility in the structural component of the LVM, as demonstrated by several studies (e.g., Woods and Thissen, 2006; Ma and Genton, 2010; Irincheeva et al., 2012; Wall et al., 2015a).

The marginal density of the observed variables can be written as

$$f(\mathbf{y}) = \int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_i \,|\, \mathbf{z}; \boldsymbol{\theta}_i) \right] p(\mathbf{z}; \boldsymbol{\Phi}) \, \mathrm{d}\mathbf{z} \tag{2.1}$$

The GLLVM assumes that the conditional distribution of each item $i = 1, ..., p$ is from the exponential family (although not necessarily the same distribution for all items). Specifically, we have $f_i(y_i \mid \mathbf{z}) = \exp\left\{\phi_i^{-1}[y_i\zeta_i - b_i(\zeta_i)] + c_i(y_i; \phi_i)\right\}$, where ζ_i is the canonical parameter, ϕ_i is a dispersion parameter, and b_i and c_i are pre-specified distribution-specific functions. In the exponential family case, we can write $\boldsymbol{\theta}_i = (\zeta_i, \phi_i)^\intercal$. For some distributions in the exponential family, the scale parameter is $\phi_i = 1$, and it is only of interest in the continuous case.

Denote the linear combination of the latent variables, also referred to as the systematic component or measurement equation, as

$$\eta_i(\mathbf{z}) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} z_j \tag{2.2}$$

In the above, we use the notation $\eta_i(\mathbf{z})$ to emphasise the functional dependence on the latent variables. We will drop the dependence on $\mathbf{z}$ for notational convenience, but it should be clear

from context. In the GLLVM, the relationship between the items and the latent variables is linear through the canonical parameter: $\zeta_i = \eta_i$. The focus in the GLLVM is on modelling the conditional mean of the items as linear functions of the latent variables, that is, $\mathbb{E}(y_i \,|\, \mathbf{z}) = b_i'(\eta_i)$. The mapping $\upsilon_i^{-1} := b_i'$ is known as the (canonical) link function, a monotonic differentiable function that connects the conditional mean of y_i with the latent variables. Note that the dispersion parameter does not depend on $\mathbf{z}$ (see Moustaki and Knott (2000); Skrondal and Rabe-Hesketh (2004) for more details).

In our proposed framework, we relax the exponential family assumption in the measurement part of the LVM, allowing $f_i(y_i|\mathbf{z};\boldsymbol{\theta}_i)$ to be any parametric distribution indexed by a D-dimensional vector of distributional parameters $\boldsymbol{\theta}_i = (\theta_i^{(1)}, ..., \theta_i^{(D)})^\intercal$. Additionally, we express the distributional parameters $\theta_i^{(d)}$ as linear functions of the latent variables. We use the sub-index (i, θ_d) to indicate that the corresponding function or regression parameter is defined for $\theta_i^{(d)} \in \boldsymbol{\theta}_i$. The measurement equation for $\theta_i^{(d)}$ is then given by:

$$\upsilon_{i,\theta_d}(\theta_i^{(d)}) = \eta_{i,\theta_d} := \alpha_{i0,\theta_d} + \sum_{j=1}^{q} \alpha_{ij,\theta_d} z_j, \quad \text{for} \quad i = 1, ..., p, \quad \text{and} \quad d = 1, ..., D; \tag{2.3}$$

Here, υ_{i,θ_d} is a distributional parameter-specific link function (e.g., identity, log, logit, etc.) chosen to ensure appropriate restrictions on the distributional parameters. As before, η_{i,θ_d} is defined as a linear combination of latent variables. Coefficients α_{i0,θ_d} are intercepts, and slopes (also called factor loadings) in (2.3) form the q-dimensional vector $\boldsymbol{\alpha}_{i,\theta_d} = (\alpha_{i1,\theta_d}, ..., \alpha_{iq,\theta_d})^\intercal$. While this measurement equation can be extended to include higher order polynomials (e.g., McDonald, 1967; Yalcin and Amemiya, 2001) or interaction and non-linear terms (Rizopoulos and Moustaki, 2008), we leave these for future consideration.

The distributional parameters in $\boldsymbol{\theta}_i$ can be categorised as location, scale, or shape parameters, depending on their role in the parameterization of f_i. Location parameters generally relate to the first-order moment of the distribution, scale parameters to the second-order moment, and shape parameters (if any) to higher-order moments[1]. To simplify notation, we denote the location parameter as $\mu := \theta_i^{(1)}$, the scale parameter as $\sigma_i := \theta_i^{(2)}$, and the shape parameters as $\nu_i := \theta_i^{(3)}$ and $\tau_i := \theta_i^{(4)}$. For many families of distributions, a maximum of four parameters (one location, one scale, and two shape parameters) is sufficient, but the model can be extended to include cases

[1]Some distributions might have alternative parametrisations with different definitions for the scale and shape parameters (see, e.g., Yee (2020) for an example of different parametrisations of the negative binomial distribution). It is recommended to work with a parameterisation that allows for easy and direct interpretation of the effects of the explanatory variables (in this case, the latent variables), where the scale and shape parameters have direct relationship with the conditional variance, skewness, and kurtosis of the observed variable.

with multiple location, scale, or shape parameters. For the remainder of the paper, we refer to the distributional parameter vector as $\boldsymbol{\theta}_i = (\mu_i, \sigma_i, \nu_i, \tau_i)^\intercal$, and use $\varphi_i \in \boldsymbol{\theta}_i$ to denote any location, scale, or shape parameter in the (conditional) distribution of item i.

Matrix notation can be convenient in some cases. Let $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_p)^\intercal$ be the vector of the same distributional parameter φ for all items. Denote $\boldsymbol{\alpha}_{0,\varphi} = (\alpha_{10,\varphi}, ..., \alpha_{p0,\varphi})^\intercal$ as a vector of intercept terms, and let A_φ be a $(q \times p)$ factor loadings matrix with rows corresponding to the vectors $\boldsymbol{\alpha}_{i,\varphi}$. Finally, let υ_φ be the vector function that applies the corresponding link function $\upsilon_{i,\varphi}$ to each entry of $\boldsymbol{\varphi}$. With this notation, we can describe the set of measurement equations for a distributional parameter φ as $\upsilon_\varphi(\boldsymbol{\varphi}) = \boldsymbol{\alpha}_{0,\varphi} + \mathrm{A}_\varphi \mathbf{z}$.

To further simplify notation, we can write a vector of parameters $\boldsymbol{\theta}^\intercal = (\boldsymbol{\mu}^\intercal, \boldsymbol{\sigma}^\intercal, \boldsymbol{\nu}^\intercal, \boldsymbol{\tau}^\intercal)$, a vector of intercepts $\boldsymbol{\alpha}_0^\intercal = (\boldsymbol{\alpha}_{0,\mu}^\intercal, \boldsymbol{\alpha}_{0,\sigma}^\intercal, \boldsymbol{\alpha}_{0,\nu}^\intercal, \boldsymbol{\alpha}_{0,\tau}^\intercal)$, and a factor loading matrix $\mathrm{A}^\intercal = [\mathrm{A}_\mu^\intercal, \mathrm{A}_\sigma^\intercal, \mathrm{A}_\nu^\intercal, \mathrm{A}_\tau^\intercal]$. Then, we can compactly write the measurement equations of a GLVM-LSS model as

$$\upsilon(\boldsymbol{\theta}) = \boldsymbol{\alpha}_0 + \mathrm{A}\mathbf{z} \tag{2.4}$$

The GLVM-LSS is the result of combining the model in (2.1) with the measurement equations for the distributional parameters in (2.4). One can easily see that many cases of the GLLVM can be derived by modelling the location parameter as a linear function of the latent variables while assuming the scale parameters are constant. This approach is also applicable to many models proposed in the LVM for specific applications. In the following section, we provide examples of how the GLVM-LSS can be used to accommodate these cases and, more importantly, extend them.

2.2.1 Some examples of GLVM-LSS

Example 1. Heteroscedastic Linear Factor Models: Heteroscedasticity can be of substantive interest. An example is the hypothesis of 'ability differentiation', which suggests that the strength of correlations between items varies with the level of latent ability (Detterman and Daniel, 1989; Deary et al., 1996). This phenomenon has been studied (e.g., Tucker-Drob, 2009; Molenaar et al., 2011), and has been linked to heteroscedastic errors in the factor model (Hessen and Dolan, 2009). Item quality control is another scenario where heteroscedasticity is important, as we expect the precision of an item to be independent of an individual's ability.

The heteroscedastic factor model (Meijer and Mooijaart, 1996; Lewin-Koh and Amemiya, 2003; Hessen and Dolan, 2009) assumes Normal distributions for items $i = 1, ..., p$, with distributional parameters $\boldsymbol{\theta}_i = (\mu_i, \sigma_i)^\intercal$ as functions of the latent variables, $y_i \,|\, \mathbf{z} \sim \mathbb{N}(\mu_i(\mathbf{z}), \sigma_i^2(\mathbf{z}))$. By choosing

the identity and the log link functions, the measurement equations are given by

$$\mu_i(\mathbf{z}) = \alpha_{i0,\mu} + \sum_{j=1}^{q} \alpha_{ij,\mu} z_j \ ,$$
$$\log(\sigma_i(\mathbf{z})) = \alpha_{i0,\sigma} + \sum_{j=1}^{q} \alpha_{ij,\sigma} z_j.$$

Thus, the GLVM-LSS framework extends existing heteroscedastic factor models in the LVM literature to accommodate multiple latent variables.

Example 2. Skew-Normal Linear Factor Models: The Skew-Normal (SN) distribution (Azzalini, 1985; Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1999) has received recent attention in the LVM literature. While multivariate SN factor models have been proposed previously (Montanari and Viroli, 2010; Liu and Lin, 2015; Asparouhov and Muthén, 2016), existing approaches only model the location parameter in terms of the latent variables. However, in the GLVM-LSS framework, we can extend the model to include the scale and shape parameters of the SN distribution. In addition, the SN distribution has been used to model skewed latent variables with both continuous and categorical items (Molenaar et al., 2010, 2011; Molenaar, 2015; Molenaar and Bolsinova, 2017).

Assume that items $i = 1, ..., p$ follow a conditional SN distribution, $y_i|\mathbf{z} \sim \text{SN}(\mu_i(\mathbf{z}), \sigma_i^2(\mathbf{z}), \nu_i(\mathbf{z}))$, in which the location ($\mu_i \in \mathbb{R}$), scale ($\sigma_i \in \mathbb{R}^+$), and shape ($\nu_i \in (0, 1)$) parameters are linear functions of the latent variables. The measurement equations for the distributional parameters $\boldsymbol{\theta}_i = (\mu_i, \sigma_i, \nu_i)^\intercal$ are given by

$$\mu_i(\mathbf{z}) = \alpha_{i0,\mu} + \sum_{j=1}^{q} \alpha_{ij,\mu} z_j \ ,$$
$$\log(\sigma_i(\mathbf{z})) = \alpha_{i0,\sigma} + \sum_{j=1}^{q} \alpha_{ij,\sigma} z_j \ ,$$
$$\text{logit}(\nu_i(\mathbf{z})) = \alpha_{i0,\nu} + \sum_{j=1}^{q} \alpha_{ij,\nu} z_j.$$

We use an alternative parameterisation of the SN distribution that is similar to the one introduced in Azzalini (1985). We refer the readers to Appendix A1.1 for further details.

Example 3. Zero-Inflated Poisson Factor Models: Excess zero responses (zero-inflation, ZI) are very common in count data collected for research in psychology and ecology. Factor models

for zero-inflated count data in the LVM literature can be classified into two classes. The first approach considers ZI as the result of having a non-susceptible or non-pathological sub-population (latent class) of individuals who respond with zero to all items (e.g., Wall et al., 2015b; Magnus and Thissen, 2017 for count data; Magnus and Liu, 2022 for ordinal items). The second approach, which we follow in this paper, is more flexible and considers ZI to be a phenomenon observed at the item level (e.g., Wang, 2010; Niku et al., 2017, 2019 for count data; Magnus and Garnier-Villarreal, 2021 for ordinal items). Regardless of the approach, ignoring ZI can lead to biased parameter estimates when fitting LVMs (Wall et al., 2015b; Magnus and Garnier-Villarreal, 2021).

In the GLVM-LSS framework, factor models for count data with excess zeros result from assuming a Zero-Inflated Poisson (ZIP) distribution for the observed items, $y_i \,|\, \mathbf{z} \sim \text{ZIP}(\lambda_i(\mathbf{z}), \pi_i(\mathbf{z}))$. Here, both the probability of a person responding a 'structural zero' for item i (π_i), and the rate parameter in the Poisson distribution (λ_i) are functions of the latent variables. Further details are found in Appendix A1.2. The measurement equations for the distributional parameters $\boldsymbol{\theta}_i = (\lambda_i, \pi_i)^\intercal$ are

$$\log(\lambda_i(\mathbf{z})) = \alpha_{i0,\lambda} + \sum_{j=1}^{q} \alpha_{ij,\lambda} z_j \,,$$

$$\text{logit}(\pi_i(\mathbf{z})) = \alpha_{i0,\pi} + \sum_{j=1}^{q} \alpha_{ij,\pi} z_j .$$

The GLVM-LSS resulting from combining the model in (2.1) and the measurement equations above is a generalisation of Wang (2010) to the multidimensional factor case. This framework can be extended to model items with excess response in other values (e.g., zero-one-inflation in Molenaar et al., 2022) or heaping around particular values of the response scale (Wall et al., 2015b). However, these examples are left for future implementation, and are not part of this Chapter.

Example 4. Beta Factor Models: Continuous data that are measured in the $(0, 1)$ interval, such as proportions, continuous response format (CRF) items, scaled Likert items, and visual analogue response scales, are of interest to quantitative social scientists. The Beta distribution is a natural distribution to model these types of items. Beta factor models have been proposed in the literature to model these items, where only the conditional mean of the items is modelled in terms of a unidimensional latent factor (Noel and Dauvier, 2007; Noel, 2014; Revuelta et al., 2022).

In this paper, we follow the location-scale parameterization of the Beta distribution proposed by Rigby et al. (2020) (see Appendix A1.1 for further details). Conditional on the latent vari-

ables, we assume that the items follow a Beta distribution, $y_i \,|\, \mathbf{z} \sim \text{Beta}(\mu_i(\mathbf{z}), \sigma_i(\mathbf{z}))$. Under this parameterisation, $\mathbb{E}(y_i \,|\, \mathbf{z}) = \mu_i$, and $\text{Var}(y_i \,|\, \mathbf{z}) = \sigma_i^2 \mu_i(1-\mu_i)$, which means that $\mu_i \in (0,1)$ acts as the location parameter, and $\sigma_i \in (0,1)$ as the scale parameter. The distributional parameters $\boldsymbol{\theta}_i = (\mu_i, \sigma_i)^\intercal$ are functions of the latent variables, with measurement equations given by

$$\text{logit}(\mu_i(\mathbf{z})) = \alpha_{i0,\mu} + \sum_{j=1}^{q} \alpha_{ij,\mu} z_j \ ,$$

$$\text{logit}(\sigma_i(\mathbf{z})) = \alpha_{i0,\sigma} + \sum_{j=1}^{q} \alpha_{ij,\sigma} z_j .$$

This GLVM-LSS model extends existing models in the Beta factor models literature in two ways. First, it allows for modelling heteroscedastic items in the $(0,1)$ interval, which is not commonly done in the literature. An exception is Verkuilen and Smithson (2012), where the authors consider random effects on the scale parameter in a Beta mixed-model. Modelling higher-order moments of Beta-distributed items is also of interest given the reasons discussed above for continuous items in the real line. Second, it generalises to the multidimensional factor case.

To summarise, the GLVM-LSS encompasses different LVMs proposed in the literature to address specific types of items or violations to the distributional assumptions in the GLLVM framework. The GLVM-LSS provides greater flexibility to applied researchers by allowing them to assume any parametric distribution for the items. The GLVM-LSS has close connections and shares similar a parameterisation with well-established regression-type models for observed variables, such as the Generalised Additive Model for Location, Scale, and Shape (Rigby and Stasinopoulos, 2005; Klein et al., 2015), and Vector Generalised Additive Models (Yee and Wild, 1996; Yee, 2015).

In the following section, we present a unified maximum marginal likelihood estimation framework, discuss inferential aspects of the GLVM-LSS, identify restrictions to avoid the rotational indeterminacy of LVMs, and introduce model selection and comparison criteria.

Model Identification

As discussed in Section 1.2, LVMs are not identified, partially because of the arbitrariness of the location and scale of the latent variables. The same is true for the GLVM-LSS framework proposed in this chapter. Assessing analytical identification for the GLVM-LSS context is a challenging task since there are few simple GLVM-LSS models with marginal log-likelihoods entirely characterised by reduced-form parameters. Additionally, analytical identification is possible only on a case-by-case basis, which precludes establishing general identification rules for the GLVM-LSS. As a result,

we rely on general rules involving restrictions on the model parameters to solve for the rotational indeterminacy. It is worth noting that these restrictions are necessary but not sufficient conditions for local identification of the model parameters, emphasising the importance of future research to establish stronger and more general conditions for the identifiability of complex models such as the GLVM-LSS.

In the unidimensional case ($q = 1$), we can assume that z_1 follows a standard Normal distribution ($z_1 \sim \mathbb{N}(0, 1)$) to address the rotational indeterminacy. However, if there is a substantial interest in estimating the variance of the latent variable, we can assume that z_1 follows a Normal distribution with variance ψ_z, which is estimated freely. In this case, we need to fix the factor loading in the measurement equation for the location parameter of the first item to be equal to one ($\alpha_{11,\mu} = 1$).

In the case of multiple latent variables ($q > 1$), the factor loadings matrix in (2.4) is not identifiable due to rotational indeterminacy, unless q^2 restrictions are imposed on the model parameters (Anderson and Rubin, 1956). In exploratory GLVM-LSS, it is often assumed that the latent variables are uncorrelated. In this case, the factor loadings matrix can be partitioned as $\mathrm{A}^\intercal = [\mathrm{A}_1^\intercal, \mathrm{A}_2^\intercal]$, where A_1 is a $(q \times q)$ matrix with entries above the diagonal fixed to zero (and thus not estimated at all), and A_2 is usually dense. In confirmatory GLVM-LSS, it is usually assumed that the latent variables are correlated. If $\mathbf{\Phi}$ is assumed to be a correlation matrix (i.e., $\mathrm{diag}(\mathbf{\Phi}) = 1$), then A_1 is fixed to be diagonal. It can be the case that we don't impose restrictions on $\mathbf{\Phi}$ (i.e., the latent variables variances/covariances are freely estimated: $\mathrm{diag}(\mathbf{\Phi}) \neq 1$), and thus A_1 is fixed an identity matrix $\mathbb{I}_q$.

For better interpretability, in both exploratory and confirmatory settings we suggest imposing these restrictions on the factor loading matrix corresponding to the measurement equations for the location parameters (i.e., A_μ). However, given the structure of the 'aggregate' factor loading matrix and the rotation of the latent variable space, the q^2 restrictions can be imposed anywhere in A or $\mathbf{\Phi}$. For further details on identification in LVMs, see Anderson and Rubin (1956) and the discussion in Section 1.2.

2.3. Estimation, Inference, and Model Selection

To estimate the GLVM-LSS model described by (2.1) and measurement equations of the type in (2.3), we use a full information marginal maximum likelihood (FIMML) estimation method (Bock and Aitkin, 1981). FIMML has been extensively used in the literature, and it has became the

norm for estimating GLLVMs (see, e.g., Bartholomew et al., 2011; Skrondal and Rabe-Hesketh, 2004). The estimation procedure involves maximising the marginal log-likelihood function which, for a random sample of size n, is given by:

$$\ell(\Theta; \mathbf{y}) = \sum_{m=1}^{n} \log \left(\int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i(\mathbf{z})) \right] \, p(\mathbf{z}; \boldsymbol{\Phi}) \, \mathrm{d}\mathbf{z} \right), \tag{2.5}$$

Here, $\mathbf{y}$ is the observed data, and Θ is a K-dimensional vector of unknown model parameters. When the latent variables are assumed to be uncorrelated (i.e., $\boldsymbol{\Phi} = \mathbb{I}_q$), a common setting in exploratory analyses, the parameter vector Θ only includes intercepts and factor loadings, i.e., $\Theta^\intercal = (\boldsymbol{\alpha}_0^\intercal, \mathrm{vec}(\mathrm{A})^\intercal)$, where 'vec' is the vectorisation operator. If the latent variables are assumed to be correlated, as it is common in confirmatory analyses, then $\Theta^\intercal = (\boldsymbol{\alpha}_0^\intercal, \mathrm{vec}(\mathrm{A})^\intercal, \mathrm{vech}(\boldsymbol{\Phi})^\intercal)$, where 'vech' is the half-vectorisation operator. Let $\Xi \subseteq \mathbb{R}^K$ be the parameter space, i.e., $\Theta \in \Xi$. The maximum likelihood estimate (MLE), denoted as $\hat{\Theta}$, is the point in the parameter space that maximises the marginal log-likelihood function:

$$\hat{\Theta} = \underset{\Theta \in \Xi}{\arg\max}\, \ell(\Theta; \mathbf{y})$$

2.3.1 Computation

In most cases, the solution for $\hat{\Theta}$ is not available in a closed form, and a combination of numerical integration and iterative optimisation algorithms is required to solve the equations $\nabla_\Theta \ell = \mathbf{0}$. To address this challenge, we propose a sequential implementation of two optimisation algorithms: the EM-algorithm (Dempster et al., 1977), in which the latent variables are treated as 'missing data'; and a (quasi-)Newton algorithm.

We propose this computational approach mostly for practical reasons. The EM-algorithm has a number of convenient properties, including a low computation cost per iteration, relative ease of implementation, guaranteed monotonic increase in the value of the objective function, and stability, particularly when the initial values are far from the mode of the log-likelihood. However, due to its (sub-)linear convergence rate (McLachlan and Krishnan, 2008), it can be slow to reach the mode. In contrast, (quasi-)Newton methods are faster due to their super-linear convergence rate and yield estimates of the information matrix as a byproduct, which can be used for the computation of standard errors. Nevertheless, these algorithms may fail to converge if initiated far from the mode and often require more intensive computational operations, such as matrix inversion. Therefore, we propose an optimisation strategy that involves using the EM-algorithm for a fixed number of

iterations and then using the resulting intermediate estimates as 'refined' starting values for the (quasi-)Newton algorithm, which is used in the direct maximisation of $\ell(\Theta; \mathbf{y})$. We provide more details about our proposed approach below.

First step: Parameter computation via the EM-algorithm

The EM-algorithm is a widely used iterative procedure for estimating LVMs. It alternates between two steps: the *E-step*, in which we approximate the expected value of the complete-data log-likelihood with respect to the posterior distribution of the latent variables, and the *M-step*, in which we optimise the expected log-likelihood obtained from the E-step. The EM-algorithm is particularly useful when the maximum likelihood estimator is not available in closed form.

Let $f(\mathbf{y}, \mathbf{z}; \Theta)$ be the joint probability function of the complete data $(\mathbf{y}, \mathbf{z})$. The complete-data log-likelihood is:

$$\begin{aligned} \ell_c(\Theta; \mathbf{y}, \mathbf{z}) &= \sum_{m=1}^{n} \log f(\mathbf{y}_m, \mathbf{z}_m; \Theta) \\ &= \sum_{m=1}^{n} \left[\left\{ \sum_{i=1}^{p} \log f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i) \right\} + \log p(\mathbf{z}_m; \boldsymbol{\Phi}) \right] \end{aligned} \tag{2.6}$$

In what follows, we use the superscript $[t]$ to indicate an estimate, gradient vector, or matrix at iteration t.

E-step: Compute the expected value of (2.6) with respect to the posterior distribution of $\mathbf{z}$ given $\mathbf{y}$, evaluated at the current parameter estimates $\Theta^{[t]}$:

$$\begin{aligned} \mathcal{Q}(\Theta; \Theta^{[t]}) &= \mathbb{E}_{\mathbf{z} \,|\, \mathbf{y}; \Theta^{[t]}} \left[\, \ell_c(\Theta; \mathbf{y}, \mathbf{z}) \, \right] \\ &= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \sum_{i=1}^{p} \log f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i) \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\ &\quad + \sum_{m=1}^{n} \int_{\mathbb{R}^q} \log p(\mathbf{z}_m; \boldsymbol{\Phi}) \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \end{aligned} \tag{2.7}$$

M-step: Update the parameter vector to $\Theta^{[t+1]} = \arg\max \mathcal{Q}(\Theta; \Theta^{[t]})$. In practice, we find $\Theta^{[t+1]}$ that increases the value of the objective function, i.e., $\mathcal{Q}(\Theta^{[t+1]}; \Theta^{[t]}) \geq \mathcal{Q}(\Theta^{[t]}; \Theta^{[t]})$. The M-step requires solving for the complete-data score vector $\mathbb{S}^{[t]}$, which is defined as the gradient of the expected complete-data log-likelihood with respect to the parameter vector, evaluated at $\Theta^{[t]}$, i.e.,

$\mathbb{S}^{[t]} := \nabla_{\Theta}\mathcal{Q}(\Theta;\Theta^{[t]}) = \mathbf{0}$.

Gradient descent (GD) is a simple yet robust update scheme used for updating parameter estimates. The parameter estimates are updated as $\Theta^{[t+1]} = \Theta^{[t]} - \omega^{[t]}\ \mathbb{S}^{[t]}$, where $\omega^{[t]} \in \mathbb{R}^{+}$ is the learning rate, which can be adaptive. The score vector contains entries for the intercepts and factor loadings and is given by:

$$\mathbb{S}^{[t]}_{[\bar{k}_{i,\varphi}]} = \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \varphi_i} \cdot \frac{\partial \varphi_i}{\partial \eta_{i,\varphi}} \cdot \frac{\partial \eta_{i,\varphi}}{\partial \boldsymbol{\alpha}_{i,\varphi}} \right] p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]})\ \mathrm{d}\mathbf{z} \tag{2.8}$$

where $\bar{k}_{i,\varphi} = \{k : \alpha_k \in \boldsymbol{\alpha}^{\intercal}_{i,\varphi},\ k = 1, ..., K\}$ is the index set of intercepts and factor loadings in the measurement equation for the distributional parameter $\varphi_i \in \boldsymbol{\theta}_i$.

An update scheme that is more efficient (but slower in terms of computation time) is the Newton-Raphson (NR) update: $\Theta^{[t+1]} = \Theta^{[t]} - (\mathbb{H}^{[t]})^{-1}\ \mathbb{S}^{[t]}$, where $\mathbb{H}^{[t]} := \nabla_{\Theta}\nabla_{\Theta^{\intercal}}\mathcal{Q}(\Theta;\Theta^{[t]})$ is the complete-data observed information matrix. Because the items are conditionally independent, the complete-data observed information matrix is block-diagonal. For the intercepts and factor loadings in the measurement equations of the distributional parameters $\varphi_i, \tilde{\varphi}_i \in \boldsymbol{\theta}_i$, the block matrices have entries following a general form:

$$\mathbb{H}^{[t]}_{[\bar{k}_{i,\varphi}, \bar{k}_{i,\tilde{\varphi}}]} = \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial^2 \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi} \partial \boldsymbol{\alpha}^{\intercal}_{i,\tilde{\varphi}}} \right] p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]})\ \mathrm{d}\mathbf{z} \tag{2.9}$$

Instead of the observed information matrix, we can also use the expected information matrix, $\mathrm{I}^{[t]} = -\mathbb{E}_{\mathbf{y}}(\mathbb{H}^{[t]})$, or its score-based approximation, $\mathrm{I}^{[t]} = \mathbb{E}_{\mathbf{y}}(\mathbb{S}^{[t]}\mathbb{S}^{[t]\intercal})$, to determine the length of the update in the direction given by $\mathbb{S}^{[t]}$. The choice between these options is problem-dependent and should be selected based on the model complexity and the available computational resources.

When selecting the update scheme, the computational intensity of score vectors and information matrices for models with complex log-likelihoods must be considered. Thus, the choice of update scheme should be determined based on the specific problem at hand. The partial derivatives in equations (2.8) and (2.9) depend on the parametric distribution assumed for the item, the choice of link function, and the specification of the measurement equation. Analytical expressions for derivatives of the distributions and the link functions implemented in this paper are provided in Appendix A2. We run the EM-algorithm step for a user-defined number of iterations (e.g., 30) or until convergence of the objective function (whichever comes first). We then use the intermediate estimates as 'refined' starting values in the direct maximisation step.

Second step: Parameter computation via direct maximisation

In the direct maximisation step, we update the parameter vector obtained in the EM-step and continue to refine the search for $\hat{\Theta}$. We solve for $\nabla_\Theta \ell(\Theta; \mathbf{y}) = \mathbf{0}$. In most cases, closed-form solutions are not available, and thus we use iterative numerical optimisation solvers to compute the MLE, such as (quasi-)Newton and trust-region algorithms[2] (see, e.g., Nocedal and Wright, 2006).

Quasi-Newton line-search methods use first-order information from the gradient of the objective function to determine the direction of the update step, and they compute an approximation of the information matrix (which would otherwise be computationally expensive to compute) to define the step size in that direction. This type of algorithm is fast but can be unstable with complex objective functions, such as non-concave functions or functions with regions that are close to flat. The score vector for the marginal log-likelihood is equivalent to the score vector for the complete-data log-likelihood in (2.8), i.e., $\nabla_\Theta \ell(\Theta; \mathbf{y})|_{\Theta=\Theta^{[t]}} \equiv \nabla_\Theta \mathcal{Q}(\Theta; \Theta^{[t]}) = \mathbb{S}^{[t]}$ (Louis, 1982). This is an important result because it shows the inherent connection between the EM-algorithm step and the direct maximisation step. There is no computational overhead in the calculation of the gradients.

Alternatively, trust-region algorithms (Nocedal and Wright, 2006, Chapter 4) use both first- and second-order information to iteratively create a quadratic approximation of $\ell(\Theta; \mathbf{y})$ around $\Theta^{[t]}$ and search for a local optimum within a certain radius from that point. These algorithms incorporate second-order information to provide a better approximation of the curvature of the marginal log-likelihood at a given point, leading to faster convergence and greater stability compared to line-search methods. However, the computation of the marginal observed information matrix, denoted as $\mathcal{H}^{[t]} = \nabla_\Theta \nabla_{\Theta^\intercal} \ell(\Theta; \mathbf{y})|_{\Theta=\Theta^{[t]}}$, is more computationally expensive than the complete-data information matrix in (2.9). For pairs of items (i, i'), and distributional parameters $\varphi_i \in \boldsymbol{\theta}_i$ and $\tilde{\varphi}_{i'} \in \boldsymbol{\theta}_{i'}$, the matrix $\mathcal{H}^{[t]}$ is made of sub-matrices of the form:

$$
\begin{aligned}
\mathcal{H}^{[t]}_{[\bar{k}_{i,\varphi}, \bar{k}_{i',\tilde{\varphi}}]} = & \sum_{m=1}^{n} \int_{\mathbb{R}^q} p(\mathbf{z} \,|\, \mathbf{y}_m) \cdot \frac{\partial^2 \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi} \partial \boldsymbol{\alpha}^\intercal_{i',\tilde{\varphi}}} \, \mathrm{d}\mathbf{z} \\
& + \sum_{m=1}^{n} \int_{\mathbb{R}^q} p(\mathbf{z} \,|\, \mathbf{y}_m) \cdot \frac{\partial \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi}} \cdot \frac{\partial \log f_{i'}(y_{i'm} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}^\intercal_{i',\tilde{\varphi}}} \, \mathrm{d}\mathbf{z} \\
& - \sum_{m=1}^{n} \int_{\mathbb{R}^q} p(\mathbf{z} \,|\, \mathbf{y}_m) \cdot \frac{\partial \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi}} \, \mathrm{d}\mathbf{z} \cdot \int_{\mathbb{R}^q} p(\mathbf{z} \,|\, \mathbf{y}_m) \frac{\partial \log f_{i'}(y_{i'm} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}^\intercal_{i',\tilde{\varphi}}} \, \mathrm{d}\mathbf{z}
\end{aligned} \tag{2.10}
$$

Note that for items $i \neq i'$, the first term in (2.10) is a null matrix. An alternative to using

[2] In our computational implementation, the analyst can choose one of the built-in `R` functions `optim` or `nlminb` (the default option), or the `trust` function from the package with the same name (Geyer, 2020).

the complete-data information matrix is to use the marginal expected information matrix, denoted as $\mathcal{I}^{[t]} = -\mathbb{E}_{\mathbf{y}}(\mathcal{H}^{[t]})$, or its score approximation. The choice of which matrix to use should be decided on a case-by-case basis. An introductory note on the trust-region algorithm is included in Appendix A3, while further details can be found in Radice et al. (2016), Marra et al. (2017), and Nocedal and Wright (2006, Chapter 4).

Because the gradient vector for the complete-data and marginal log-likelihoods are equivalent, the MLE obtained under both algorithms are nearly identical, differing only in computational accuracy. Thus, the two-step estimation algorithm proposed in this paper is mainly a matter of computational efficiency. In practice, the analyst can choose to estimate Θ only through the EM-algorithm (by setting the number of iterations of the EM-step to a reasonably large number, e.g., 5000, and limiting the number of iterations of the optimisation solver to zero), or only through the direct maximisation of the marginal log-likelihood (by fixing the number of iterations in the EM-step to zero). A similar two-stage computational strategy is implemented in the popular commercial software M*plus* (Muthén and Muthén, 1998 2017). The difference is that M*plus* relies mostly on EM updates, and only performs a single quasi-Newton update wen consecutive EM iterations fail to result in insufficient increase of the marginal log-likelihood.

An additional benefit of using a FIMML estimation method is that it protects against potential bias in the presence of item non-response arising from ignorable missing data mechanisms. To address missing values, we use an indicator matrix with binary entries that indicate whether an observation is missing from the data or not. In such cases, the likelihood is computed based on the observed data corresponding to the non-null entries in the missing data indicator matrix (see, e.g., O'Muircheartaigh and Moustaki, 1999).

For computational simplicity, we use a Gauss-Hermite (GH) rule with fixed quadrature points (e.g., Moustaki and Knott, 2000) to numerically evaluate the integrals in (2.8), (2.9), and (2.10). The use of a fixed-point GH quadrature rule is possible due to the asymptotic normality of $p(\mathbf{z} \mid \mathbf{y})$ (see, e.g., Chang and Stout, 1993; Chang, 1996; Kornely and Kateri, 2022). See Appendix A4 for further details. Alternative methods include adaptive GH quadratures (e.g., Schilling and Bock, 2005; Skrondal and Rabe-Hesketh, 2004), Laplace approximations (e.g., Huber et al., 2004; Bianconcini and Cagnone, 2012), or Monte Carlo approximations (e.g., Sammel et al., 1997; Shi and Lee, 2000; Cai, 2010). These methods lead to approximate MLE solutions, with approximation bias decreasing with sample size, test length, and number of quadrature points (see, e.g., Cagnone et al., 2009; Bianconcini, 2014; Jin and Andersson, 2020).

Once we have obtained an MLE for the factor loading matrix, $\hat{\mathbf{A}}$, we can apply an orthogonal

or oblique rotation to obtain a more interpretable or sparse solution, if needed (see, e.g., Jennrich, 2004, 2006, 2007; Liu et al., 2023).

Computation of factor correlations

When the latent variables are of substantive interest, the estimation of the relationships between latent variables is often of interest. If we assume that $\mathbf{\Phi}$ is a correlation matrix, the estimation of the correlation coefficients requires special consideration to ensure positive semi-definiteness and diagonal entries equal to one. To handle these constraints, we can reparameterise the factor correlation matrix through a Cholesky decomposition, $\mathbf{\Phi} = \mathrm{L}\mathrm{L}^\intercal$, where L is a lower triangular matrix with a fixed entry $\mathrm{L}_{[1,1]} = 1$. This approach allows us to include the $q \times (q+1)/2 - 1$ elements of L in the parameter vector Θ, instead of the $q \times (q-1)/2$ non-redundant correlation coefficients in $\mathbf{\Phi}$.

Let L_j be the j^{th} row of L, and $\mathrm{L}_{j,[k]}$ be the k^{th} element of L_j. Since $\mathbf{\Phi}$ is a correlation matrix, we have $||\mathrm{L}_j|| = 1$, for $j = 1, ..., q$. Therefore, solving for L becomes a constrained optimisation problem, which can be easily handled by a (quasi-)Newton proximal algorithm (Parikh and Boyd, 2014; Lee et al., 2014; Zhang and Chen, 2022) in both the M-step of the EM-algorithm and in the direct maximisation problem.

The gradients of the complete-data log-likelihood function and the marginal log-likelihood with respect to L_j are equivalent, i.e., $\nabla_{\mathrm{L}_j} \mathcal{Q}(\Theta; \Theta^{[t]}) \equiv \nabla_{\mathrm{L}_j} \ell(\Theta; \mathbf{y})|_{\Theta=\Theta^{[t]}} = \mathbb{S}_{\mathrm{L}_j}^{[t]}$. Consequently, the two algorithms yield identical solutions, making the two-step estimation algorithm computationally efficient. The vector $\mathbb{S}_{\mathrm{L}_j}^{[t]}$ at iteration t of the EM-algorithm or the (quasi-)Newton solver has entries that follow a general format:

$$\begin{aligned}
\mathbb{S}_{\mathrm{L}_{j,[k]}}^{[t]} &= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial}{\partial \mathrm{L}_{j,[k]}} \log p(\mathbf{z}_m; \mathrm{L}) \right] \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&= -n \cdot \mathrm{tr}\left(\mathrm{L}^\intercal (\mathrm{L}\mathrm{L}^\intercal)^{-1} \mathrm{D}_{jk} \right) + \sum_{m=1}^{n} \left[\mathrm{tr}\left(\mathrm{G}_{jk} \mathbb{V}_m^{[t]} \right) + \breve{\mathbf{z}}_m^{[t]\intercal} \mathrm{G}_{jk} \breve{\mathbf{z}}_m^{[t]} \right]
\end{aligned} \tag{2.11}$$

Here, $\mathrm{D}_{jk} = \partial \mathrm{L} / \partial \mathrm{L}_{j,[k]}$ is a square matrix of dimension q, with a value of 1 in the $[j,k]$ position and zero elsewhere. The matrix $\mathrm{G}_{jk} = (\mathrm{L}\mathrm{L}^\intercal)^{-1} \mathrm{D}_{jk} \mathrm{L}^\intercal (\mathrm{L}\mathrm{L}^\intercal)^{-1}$, the conditional mean $\breve{\mathbf{z}}_m^{[t]} = \mathbb{E}(\mathbf{z} | \mathbf{y}_m; \Theta^{[t]})$, and conditional variance $\mathbb{V}_m^{[t]} = \mathbb{E}((\mathbf{z} - \breve{\mathbf{z}}_m^{[t]})(\mathbf{z} - \breve{\mathbf{z}}_m^{[t]})^\intercal | \mathbf{y}_m; \Theta^{[t]})$, are computed using the properties of the trace operator and the linearity of the conditional expectation. The derivations for (2.11) and the equivalence of $\mathbb{S}$ between the EM-algorithm and the direct maximisation approach are detailed in Appendix A2. To simplify the computation of the matrix in the NR update rule

in the M-step of the EM-algorithm and in the direct maximisation, we use the BFGS updating formula (Nocedal and Wright, 2006, Chapter 6).

We denote the updated value for column L_j obtained from either the EM-algorithm or direct optimisation solver at iteration $t+1$ as $\tilde{\mathrm{L}}_j^{[t+1]}$. To ensure that $\mathrm{L}_j^{[t+1]}$ is a unit-norm vector, we project $\tilde{\mathrm{L}}_j^{[t+1]}$ onto the feasible region of the constrained optimisation problem:

$$\mathrm{L}_j^{[t+1]} = \underset{\mathrm{L}_j : ||\mathrm{L}_j||=1}{\arg\min} ||\mathrm{L}_j - \tilde{\mathrm{L}}_j^{[t+1]}|| = \frac{1}{||\tilde{\mathrm{L}}_j^{[t+1]}||}\tilde{\mathrm{L}}_j^{[t+1]}, \qquad \text{for } j = 1, ..., q$$

Zhang and Chen (2022) present a thorough explanation of proximal (quasi-)Newton algorithms in the context of estimation in LVMs.

It is important to note that the trust-region implementation `trust` is unsuitable for constrained maximisation problems and that the L-BFGS-B solver in `optim` does not produce an estimate $\hat{\boldsymbol{\Phi}}$ that satisfies the positive semi-definite requirement of the correlation matrix. Thus, we propose an alternating procedure in the direct maximisation step. First, we obtain $(\hat{\boldsymbol{\alpha}}_0^\intercal, \mathrm{vec}(\hat{\mathrm{A}})^\intercal)$ using any of the implemented algorithms, treating $\hat{\boldsymbol{\Phi}} = \hat{\mathrm{L}}\hat{\mathrm{L}}^\intercal$ (either from the EM-step or previous direct maximisation steps) as fixed. Second, we update $\hat{\boldsymbol{\Phi}}$ while treating $(\hat{\boldsymbol{\alpha}}_0^\intercal, \mathrm{vec}(\hat{\mathrm{A}})^\intercal)$ from the direct maximisation as fixed. We repeat these steps until the marginal log-likelihood converges.

2.3.2 Asymptotic Properties of the Maximum Likelihood Estimator

Assuming suitable regularity conditions and a correct specification of the GLVM-LSS model, the maximum likelihood estimator is asymptotically unbiased, maximally efficient, and normally distributed with variance given by the inverse of the expected information matrix (see, e.g., van der Vaart, 1998):

$$\sqrt{n}(\hat{\Theta} - \Theta^*) \xrightarrow{d} \mathbb{N}(\mathbf{0}, n\mathcal{I}(\Theta^*)^{-1}),$$

where Θ^* denotes the true value of the parameter vector and $\mathcal{I}(\Theta^*)$ is the expected information matrix evaluated at Θ^*. Proofs are included in Appendix A5.

2.3.3 Goodness of fit and Model Selection

Model selection and assessing goodness of fit in GLVM-LSS models require careful consideration of and comparison between several different scenarios: i) nested models with the same number of latent variables, ii) models with different numbers of latent variables, and iii) models with different

parametric distributions for the manifest variables. These scenarios may also be combined in various ways.

Formally, let $\mathcal{M} = \{\Theta,\ \mathbf{z},\ \mathcal{F}\}$ denote an GLVM-LSS model, with $\Theta \in \Xi$ representing the model parameters, $\mathbf{z} \in \mathbb{R}^q$ the latent variables, and $\mathcal{F} = \{f_1(\cdot \mid \mathbf{z}; \boldsymbol{\theta}_1), ..., f_p(\cdot \mid \mathbf{z}; \boldsymbol{\theta}_p)\}$ the set of p parametric distributions assumed for the observed variables. Comparison between nested models (scenario i), such as $\mathcal{M}_0 = \{\Theta \in \Xi_0, \mathbf{z}, \mathcal{F}\}$ and $\mathcal{M}_1 = \{\Theta \in \Xi_1, \mathbf{z}, \mathcal{F}\}$, where Ξ_1 is obtained by imposing restrictions on Ξ_0 (usually in the form of fixed values), can be evaluated using the likelihood-ratio test (LRT) for normal linear factor models, and the Pearson-χ^2-test or G^2-test for binary and/or polytomous items. However, one should proceed with caution when using these tests because they are sensitive to departures from distributional assumptions and may suggest selecting an overly complex model, as there is no penalty for over-parametrisation (Akaike, 1987). Additionally, comparing non-nested LVMs via LRT can result in inflated type-I errors for the test statistic, as some of the regularity conditions necessary for the correct asymptotic distribution of the LRT statistic are not satisfied in this case (see Chen et al., 2020).

For comparison between non-nested modes (scenarios ii and iii), such as $\mathcal{M}_0 = \{\Theta \in \Xi_0, \mathbf{z} \in \mathbb{R}^{q_0}, \mathcal{F}_0\}$ and $\mathcal{M}_1 = \{\Theta \in \Xi_1, \mathbf{z} \in \mathbb{R}^{q_1}, \mathcal{F}_1\}$, where $\Xi_0 \neq \Xi_1$, $q_0 \neq q_1$, and/or $\mathcal{F}_0 \neq \mathcal{F}_1$, we recommend using information criteria. Information criteria involve adding a penalty term that is proportional to the number of parameters to the fitted deviance, which is defined as $-2\ell(\Theta; \mathbf{y})$. In general, information criteria can be expressed as $\text{IC}(\hat{\Theta}) = -2\ell(\Theta; \mathbf{y}) + \kappa K$, where $\kappa > 0$ is a constant that defines the weight assigned to the penalty on model complexity. Two popular criteria are the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978). AIC uses a penalty of $2K$ to select the model with the most accurate prediction, while BIC uses a penalty of $\log(n)K$, where n is the sample size, to achieve consistent model selection. The model with the lowest information criteria provides the best fit and is thus preferred.

2.4. Simulation Studies

In this section, we conduct simulation studies to evaluate the performance of the proposed GLVM-LSS framework under finite sample settings. To assess the accuracy of the proposed two-step ML estimator, we compute the mean squared error (MSE) and absolute bias (AB) for each parameter in the model. For ease of comparison, we report the average MSE (AvMSE) and average AB (AvAB) for intercepts and factor loadings, separately for each distributional parameter in the corresponding distribution. We also calculate similar measures for the factor correlations, if they are included in

the model. All simulations were conducted using R version 4.2.2 (R Core Team, 2022), and the code and replication files are available at https://github.com/ccardehu/GLVM-LSS. See Appendix A6 for details on the software implementation.

2.4.1 Simulation Study I: Exploratory LVM-LSS models

In the first simulation study, we aim to evaluate the performance of the proposed GLVM-LSS framework in various exploratory settings. To this end, we consider four different sample sizes, namely $n = \{200, 500, 1000, 5000\}$, and three different test lengths, with $p = \{5, 10, 20\}$ items, resulting in a total of 12 conditions. We simulate $L = 1000$ independent datasets and compute the AvMSE and AvAB for the estimated parameters.

Case I: Heteroscedastic Factor Model: We first consider a heteroscedastic factor model, where the location (μ_i) and scale (σ_i) parameters of Normal items are modelled as linear functions of the latent variables $\mathbf{z}$, i.e., $y_i \,|\, \mathbf{z} \sim \mathbb{N}(\mu_i(\mathbf{z}),\ \sigma_i^2(\mathbf{z}))$. We assume two uncorrelated latent variables, $(z_1, z_2)^\intercal \sim \mathbb{N}(\mathbf{0}, \mathbb{I}_2)$. This GLVM-LSS is a multidimensional extension of the model proposed in Hessen and Dolan (2009).

For the location measurement equations, we sampled the intercepts and slopes from $\alpha_{i0,\mu} \sim \text{Unif}(1.0, 2.0)$ and $\alpha_{ij,\mu} \sim \text{Unif}(0.5, 1.5)$, respectively. The sign of the $\alpha_{ij,\mu}$'s was randomly assigned with a probability of 0.5. The scale measurement equation parameters were generated from $(\alpha_{i0,\sigma}, \alpha_{ij,\sigma}^\intercal)^\intercal \sim \text{Unif}(0.1, 0.4)$. We impose appropriate restrictions on the factor loading matrix to avoid rotational indeterminacy. The $L = 1000$ datasets were randomly generated using the same set of factor loadings. We used the GH rule with 25 quadrature points in each latent dimension (625 in total) for numerical integration. Table 2.1 summarises the results, which show that both the AvMSE and AvAB decrease as the sample size increases, consistent with ML theory.

Case II: Zero-Inflated Poisson Items: In the second setting, we examine a GLVM-LSS model for count data with zero-inflation. Estimating the zero-inflated Poisson (ZIP) GLVM-LSS model requires additional considerations, which are discussed in Appendix A1.2. The mixing probability and rate parameter of the items depend linearly on the latent variables, i.e., $y_i \,|\, \mathbf{z} \sim \text{ZIP}(\lambda_i(\mathbf{z}), \pi_i(\mathbf{z}))$. For simplicity, we consider two uncorrelated latent variables, making this GLVM-LSS a multidimensional extension of the model proposed by Wang (2010).

The parameters in the rate measurement equation were drawn from $\alpha_{i0,\lambda} \sim \text{Unif}(2, 3)$, and

		Average MSE (AvMSE)				Average AB (AvAB)			
p	n	Inter. ($\hat{\alpha}_{i0,\mu}$)	Load. ($\hat{\boldsymbol{\alpha}}_{i,\mu}$)	Inter. ($\hat{\alpha}_{i0,\sigma}$)	Load. ($\hat{\boldsymbol{\alpha}}_{i,\sigma}$)	Inter. ($\hat{\alpha}_{i0,\mu}$)	Load. ($\hat{\boldsymbol{\alpha}}_{i,\mu}$)	Inter. ($\hat{\alpha}_{i0,\sigma}$)	Load. ($\hat{\boldsymbol{\alpha}}_{i,\sigma}$)
5	200	0.0176	0.0991	0.0465	0.0312	0.0047	0.0391	0.0813	0.0204
	500	0.0056	0.0068	0.0037	0.0033	0.0023	0.0026	0.0129	0.0019
	1000	0.0023	0.0070	0.0048	0.0030	0.0010	0.0028	0.0126	0.0040
	5000	0.0007	0.0015	0.0004	0.0004	0.0006	0.0007	0.0011	0.0006
10	200	0.0166	0.0196	0.0067	0.0068	0.0054	0.0088	0.0232	0.0044
	500	0.0059	0.0057	0.0021	0.0024	0.0025	0.0048	0.0095	0.0018
	1000	0.0033	0.0031	0.0009	0.0010	0.0008	0.0013	0.0038	0.0010
	5000	0.0006	0.0005	0.0002	0.0002	0.0004	0.0012	0.0008	0.0003
20	200	0.0162	0.0115	0.0044	0.0045	0.0039	0.0058	0.0189	0.0058
	500	0.0079	0.0060	0.0016	0.0016	0.0027	0.0062	0.0072	0.0011
	1000	0.0040	0.0036	0.0008	0.0008	0.0022	0.0056	0.0035	0.0010
	5000	0.0007	0.0010	0.0002	0.0002	0.0006	0.0030	0.0006	0.0012

Table 2.1: Simulation Study I, Case I: Average Mean Squared Error (AvMSE) and Average Absolute Bias (AvAB) for the MLE of a linear factor model with heteroscedastic items, by number of items and sample size. The performance measures are computed for the estimated parameters $\hat{\alpha}_k$ in the location loading matrix ($\hat{\mathrm{A}}_\mu$) and scale loading matrix ($\hat{\mathrm{A}}_\sigma$).

$\alpha_{ij,\lambda} \sim \text{Unif}(0.2, 0.6)$. We assign the signs of $\alpha_{ij,\lambda}$'s at random with a probability of 0.5. For the mixing probability, the parameters were sampled from $\alpha_{i0,\pi} \sim \text{Unif}(-2, -1)$ and $\alpha_{ij,\pi} \sim \text{Unif}(1.5, 2.5)$. Appropriate parameter restrictions are imposed to avoid rotational indeterminacy. The $L = 1000$ datasets were randomly generated using the same set of factor loadings. We use the GH rule for numerical integration with 25 quadrature points for each latent dimension (625 in total). Table 2.2 shows the AvMSE and AvAB for this model. As expected, the AvMSE and AvAB decrease with the sample size.

Case III: Items on the $(0, 1)$ **Interval:** In the third setting, we consider a GLVM-LSS model with items that follow a location-scale parameterization of the Beta distribution, conditioned on a unidimensional latent variable. In this model, the location and scale parameters are linearly dependent on the latent factor, i.e., $y_i \,|\, \mathbf{z} \sim \text{Beta}(\mu_i(\mathbf{z}), \sigma_i(\mathbf{z}))$.

The population values for the parameters in the location measurement equation are drawn from $(\alpha_{i0,\mu}, \alpha_{i1,\mu})^\intercal \sim \text{Unif}(0, 1)$. The signs of the $\alpha_{i1,\mu}$'s are randomly determined with a probability of 0.5. The parameters in the scale measurement equation are sampled from $(\alpha_{i0,\sigma}, \alpha_{i1,\mu})^\intercal \sim \text{Unif}(-1, 0.2)$. We generate the true parameters in this way to ensure that the conditional densities

		Average MSE (AvMSE)				Average AB (AvAB)			
p	n	Inter. $(\hat{\alpha}_{i0,\lambda})$	Load. $(\hat{\boldsymbol{\alpha}}_{i,\lambda})$	Inter. $(\hat{\alpha}_{i0,\pi})$	Load. $(\hat{\boldsymbol{\alpha}}_{i,\pi})$	Inter. $(\hat{\alpha}_{i0,\lambda})$	Load. $(\hat{\boldsymbol{\alpha}}_{i,\lambda})$	Inter. $(\hat{\alpha}_{i0,\pi})$	Load. $(\hat{\boldsymbol{\alpha}}_{i,\pi})$
5	200	0.0024	0.0033	0.2299	0.4826	0.0034	0.0050	0.1212	0.1457
	500	0.0009	0.0008	0.0602	0.0969	0.0006	0.0012	0.0410	0.0582
	1000	0.0004	0.0004	0.0277	0.0459	0.0007	0.0007	0.0131	0.0206
	5000	0.0001	0.0001	0.0048	0.0093	0.0001	0.0005	0.0025	0.0048
10	200	0.0022	0.0018	0.1158	0.1783	0.0021	0.0021	0.0566	0.0856
	500	0.0008	0.0007	0.0478	0.0754	0.0009	0.0007	0.0277	0.0446
	1000	0.0005	0.0004	0.0212	0.0307	0.0006	0.0006	0.0074	0.0153
	5000	0.0001	0.0001	0.0048	0.0063	0.0003	0.0009	0.0037	0.0018
20†	200	0.0024	0.0016	0.1305	0.1707	0.0078	0.0026	0.0343	0.0832
	500	0.0012	0.0009	0.0564	0.0639	0.0041	0.0035	0.0556	0.0347
	1000	0.0006	0.0004	0.0233	0.0298	0.0013	0.0024	0.0165	0.0185
	5000	0.0001	0.0001	0.0057	0.0065	0.0010	0.0024	0.0053	0.0073

Table 2.2: Simulation Study I, Case II: Average Mean Squared Error (AvMSE) and Average Absolute Bias (AvAB) for the MLE of a LVM with ZIP items, by number of items and sample size. These performance measures are computed for the estimated parameters $\hat{\alpha}_k$ in the loading matrices $\hat{\mathrm{A}}_\lambda$ and $\hat{\mathrm{A}}_\pi$. † Note: For computational reasons, we ran $L = 100$ simulations in this setting.

$f_i(y_i|\mathbf{z})$ are uni-modal. The $L = 1000$ datasets were randomly generated using the same set of factor loadings. Although the Beta distribution allows for bimodal densities for certain combinations of μ_i and σ_i, this is not common in the applications of interest (see Noel, 2014 for a unidimensional Beta factor model that handles the bi-modality of items). We use 50 quadrature points in the GH rule for numerical integration. The simulation results are presented in Table 2.3. As expected, the AvMSE and the AvAB for the factor loadings in the measurement equations for the location and scale parameters decrease with the sample size.

Case IV: Skew-Normal items: In our final setting, we consider a GLVM-LSS model with continuous items that follow a reparameterization of the Skew-Normal (SN) distribution (see Appendix A1.1), conditional on a single latent variable. The location ($\mu \in \mathbb{R}$), scale ($\sigma \in \mathbb{R}^+$), and shape ($\nu \in (0, 1)$) parameters depend linearly on the latent factor, i.e., $y_i \,|\, \mathbf{z} \sim \mathrm{SN}(\mu_i(\mathbf{z}), \sigma_i^2(\mathbf{z}), \nu_i(\mathbf{z}))$.

The true values for the parameters in the location measurement equation are drawn from $\alpha_{0i,\mu} \sim \mathrm{Unif}(-1, 1)$ and $\alpha_{i1,\mu} \sim \mathrm{Unif}(0.5, 1.5)$. The signs of the slope parameters are randomly defined with probability 0.5. The parameters in the scale measurement equation are sampled from $(\alpha_{i0,\sigma}, \alpha_{i1,\sigma})^\intercal \sim \mathrm{Unif}(0.2, 0.4)$, while the parameters in the shape measurement equation are drawn

		Average MSE (AvMSE)				Average AB (AvAB)			
p	n	Inter. ($\hat{\alpha}_{i0,\mu}$)	Load. ($\hat{\alpha}_{i1,\mu}$)	Inter. ($\hat{\alpha}_{i0,\sigma}$)	Load. ($\hat{\alpha}_{i1,\sigma}$)	Inter. ($\hat{\alpha}_{i0,\mu}$)	Load. ($\hat{\alpha}_{i1,\mu}$)	Inter. ($\hat{\alpha}_{i0,\sigma}$)	Load. ($\hat{\alpha}_{i1,\sigma}$)
5	200	0.0045	0.0042	0.0088	0.0079	0.0035	0.0036	0.0165	0.0079
	500	0.0022	0.0024	0.0036	0.0029	0.0014	0.0015	0.0078	0.0017
	1000	0.0009	0.0007	0.0016	0.0015	0.0010	0.0007	0.0030	0.0012
	5000	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0010	0.0005
10	200	0.0051	0.0043	0.0067	0.0063	0.0030	0.0047	0.0143	0.0034
	500	0.0017	0.0013	0.0026	0.0024	0.0020	0.0038	0.0065	0.0053
	1000	0.0012	0.0009	0.0013	0.0011	0.0017	0.0022	0.0020	0.0030
	5000	0.0003	0.0002	0.0003	0.0002	0.0010	0.0014	0.0012	0.0021
20	200	0.0045	0.0029	0.0065	0.0055	0.0018	0.0064	0.0111	0.0048
	500	0.0026	0.0019	0.0026	0.0022	0.0022	0.0033	0.0065	0.0030
	1000	0.0011	0.0008	0.0013	0.0015	0.0030	0.0040	0.0040	0.0072
	5000	0.0002	0.0002	0.0003	0.0003	0.0013	0.0022	0.0013	0.0038

Table 2.3: Simulation Study I, Case III: Average Mean Squared Error (AvMSE) and Average Absolute Bias (AvAB) for the MLE of a LVM with Beta distributed items, by number of items and sample size. These performance measures are computed for the estimated parameters $\hat{\alpha}_k$ in the location loading matrix ($\hat{\mathrm{A}}_\mu$) and scale loading matrix ($\hat{\mathrm{A}}_\sigma$).

from $\alpha_{0,\nu} \sim \text{Unif}(-2, 2)$ and $\alpha_{i1,\nu} \sim \text{Unif}(0.2, 0.5)$, with the slope signs treated similarly as before. The $L = 1000$ datasets were randomly generated using the same set of factor loadings. We use 100 quadrature points in the GH rule for numerical integration. The simulation results are presented in Table 2.4.

Based on the simulation results presented in Table 2.4, we observe a high variance in the parameter estimates of the shape parameter measurement equation, particularly for low sample sizes and number of observed variables. This is consistent with the slow convergence of the skewness parameter reported in Chiogna (2005). Our analysis reveals that the maximum likelihood estimate of the reparameterised skewness parameter ($\hat{\nu}$) has an asymptotic variance proportional to $23.77 \cdot n^{-1}$ in a neighbourhood of zero (see Appendix A1.1). To ensure numerical stability and obtain accurate results, in simulations and real-world applications we recommend fitting the model when the sample size is large, as this is when the asymptotic properties of the estimates hold (see, e.g., Monti, 2003; Eberl and Klar, 2020; Jiang and Xu, 2022).

		Average MSE (AvMSE)						Average AB (AvAB)					
p	n	Inter. ($\hat{\alpha}_{i0,\mu}$)	Load. ($\hat{\alpha}_{i1,\mu}$)	Inter. ($\hat{\alpha}_{i0,\sigma}$)	Load. ($\hat{\alpha}_{i1,\sigma}$)	Inter. ($\hat{\alpha}_{i0,\nu}$)	Load. ($\hat{\alpha}_{i1,\nu}$)	Inter. ($\hat{\alpha}_{i0,\mu}$)	Load. ($\hat{\alpha}_{i1,\mu}$)	Inter. ($\hat{\alpha}_{i0,\sigma}$)	Load. ($\hat{\alpha}_{i1,\sigma}$)	Inter. ($\hat{\alpha}_{i0,\nu}$)	Load. ($\hat{\alpha}_{i1,\nu}$)
5	200	0.0134	0.0146	0.0061	0.0062	9.6714	4.0814	0.0033	0.0097	0.0131	0.0045	0.6447	0.3681
	500	0.0055	0.0052	0.0022	0.0023	1.3861	0.6242	0.0018	0.0022	0.0071	0.0021	0.1786	0.0939
	1000	0.0027	0.0026	0.0010	0.0011	0.2658	0.1627	0.0005	0.0018	0.0028	0.0004	0.0638	0.0383
	5000	0.0005	0.0005	0.0002	0.0002	0.0265	0.0198	0.0006	0.0003	0.0005	0.0003	0.0125	0.0057
10	200	0.0151	0.0131	0.0041	0.0046	0.9162	0.9612	0.0043	0.0126	0.0094	0.0032	0.1816	0.0939
	500	0.0059	0.0047	0.0016	0.0017	0.1471	0.1818	0.0020	0.0025	0.0041	0.0011	0.0473	0.0160
	1000	0.0031	0.0023	0.0008	0.0008	0.0558	0.0634	0.0016	0.0016	0.0022	0.0006	0.0208	0.0115
	5000	0.0006	0.0005	0.0002	0.0002	0.0099	0.0103	0.0003	0.0010	0.0004	0.0002	0.0053	0.0035
20	200	0.0151	0.0155	0.0040	0.0039	0.4441	0.6405	0.0067	0.0179	0.0064	0.0028	0.1456	0.0301
	500	0.0059	0.0043	0.0014	0.0014	0.1085	0.1356	0.0027	0.0038	0.0039	0.0011	0.0422	0.0180
	1000	0.0030	0.0021	0.0007	0.0007	0.0488	0.0528	0.0012	0.0025	0.0019	0.0010	0.0190	0.0118
	5000	0.0006	0.0004	0.0001	0.0001	0.0088	0.0091	0.0006	0.0011	0.0005	0.0003	0.0034	0.0030

Table 2.4: Simulation Study I, Case IV: Average Mean Squared Error (AvMSE) and Average Absolute Bias (AvAB) for the MLE of a LVM with Skew Normal distributed items, by number of items and sample size. The performance measures are computed for the estimated parameters $\hat{\alpha}_k$ in the location loading matrix ($\hat{\mathrm{A}}_\mu$), scale loading matrix ($\hat{\mathrm{A}}_\sigma$), and shape loading matrix ($\hat{\mathrm{A}}_\nu$).

2.4.2 Simulation Study II: A Confirmatory GLVM-LSS model with Binary and Skew-Normal items

In the second study, we consider a confirmatory GLVM-LSS model based on the first example in the applications section. The model assumes two latent variables ($q = 2$) and 12 items, with the first six items distributed Bernoulli, conditional on the first factor, $y_i \mid z_1 \sim \text{Bernoulli}(\pi_i(z_1))$ for $i = 1, ..., 6$, and the remaining items distributed Skew-Normal, conditional on the second factor, $y_i \mid z_2 \sim \text{SN}(\mu_i(z_2), \sigma_i(z_2), \nu_i(z_2))$ for $i = 7, ..., 12$. The latent variables follow a multivariate standard Normal distribution, $(z_1, z_2)^\intercal \sim \mathbb{N}(\mathbf{0}, \mathbf{\Phi})$, with $\mathbf{\Phi}$ a correlation matrix (i.e., $\text{diag}(\mathbf{\Phi}) = 1$) with off-diagonal entries denoted by $\phi_{12} = \phi_{21} = \phi_{\mathbf{z}}$. A path diagram representation of the model is shown in Figure 2.1.

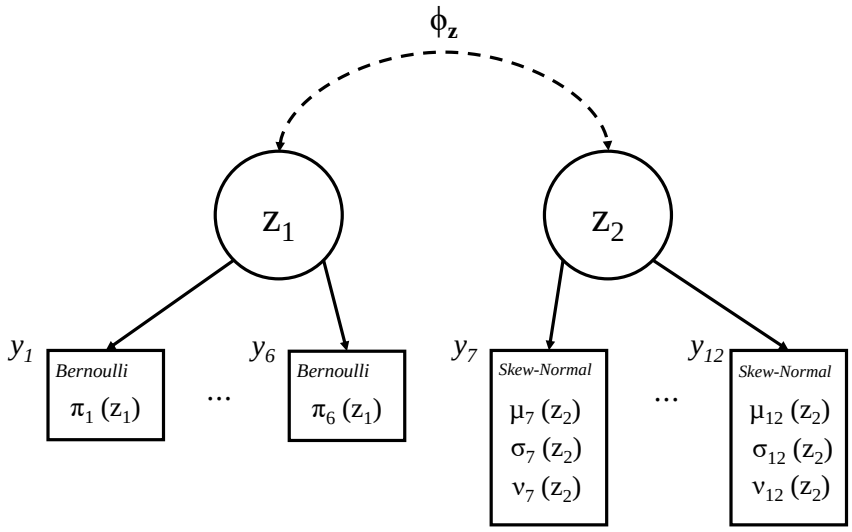


Figure 2.1: Path Diagram of Confirmatory GLVM-LSS simulation study.

The sample sizes considered are $n = \{500, 1000, 5000\}$, which are common in psychometrics research. All the intercepts and factor loadings in the model were randomly generated by sampling from uniform distributions. For each item i, the parameters in the location measurement equations were sampled from $(\alpha_{0i,\pi}, \alpha_{0i,\mu})^\intercal \sim \text{Unif}(-1.5, 1.5)$, $\alpha_{i1,\pi} \sim \text{Unif}(1, 2)$, and $\alpha_{i2,\mu} \sim \text{Unif}(0.5, 1.5)$. The scale measurement equation parameters for SN items were generated by sampling $\alpha_{i0,\sigma} \sim \text{Unif}(-1.5, -0.5)$, and $\alpha_{i2,\sigma} \sim \text{Unif}(0.3, 0.6)$. The shape measurement equation parameters for SN items were generated by sampling $\alpha_{i0,\nu} \sim \text{Unif}(-1.5, -0.5)$, and $\alpha_{i2,\nu} \sim \text{Unif}(0.3, 1)$. The signs for $\alpha_{i2,\sigma}$'s, $\alpha_{i0,\nu}$'s, and $\alpha_{i2,\nu}$'s were set at random with a probability of 0.5. The correlation between latent variables was set at $\phi_{\mathbf{z}} = 0.3$, and 300 independent datasets were simulated to compute the AvMSE and AvAB for the estimated parameters. The $L = 300$ datasets were randomly generated using the same set of factor loadings and the same factor correlation. The true parameter vector

was used as the starting point for estimation in the EM-algorithm for simplicity, but it is important to note that different starting points should be used in practical applications. The results by sample size are presented in Table 2.5.

	Average MSE (AvMSE)					Average AB (AvAB)				
n	$\hat{\mathrm{A}}_{\pi}$	$\hat{\mathrm{A}}_{\mu}$	$\hat{\mathrm{A}}_{\sigma}$	$\hat{\mathrm{A}}_{\nu}$	$\hat{\phi}_{\mathbf{z}}$	$\hat{\mathrm{A}}_{\pi}$	$\hat{\mathrm{A}}_{\mu}$	$\hat{\mathrm{A}}_{\sigma}$	$\hat{\mathrm{A}}_{\nu}$	$\hat{\phi}_{\mathbf{z}}$
500	0.0294	0.0018	0.0021	0.3431	0.0030	0.1301	0.0314	0.0355	0.4081	0.0432
1000	0.0138	0.0010	0.0010	0.1162	0.0013	0.0900	0.0228	0.0246	0.2514	0.0295
3000	0.0044	0.0005	0.0004	0.0290	0.0005	0.0509	0.0159	0.0151	0.1303	0.0176

Table 2.5: Simulation Study II: Average Mean Squared Error (AvMSE) and Average Absolute Bias (AvAB) for the MLE of a confirmatory GLVM-LSS with Bernoulli and Skew-Normal distributed items, by sample size.

For simplicity, we present the aggregate results for intercepts and factor loadings in each matrix $\hat{\mathrm{A}}_{\varphi}$, $\varphi \in \boldsymbol{\theta}$. In all cases, the AvMSE and AvAB decrease with sample size, as expected[3]. This simulation study shows that, under confirmatory settings, the factor correlation $\phi_{\mathbf{z}}$ is consistently estimated.

2.5. Empirical Applications

2.5.1 PISA 2018: A joint model for item response and response times

We present an empirical example of a confirmatory GLVM-LSS model for binary item responses and continuous response times. This type of joint analysis has been extensively studied in educational testing literature, and is particularly valuable because response times provide information about a student's ability and test-taking strategies, as well as aiding item calibration and test design (van der Linden, 2007, 2008; van der Linden and Guo, 2008; van der Linden et al., 2010). A comprehensive framework for this type of analysis is the hierarchical model for speed and accuracy on test items, originally proposed by van der Linden (2007, 2009), and subsequently extended by others (Molenaar et al., 2015; Bolsinova et al., 2017; Bolsinova and Molenaar, 2018). For a review of models involving items and response times, see De Boeck and Jeon (2019).

[3]The shape parameter factor loadings for the Skew-Normal items show higher AvMSE and AvAB than those for the location and scale parameters. The higher AvMSE is expected, due to the asymptotic behaviour of the MLEs for the Skew-Normal distribution explained in Appendix A1.1. We expect lower figures for the AvAB by increasing the accuracy of the numerical integration using a higher number of quadrature points in the Gaussian Hermite quadrature. We do not pursue this further here, due to the increased computational demand of this simulation study. In spite of this, it should be noted that both the AvMSE and AvAB for the factor loadings in the shape parameter measurement equations decrease as the sample size increases, as expected.

The hierarchical model proposed by van der Linden (2007, 2009) consists of two connected models: (i) an IRT model for the observed items, where the probability of responding correctly depends on the individual's latent ability; and (ii) a normal linear factor model for the logarithm of the response times (log-RT), where individuals with a higher latent speed factor will tend to respond more quickly. The precision parameter, defined as the reciprocal of the standard deviation of the log-RT, plays an important role in identifying items with high (or low) heterogeneity in their (log-)RTs. Items with high log-RT precision (i.e., low variance, homogeneous log-RTs) will discriminate better between individuals with different levels of the speed factor better than items with low log-RT precision (i.e., high variance, heterogeneous log-RTs). However, in this model, the precision (i.e., scale parameter) does not depend on the latent speed factor. The latent ability and latent speed factors are jointly assumed to be normal and correlated, in order to capture what is referred to in the literature as the 'speed-accuracy trade-off' (Zimmerman, 2011). Appropriate restrictions are imposed to ensure identifiability and interpretability. The author proposes a fully Bayesian estimation approach using a Gibbs sampler.

To showcase the proposed modelling framework, we elaborate on the baseline model discussed above by assuming that the log-RT follows a Skew-Normal (SN) distribution conditional on the speed factor. The SN distribution allows for modelling not only varying heterogeneity (variance), but also varying skewness in the log-RTs along the latent speed trait scale. Higher order moments of RTs can give valuable insight on students' test-taking strategies and thought processing during high-stakes standardised tests, as well as information on item quality.

We employ data from the 2018 PISA computer-based mathematics exam and focus on a sample of Brazilian students who answered 9 binary items from the first testing cluster. For simplicity, we only consider individuals who provided complete answers, yielding a sample size of 1280 students. However, as discussed in Section 2.3.1, the FIMML estimation procedure can deal with units with partially incomplete responses. The response times (in milliseconds) were transformed into log-minutes. The empirical marginal distributions of the response times in log-minutes are displayed in Figure 2.2. A majority of the log-RT exhibit some degree of skewness, and we observe improved model fit when we assume the log-RT to follow a SN distribution instead of a Normal distribution. The solid lines represent marginal distributions assuming SN-distributed log-RT, while the dashed ones correspond to the marginals under a Normal distribution.

Formally, the proposed GLVM-LSS model posits that item responses (IR), denoted by $y_1, \ldots, y_9$, follow a conditional Bernoulli distribution, with the probability of answering correctly (location parameter) represented as a function of the latent ability (z_1). That is, $y_i \mid z_1 \sim \text{Bernoulli}(\pi_i(z_1))$

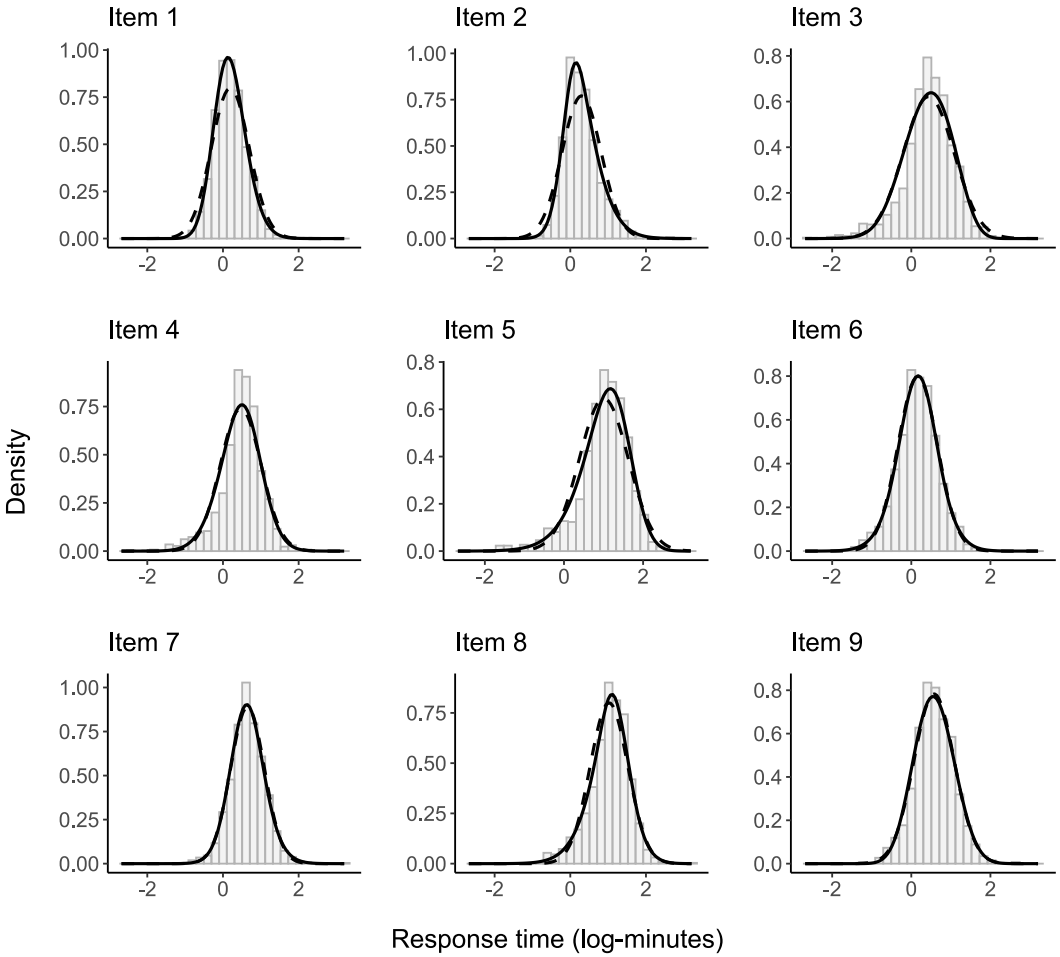


Figure 2.2: PISA 2018: Empirical and model-implied marginal distributions for response times (in log-minutes). The solid line (——) is the SN model and the dashed line (- - -) the Normal model.

for items $i = 1, \ldots, 9$. The measurement equations for the IRs are expressed as:

$$\text{logit}(\pi_i) = \alpha_{i0,\pi} + \alpha_{i1,\pi} z_1 \qquad i = 1, \ldots, 9 \tag{2.12}$$

where $\alpha_{i0,\pi}$ and $\alpha_{i1,\pi}$ denote item i's difficulty and discrimination parameters, respectively. The log-RT follow a conditional Skew-Normal (SN) distribution, with location, scale, and shape parameters potentially modelled in terms of the speed factor (z_2), i.e., $\log(t_i) \mid z_2 \sim \text{SN}(\mu_i(z_2), \sigma_i^2(z_2), \nu_i(z_2))$, with $t_1, \ldots, t_9$ representing the RTs in minutes. We assume that $(z_1, z_2)^\intercal \sim \mathbb{N}_2(\mathbf{0}, \mathbf{\Phi})$, where $\mathbf{\Phi}$ is a correlation matrix. The measurement equations for the distributional parameters of log-RT are:

$$\mu_i = \alpha_{i0,\mu} + \alpha_{i1,\mu} z_2 \tag{2.13}$$

$$\log(\sigma_i) = \alpha_{i0,\sigma} + \alpha_{i1,\sigma} z_2 \tag{2.14}$$

$$\text{logit}(\nu_i) = \alpha_{i0,\nu} + \alpha_{i1,\nu} z_2 \tag{2.15}$$

We estimated seven models of incremental complexity on our PISA 2018 dataset using the two-step full-information maximum likelihood procedure described in Section 2.3. To ensure robustness, we tested different starting values to check that the solution did not correspond to a local minimum. The initial parameter values resulted from a warm start, in which we first performed PCA on the matrix of observed variables, retained $q = 2$ principal components, and used them as explanatory variables in a series of independent GAMLSS regressions where the items are the response variables.

In all cases, the IRT model for the IR is the same as described above, with the measurement equation given in (2.12). For the log-RT model, the baseline model (Model 1) is similar to the hierarchical model in van der Linden (2007). In Model 1, we assume that the log-RT follow Normal distributions conditional on the speed factor, with factor loadings fixed to a value of -1 (as in van der Linden's paper). This means that an individual with higher values of the speed factor will tend to have shorter response times. We freely estimated the variance of the speed factor, and assumed the scale parameter to depend only on the constant term. Model 2 results from freely estimating the factor loadings in the log-RT model (as in the 'unrestricted model' in Molenaar et al. 2015), but fixing $\text{var}(z_2) = 1$ for identification purposes. Finally, Model 3 is the heteroscedastic version of Model 2.

Model 4 corresponds to a homoscedastic SN model, where the log-RT follow a SN distribution with a location parameter (μ) that depends on the speed factor, and with constant scale (σ) and shape (ν) parameters. In Models 5 and 6, we model $(\mu, \sigma)^\intercal$ and $(\mu, \nu)^\intercal$ as functions of z_2, respectively. Model 7, the full-SN model, allows all distributional parameters to depend on the speed factor. Results are presented in Table 2.6. Notably, all models that consider a SN distribution for the log-RT outperform the models with Normal distributions in terms of model fit. Model 7 provides the best fit based on its AIC and BIC values.

Table 2.7 presents the estimated intercepts, loadings, and factor correlation of Model 7. The interpretation of the intercepts and slopes in the measurement equations for the location parameter of the IR (π_i's) and log-RT (μ_i's) is straightforward. Specifically, the $\hat{\alpha}_{i0,\pi}$'s and $\hat{\alpha}_{i1,\pi}$'s represent the difficulty and discrimination parameters for the IR, respectively. More difficult items are those with more negative $\hat{\alpha}_{i0,\pi}$'s, and items with more discrimination power are those with higher $\hat{\alpha}_{i1,\pi}$'s. Thus, individuals with higher latent ability (i.e., higher values of z_1) will tend to have a higher probability of responding correctly to any given item.

For the response times, the $\hat{\alpha}_{i0,\mu}$'s represent the average response times (also known as the item intensity, as described in van der Linden, 2007), and the $\hat{\alpha}_{i1,\mu}$'s define the linear relationship between the speed factor and the log-RT. Note how all of the slopes in the measurement equation for

Model		AIC	BIC	K
1.	Bernoulli (π) + Normal (μ, fixed $\alpha_{i1,\mu}$)	26173.08	26368.96	38
2.	Bernoulli (π) + Normal (μ)	25908.67	26145.79	46
3.	Bernoulli (π) + Normal (μ, σ)	25754.91	26038.42	55
4.	Bernoulli (π) + Skew-Normal (μ)	25326.02	25609.53	55
5.	Bernoulli (π) + Skew-Normal (μ, σ)	25281.41	25611.30	64
6.	Bernoulli (π) + Skew-Normal (μ, ν)	25232.80	25562.70	64
7.	Bernoulli (π) + Skew-Normal (μ, σ, ν)	**25171.90**	**25548.18**	73

Table 2.6: PISA 2018: AIC and BIC for GLVM-LSS for the joint modelling of item responses and response times. In parenthesis: the distributional parameters modelled in terms of the latent variables, e.g., Bernoulli (π) + Skew-Normal (μ, σ, ν) means the probability of answering correctly depends on z_1; while the location, scale, and shape parameters of the SN distribution depend on z_2. $K = \dim(\hat{\Theta})$ is the number of parameters in the corresponding model.

the location parameter of the log-RTs are negative. In that sense, individuals with higher (positive) latent speed trait will have a tendency to respond faster to any given item, while individuals with lower (negative) latent speed trait will, on average, take longer to respond.

The estimated correlation between the latent ability and the speed factor is -0.28 (SE 0.025)[4], suggesting that test takers with higher latent ability generally take longer times to respond. This result aligns with previous studies on the speed-accuracy trade-off, which indicates that individuals who respond slowly make fewer mistakes compared to those who respond quickly and make more mistakes (see, e.g., van der Linden (2007), and Heitz (2014) for a general overview on the subject). Indeed, the estimated correlation coefficient between the total (sum) score and the average response times (in log-minutes) is 0.192 (95% confidence interval: 0.14, 0.24). Previous studies have found correlations between the latent ability and the latent speed trait of similar magnitude in the context of large scale educational testing of quantitative subjects (see, e.g., van der Linden and Guo, 2008).

The GLVM-LSS framework includes measurement equations for the scale (standard deviation) and shape (skewness) parameters of the log-RT. The estimates $\hat{\alpha}_{i1,\sigma}$ and $\hat{\alpha}_{i1,\nu}$ reveal that some items exhibit heteroscedasticity (items 2, 3, 4, 5, 8) and/or varying skewness (items 2, 3, 4, 6, 7, 8, 9) in their log-RT along the speed factor dimension. Figures 2.3 and 2.4 display the item characteristic curves (ICC) for the item responses and the fitted Skew-Normal conditional distributions for the log-RT (parameterized by the coefficients in Table 2.7) for selected items. For the log-RTs, we plot the (conditional) mean, median, and percentiles (0.025, 0.10, 0.25, 0.75, 0.90, and 0.975) to

[4]The estimated correlation and estimated standard error are similar for all 7 models: -0.28 (transformed from a covariance of -0.09 with SE equal to 0.008) for Model 1, -0.31 (SE 0.024) for Model 2, -0.29 (SE 0.024) for Model 3, -0.28 (SE 0.025) for Model 4, -0.29 (SE 0.025) for Model 5, and -0.28 (SE 0.025) for Model 6.

Item	Estimated coefficients in the equation for item responses (IR)				Estimated coefficients in the equations for (log-)response times (log-RT)											
	Location parameter (π_i)				Location parameter (μ_i)				Scale parameter (σ_i)				Shape parameter (ν_i)			
	$\hat{\alpha}_{i0,\pi}$		$\hat{\alpha}_{i1,\pi}$		$\hat{\alpha}_{i0,\mu}$		$\hat{\alpha}_{i1,\mu}$		$\hat{\alpha}_{i0,\sigma}$		$\hat{\alpha}_{i1,\sigma}$		$\hat{\alpha}_{i0,\nu}$		$\hat{\alpha}_{i1,\nu}$	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Item 1	0.64	(0.06)	0.79	(0.09)	0.19	(0.01)	-0.17	(0.01)	-0.95	(0.02)	-0.03	(0.02)	0.63	(0.13)	-0.03	(0.15)
Item 2	-0.47	(0.07)	1.03	(0.10)	0.30	(0.01)	-0.24	(0.01)	-0.89	(0.02)	-0.10	(0.02)	1.56	(0.15)	-0.78	(0.18)
Item 3	-0.04	(0.08)	1.95	(0.20)	0.42	(0.02)	-0.25	(0.01)	-0.61	(0.02)	-0.08	(0.02)	-1.07	(0.14)	0.81	(0.16)
Item 4	-0.69	(0.07)	0.96	(0.10)	0.45	(0.01)	-0.34	(0.01)	-0.87	(0.01)	-0.05	(0.02)	-1.45	(0.12)	-0.33	(0.15)
Item 5	-2.84	(0.21)	2.28	(0.24)	1.00	(0.02)	-0.36	(0.02)	-0.68	(0.02)	0.14	(0.02)	-1.04	(0.13)	-0.32	(0.20)
Item 6	-0.91	(0.06)	0.32	(0.08)	0.16	(0.01)	-0.36	(0.01)	-0.97	(0.02)	0.03	(0.03)	0.11	(0.11)	-0.88	(0.21)
Item 7	-4.79	(0.42)	2.49	(0.32)	0.65	(0.01)	-0.33	(0.01)	-1.15	(0.02)	-0.04	(0.02)	0.35	(0.14)	-0.58	(0.16)
Item 8	-3.67	(0.30)	2.39	(0.28)	1.02	(0.01)	-0.39	(0.01)	-1.02	(0.02)	0.12	(0.02)	-1.22	(0.17)	-1.60	(0.26)
Item 9	-2.73	(0.16)	1.46	(0.16)	0.58	(0.01)	-0.30	(0.01)	-0.90	(0.02)	-0.00	(0.02)	0.30	(0.10)	0.36	(0.13)

Estimated latent correlation (z_1, z_2): $\hat{\phi}_{\mathbf{z}} = -0.28$ (SE: 0.025)

Table 2.7: PISA 2018: Estimated coefficients (Est.) and Standard Errors (SE) for joint model of item responses and response times (Model 7).

demonstrate how the distribution's shape changes across the speed factor dimension.

Figures 2.3a and 2.3b for item 2, and Figures 2.3c and 2.3d for item 3, show how the (conditional) variance of $\log(t_2)$ and $\log(t_3)$ drops as we move along the latent speed factor dimension. However, the conditional skewness changes in opposite directions. Specifically, $\log(t_2)$ is positively (right) skewed for individuals in the left tail of the latent speed factor, while $\log(t_3)$ is negatively (left) skewed for the same group of students. This might suggest differences on the items' characteristics (e.g., wording, task, difficulty) and/or the cognitive processes required for their completion. On the other hand, the distributions for the response times are very much symmetric for individuals on the right tail of the speed factor dimension.

For Items 5 (Figures 2.4a and 2.4b) and 8 (Figures 2.4c and 2.4d), the estimated positive slope for the scale parameter suggests that response times will be more heterogeneous for individuals on the upper tail of the latent speed factor distribution. These items are among the most difficult ones (higher $\hat{\alpha}_{i0,\pi}$'s) and also require more time on average (higher $\hat{\alpha}_{i0,\mu}$'s). Moreover, they also exhibit varying skewness parameters. For example, for Item 8, the direction of the skewness changes depending on the location along the latent speed factor scale.

2.5.2 ANES 2020: Thermometer items

In our second empirical example, we use data from the American National Election Study (ANES)[5]. The data set includes thirteen post-election feeling thermometer items from the ANES 2020 survey. Participants were asked to rate their feelings towards different social, religious, gender and sexuality, and economic groups or collectives on a scale of 0 to 100 (degrees). Higher ratings indicate more favourable attitudes towards the group, while lower ratings correspond to an unfavourable position. Ratings around 50 are indicative of more neutral attitudes. The selected items cover a broad range of topics, including religious groups (Christian fundamentalists, Christians, Muslims, Jews), sexual orientation and gender identity groups (Gay men and Lesbians, Transgender people), social and political movements (Feminists, #MeToo and BLM movements), and groups related to economic matters and professions (labour unions, big businesses, journalists, scientists).

While ANES thermometer items have been used in the literature as proxy variables for political orientation and measures of personal and societal values (see, e.g., Abelson et al., 1982; Krasa and Polborn, 2014; Guth, 2019), we emphasise that these items were not part of a psychomet-

[5] American National Election Studies, 2021. (www.electionstudies.org). Full Release (dataset and documentation). July 19, 2021 version. These materials are based on work supported by the National Science Foundation under grant numbers SES-1444721, 2014-2017, the University of Michigan, and Stanford University.

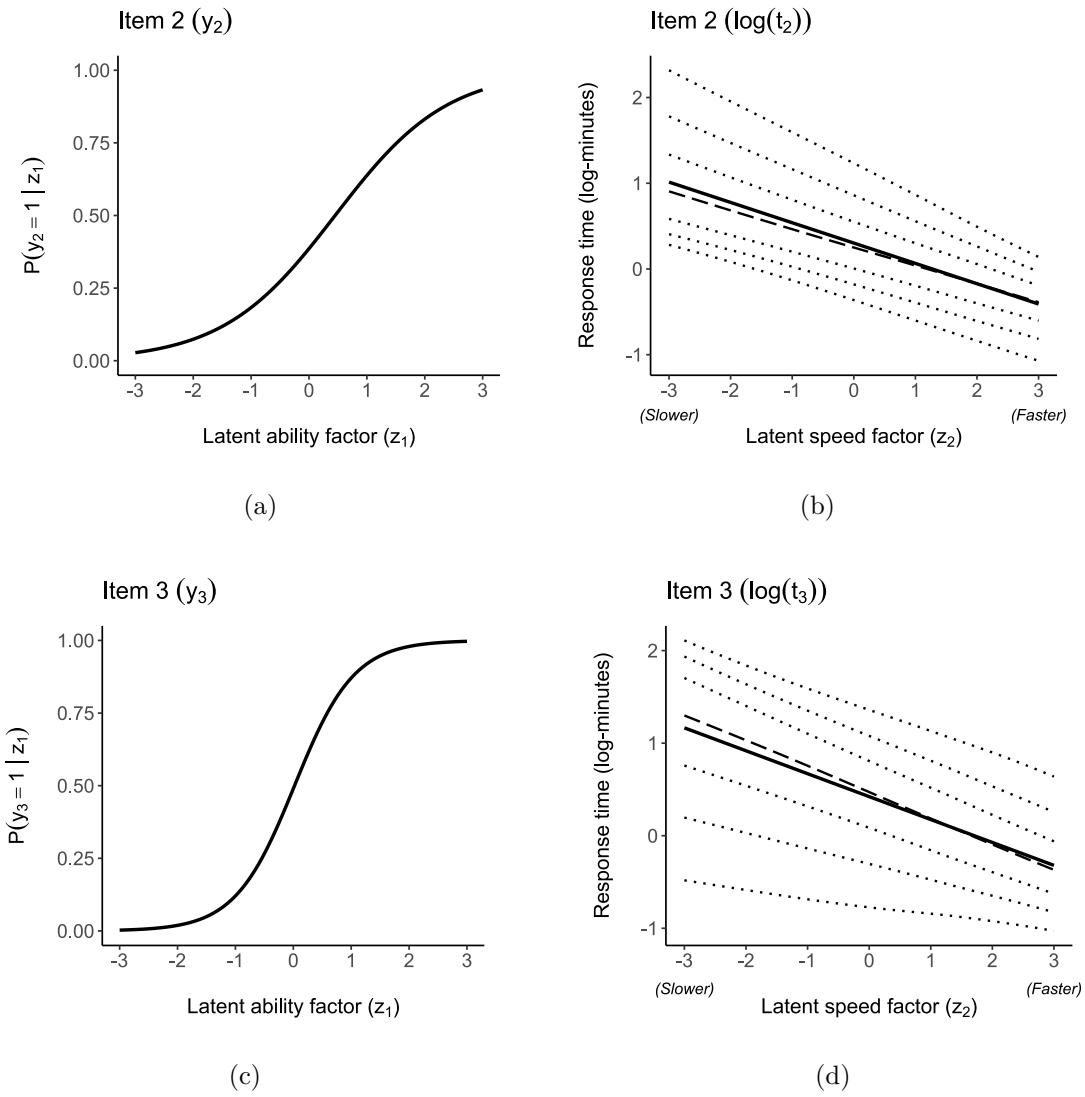


Figure 2.3: PISA 2018: Fitted conditional expected values (solid line, ——), median (dashed line, - - -), and percentiles (dotted lines, ·····) for IR and log-RT for items 2 and 3.

rics test and are not intended to measure any specific latent construct(s). Nonetheless, we believe that these questions can provide valuable information about an individual's position on a *conservative-progressive* scale. Therefore, we only present results for unidimensional Beta latent variable models[6].

Given the nature of the data, we assume a Beta distribution for the items, conditional on the latent factor: $y_i \mid \mathbf{z} \sim \text{Beta}(\mu_i(\mathbf{z}), \sigma_i(\mathbf{z}))$. We scale the items by a factor of $1/100$. As is customary in the literature (e.g., Noel, 2014), we replace extreme responses on the boundaries of the interval with numerical values arbitrarily close to 0 and 1 (e.g., 1^{-3} and $(1 - 1^{-3})$, respectively) to ensure that the values are within the interval $(0, 1)$. We exclude individuals with no post-election data,

[6]We also explored two-dimensional Beta factor models, but careful analysis of the expected information matrix (evaluated at the MLE) reveals that the heteroscedastic Beta factor model with $q = 2$ is not of full rank, and thus, following the results in Section 1.2, the model is not identified.

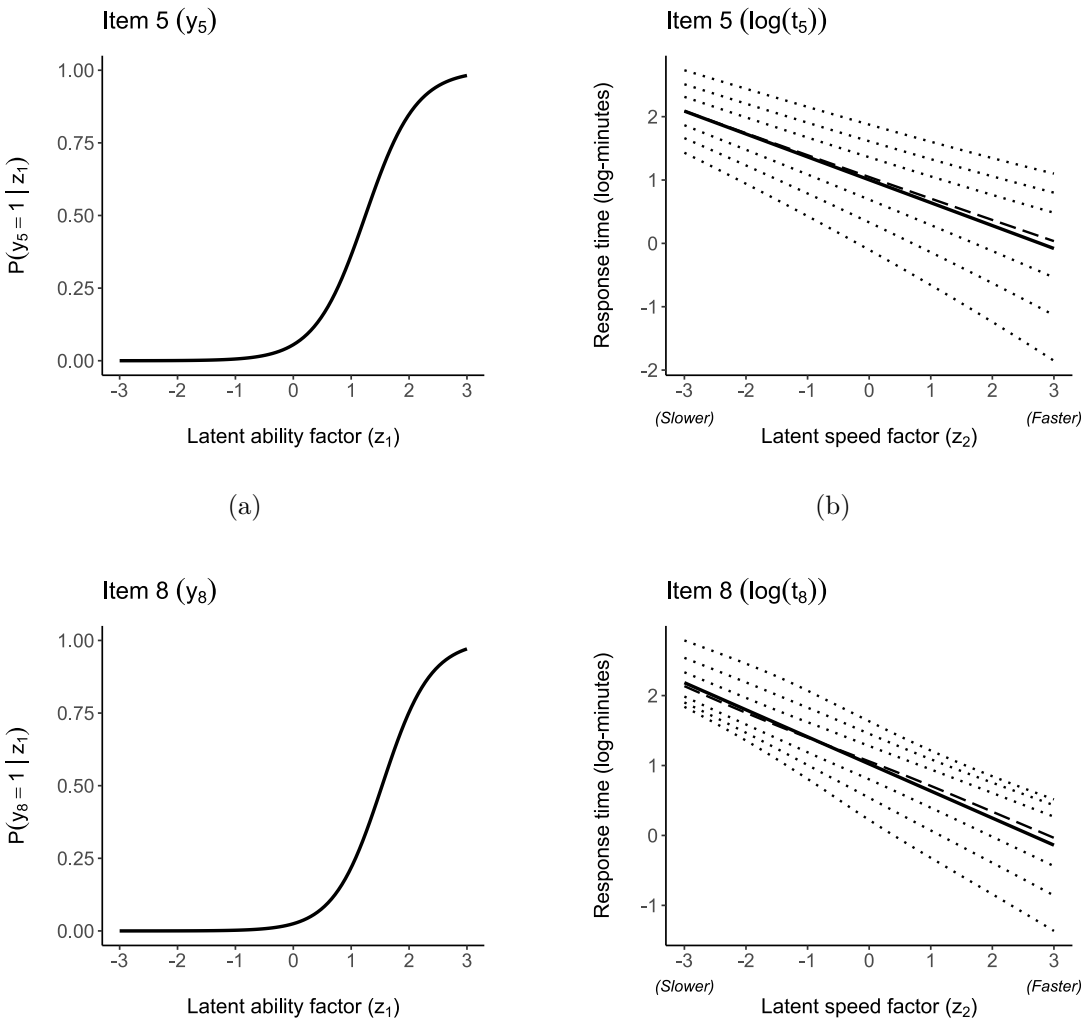


Figure 2.4: PISA 2018: Fitted conditional expected values (solid line, ——), median (dashed line, - - -), and percentiles (dotted lines, ·····) for IR and log-RT for items 5 and 8.

incomplete interviews, or technical errors in their answers from the analysis. Moreover, we treat responses of *Don't know'*, *Don't recognise'*, and *'Refuse'* as missing data. After applying these criteria, the resulting sample consists of 7253 respondents.

Item descriptive statistics for the ANES 2020 dataset are presented in Table 2.8. Most items have negatively skewed marginal empirical distributions and negative excess kurtosis, with the exception of item *Scientists*. The empirical cumulative distribution functions (ECDF) for the items are displayed in Figure 2.5. Although the thermometer ratings are measured on a continuous scale, respondents tend to round their answers to the nearest 5 or 10, resulting in a stepped appearance in the ECDFs. We observe no substantial inflation of extreme responses, but some items exhibit a higher proportion of responses around the middle point of the thermometer (around 0.5). This could be due to respondents being unwilling to commit to an opinion that places them on either

Item	Count	Mean	SD	SK	KU
Christian fundamentalists	7040	0.46	0.29	0.00	-0.73
Christians	7175	0.72	0.25	-0.69	-0.11
Muslims	7126	0.59	0.25	-0.26	0.03
Jews	7127	0.74	0.22	-0.47	-0.24
Gay men and Lesbians	7149	0.66	0.27	-0.51	-0.17
Transgender people	7139	0.60	0.28	-0.34	-0.34
Feminists	7159	0.59	0.27	-0.35	-0.40
#MeToo movement	6030	0.59	0.30	-0.45	-0.63
BLM movement	7176	0.53	0.36	-0.26	-1.30
Labour unions	7148	0.58	0.24	-0.29	-0.09
Big businesses	7168	0.48	0.23	-0.18	-0.13
Journalists	7196	0.51	0.29	-0.26	-0.92
Scientists	7193	0.79	0.20	-1.01	0.94

Table 2.8: ANES 2020: Item descriptive statistics. *Count* is the number of observed responses for each item, *SD* is the standard deviation, *SK* is the skewness, and *KU* is the excess kurtosis.

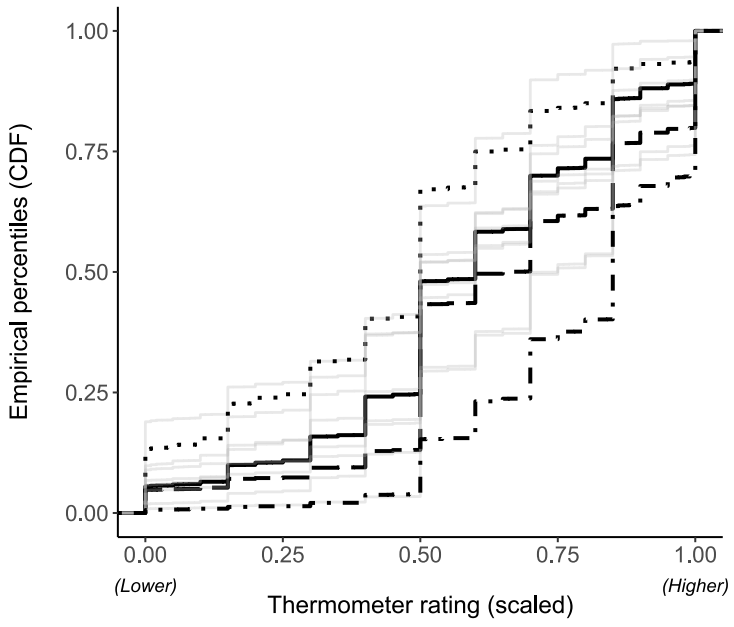


Figure 2.5: ANES 2020: Empirical cumulative distribution function (ECDF). Highlighted items: *Feminists* (solid line, ——), *Gay men and Lesbians* (dashed line, - - -), *Christian fundamentalists* (dotted line, ·····), and *Scientists* (dash-dot line, -·-·)

side of the *conservative-progressive* scale, or because the items relate to sensitive topics, such as religion or politics. In such cases, respondents may choose to respond 50%' instead of selecting an uninformative option, such as '*Don't know*' or '*Refuse*'. A potential avenue for future research could be the incorporation of latent classes associated with different response strategies, building on approaches proposed for latent variable models with binary items (e.g., Moustaki and Knott, 2014).

We estimated two Beta factor models to compare their performance. The starting values for the parameters in the estimation algorithm were set using the warm start strategy described in the previous section. We tried different starting values (warm start strategy plus noise coming from Unif$(-1, 1)$) to explore if the solution corresponded to a local minimum, but obtained similar results in each case (up to numerical precision). In the baseline (homoscedastic) model, we assumed a constant scale parameter, while in the alternative (heteroscedastic) model, we allowed the scale parameter to depend on the latent factor. We assessed model fit using both AIC and BIC, and the results are presented in Table 2.9. The heteroscedastic model showed better model fit, suggesting that modelling heteroscedasticity is important in this dataset. Table 2.10 provides parameter estimates and their standard errors for the heteroscedastic model. These results confirm that thermometer items measure individuals' beliefs along a '*conservative-progressive*' scale.

Model	AIC	BIC	K
Beta (μ)	-95075.12	-94806.44	39
Beta (μ, σ)	**-96805.52**	**-96447.28**	52

Table 2.9: ANES 2020: AIC and BIC for the Beta factor models. In parenthesis: the distributional parameters modelled in terms of the latent variables, e.g., Beta (μ, σ) means both the location and scale parameters depend on the latent *conservative-progressive* factor. $K = \dim(\hat{\Theta})$ is the number of parameters in the corresponding model.

The estimated factor loadings on the location measurement equation, $\hat{\alpha}_{i1,\mu}$, indicate that most items have positive loadings, suggesting that individuals who are more progressive (conservative) tend to rate these groups higher (lower) on average. However, items related to *Christian fundamentalists*, *Christians*, and *Big businesses* have negative loadings with lower magnitudes. Thus, more progressive (conservative) individuals tend to rate them lower (higher) on average. However, the discrimination power of these negative items is lower than that of other items, as indicated by the absolute value of their factor loadings. This implies that individuals with different positions on the latent dimension could still report similar ratings for these items. Furthermore, the intercepts for the location parameter ($\hat{\alpha}_{i0,\mu}$'s) have the same interpretation as an IRT difficulty parameter.

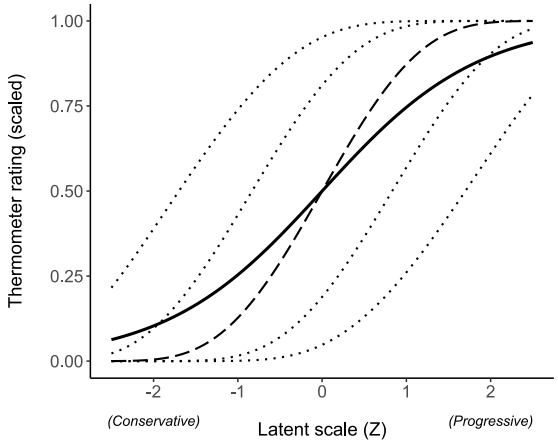
Item	Location parameter (μ) measurement equation				Scale parameter (σ) measurement equation			
	$\alpha_{i0,\mu}$		$\alpha_{i1,\mu}$		$\alpha_{i0,\sigma}$		$\alpha_{i1,\sigma}$	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Christian fundament.	-0.19	(0.02)	-0.47	(0.02)	0.67	(0.01)	0.05	(0.01)
Christians	0.96	(0.02)	-0.26	(0.02)	0.59	(0.01)	0.07	(0.01)
Muslims	0.41	(0.01)	0.98	(0.02)	0.01	(0.01)	-0.10	(0.01)
Jews	1.15	(0.02)	0.51	(0.02)	0.29	(0.01)	-0.16	(0.01)
Gay men and Lesbians	0.90	(0.02)	1.31	(0.02)	-0.06	(0.01)	-0.27	(0.01)
Transgender people	0.55	(0.01)	1.37	(0.02)	-0.12	(0.02)	-0.17	(0.01)
Feminists	0.45	(0.01)	1.21	(0.02)	-0.10	(0.01)	-0.16	(0.01)
#MeeToo movement	0.41	(0.02)	1.26	(0.02)	0.15	(0.02)	-0.28	(0.01)
BLM movement	0.21	(0.02)	1.23	(0.02)	0.53	(0.01)	-0.33	(0.01)
Labour Unions	0.39	(0.01)	0.62	(0.01)	0.21	(0.01)	-0.13	(0.01)
Big Businesses	-0.17	(0.01)	-0.14	(0.01)	0.26	(0.01)	0.05	(0.01)
Journalists	0.02	(0.01)	0.93	(0.02)	0.23	(0.01)	-0.17	(0.01)
Scientists	1.58	(0.02)	0.82	(0.02)	0.02	(0.01)	-0.25	(0.01)

Table 2.10: ANES 2020: Estimated (Est.) coefficients and their standard errors (SE) for the heteroscedastic Beta factor model.

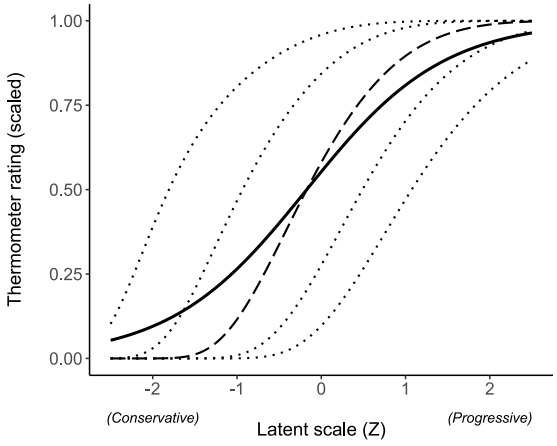
An important feature of this dataset is that the scale parameter σ_i, which is related to the conditional variance of the item, also varies along the latent scale. Items with positive $\alpha_{i1,\mu}$'s have negative $\alpha_{i1,\sigma}$'s, indicating that individuals on the 'progressive' side of the latent scale will tend to show less variance in their ratings when compared to individuals on the 'conservative' side of the scale, for these items. This is an interesting result, suggesting that individuals on the 'conservative' side may be more diverse in their views, attitudes, or beliefs, when it comes to certain topics, while individuals on the 'progressive' side may hold more homogeneous views, at least for some items. For the 'negative' items, the factor loadings for the scale parameter are much lower (in absolute value), suggesting very low (to no) heteroscedasticity.

Figure 2.6 compares the fitted (conditional) Beta distribution implied by the homoscedastic and heteroscedastic models for selected items. Apart from the fitted mean, note how the median and percentiles (0.10, 0.25, 0.75, 0.90) of the items distributions change along the latent scale. The homoscedastic Beta factor model (Figures 2.6a and 2.6c) does not capture the asymmetries in the items' conditional distributions along the latent scale that the heteroscedastic model does (Figures 2.6b and 2.6d). For items that tend to be homoscedastic the differences between the two fitted models are not significant (e.g, ratings for *Christians*, Figures 2.6e and 2.6f).

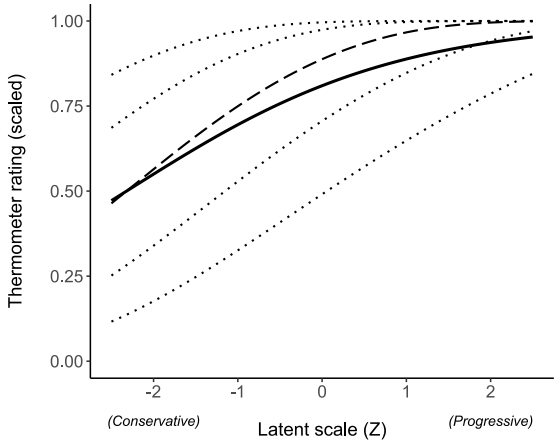
Using the estimated parameters, we also calculate factor scores using the empirical Bayes (EB)



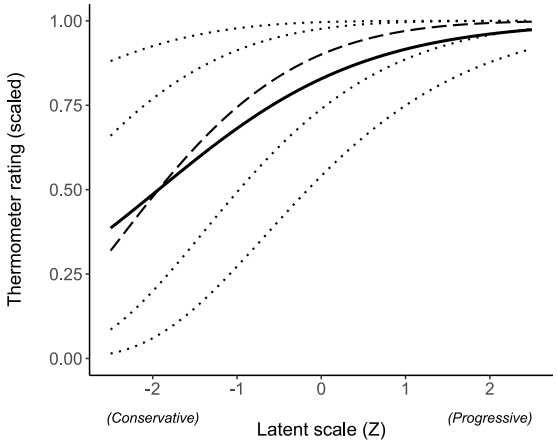
(a) Item: BLM (homoscedastic model)



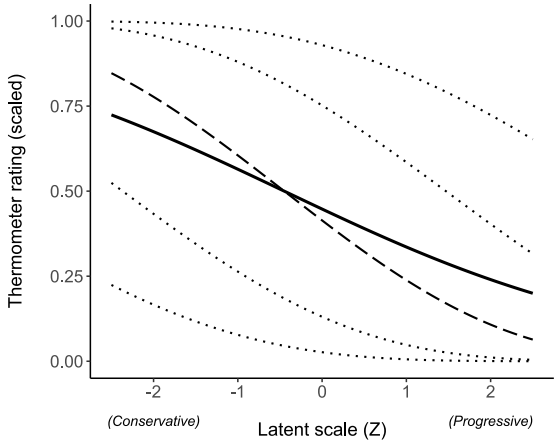
(b) Item: BLM (heteroscedastic model)



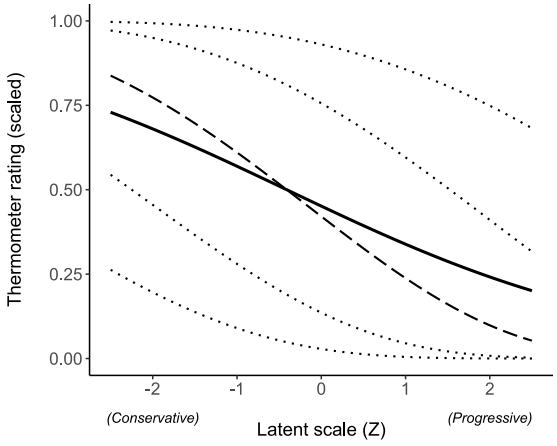
(c) Item: Scientists (homoscedastic model)



(d) Item: Scientists (heteroscedastic model)

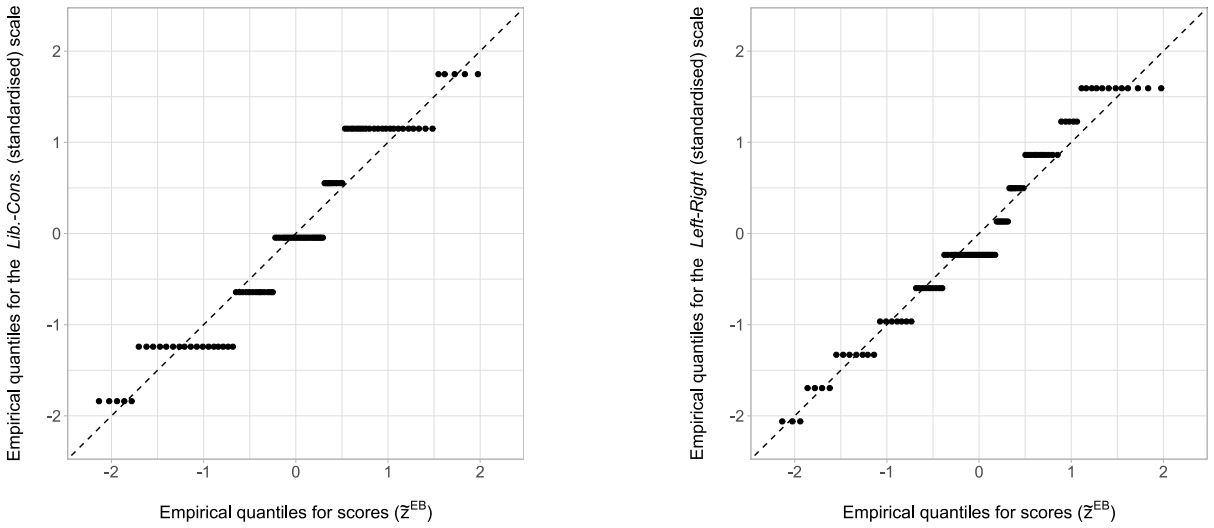


(e) Item: Christians (homoscedastic model)



(f) Item: Christians (heteroscedastic model)

Figure 2.6: ANES 2020: Fitted conditional expected values (solid line, ——), median (dashed line, - - -), and percentiles (dotted lines, ·····).



(a) *Liberal-Conservative* scale

(b) *Left-Right* scale

Figure 2.7: ANES 2020: Empirical QQ-plots of (standardised) political orientation scales against Empirical Bayes factor scores (sign reversed).

method described in Section 1.4. To ensure the consistency of the factor scores, we cross-examine them with two self-reported measures of individuals' political orientation. The first measure is a 7-point *liberal-conservative* scale, which reads: *"We hear [...] about liberals and conservatives. Here is a seven-point scale on [...] political views. Where would you place yourself on this scale, or haven't you thought much about this?"*. The scale ranges from 1 (extremely liberal) to 7 (extremely conservative), and we consider the 'haven't thought about this' response as missing values, resulting in 985 observations. The second measure is an 11-point *left-right* scale, which reads: *"In politics people sometimes talk of left and right. Where would you place yourself on a scale from 0 to 10 where 0 means the left and 10 means the right?"*. Technical errors in the interview cause 17 missing values for this measure. For better comparison, we switch the sign of the factor scores in the subsequent analysis. The Pearson correlations between the factor scores and the *liberal-conservative* and *left-right* scales are 0.65 (95% confidence interval: 0.64, 0.67) and 0.56 (95% confidence interval: 0.54, 0.57), respectively. The empirical QQ-plots for these scales and the factor scores are displayed in Figure 2.7. The plots indicate that the factor scores obtained from the heteroscedastic Beta factor model are consistent with the (standardised) self-reported measures of political orientation.

2.6. Discussion

In this Chapter, we propose a distributional approach to latent variable modelling. We present a class of Generalised Latent Variable Models for Location, Scale, and Shape parameters (GLVM-

LSS). In this framework, we model the distributional parameters characterising the items' conditional distributions as linear functions of the latent variables. By modelling the whole conditional distribution in terms of the latent variables, rather than just the conditional mean, the GLVM-LSS framework captures a wider range of characteristics of the data, making it more flexible and comprehensive than traditional LVMs. Our approach allows for modelling distributions beyond the exponential family and is useful in real-world applications where the items display heteroscedasticity, skewness, kurtosis, zero/one/maximum value inflation, and heaping, truncation or censoring. The GLVM-LSS class can be viewed as an umbrella class of LVMs that includes previous works addressing these issues as particular cases.

The parameters of the GLVM-LSS model are estimated via full information maximum likelihood using a computationally efficient two-stage optimisation procedure that combines the EM-algorithm with a direct maximisation of the marginal log-likelihood through (quasi-)Newton methods. To demonstrate the effectiveness of our approach, we present simulation studies and empirical analyses of real-world data in psychometrics and public opinion research. Our proposed method is implemented in `R`, with code and replication files available online. Our current software implementation allows the analyst to estimate GLVM-LSS with mixed data from the Bernoulli, Poisson, Normal, Log-Normal, Skew-Normal, Gamma, Beta, and Zero-Inflated Poisson distributions. Future extensions will include distributions for categorical items (e.g., Moustaki, 2003), survival data (e.g., Moustaki and Steele, 2005), and continuous and discrete items displaying heaping or truncation/censoring (see, e.g., Dolan et al., 2002; Wall et al., 2015b; Magnus and Thissen, 2017). Furthermore, we could include covariate effects on the manifest and/or latent variables, as this would help to explore how the distributional parameters vary with observed covariates and links with important concepts in psychometrics such as differential item functioning and measurement invariance.

While the GLVM-LSS provides greater flexibility and generalisability compared to previous models, there are still some aspects that can be improved in future research. One challenge of the proposed framework is that as the models become more complex, they can be harder to interpret. The increased flexibility of modelling distributional parameters in terms of latent variables also leads to a significant increase in the number of parameters to be estimated, which can pose a problem in exploratory settings. To address this, we suggest exploring regularised maximum likelihood estimation of the GLVM-LSS to produce more sparse and interpretable factor loading matrices. This approach can also perform model selection by appropriately selecting the regularisation parameter (see, e.g., Geminiani et al., 2021).

Another limitation is the numerical integration method used. While the Gauss-Hermite (GH)

rule with fixed quadrature points is easy to implement and works well for problems with a few latent variables, it may not be adequate for higher dimensional settings. Adaptive GH rules (Schilling and Bock, 2005; Skrondal and Rabe-Hesketh, 2004) or Laplace approximations (Huber et al., 2004; Bianconcini and Cagnone, 2012) can be considered as alternatives, as they have been shown to produce fast and accurate solutions.

Lastly, local model misfit is an important issue that needs to be addressed in future research. This is particularly relevant as assumptions about the items' distributions and the linearity in the measurement equations can be easily violated. To identify local misfit, developing residual-based diagnostic tools can be a reasonable approach (see, e.g., Pan and Lin, 2005; Sánchez et al., 2009).

Chapter 3

Penalised Marginal Maximum Likelihood Estimation with Automatic Selection of Tuning Parameters for Generalised Latent Variable Models for Location, Scale, and Shape parameters

In this chapter, we propose a penalised marginal maximum likelihood estimation procedure for the Generalised Latent Variable Model for Location, Scale, and Shape parameters (GLVM-LSS). In some applications, the GLVM-LSS can be unnecessarily complex and end up over-fitting observed data. For example, in a given test, some items might have scale and shape parameters that depend on the latent variables, while other items might not. Without addressing this issue, these factor loadings would be estimated regardless, unnecessarily adding to the model complexity and complicating interpretation. To address these issues and achieve simpler and more interpretable factor loading matrices, we estimate the model parameters via penalised marginal maximum likelihood. The amount of penalisation is determined by a tuning parameter, which controls the sparsity level of the penalised maximum likelihood solution. Typically, selecting the optimal value for the tuning parameter involves computationally intensive techniques like grid-search or cross-validation, followed by choosing the value that yields the lowest information criteria. In the GLVM-LSS we

penalise the location, scale, and shape measurement equations separately, and thus the tuning parameter becomes a vector rather than a scalar. Consequently, conventional approaches such as grid-search or cross-validation become impractical. We propose an automatic and data-driven procedure that gives the optimal value for the tuning parameter vector by minimising an approximation of an information criteria with a theoretically grounded definition of model complexity. The properties of the proposed estimation framework are demonstrated through simulation studies and empirical applications in educational testing.

3.1. Introduction

In Chapter 2, we introduced the Generalised Latent Variable Model for Location, Scale, and Shape parameters (GLVM-LSS). The proposed framework extends the traditional Generalised Linear Latent Variable Model (GLLVM, Skrondal and Rabe-Hesketh, 2004; Bartholomew et al., 2011) by explicitly modelling the location, scale, and shape parameters characterising the items' conditional distributions as linear functions of the latent variables. By modelling the whole conditional distribution in terms of the latent variables, rather than just the conditional mean, the GLVM-LSS captures a wider range of characteristics of the data, making it more flexible and comprehensive than traditional LVMs. The GLVM-LSS allows for modelling distributions beyond the exponential family and is useful in real-world applications where the items exhibit characteristics such as heteroscedasticity, skewness, kurtosis, inflation at specific values (e.g., zeros, ones, or maximums), and phenomena like heaping, truncation or censoring. The GLVM-LSS can be viewed as an umbrella class of LVMs that includes previous works addressing these issues as particular cases.

In most LVM applications, it is important to obtain sparse factor loading matrices that exhibit a 'simple structure' (Thurstone, 1947). This means that observed items should primarily load on one or a few latent factors with high absolute values, known as primary loadings. The remaining loadings on other latent variables, referred to as cross-loadings, should be close to zero or very small. Latent variables are ideally measured by distinct subsets of observed variables with high primary loadings and low cross-loadings. Solutions following a simple structure make the interpretation of the latent variables easier and are preferred when aiming for a concise explanation of the substantive hypothesis under study.

There are two main approaches for obtaining sparse factor loading matrices in the LVM literature: rotation methods and penalised estimation methods. Rotation methods involve a two-step process. First, an estimate of the factor loading matrix is obtained, typically using maximum likelihood estimation, while imposing appropriate identification constraints to address rotational

indeterminacy (discussed in Section 1.2). In the second step, the estimated factor loading matrix is rotated to minimise a loss function related to sparsity. A lower value of the loss function indicates a more interpretable solution. Various loss functions have been proposed in the literature (Mulaik, 2009, Chapters 10-12). Rotation methods can be classified as orthogonal when latent variables are required to be uncorrelated (e.g., Jennrich, 2001, 2004), or as oblique when latent variables are allowed to be correlated (e.g., Jennrich, 2002, 2006; Liu et al., 2023). It has been proven that rotation methods with L_1-norm loss functions can produce perfect simple structures if they exist in the true factor loading matrix (Jennrich, 2006, Theorem 1). However, in practice, the rotated solutions often end up being dense and it is then left to the analyst to decide which parameters should be considered as zero based on a subjective thresholding criterion. For example, factor loadings below a fixed threshold (usually set between $\pm$ 0.1 or $\pm$ 0.3) are set to zero using a hard-thresholding approach (Hair et al., 2010). This approach is not only subjective in nature, but also affects the effective degrees of freedom of the model, which are used to evaluate model fit. Additionally, different rotation techniques employ different objective functions in their optimisation problems, leading to different factor structures that may not always be directly comparable.

The second approach for obtaining sparse factor loading matrices involves penalised estimation methods. In these methods, a sparsity-inducing L_1-penalty term is introduced into the objective function of the estimation problem. Penalised methods simultaneously estimate the model parameters and generate a sparse solution. Unlike rotation methods, simple structures are not imposed on the factor loading matrix and are only obtained if supported by the data. For multivariate Normal items, penalised factor models have been proposed by Choi et al. (2010); Hirose and Yamamoto (2014, 2015); Trendafilov et al. (2017) and Jin et al. (2018). Penalised Structural Equation Models (PSEM) have been studied by Jacobucci et al. (2016); Huang et al. (2017) and Huang (2018). For binary and categorical items, related works include Chen et al. (2015); Sun et al. (2017) and Battauz (2020). Huang (2020) extended the PSEM to ordinal responses. Count data has been addressed in the works of Hui et al. (2018) and Lee and Park (2021). The level of sparsity is determined by a non-negative tuning (or regularisation) parameter, which serves as the weight of the penalty term in the estimation objective function. A higher tuning parameter produces sparser solutions but leads to a loss of model fit due to the introduction of an asymptotically decaying bias term. The optimal penalised solution is obtained by fitting a sequence of candidate models with different values of the tuning parameter, typically selected from a predefined grid. The final model is chosen based on information criteria or cross-validation error assessment. However, for complex models with multiple tuning parameters, this process can be computationally expensive and time-consuming, which limits the practicality of these methods in applied research. Therefore, more efficient alternatives are needed to determine the appropriate value of the tuning parameter

that balances fidelity to the data (goodness of fit) and interpretability (sparsity).

The issue of selecting optimal tuning parameters has been addressed by Geminiani et al. (2021). They proposed a penalised estimation framework for single- and multiple-group Normal linear factor models and introduced an efficient data-driven procedure for selecting the optimal value of the tuning parameter. This procedure is based on minimising an approximate unbiased risk estimate (Stein, 1981), which is proportional to the AIC. See also Wood (2004); Marra et al. (2017). To solve the optimisation problems required by the procedure, Geminiani et al. (2021) employed smooth approximations of the L_1-penalty functions. These approximations behave locally as the original L_1-norm penalties, and as a result, there is minimal loss in the sparsity-inducing properties of the approximate penalty functions. Their framework provides a unified estimation and inferential framework for penalised estimation of Normal linear factor models.

In the GLVM-LSS framework, achieving sparse factor loading matrices and having an efficient method for selecting the optimal tuning parameters are two crucial methodological aspects to bear in mind due to the following reasons. Firstly, sparse solutions with 'simple structures' enhance the interpretation of factor loading matrices by capturing the relationship between the latent factors and the higher order moments of observed items. Secondly, incorporating a penalty term helps prevent over-fitting in the GLVM-LSS, which is important considering the increased complexity resulting from modelling the entire conditional distribution rather than just the conditional mean. Following the logic of the bias-variance trade-off (Hastie et al., 2009), the penalised estimates lead to a more 'generalisable' model for prediction and assigning factor scores to new samples. Moreover, the penalised estimation can be interpreted as a model selection technique, and thus the penalised solution will distinguish between items with constant scale and shape parameters, and items whose distributional parameters and higher order moments vary with the latent variables. Lastly, it is desirable to have independent control over the amount of penalisation for the location, scale, and shape factor loading matrices in the items' conditional distributions. This requires a vector of tuning parameters, unlike the scalar approach commonly used in penalised estimation in the LVM literature. Consequently, grid-search methods become impractical for determining the optimal tuning parameters in practice.

In this chapter, we introduce a penalised estimation framework with automatic tuning parameter selection for the class of Generalised Latent Variable Models for Location, Scale, and Shape parameters (GLVM-LSS). Our framework can be seen as an extension of the approach in Geminiani et al. (2021) to LVMs that do not have closed-form solutions. Specifically, we propose a penalised marginal maximum likelihood estimation method using locally approximated L_1 penalties. The optimal tuning parameter vector, which controls the level of sparsity in the model parameters, is

selected by minimising an approximate AIC criterion. The remainder of this chapter is organised as follows. In Section 3.2 we present a brief description of the GLVM-LSS framework, previously introduced in Chapter 2. In Section 3.3 we discuss the penalised estimation framework. In particular, Section 3.3.1 describes common convex L_1-norm based penalty functions, and Section 3.3.2 discusses how by using local approximations of the penalty terms, we are able to adapt the two-step maximum likelihood estimation strategy introduced in Section 2.3 to the penalised case. Section 3.4 presents a Generalised Information Criterion for model selection and assessing goodness-of-fit. Section 3.5 discusses the automatic, data-driven procedure for estimating the optimal value of the tuning parameter vector. We finish by presenting simulation studies in Section 3.6 and empirical applications to educational testing in Section 3.7.

3.2. Model description

Consider the Generalised Latent Variable Model for Location, Scale, and Shape parameters (GLVM-LSS) framework introduced in Chapter 2. We denote the vector of observed variables as $\mathbf{y} = (y_1, ..., y_p)^\intercal \in \mathbb{R}^p$, and the vector of latent variables as $\mathbf{z} = (z_1, ..., z_q)^\intercal \in \mathbb{R}^q$, where $q \ll p$. The density of the observed data is $f(\mathbf{y}) = \int_{\mathbb{R}_q} f(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}$. The conditional distribution $f(\mathbf{y} \mid \mathbf{z})$, referred to as the measurement component of the LVM, describes the relationship between the observed variables $\mathbf{y}$ and the latent variables $\mathbf{z}$. The structural component $p(\mathbf{z})$ specifies the relationships among the latent variables. Assuming conditional independence, we consider the observed variables to be conditionally independent given $\mathbf{z}$. We also assume that the conditional distributions of the items follow a known parametric form parameterised by $\boldsymbol{\theta}_i = (\mu_i, \sigma_i, \nu_i, \tau_i)^\intercal$, representing the location, scale, and shape parameters for item i. For the structural model, we adopt a multivariate Normal distribution, $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Phi})$, which is commonly used in the LVM literature due to its mathematical and computational convenience (see Bartholomew et al., 2011, Chapter 2). The marginal density of the observed data follows is (as in eq. 2.1):

$$f(\mathbf{y}) = \int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_i \mid \mathbf{z}; \boldsymbol{\theta}_i) \right] p(\mathbf{z}; \boldsymbol{\Phi}) \, \mathrm{d}\mathbf{z}$$

In the GLVM-LSS framework, the distributional parameters φ_i in $\boldsymbol{\theta}_i$ are modelled as linear functions of the latent variables. Specifically, for an arbitrary location, scale, or shape parameter, we define a corresponding measurement equation:

$$\upsilon_{i,\varphi}(\varphi_i) = \alpha_{i0,\varphi} + \sum_{j=1}^{q} \alpha_{ij,\varphi} z_j,$$

where $\upsilon_{i,\varphi}$ represents a parameter-specific link function (e.g., identity, log, logit, etc.) chosen to ensure appropriate restrictions on the distributional parameters. The intercepts in the measurement equation are denoted by $\alpha_{i0,\varphi}$, and the factor loadings (slopes) by $\alpha_{ij,\varphi}$, where $j = 1, ..., q$. The vector of factor loadings for parameter φ_i is denoted as $\boldsymbol{\alpha}_{i,\varphi} = (\alpha_{i1,\varphi}, ..., \alpha_{iq,\varphi})^\intercal$. The sub-index (i, φ) indicates that the corresponding function or regression parameter is defined for $\varphi_i \in \boldsymbol{\theta}_i$. It is important to note that, in sparse settings, some of the factor loadings in $\boldsymbol{\alpha}_{i,\varphi}$ can be equal to zero.

In matrix notation, the set of measurement equations for $\varphi \in \boldsymbol{\theta}$ can be written as $\upsilon_\varphi(\boldsymbol{\varphi}) = \boldsymbol{\alpha}_{0,\varphi} + \mathrm{A}_\varphi \mathbf{z}$, where $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_p)^\intercal$ is the vector containing the same distributional parameter φ for all items, $\boldsymbol{\alpha}_{0,\varphi} = (\alpha_{10,\varphi}, ..., \alpha_{p0,\varphi})^\intercal$ is a vector of intercept terms, A_φ is a sparse $(q \times p)$ matrix with rows corresponding to the factor loading vectors $\boldsymbol{\alpha}_{i,\varphi}$, and υ_φ is the vector function that applies the corresponding link function $\upsilon_{i,\varphi}$ to each entry of $\boldsymbol{\varphi}$.

To further simplify notation, we can express the system of all location, scale, and shape measurement equations as $\upsilon(\boldsymbol{\theta}) = \boldsymbol{\alpha}_0 + \mathrm{A}\mathbf{z}$, where $\boldsymbol{\theta}^\intercal = (\boldsymbol{\mu}^\intercal, \boldsymbol{\sigma}^\intercal, \boldsymbol{\nu}^\intercal, \boldsymbol{\tau}^\intercal)$ represents the vector of all distributional parameters, $\boldsymbol{\alpha}_0^\intercal = (\boldsymbol{\alpha}_{0,\mu}^\intercal, \boldsymbol{\alpha}_{0,\sigma}^\intercal, \boldsymbol{\alpha}_{0,\nu}^\intercal, \boldsymbol{\alpha}_{0,\tau}^\intercal)$ is the vector of intercepts, and $\mathrm{A}^\intercal = [\mathrm{A}_\mu^\intercal, \mathrm{A}_\sigma^\intercal, \mathrm{A}_\nu^\intercal, \mathrm{A}_\tau^\intercal]$ represents the sparse factor loadings matrix. This notation allows for a more compact representation of the measurement equations describing the relationships between the latent variables and the distributional parameters in the GLVM-LSS framework.

3.3. Estimation

In the GLVM-LSS framework, the model complexity depends not only on the number of items and latent factors, but also on the number of location, scale, and shape parameters characterising the items distributions in the measurement model. We propose a penalised full-information marginal maximum likelihood (PMML) estimation method to estimate the model parameters in a way that promotes sparsity and interpretability of the estimated factor loading matrices. PMML has been extensively used in the literature of penalised estimation for LVMs (see, e.g., Chen et al., 2015; Battauz, 2020). The penalised estimation procedure involves maximising the penalised marginal log-likelihood function, which is given by:

$$\begin{aligned}\ell_p(\Theta; \mathbf{y}) &= \sum_{m=1}^{n} \log \left(\int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i(\mathbf{z})) \right] \, p(\mathbf{z}; \Phi) \, \mathrm{d}\mathbf{z} - \mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, w) \right) \\ &= \ell(\Theta; \mathbf{y}) - n\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w}) \end{aligned} \tag{3.1}$$

where $\ell(\Theta; \mathbf{y})$ is the marginal log-likelihood as defined in equation (2.5). The vector $\Theta^\intercal = (\boldsymbol{\alpha}_0^\intercal, \text{vec}(\text{A})^\intercal, \text{vech}(\boldsymbol{\Phi})^\intercal)$ represents the unknown model parameters, and K is the total number of model parameters. The term $\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w})$ is a non-negative scalar-valued function that introduces sparsity in the factor loadings. The levels of sparsity in the location, scale, and shape measurement equations are controlled independently by their corresponding non-negative tuning parameter in $\boldsymbol{\lambda} = (\lambda_\mu, \lambda_\sigma, \lambda_\tau, \lambda_\nu)^\intercal$. The sub-index $\boldsymbol{\lambda}$ is just to emphasise the dependence of the penalty term on the tuning parameter vector. Additionally, the penalty term can depend on a vector of loading-specific weights $\boldsymbol{w} = (w_1, ..., w_K)^\intercal$, where higher (lower) weights indicate stronger (weaker) penalisation and more (less) sparsity. Weights are useful, for example, when we want large parameters (in absolute value) to be subject to weak penalty and small parameters, around zero, to be penalised heavier. Moreover, weights are usually pre-specified and depend on the specific functional form of the penalty term, as described in Section 3.3.1.

For a fixed value of the tuning parameter vector $\boldsymbol{\lambda}$, the penalised maximum likelihood estimate (PMLE), denoted by $\hat{\Theta} = \hat{\Theta}(\boldsymbol{\lambda})$, is the value that maximises the penalised marginal log-likelihood function:

$$\hat{\Theta} = \underset{\Theta \in \Xi}{\arg\max}\, \ell_p(\Theta; \mathbf{y})$$

where $\Xi \subseteq \mathbb{R}^K$ represents the parameter space. The tuning parameter vector $\boldsymbol{\lambda}$ controls the level of sparsity in the estimated parameters $\hat{\Theta}$. When $\boldsymbol{\lambda} = \mathbf{0}$, the penalty term is effectively removed, and the estimation reduces to the marginal maximum likelihood estimation presented in Chapter 2. The components of the penalised marginal log-likelihood function, $\ell(\Theta; \mathbf{y})$ and $\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w})$, serve different purposes in the estimation process. The log-likelihood $\ell(\Theta; \mathbf{y})$ controls model fit, while the penalty term $\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w})$ controls for model complexity. By adjusting the values of the tuning parameter vector $\boldsymbol{\lambda}$, we can control the trade-off between model fit and model complexity.

3.3.1 Sparsity Inducing Penalties

The mapping $\mathcal{P}_{\boldsymbol{\lambda}} : \Xi \to (\mathbb{R}^+ \cup \{0\})$ is used to introduce sparsity in the estimated parameters. However, not all parameters in Θ are penalised equally, and it is often the case that only a subset of parameters is subject to penalisation. For example, it is common that intercept terms are not penalised, as and it is only the factor loadings (which give information about the relationship between the observed items and the latent variables) that are subject to penalisation. To describe this, we modify the notation as follows. Let $\Theta = (\alpha_1, ..., \alpha_{k^*}, \alpha_{k^*+1}, ..., \alpha_K)^\intercal$ be the vector of all model parameters. Denote $\Theta_p = (\alpha_1, ..., \alpha_{k^*})^\intercal$ as the K_p-dimensional vector of model parameters subject to penalisation (e.g., factor loadings), and $\Theta_u = (\alpha_{k^*+1}, ..., \alpha_K)^\intercal$ as the K_u-dimensional

vector of model parameters that are not penalised (e.g., intercepts, free factor loadings, factor correlations). Note that $K_p + K_u = K$.

Moreover, define the index set $\mathcal{A} = \{k : \alpha_k = 0, \alpha_k \in \Theta^*\}$, which represents the indices of model parameters that are zero in the true parameter vector Θ^*. Correspondingly, $\Theta^*_{\mathcal{A}} = \mathbf{0}$ represents the vector of true model parameters indexed by $\mathcal{A}$. Conversely, we define $\mathcal{A}^c$ as the indices of true non-zero parameters in Θ^*, and $\Theta^*_{\mathcal{A}^c}$ as the vector containing the true non-zero parameters.

The penalty term $\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w})$ in (3.1) can be expressed in a general form as the sum of K individual L_1-norm-based penalties:

$$\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, w) = \sum_{k=1}^{K} \mathcal{P}_{\lambda,k}(||\mathrm{R}_k\Theta||_1; \lambda_{\varphi,k}, w_k)$$

Here, $||\cdot||_1$ denotes the L_1-norm, and R_k is a $K \times K$ matrix. For $k = 1, ..., k^*$ (i.e., model parameters subject to penalisation), the diagonal elements of R_k are zero except for the $[k, k]^{\text{th}}$ element, which is equal to 1, and thus $||\mathrm{R}_k\Theta||_1 = |\alpha_k|$. For $k = k^* + 1, \ldots, K$ (i.e., unpenalised parameters), $\mathrm{R}_k = \mathbf{0}$, and thus $||\mathrm{R}_k\Theta||_1 = 0$. The amount of shrinkage on the parameters $\alpha_k \in \Theta_p$ is controlled by the corresponding tuning parameter $\lambda_{\varphi,k}$, which depends on whether the parameter belongs to a location, scale, or shape measurement equation. For example, if α_k belongs to a location measurement equation, then $\lambda_{\varphi,k} = \lambda_\mu$. Similarly, for other scale and shape parameters, the corresponding tuning parameters are used (i.e., λ_σ and λ_τ, respectively). For parameters $\alpha_k \in \Theta_u$, $\lambda_{\varphi,k} = 0$, meaning no penalty is imposed on those parameters.

In the context of model selection, two popular convex L_1-penalties are commonly used[1]: the Lasso (Tibshirani, 1996) and the adaptive Lasso (Alasso, Zou, 2006). Under certain conditions, these penalties have been shown to provide consistent variable selection (Zhao and Yu, 2006; Zou, 2006). Under the general notation above, the Lasso can be expressed as:

$$\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{k=1}^{k^*} \lambda_{\varphi,k} \cdot |\alpha_k| \tag{3.2}$$

The Lasso has an important limitation in that it penalises all parameters equally, with $w_k = 1$ for $k = 1, \ldots, k^*$. This penalisation scheme can lead to biased estimates for large coefficients, making the Lasso sub-optimal in terms of estimation risk. In contrast, the Alasso addresses this limitation by incorporating parameter-specific weights. By assigning different weights to each

[1]Non-convex alternatives, such as the SCAD (Fan and Li, 2001) and the MCP (Zhang, 2010) are also considered in Appendix B1.

parameter, the Alasso introduces lower penalisation for larger coefficients and higher penalisation for weaker effects. The Alasso in general notation is:

$$\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{k=1}^{k^*} \lambda_{\varphi,k} \cdot w_k \cdot |\alpha_k| = \sum_{k=1}^{k^*} \frac{\lambda_{\varphi,k}|\alpha_k|}{|\hat{\alpha}_k|^a}, \qquad \text{for } a > 0. \tag{3.3}$$

where $\hat{\alpha}_k$ represents a consistent estimate of the parameter $\alpha_k \in \Theta_p$, such as the (potentially rotated) maximum likelihood estimate (MLE) of the factor loading. The parameter-specific Alasso weight is $w_k = |\hat{\alpha}_k|^{-a}$. The additional parameter $a > 0$ controls the influence of w_k on the penalty, and it is typically fixed at a value not exceeding 2. The Alasso is preferred over the Lasso due to its 'oracle property' (Fan and Li, 2001). This property implies two key asymptotic characteristics: i) sparsity, which means $\hat{\Theta}_{\mathcal{A}} = \mathbf{0}$, indicating that the estimated parameters are exactly zero for the true zero parameters; and ii) asymptotic Normality of the true non-zero parameter estimates, $\sqrt{n}(\hat{\Theta}_{\mathcal{A}^c} - \Theta^*_{\mathcal{A}^c}) \xrightarrow{d} \mathbb{N}(\mathbf{0}, \mathcal{I}_{\mathcal{A}^c})$. Here, the asymptotic covariance of the model parameters $\mathcal{I}_{\mathcal{A}^c} = \mathcal{I}(\Theta^*_{\mathcal{A}^c}; \Theta^*_{\mathcal{A}} = \mathbf{0})$ is the expected information matrix for the true non-zero parameters knowing that $\Theta^*_{\mathcal{A}} = \mathbf{0}$.

3.3.2 Computation

The presence of the non-differentiable L_1-penalty term in equation (3.1) poses challenges for gradient-based iterative optimisation algorithms that use up to second-order information to compute the MLE, such as the ones presented in Section 2.3. These algorithms cannot be directly applied to compute the PMLE due to the non-differentiability of the penalty term. Several algorithms have been proposed in the literature (Efron et al., 2004; Friedman et al., 2007, 2010) to solve penalised estimation problems efficiently. However, these algorithms may encounter difficulties when dealing with correlated covariates (latent factors in our case) and can face convergence issues when the objective function is non-smooth. Proximal algorithms (Parikh and Boyd, 2014; Lee et al., 2014) also provide computationally efficient and theoretically solid ways of dealing with optimisation problems with non-smooth objective functions. Stochastic proximal algorithms have been explored in Zhang and Chen (2022) for the estimation of LVM.

In this study, we address this challenge by using local approximations of the Lasso and Alasso penalties. These approximations result in quadratic functions that are twice-differentiable everywhere. This enables us to adapt computational framework introduced in Section 2.3 to the penalised estimation problem. For the computation of the model parameters, we propose a two-step iterative estimation procedure that combines the flexibility and simplicity of the EM-algorithm with the robustness of (quasi-)Newton algorithms used for the direct maximisation of the marginal

penalised log-likelihood.

Local Approximations of Sparsity Inducing Penalties

Local approximations for L_1-penalties have been widely used to reduce the computational burden associated with penalised estimation problems (see, e.g., Ulbricht, 2010; Fan and Li, 2001; Filippou et al., 2017 in the regression context; and Battauz, 2020; Geminiani et al., 2021 in the LVM context). One popular approximation is (Koch, 1996):

$$|x| \approx e_{\bar{c}}(x) := (x^2 + \bar{c})^{1/2}\,, \quad x \in \mathbb{R}\,, \bar{c} > 0$$

where $e_{\bar{c}}(x)$ is a twice-continuously differentiable function for a fixed constant $\bar{c}$ that controls the approximation's closeness to the L_1-norm. Note that $\lim_{\bar{c}\to 0} e_{\bar{c}}(x) \to |x|$. In the current context, this approximation yields $||\mathrm{R}_k\Theta||_1 \approx ((\mathrm{R}_k\Theta)^\intercal(\mathrm{R}_k\Theta)+\bar{c})^{1/2}$. Let $\xi_k = \mathrm{R}_k\Theta$, where the k-th element in $\xi_k = (0,\dots,0,\alpha_k,0,\dots,0)^\intercal$ corresponds to the k-th parameter in Θ. It is important to note that $\nabla_{\xi_k}||\xi_k||_1 = \frac{\partial||\xi_k||_1}{\partial\xi_k} = (\xi_k^\intercal\xi_k + \bar{c})^{-1/2}\xi_k$ is defined and well-behaved everywhere.

To simplify notation, we write $\mathcal{P}_{\boldsymbol{\lambda}} := \mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w})$. Assume $\tilde{\Theta}$ is a point in the neighbourhood of Θ. The penalty function in (3.1) can be approximated by a first-order Taylor expansion around $\tilde{\Theta}$

$$\mathcal{P}_{\boldsymbol{\lambda}}(\Theta) \approx \mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta}) + \nabla_\Theta\mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta})^\intercal(\Theta - \tilde{\Theta})\,,$$

After some manipulation (Appendix B2), the approximation above can be expressed as:

$$\mathcal{P}_{\boldsymbol{\lambda}}(\Theta) \approx \frac{1}{2}\Theta^\intercal\left[\sum_{k=1}^{k^*}\frac{\partial\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta}) + \bar{c}\right]^{-1/2}\mathrm{R}_k^\intercal\mathrm{R}_k\right]\Theta = \frac{1}{2}\Theta^\intercal\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\,, \tag{3.4}$$

where $\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})$ is a $K \times K$ block diagonal matrix of the form

$$\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta}) = \begin{bmatrix} S_{\boldsymbol{\lambda}}(\tilde{\Theta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\,,$$

with null sub-matrices $\mathbf{0}$ of dimension $(K-k^*)\times(K-k^*)$, and a $k^*\times k^*$ diagonal $S_{\boldsymbol{\lambda}}(\tilde{\Theta})$ matrix with entries

$$S_{\boldsymbol{\lambda}}(\tilde{\Theta})_{[k,k]} = \frac{\partial\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta}) + \bar{c}\right]^{-1/2} \qquad \text{for} \quad k = 1, ..., k^*$$

that define the amount of penalisation on α_k given $(\lambda_{\varphi,k}, w_k)$, and are determined by the functional

form of the local approximations of the Lasso in (3.2) and the Alasso in (3.3). These approximations, derived in Appendix B2, take the form

Lasso: $S_{\boldsymbol{\lambda}}(\tilde{\Theta})_{[k,k]} = \lambda_{\varphi,k} \cdot (\tilde{\alpha}_k^2 + \bar{c})^{-1/2}$;

Alasso: $S_{\boldsymbol{\lambda}}(\tilde{\Theta})_{[k,k]} = \lambda_{\varphi,k} \cdot (|\hat{\alpha}_k|^a \cdot (\tilde{\alpha}_k^2 + \bar{c})^{1/2})^{-1}$;

By using the approximation of the penalty term in (3.4), the penalised marginal log-likelihood in (3.1) can be expressed as:

$$\ell_p(\Theta; \mathbf{y}, \lambda) = \ell(\Theta; \mathbf{y}) - \frac{n}{2}\Theta^\intercal \mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta \tag{3.5}$$

The selection of the penalty function is not trivial. Penalties that possess the oracle property, such as the Alasso, are generally preferred over the more biased Lasso (see, e.g., Choi et al., 2010; Hirose and Yamamoto, 2014; Huang et al., 2017). However, simulation results presented in Geminiani et al. (2021) suggest that once an optimal tuning vector is chosen, the specific choice of penalty has little impact on the true sparsity recovery. Moreover, these differences become even smaller as sample sizes increase. The practical importance lies in selecting an estimated $\hat{\boldsymbol{\lambda}}$ vector that minimises an information or cross-validation criterion, rather than focusing solely on the choice of the penalty function, in order to achieve consistent model selection and optimal sparsity recovery. The problem of tuning parameter selection is addressed in Section 3.5.

Computation of the Penalised Parameter Estimates

The introduction of the twice-differentiable local approximation to the penalty term allows for adapting the two-step maximum likelihood estimation strategy introduced in Section 2.3 to the penalised case. Specifically, for a given vector of tuning parameters $\boldsymbol{\lambda}$, we compute the PMLE $\hat{\Theta} = \arg\max_\Theta \ell_p(\Theta; \mathbf{y})$ by sequentially applying an EM-algorithm and a direct maximisation of the penalised marginal log-likelihood using a (quasi-)Newton solver.

This implementation is based on purely practical considerations, as it aims to capitalise on the advantages of both methods. On one hand, the EM-algorithm has low computation cost per iteration and it is relatively easy to implement. It is also robust to starting values far from the mode and it guarantees monotone increments of the objective function at each iteration. The EM-algorithm has been extensively used in the LVM literature for penalised estimation problems with no closed-form solutions (see, e.g., Sun et al., 2017). However, it is also known for its (sub-)linear convergence rate, which makes it slow to reach a local mode (McLachlan and Krishnan, 2008).

On the other hand, (quasi-)Newton algorithms exhibit (super-)linear convergence rates, making them faster in reaching a local mode compared to the EM-algorithm. Furthermore, these methods often provide estimates of the information matrix, which is essential for computing standard errors. However, (quasi-)Newton algorithms can encounter convergence issues when initialised far from the mode and typically require more computationally intensive operations, such as matrix inversions.

In our approach, we propose an optimisation strategy that combines the advantages of both methods. We first use the EM-algorithm for a fixed number of iterations to obtain intermediate estimates. These estimates serve as refined starting values for the (quasi-)Newton algorithm, which performs the direct maximisation of $\ell_p(\Theta; \mathbf{y})$. This sequential implementation leverages the computational efficiency and robustness of the EM-algorithm while benefiting from the faster convergence of (quasi-)Newton algorithms. Our proposed framework can be seen as a generalisation of the estimation strategy presented in Geminiani et al. (2021) to LVMs with no closed-form solutions. In the following sub-sections, we provide further details and elaboration on our approach.

First Step: Parameter computation via penalised EM-algorithm

The EM-algorithm (Dempster et al., 1977) is an iterative procedure that consists of two main steps: the *E-step* and the *M-step*. In the E-step, we compute the expected value of the complete-data penalised log-likelihood with respect to the posterior distribution of the latent factors. In the M-step, we maximise the expected penalised log-likelihood obtained in the E-step.

Let $f(\mathbf{y}, \mathbf{z}; \Theta)$ denote the joint probability function of the complete data $(\mathbf{y}, \mathbf{z})$. Building upon Equation (2.6) and utilising the approximate penalty term introduced in Equation (3.4), the complete-data penalised log-likelihood for a given tuning parameter vector $\boldsymbol{\lambda}$ is:

$$\begin{aligned}\ell_{cp}(\Theta; \mathbf{y}, \mathbf{z}) &= \sum_{m=1}^{n} \left[\left\{ \sum_{i=1}^{p} \log f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i(\mathbf{z})) \right\} + \log p(\mathbf{z}_m; \Phi) \right] - \frac{n}{2} \Theta^\intercal \mathcal{S}_\lambda(\tilde{\Theta}) \Theta \\ &= \ell_c(\Theta; \mathbf{y}, \mathbf{z}) - \frac{n}{2} \Theta^\intercal \mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta}) \Theta \end{aligned} \tag{3.6}$$

E-step: In the E-step, we compute the expected value of (3.6) with respect to the posterior distribution of $\mathbf{z}$ given $\mathbf{y}$ and the current estimates $\Theta^{[t]}$. The locally approximated penalty term is evaluated at $\Theta^{[t]}$. The objective function in the E-step, denoted as $\mathcal{Q}p(\Theta; \Theta^{[t]})$, is given by:

$$\mathcal{Q}_p(\Theta; \Theta^{[t]}) = \mathbb{E}_{\mathbf{z} \,|\, \mathbf{y}; \Theta^{[t]}} \left[\, \ell_{cp}(\Theta; \mathbf{y}, \mathbf{z}) \, \right] = \mathcal{Q}(\Theta; \Theta^{[t]}) - \frac{n}{2} \Theta^\intercal \mathcal{S}_\lambda(\Theta^{[t]}) \Theta \tag{3.7}$$

where $\mathcal{Q}(\Theta;\Theta^{[t]}) = \mathbb{E}_{\mathbf{z}\mid\mathbf{y};\Theta^{[t]}}[\ell_c(\Theta;\mathbf{y},\mathbf{z})]$. In practice, the posterior expectation is often not available in closed form and thus numerical techniques, such as the Gaussian-Hermite quadrature or other methods, are used to evaluate the multivariate integrals (see further details in Appendix A4).

M-step: In the M-step, we update the parameter vector to $\Theta^{[t+1]} = \arg\max \mathcal{Q}_p(\Theta;\Theta^{[t]})$. In practice, it suffices to find $\Theta^{[t+1]}$ such that $\mathcal{Q}_p(\Theta^{[t+1]};\Theta^{[t]}) \geq \mathcal{Q}_p(\Theta^{[t]};\Theta^{[t]})$. We can achieve this update by solving for the complete-data penalised score vector $\mathbb{S}_p^{[t]} := \nabla_\Theta \mathcal{Q}_p(\Theta;\Theta^{[t]}) = \mathbf{0}$. There are different update rules that can be used, such as a gradient descent (GD) type update with a (possibly adaptive) step size $\omega^{[t]}$, $\Theta^{[t+1]} = \Theta^{[t]} - \omega^{[t]}\mathbb{S}_p^{[t]}$; or a Newton-Raphson (NR) type update, $\Theta^{[t+1]} = \Theta^{[t]} - (\mathbb{H}_p^{[t]})^{-1}\ \mathbb{S}_p^{[t]}$, where the penalised score and the penalised observed and expected information matrices are:

$$\mathbb{S}_p^{[t]} := \nabla_\Theta \mathcal{Q}_p(\Theta;\Theta^{[t]}) = \mathbb{S}^{[t]} - n\mathcal{S}_\lambda(\Theta^{[t]})\Theta \tag{3.8}$$

$$\mathbb{H}_p^{[t]} := \nabla_\Theta \nabla_{\Theta^\intercal} \mathcal{Q}_p(\Theta;\Theta^{[t]}) = \mathbb{H}^{[t]} - n\mathcal{S}_\lambda(\Theta^{[t]}) \tag{3.9}$$

$$\mathrm{I}_p^{[t]} := -\mathbb{E}_{\mathbf{y}}\left(\mathbb{H}_p^{[t]}\right) = \mathrm{I}^{[t]} + n\mathcal{S}_\lambda(\Theta^{[t]}) \tag{3.10}$$

which are derived from the score vector with entries described by (2.8), and the observed (or expected) information matrix with entries given by (2.9). We repeat the E-step and M-step iteratively until convergence or until a predefined number of iterations is reached. It is important to note that the EM-algorithm may terminate earlier if the complete-data penalised observed (or information) matrix evaluated at the current parameter value is not semi-definite positive due to the approximation error from the numerical integration.

Second Step: Parameter computation via penalised direct maximisation

In the direct maximisation step, we refine the parameter estimates obtained from the EM-step by solving for $\nabla_\Theta \ell_p(\Theta;\mathbf{y}) = \mathbf{0}$, where $\ell_p(\Theta;\mathbf{y})$ is the penalised marginal log-likelihood. However, since closed-form solutions are often not available, iterative numerical optimisation solvers are used to compute the PMLE.

Quasi-Newton and trust-region algorithms use first- and second-order information from the penalised marginal log-likelihood. The score vector for the penalised marginal-likelihood, which is used to determine the search direction in the parameter space, is equivalent to the score vector of the complete-data penalised log-likelihood, $\nabla_\Theta \ell_p(\Theta;\mathbf{y})|_{\Theta=\Theta^{[t]}} \equiv \nabla_\Theta \mathcal{Q}_p(\Theta;\Theta^{[t]}) = \mathbb{S}_p^{[t]}$. The second-order information in the observed and/or expected information matrices is used to better

approximate of the curvature of the penalised marginal log-likelihood evaluated at $\Theta^{[t]}$. These matrices are given by:

$$\mathcal{H}_p^{[t]} := \nabla_\Theta \nabla_{\Theta^\intercal} \ell_p(\Theta; \mathbf{y})|_{\Theta=\Theta^{[t]}} = \mathcal{H}^{[t]} - n\mathcal{S}_\lambda(\Theta^{[t]}) \tag{3.11}$$

$$\mathcal{I}_p^{[t]} := -\mathbb{E}_\mathbf{y}\left(\mathcal{H}_p^{[t]}\right) = \mathcal{I}^{[t]} + n\mathcal{S}_\lambda(\Theta^{[t]}) \tag{3.12}$$

which follow from the expression in (2.10). The factor correlations are not penalised, and thus their computation is through the (quasi-)Newton proximal algorithm described in Section 2.3.1.

3.4. Goodness-of-fit and Model selection

Selecting the best model requires finding the right balance between two competing objectives: model fit and model complexity. This balance becomes particularly important within a PMML estimation framework, as the tuning vector $\boldsymbol{\lambda}$ influences both goals. Less complex models tend to favour sparsity and offer better interpretability, but may fit the data worse. On the other hand, more complex models may fit the data better but are often less interpretable. To address this trade-off, information criteria, such as the Akaike Information Criterion (AIC, Akaike, 1974) or the Bayesian Information Criterion (BIC, Schwarz, 1978), balance a measure of model fit, typically the log-likelihood, with a measure of model complexity represented by the degrees of freedom (dof), which is the number of uniquely estimated parameters in the model. These information criteria have been extensively used in the LVM literature for model selection and assessing goodness-of-fit. However, the AIC and BIC are derived under the assumption that model parameters are estimated via maximum likelihood, which is not the case in our current setting. Consequently, the dof are no longer an appropriate measure of model complexity, and relying on the number of non-zero parameters as an indicator of model complexity can significantly affect the assessment of model fit (Jacobucci et al., 2016).

We use the Generalised Information Criterion (GIC, Konishi and Kitagawa, 1996, 2008) as our model selection criterion and to assess model fit. The model complexity term in the GIC is a theoretically founded definition for the *effective degrees of freedom* (edf hereafter) of the penalised GLVM-LSS model. Given a vector of parameter estimates $\hat{\Theta} = \hat{\Theta}(\boldsymbol{\lambda})$ obtained through PMML estimation with a fixed tuning vector $\boldsymbol{\lambda}$, the GIC is defined as:

$$\text{GIC}(\hat{\Theta}, \boldsymbol{\lambda}) = -2\ell(\hat{\Theta}) + 2 \cdot \text{tr}\left(\mathcal{H}_p(\hat{\Theta})^{-1}\mathcal{H}(\hat{\Theta})\right) \tag{3.13}$$

where $\ell(\hat{\Theta})$ represents the log-likelihood, and the observed information matrix evaluated at the PMLE, $\mathcal{H}(\hat{\Theta})$, is given in (2.10). The theoretical derivations of the GIC and of the edf can be found in Appendix B3. The GIC extends the AIC and shares its tendency to favour overly complex models (Shao, 1997). An approximate Generalised Bayesian Information Criterion (GBIC) is obtained by giving a weight of $\log(n)$ to the complexity term in (3.13):

$$\text{GBIC}(\hat{\Theta}, \boldsymbol{\lambda}) = -2\ell(\hat{\Theta}) + \log(n) \cdot \text{tr}\left(\mathcal{H}_p(\hat{\Theta})^{-1}\mathcal{H}(\hat{\Theta})\right) \tag{3.14}$$

It is worth noting that for an unpenalised model ($\boldsymbol{\lambda} = \mathbf{0}$), the observed penalised information matrix $\mathcal{H}_p(\hat{\Theta})$ is equal to $\mathcal{H}(\hat{\Theta})$, resulting in the edf being equal to the number of estimated parameters in the model, K. In this case, the complexity term in the GIC corresponds to the number of parameters.

For the penalised case, the complexity term is given by $\text{edf} = \text{tr}(\mathcal{H}_p(\hat{\Theta})^{-1}\mathcal{H}(\hat{\Theta}))$. As $\boldsymbol{\lambda} \to \mathbf{0}$, the edf tends to K, indicating that the model becomes less penalised and resembles the unpenalised case. On the other hand, as $\boldsymbol{\lambda} \to \infty$, the edf tends to $K - k^*$, where k^* is the number of penalised parameters in the model. Thus, the edf ranges between $K - k^*$ and K for $0 < \boldsymbol{\lambda} < \infty$. The edf can be interpreted as the sum of individual contributions from each parameter $\hat{\alpha}_k \in \hat{\Theta}$ for which $|\hat{\alpha}_k| > 0$. Each estimated parameter adds a contribution to the edf in the range of $[0, 1]$, inversely related to the amount of penalisation the parameter has been subject to.

Many works on penalised estimation of LVMs compute the degrees of freedom as the count of estimated non-zero parameters, based on results that show that, for lasso-penalised linear models, this number is an unbiased estimate of the total degrees of freedom (Zou et al., 2007). However, the edf provides a better-calibrated measure of model complexity. The edf is theoretically related to the estimated bias term in the GIC and offers a more accurate assessment of the complexity of the penalised model.

3.5. Selection of Tuning Parameters

Selecting the optimal value for $\boldsymbol{\lambda} = (\lambda_\mu, \lambda_\sigma, \lambda_\nu, \lambda_\tau)^\intercal$ can be viewed as a model selection problem, where we compare candidate models estimated using different tuning parameter vectors. The BIC, known for its consistency in model selection (Davison, 2003, Chapter 4.7), guides the choice of the

optimal tuning parameter vector $\hat{\boldsymbol{\lambda}}$ by minimising the GBIC value:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} \in [0,\infty)^S}{\arg\min}\, \text{GBIC}(\hat{\Theta}, \boldsymbol{\lambda}) \tag{3.15}$$

where $S = \dim(\boldsymbol{\lambda})$ represents the number of different tuning parameters for the location, scale, or shape parameters indexing the conditional distributions in the measurement part of the GLVM-LSS. Typically, $\hat{\boldsymbol{\lambda}}$ is chosen from a set of candidate values through a grid-search procedure in $[0, \infty)^S$. However, in the GLVM-LSS case this approach can be time-consuming and computationally expensive due to the high dimensionality of the grid.

An alternative method is to estimate $\hat{\boldsymbol{\lambda}}$ using an automatic, data-driven procedure as proposed by Geminiani et al. (2021) for the case of the Normal linear factor model[2]. The procedure draws upon the literature from generalised additive models (Wood, 2004; Marra et al., 2017) and involves minimising the mean squared error (MSE) of the estimated model parameters. This quantity serves as an approximate unbiased risk estimate (UBRE) and an approximate AIC. Further details are presented in Appendix B4. Simulation results in Geminiani et al. (2021) demonstrate that the automatic procedure yields lower GBIC values compared to grid-search for various convex and non-convex penalties, including Lasso and Alasso. A brief explanation of the methods is discussed below.

Let $\boldsymbol{\lambda}_0$ represent the (initial) fixed value for the tuning parameter vector. In the neighbourhood of the (local) mode, the update rules of the quasi-Newton and/or trust-region algorithms behave similarly to the classic unconstrained Newton-Raphson update rule (Nocedal and Wright, 2006, Chapter 4). Specifically, at iteration $t+1$, the score vector $\mathbb{S}_p^{[t+1]}$ is close to zero, allowing us to express the NR update step as:

$$0 \approx \mathbb{S}_p^{[t+1]} \approx \mathbb{S}_p^{[t]} + \mathcal{H}_p^{[t]}(\Theta^{[t+1]} - \Theta^{[t]})$$

Solving for $\Theta^{[t+1]}$ yields (see Appendix B4.1):

$$\Theta^{[t+1]} = \left[\mathbb{J}^{[t]} + n\mathcal{S}_{\boldsymbol{\lambda}_0}^{[t]}\right]^{-1} \sqrt{\mathbb{J}^{[t]}}^{\intercal}\, \mathbb{K}^{[t]}$$

where $\mathbb{J}^{[t]} = -\mathcal{H}^{[t]}$, the vector $\mathbb{K}^{[t]} = \mu_{\mathbb{K}}^{[t]} + \varepsilon^{[t]}$, with $\mu_{\mathbb{K}}^{[t]} = \sqrt{\mathbb{J}^{[t]}}\Theta^{[t]}$, and $\varepsilon^{[t]} = \sqrt{\mathbb{J}^{[t]}}^{-\intercal}\mathbb{S}^{[t]}$. For brevity, let $\mathcal{S}_{\boldsymbol{\lambda}_0}^{[t]} = \mathcal{S}_{\boldsymbol{\lambda}_0}(\Theta^{[t]}; \boldsymbol{\lambda}, \boldsymbol{w})$. The squared root $\sqrt{\mathbb{J}}$ and its inverse can be obtained

[2]This procedure is only valid for Lasso and Alasso penalties, as the local approximations from the vector $\boldsymbol{\lambda}$ are separable (see Appendix B2). The local approximations for SCAD and MCP penalties are non-separable and require a grid-search approach for selecting the optimal tuning parameter $\hat{\boldsymbol{\lambda}}$.

through an eigenvalue decomposition of $\mathcal{H}$. According to standard likelihood theory, we know that $\varepsilon \sim \mathbb{N}(0, \mathbb{I}_K)$. Therefore, $\mathbb{K} \sim \mathbb{N}(\mu_{\mathbb{K}}, \mathbb{I}_K)$, where $\mu_{\mathbb{K}} = \sqrt{\mathbb{J}}\Theta^*$ and Θ^* represents the true parameter vector.

At convergence, the PMLE can be written as

$$\hat{\Theta}(\boldsymbol{\lambda}_0) \equiv \hat{\Theta} = \left[\hat{\mathbb{J}} + n\hat{\mathcal{S}}_{\boldsymbol{\lambda}_0}\right]^{-1} \sqrt{\hat{\mathbb{J}}}^{\mathsf{T}} \hat{\mathbb{K}} \tag{3.16}$$

where the 'hat' notation denotes the corresponding matrix evaluated at the PMLE, e.g., $\hat{\mathbb{J}} = \mathbb{J}(\hat{\Theta})$, etc. Using (3.16), we have that:

$$\hat{\mu}_{\mathbb{K}} = \sqrt{\hat{\mathbb{J}}}\left[\hat{\mathbb{J}} + n\hat{\mathcal{S}}_{\boldsymbol{\lambda}_0}\right]^{-1} \sqrt{\hat{\mathbb{J}}}^{\mathsf{T}}\ \hat{\mathbb{K}} = \hat{\mathbb{A}}_{\boldsymbol{\lambda}_0}\hat{\mathbb{K}}$$

where $\hat{\mathbb{A}}_{\boldsymbol{\lambda}_0} = \sqrt{\hat{\mathbb{J}}}\left[\hat{\mathbb{J}} + n\hat{\mathcal{S}}_{\boldsymbol{\lambda}_0}\right]^{-1} \sqrt{\hat{\mathbb{J}}}^{\mathsf{T}}$ is the projection matrix of the fitting problem, and depends on the tuning parameter vector through $\hat{\mathcal{S}}_{\boldsymbol{\lambda}_0}$. The matrix $\hat{\mathbb{K}}$ is linked to the unpenalised part of the model.

When updating the tuning parameter vector, our objective is to reduce the model complexity (in terms of edf) that is not supported by the data. The tuning parameter vector $\boldsymbol{\lambda}_0$ is updated by minimising the mean squared error (MSE) of $\hat{\mu}_{\mathbb{K}}$ (see Appendix B4.1):

$$\mathcal{V}(\boldsymbol{\lambda}; \hat{\Theta}) := \mathbb{E}\left(||\mu_{\mathbb{K}} - \hat{\mu}_{\mathbb{K}}||_2^2\right) = \mathbb{E}\left(||\mathbb{K} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2\right) + 2\text{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}) - K \tag{3.17}$$

where $\text{tr}(\mathbb{A}_{\boldsymbol{\lambda}}) = \text{tr}([\mathbb{J} + n\mathcal{S}_{\boldsymbol{\lambda}}]^{-1}\mathbb{J}) = \text{tr}(\mathcal{H}_p^{-1}\mathcal{H})$ corresponds to the edf and is also equivalent to the bias term of the GIC in (3.13). However, the MSE in (3.17) depends on the unknown Θ^* (through $\mathbb{K}$), so we compute an estimate of it using the available PMLE $\hat{\Theta}(\boldsymbol{\lambda}_0)$:

$$\hat{\mathcal{V}}(\boldsymbol{\lambda}; \hat{\Theta}) \propto ||\hat{\mathbb{K}} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2 + 2\text{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}) \tag{3.18}$$

This expression approximates an unbiased risk estimator (UBRE, Wood, 2017, Chapter 6) and serves as an approximate AIC (see Appendix B4.2). For a given $\hat{\Theta}(\boldsymbol{\lambda}_0)$, the optimal $\hat{\boldsymbol{\lambda}}$ is given by:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} \in [0,\infty)^S}{\arg\min}\, \hat{\mathcal{V}}(\boldsymbol{\lambda}; \hat{\Theta}) = \underset{\boldsymbol{\lambda} \in [0,\infty)^S}{\arg\min} \left\{||\hat{\mathbb{K}} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2 + 2\text{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}})\right\} \tag{3.19}$$

To solve this optimisation problem, we resort to iterative numerical methods such as the quasi-Newton or trust-region algorithms, which are also employed in the estimation step. The analytical expressions for the first- and second-order derivatives of (3.19) are provided in Appendix B4.3.

Once we obtain an updated value for the tuning parameter vector, denoted as $\hat{\boldsymbol{\lambda}}_1$, we update the model parameters to $\hat{\Theta}(\hat{\boldsymbol{\lambda}}_1)$. This two-step procedure of model parameter estimation and tuning parameter update is repeated iteratively until convergence is achieved. Ultimately, we obtain the model parameters and the optimal tuning parameter vector as $(\hat{\Theta}(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}^*)^\intercal$. In practice, we initialise the initial tuning parameter vector with a value arbitrarily close to zero (e.g., 10^{-8}).

3.5.1 Influence factor

The tuning parameter vector $\hat{\boldsymbol{\lambda}}$ obtained by minimising the approximate UBRE (and approximate AIC), as by equation (3.19), will often differ from the one obtained by minimising the GBIC in the $[0, \infty)^S$-dimensional grid, as given by equation (3.15). In certain situations, the final model obtained through (3.19) can be overly dense. To address this issue, we introduce an additional parameter $\gamma \geq 1$, referred to as the *influence factor* (Wood, 2017). This factor multiplies the term $2\text{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}})$ in equation (3.18), thereby increasing the importance of the edf (model complexity) in the UBRE. By increasing γ, we encourage sparser models. Consequently, the modified UBRE in the optimisation problem (3.19) becomes:

$$\hat{\mathcal{V}}(\boldsymbol{\lambda}; \hat{\Theta}) \propto ||\hat{\mathbb{K}} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2 + \gamma \cdot 2\text{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}) \tag{3.20}$$

The sparsity-inducing *influence factor* plays an important role in achieving a better balance between model fit and model complexity. It allows for sensitivity analysis to examine how the sparsity of the factor loading matrices changes with different values of γ. In the literature on regression splines it has been observed that choosing $\gamma \approx 1.4$ corrects the tendency of over-fitting when using prediction error criteria (Kim and Gu, 2004). A value of $\gamma \approx 1.5$ finds a justification from the viewpoint of double cross-validation (Wood, 2017, Chapter 6). However, in the current context of model selection, higher values of γ tend to produce sparser solutions by increasing the relative importance of the bias term in the GIC. From a practical standpoint, we recommend adopting a sensitivity analysis approach by exploring various candidate values for γ within a user-defined range. Subsequently, the fitted model with the lowest GBIC can be selected as the preferred choice.

It should be noted that, in some way, the influence factor acts as a tuning parameter that influences the resulting sparsity of the factor loading matrix solution. However, the automatic selection procedure described in this section simplifies the problem by reducing the amount of 'tuning parameters' in the penalised estimation problem. A promising future research avenue

includes exploring ways of automatically selecting the value of γ, for example, in the spirit of Gu (1992) with 'performance' (oriented) iterations.

3.6. Simulation Studies

In this section, we assess the parameter recovery properties of the proposed PMML estimation framework with automatic selection of tuning parameters across various simulation designs. Previous studies have indicated that penalised LVMs tend to exhibit better model fit compared to their unpenalised counterparts when the true data generating factor loading matrices are sparse (see, e.g., Choi et al., 2010; Hirose and Yamamoto, 2014; Jin et al., 2018 and Geminiani et al., 2021 for continuous items; and Sun et al., 2017; Battauz, 2020 and Huang, 2020 for categorical data). To demonstrate the effectiveness of the automatic selection procedure outlined in Section 3.5, we use the Alasso penalty, which possesses the oracle property. In our analysis, we employ the MLEs as weights in the Alasso penalty and as a benchmark for comparative analysis on performance measures. The MLEs are obtained following the (unpenalised) estimation procedure described in Chapter 2.

We consider different combinations of test lengths ($p = 10, 20$), sample sizes ($n = 200, 500, 1000$), and influence factor values ($\gamma = 1, 2, 3, 4, 5$), resulting in a total of 30 distinct simulation settings. For each setting, we generate $L = 300$ datasets. All simulations are performed in `R` v.4.2.2 (R Core Team, 2022). The code can be accessed at `https://github.com/ccardehu/GLVM-LSS`. See Appendix B6 for details on the software implementation.

3.6.1 Performance Evaluation Criteria

To evaluate the overall performance and estimation quality of the proposed penalised estimation framework, we compute the mean squared error (MSE) and absolute bias (AB) for the estimated model parameters. Let $\alpha_k \in \Theta^*$ be the true value for the model parameter, and $\hat{\alpha}_k^{(l)} \in \hat{\Theta}^{(l)}$ be its estimate using the l^{th} generated sample, where $l = 1, \ldots, L$. The mean squared error (MSE) of $\hat{\alpha}_k$ is calculated as:

$$\text{MSE}(\hat{\alpha}_k) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\alpha}_k^{(l)} - \alpha_k)^2, \quad k = 1, ..., K$$

Similarly, the absolute bias (AB) of $\hat{\alpha}_k$ is given by:

$$\text{AB}(\hat{\alpha}_k) = |\bar{\hat{\alpha}}_k - \alpha_k| = \left|\frac{1}{L}\sum_{l=1}^{L}\hat{\alpha}_k^{(l)} - \alpha_k\right|, \quad k = 1, ..., K$$

For comparison purposes, we compute the average MSE (AvMSE) and average AB (AvAB) across the estimated parameters. We compute these measures separately for the penalised parameters (factor loadings) and the unpenalised parameters (intercepts, free loadings, and factor correlations) for different location, scale, and shape measurement equations.

Let $\mathcal{K}_{p,\varphi} = \{k : \hat{\alpha}_{k,\varphi} \in \hat{\Theta}_{p,\varphi},\ \varphi \in \boldsymbol{\theta}\}$ be the index set for the penalised parameters in the measurement equations for the distributional parameter $\varphi \in \boldsymbol{\theta}$. Similarly, $\mathcal{K}_{u,\varphi} = \{k : \hat{\alpha}_{k,\varphi} \in \hat{\Theta}_{u,\varphi},\ \varphi \in \boldsymbol{\theta}\}$ is the index set for the unpenalised parameters in the measurement equations for distributional parameter $\varphi \in \boldsymbol{\theta}$. The number of elements in these index sets are denoted as $\text{card}(\mathcal{K}_{p,\varphi}) = K_{p,\varphi}$ and $\text{card}(\mathcal{K}_{u,\varphi}) = K_{u,\varphi}$, respectively, with $\sum_{x\in\{p,u\}}\sum_{\varphi\in\boldsymbol{\theta}} K_{x,\varphi} = \dim(\hat{\Theta}) = K$. For each location, scale, or shape parameter $\varphi \in \boldsymbol{\theta}$, we compute the average MSE (AvMSE) for the penalised ($\hat{\Theta}_{p,\varphi}$) and unpenalised ($\hat{\Theta}_{u,\varphi}$) parameter estimates as:

$$\text{AvMSE}(\hat{\Theta}_{x,\varphi}) = \frac{1}{K_{x,\varphi}} \sum_{k\in\mathcal{K}_{x,\varphi}} \text{MSE}(\hat{\alpha}_k), \quad \text{for } x = \{p,u\},\ \varphi \in \{\mu,\sigma,\tau,\nu\},$$

Similarly, we compute the average absolute bias (AvAB) as:

$$\text{AvAB}(\hat{\Theta}_{x,\varphi}) = \frac{1}{K_{x,\varphi}} \sum_{k\in\mathcal{K}_{x,\varphi}} \text{AB}(\hat{\alpha}_k), \quad \text{for } x = \{p,u\},\ \varphi \in \{\mu,\sigma,\tau,\nu\}.$$

We also evaluate the performance of the proposed method in terms of sparsity recovery. For each distributional parameter $\varphi \in \boldsymbol{\theta}$, define an indicator vector $\mathbb{T}$ of dimension $K_{p,\varphi}$ with entries given by $\mathbb{t}_k = \mathbb{1}(\alpha_k \neq 0)$ for $k \in \mathcal{K}_{p,\varphi}$. This vector indicates whether the k^{th} penalised factor loading in $\Theta_{p,\varphi}$ is different from zero ($\mathbb{t}_k = 1$) or equal to zero ($\mathbb{t}_k = 0$). Similarly, for each generated sample $l = 1, ..., L$, define the corresponding indicator vector $\hat{\mathbb{T}}^{(l)}$ for the estimated penalised factor loadings $\hat{\Theta}^{(l)}_{p,\varphi}$, with entries $\hat{\mathbb{t}}^{(l)}_k = \mathbb{1}(\hat{\alpha}^{(l)}_k \neq 0)$ for $k \in \mathcal{K}_{p,\varphi}$.

To assess sparsity recovery, we compute the correct estimation rate (CER):

$$\text{CER}(\hat{\Theta}^{(l)}_{p,\varphi}) = \left(\frac{1}{K_{p,\varphi}} \sum_{k\in\mathcal{K}_{p,\varphi}} \mathbb{1}(\hat{\mathbb{t}}^{(l)}_k = \mathbb{t}_k)\right), \quad \text{for } l = 1, ..., L \text{ and } \varphi \in \{\mu,\sigma,\tau,\nu\}.$$

The $\text{CER}(\hat{\Theta}^{(l)}_{p,\varphi})$ represents the proportion of estimated factor loadings subject to penalisation that are correctly identified as being either different from, or equal to zero. Values closer to 1.0 suggest better recovery of the true sparsity. For comparison across simulations, we compute the average CER (AvCER) :

$$\text{AvCER}(\hat{\Theta}_{p,\varphi}) = \frac{1}{L}\sum_{l=1}^{L} \text{CER}(\hat{\Theta}^{(l)}_{p,\varphi}), \quad \text{for } \varphi \in \{\mu,\sigma,\tau,\nu\}$$

Additionally, we aim for the proposed method to accurately recover the true sparsity structure by effectively preserving the non-zero parameters and penalising the zero parameters. To evaluate this, we compute the true positive rate (TPR) and the false positive rate (FPR).

For each distributional parameter $\varphi \in \boldsymbol{\theta}$, define $\mathcal{T}_\varphi$ as the index set for the 'true non-zero' parameters in $\Theta_{p,\varphi}$, that is $\mathcal{T}_\varphi = \{k : \alpha_k \neq 0, k \in \mathcal{K}_{p,\varphi}\}$. Note that $\text{card}(\mathcal{T}_\varphi) \leq K_p$, and that $\mathcal{T}_\varphi$ is a subset of $\mathcal{K}_{p,\varphi}$. Similarly, define $\mathcal{T}_\varphi^c$ as the index set for the 'true zero' parameters in $\Theta_{p,\varphi}$, i.e., $\mathcal{T}_\varphi^c = \{k : \alpha_k = 0, k \in \mathcal{K}_{p,\varphi}\}$. It follows that $\mathcal{T}_\varphi \cup \mathcal{T}_\varphi^c = \mathcal{K}_{p,\varphi}$ and, consequently, $\text{card}(\mathcal{T}_\varphi) + \text{card}(\mathcal{T}_\varphi^c) = K_{p,\varphi}$. The TPR is computed as follows:

$$\text{TPR}(\hat{\Theta}_{p,\varphi}) = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{\text{card}(\mathcal{T}_\varphi)}\sum_{k\in\mathcal{T}_\varphi}\mathbb{1}(\hat{\alpha}_k^{(l)} \neq 0)\right), \quad \text{for } \varphi \in \{\mu, \sigma, \tau, \nu\}$$

The TPR measures the proportion of correctly identified non-zero parameters. Values closer to 1.0 are desired. The FPR is computed as:

$$\text{FPR}(\hat{\Theta}_{p,\varphi}) = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{\text{card}(\mathcal{T}_\varphi^c)}\sum_{k\in\mathcal{T}_\varphi^c}\mathbb{1}(\hat{\alpha}_k^{(l)} \neq 0)\right), \quad \text{for } \varphi \in \{\mu, \sigma, \tau, \nu\}$$

The FPR, on the other hand, quantifies the proportion of true zero parameters that are incorrectly identified as non-zero. Lower values of FPR are desirable as they indicate a more accurate identification of zero parameters. When computing the AvCER, TPR, and FPR, we round the estimated parameter values to the nearest decimal point. For comparison purposes, we also calculate these performance measures for the unpenalised model. In the unpenalised case, we apply a hard threshold of 0.1 and set all estimated factor loadings with absolute values smaller than this threshold to zero. This thresholding procedure is commonly employed in the LVM literature to obtain sparse factor loading matrices (Hair et al., 2010). For more detailed information on parameter initialisation, factor loading identification strategy, and the generation of sparse factor loading matrices, please refer to Appendix B5.

3.6.2 Simulation Study I: Normal linear factor model with heteroscedastic items

In our first simulation study, we consider a heteroscedastic Normal factor model with sparsity in both the location and scale factor loading matrices. Formally, we assume a GLVM-LSS with Normally distributed items conditional on the latent variables, $y_i \,|\, \mathbf{z} \sim \mathbb{N}(\mu_i(\mathbf{z}), \sigma_i^2(\mathbf{z}))$. For simplicity,

the latent variables are uncorrelated. The measurement equations are of the form:

$$\mu_i(\mathbf{z}) = \alpha_{i0,\mu} + \sum_{j=1}^{2} \alpha_{ij,\mu} z_j$$
$$\log(\sigma_i(\mathbf{z})) = \alpha_{i0,\sigma} + \sum_{j=1}^{2} \alpha_{ij,\sigma} z_j$$

In the measurement equations for the location parameter (μ), the intercept parameters are generated from a uniform distribution as $\alpha_{i0,\mu} \sim \text{Unif}(1.0, 2.0)$, while the factor loadings are drawn from $\alpha_{ij,\mu} \sim \text{Unif}(0.5, 1.5)$. The sign of the factor loadings for μ is randomly determined with a probability of 0.5. To ensure identifiability, we fix $\alpha_{12,\mu} = 0$. For the scale parameter (σ), the parameters (intercepts and slopes) in the measurement equations are generated from a uniform distribution, as $\boldsymbol{\alpha}_{i,\sigma} = (\alpha_{i0,\sigma}, \alpha_{ij,\sigma})^\intercal \sim \text{Unif}(0.1, 0.4)$.

We induce sparsity in the factor loading matrices A_μ and A_σ by following the procedure described in Appendix B5. Homoscedastic items, where the variances are constant, are obtained when $\alpha_{ij,\sigma} = 0$ for $j = 1, 2$. On the other hand, items exhibiting heteroscedasticity have one or both non-zero factor loadings in the scale measurement equation. The $L = 300$ datasets were randomly generated using the same set of factor loadings.

Table 3.1 shows the simulation results for $L = 300$ replications, by test length (p), sample size (n), and five different penalised estimations corresponding to various values of the influence factor $\gamma = \{1, 2, 3, 4, 5\}$. In each case, the tuning parameters $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_\mu, \hat{\lambda}_\sigma)^\intercal$ were automatically estimated using the procedure outlined in Section 3.5. For comparison, the results of the unpenalised MML estimation are also included (first row in each case). Several conclusions can be drawn from the results:

- As expected, the proposed PMML estimation framework consistently yields lower average GBIC values compared to the (unpenalised) MMLE in all scenarios.
- The PMLEs exhibit slightly higher average absolute bias (AvAB) than the MLEs. However, this bias diminishes as the sample size increases.
- Conversely, the PMLEs display lower average mean squared error (AvMSE) compared to the MLEs. This difference is particularly pronounced for the factor loadings in the location measurement equations (matrix A_μ), where the non-zero entries are larger in absolute value and the matrix is denser. For the factor loadings in the scale measurement equations (matrix A_σ), especially in smaller sample size scenarios (200, 500), the MSE is predominantly influ-

enced by the bias term, resulting in higher AvMSE for the PMLEs. However, this difference diminishes as the sample size increases.

- Across all cases, higher values of γ consistently lead to better results and more consistent model selection. For small sample sizes (n = 200), a value of $\gamma = 2$ yields the lowest GBIC, while for medium and large sample sizes (n = 500, 1000), the lowest GBIC corresponds to $\gamma = 3$.

- In terms of sparsity recovery, the correct estimation rate (CER) and true positive rate (TPR) approach 1.0, while the false positive rate (FPR) approaches 0.0 for the PMLEs as the sample size increases. Conversely, these measures are considerably worse for the MLEs when using a threshold of ± 0.1 for sparsity recovery.

- Within a given sample size and test length, the average optimal values of the tuning parameters $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_\mu, \hat{\lambda}_\sigma)^\intercal$ increase with γ.

This framework provides flexibility for model selection and testing, particularly when the presence of heteroscedasticity is unknown in advance. It allows us to evaluate the performance of the proposed method in terms of correctly identifying the sparsity structure and estimating the factor loadings accurately.

				Location Parameter (μ)								Scale Parameter (σ)							
p	n	Influence factor (γ)	Avg. GBIC	Intercepts ($\hat{\alpha}_{0,\mu}$)		Loadings ($\hat{\boldsymbol{\alpha}}_{i,\mu}$)						Intercepts ($\hat{\alpha}_{0,\sigma}$)		Loadings ($\hat{\boldsymbol{\alpha}}_{i,\sigma}$)					
				AvMSE	AvAB	AvMSE	AvAB	AvCER	TPR	FPR	Avg. λ_μ	AvMSE	AvAB	AvMSE	AvAB	AvCER	TPR	FPR	Avg. λ_σ
10	200	-	7529.97	0.0144	0.0065	0.0193	0.0094	0.7960	1.000	0.485	-	0.0074	0.0259	0.0065	0.0046	0.816	0.834	0.208	-
		1.0	7456.88	0.0139	0.0064	0.0122	0.0162	0.9263	0.999	0.174	0.0026	0.0072	0.0202	0.0057	0.0140	0.848	0.848	0.153	0.0017
		2.0	7450.73	0.0139	0.0079	0.0121	0.0254	0.9632	0.998	0.085	0.0047	0.0070	0.0174	0.0055	0.0181	0.866	0.831	0.087	0.0023
		3.0	7451.67	0.0141	0.0081	0.0136	0.0356	0.9785	0.996	0.046	0.0080	0.0070	0.0141	0.0065	0.0286	0.855	0.778	0.039	0.0036
		4.0	7451.17	0.0143	0.0094	0.0146	0.0377	0.9805	0.996	0.040	0.0125	0.0070	0.0132	0.0073	0.0349	0.838	0.741	0.029	0.0054
		5.0	7453.33	0.0145	0.0115	0.0168	0.0462	0.9866	0.994	0.023	0.0166	0.0071	0.0125	0.0082	0.0410	0.827	0.715	0.019	0.0078
	500	-	18510.84	0.0052	0.0026	0.0070	0.0069	0.8970	1.000	0.245	-	0.0020	0.0097	0.0022	0.0020	0.907	0.871	0.043	-
		1.0	18428.25	0.0051	0.0025	0.0043	0.0073	0.9526	1.000	0.113	0.0009	0.0020	0.0078	0.0017	0.0044	0.936	0.953	0.088	0.0004
		2.0	18421.82	0.0050	0.0031	0.0041	0.0102	0.9811	1.000	0.045	0.0019	0.0020	0.0069	0.0016	0.0061	0.949	0.944	0.045	0.0006
		3.0	18420.50	0.0051	0.0039	0.0043	0.0156	0.9926	1.000	0.018	0.0030	0.0019	0.0058	0.0017	0.0097	0.948	0.926	0.023	0.0011
		4.0	18421.96	0.0051	0.0045	0.0046	0.0189	0.9960	1.000	0.010	0.0044	0.0020	0.0056	0.0021	0.0145	0.938	0.899	0.007	0.0015
		5.0	18422.25	0.0051	0.0046	0.0045	0.0169	0.9967	1.000	0.008	0.0058	0.0020	0.0054	0.0024	0.0177	0.927	0.877	0.003	0.0020
	1000	-	36763.82	0.0027	0.0030	0.0035	0.0029	0.9511	1.000	0.116	-	0.0010	0.0047	0.0010	0.0015	0.936	0.893	0.005	-
		1.0	36672.87	0.0027	0.0023	0.0020	0.0035	0.9635	1.000	0.087	0.0004	0.0009	0.0037	0.0007	0.0024	0.976	0.986	0.038	0.0002
		2.0	36667.16	0.0026	0.0025	0.0019	0.0053	0.9804	1.000	0.047	0.0009	0.0009	0.0033	0.0007	0.0030	0.984	0.988	0.021	0.0003
		3.0	36665.37	0.0026	0.0028	0.0018	0.0078	0.9914	1.000	0.020	0.0014	0.0009	0.0028	0.0007	0.0046	0.986	0.983	0.010	0.0004
		4.0	36665.58	0.0026	0.0031	0.0019	0.0105	0.9967	1.000	0.008	0.0021	0.0009	0.0026	0.0008	0.0064	0.985	0.978	0.005	0.0006
		5.0	36666.89	0.0026	0.0033	0.0020	0.0119	0.9981	1.000	0.005	0.0028	0.0009	0.0028	0.0009	0.0087	0.977	0.961	0.001	0.0008
20	200	-	14652.08	0.0169	0.0159	0.2552	0.2030	0.6435	0.970	0.826	-	0.0050	0.0228	0.0241	0.0678	0.752	0.871	0.516	-
		1.0	14551.39	0.0166	0.0119	0.1940	0.1695	0.7518	0.934	0.510	0.0038	0.0047	0.0165	0.0184	0.0592	0.793	0.839	0.309	0.0028
		2.0	14542.33	0.0161	0.0129	0.1946	0.1750	0.7769	0.927	0.439	0.0058	0.0047	0.0164	0.0182	0.0610	0.799	0.843	0.300	0.0029
		3.0	14544.49	0.0167	0.0201	0.1869	0.1852	0.7825	0.915	0.408	0.0093	0.0047	0.0159	0.0184	0.0656	0.793	0.817	0.260	0.0038
		4.0	14548.25	0.0173	0.0247	0.1895	0.1870	0.7786	0.910	0.410	0.0146	0.0048	0.0140	0.0193	0.0711	0.776	0.780	0.234	0.0054
		5.0	14552.60	0.0178	0.0274	0.1885	0.1866	0.7836	0.907	0.393	0.0194	0.0048	0.0120	0.0199	0.0760	0.769	0.762	0.216	0.0078
	500	-	35920.40	0.0065	0.0125	0.0932	0.1050	0.7362	0.983	0.618	-	0.0019	0.0109	0.0079	0.0334	0.909	0.951	0.186	-
		1.0	35801.68	0.0063	0.0043	0.0592	0.0531	0.9032	0.978	0.205	0.0013	0.0018	0.0069	0.0051	0.0170	0.939	0.965	0.120	0.0005
		2.0	35788.92	0.0062	0.0057	0.0588	0.0608	0.9320	0.976	0.132	0.0025	0.0018	0.0069	0.0051	0.0189	0.948	0.965	0.090	0.0008
		3.0	35787.43	0.0061	0.0088	0.0583	0.0679	0.9401	0.976	0.112	0.0037	0.0018	0.0073	0.0052	0.0230	0.948	0.960	0.079	0.0012
		4.0	35790.11	0.0062	0.0122	0.0600	0.0733	0.9432	0.975	0.103	0.0051	0.0018	0.0077	0.0055	0.0273	0.945	0.951	0.070	0.0016
		5.0	35792.85	0.0063	0.0141	0.0599	0.0759	0.9415	0.971	0.101	0.0068	0.0018	0.0075	0.0057	0.0302	0.938	0.939	0.066	0.0021
	1000	-	71359.53	0.0032	0.0073	0.0323	0.0591	0.8111	0.997	0.456	-	0.0009	0.0045	0.0031	0.0181	0.971	0.979	0.049	-
		1.0	71183.06	0.0030	0.0036	0.0156	0.0187	0.9553	0.993	0.099	0.0006	0.0008	0.0029	0.0016	0.0056	0.985	0.993	0.034	0.0002
		2.0	71170.20	0.0030	0.0033	0.0154	0.0222	0.9757	0.993	0.049	0.0012	0.0008	0.0028	0.0015	0.0071	0.989	0.994	0.022	0.0003
		3.0	71167.77	0.0030	0.0048	0.0155	0.0263	0.9837	0.993	0.029	0.0019	0.0008	0.0031	0.0016	0.0096	0.989	0.993	0.019	0.0005
		4.0	71168.79	0.0030	0.0061	0.0156	0.0294	0.9859	0.992	0.024	0.0026	0.0008	0.0034	0.0016	0.0117	0.989	0.992	0.018	0.0007
		5.0	71170.29	0.0030	0.0072	0.0156	0.0308	0.9865	0.992	0.022	0.0033	0.0008	0.0035	0.0017	0.0134	0.986	0.988	0.017	0.0009

Table 3.1: Simulation Study I: Performance measures for the MML (first row for each combination of number of items and sample size) and the PMML estimation of a heteroscedastic Normal linear factor model with sparse factor loadings matrices for the location (μ) and scale (σ) parameters. Results by number of items (p), sample size (n), and influence factor (γ). AvMSE stands for the average Mean Squared Error across simulations, AvAB for the average Absolute Bias across simulations, AvCER for average Correct Estimation Rate across simulations, TPR for True Positive Rate, and FPR for False Positive Rate. Results for the Alasso penalty with automatic selection of the tuning parameter vector $\lambda = (\lambda_\mu, \lambda_\sigma)$, with additional parameter $a = 2$.

3.6.3 Simulation Study II: Heteroscedastic Beta factor model

In our second simulation study, we focus on a one-factor model with continuous items in the interval $(0, 1)$. Some of these items exhibit heteroscedasticity along with the latent variable. A related framework was proposed by Verkuilen and Smithson (2012), who introduced a model with random effects on the scale parameter. In our approach, we generate items that, conditional on the latent variable $\mathbf{z}$, follow a location-scale parameterization of the Beta distribution: $y_i \mid \mathbf{z} \sim \text{Beta}(\mu_i(\mathbf{z}), \sigma_i(\mathbf{z}))$. Here, $\mu_i \in (0, 1)$ is a location parameter, and $\sigma_i \in (0, 1)$ is a scale parameter (refer to Section 2.4.1 and Appendix A1.1 for more details). Under this parameterization, $\mathbb{E}(y_i|\mathbf{z}) = \mu_i$, and the conditional variance is $\text{var}(y_i \mid \mathbf{z}) = \sigma_i^2 \mu_i(1 - \mu_i)$. While this model can be extended to include multiple factors, for computational simplicity, we maintain it as unidimensional in the latent variable space. The measurement equations for μ_i and σ_i are given by:

$$\text{logit}(\mu_i(\mathbf{z})) = \alpha_{i0,\mu} + \alpha_{i1,\mu} z_1$$

$$\text{logit}(\sigma_i(\mathbf{z})) = \alpha_{i0,\sigma} + \alpha_{i1,\sigma} z_1$$

The population intercepts and slopes in the measurement equation for μ_i are randomly drawn from uniform distributions: $\alpha_{i0,\mu} \sim \text{Unif}(-1.5, 1.5)$ and $\alpha_{i1,\mu} \sim \text{Unif}(0.5, 1.5)$, respectively. The signs of the factor loadings $\alpha_{i1,\mu}$ are assigned at random with probability 0.5. Similarly, the population intercepts and slopes in the measurement equation for σ_i are generated from uniform distributions: $\alpha_{i0,\sigma} \sim \text{Unif}(-2.5, -0.5)$ and $\alpha_{i1,\sigma} \sim \text{Unif}(0.3, 0.6)$. To create homoscedastic items, where the scale parameter remains constant along the latent scale, we randomly set some factor loadings in the measurement equation for the scale parameter to zero. The signs of the non-zero factor loadings are assigned randomly as described earlier. We generate the true parameters in this way to ensure that the conditional densities $f_i(y_i \mid \mathbf{z})$ are uni-modal. The $L = 300$ datasets were randomly generated using the same set of factor loadings. While the Beta distribution allows for bimodal densities under certain combinations of μ_i and σ_i, this is not common in the applications of interest for this study. Noel (2014) proposed a unidimensional Beta factor model that can handle the bi-modality of items if necessary, but the scale parameter is assumed to be constant along the latent scale.

The simulation results are summarised in Table 3.2, and the first row of each section presents the results of the (unpenalised) MML estimation for comparison. Based on the results, the following conclusions are drawn:

- The proposed penalised estimation framework consistently outperforms the MML estimation

benchmark in terms of model selection. In all cases, the penalised estimation yields lower GBIC values compared to the unpenalised estimation.

- Generally, higher values of the influence factor lead to better estimation results. This is supported by the lower GBIC values, higher CERs and TPRs, and lower FPRs for $\gamma = 4$ with small sample sizes and $\gamma = 5$ for medium and large sample sizes.

- Regarding sparsity recovery, the automatic selection of tuning parameters $\boldsymbol{\lambda}$ leads to TPR values close to 1.0, indicating that the non-zero factor loadings in the scale measurement equation are correctly estimated in most cases. The FPR decreases and converges to zero as the sample size increases, as expected. However, for the MLEs, the FPR is considerably worse for $n = 200$ (computed using a threshold of $\pm$ 0.1) but improves with larger sample sizes.

- For a given sample size and test length, the average optimal value of $\hat{\lambda}_\sigma$ increases with γ. This can be attributed to the higher importance of the effective degrees of freedom (edf) term, which captures the model complexity, in the optimisation problem (3.19).

				Location Parameter (μ)								Scale Parameter (σ)							
p	n	Influence factor (γ)	Avg. GBIC	Intercepts ($\hat{\alpha}_{0,\mu}$)		Loadings ($\hat{\boldsymbol{\alpha}}_{i,\mu}$)						Intercepts ($\hat{\alpha}_{0,\sigma}$)		Loadings ($\hat{\boldsymbol{\alpha}}_{i,\sigma}$)					
				AvMSE	AvAB	AvMSE	AvAB	AvCER	TPR	FPR	Avg. λ_μ	AvMSE	AvAB	AvMSE	AvAB	AvCER	TPR	FPR	Avg. λ_σ
10	200	-	-3405.47	0.0060	0.0051	0.0040	0.0081	1.0000	1.000	0.000	-	0.0056	0.0140	0.0051	0.0042	0.949	0.999	0.126	-
		1.0	-3423.68	0.0060	0.0023	0.0047	0.0282	1.0000	1.000	0.000	0.0045	0.0055	0.0138	0.0043	0.0145	0.986	0.999	0.033	0.0027
		2.0	-3423.97	0.0059	0.0013	0.0053	0.0366	1.0000	1.000	0.000	0.0084	0.0055	0.0138	0.0044	0.0182	0.985	0.999	0.035	0.0036
		3.0	-3424.26	0.0060	0.0019	0.0054	0.0356	1.0000	1.000	0.000	0.0128	0.0055	0.0139	0.0050	0.0220	0.992	0.998	0.018	0.0056
		4.0	-3424.98	0.0060	0.0027	0.0050	0.0305	1.0000	1.000	0.000	0.0170	0.0055	0.0139	0.0052	0.0222	0.993	0.996	0.010	0.0080
		5.0	-3424.91	0.0060	0.0024	0.0056	0.0370	1.0000	1.000	0.000	0.0214	0.0055	0.0138	0.0057	0.0264	0.993	0.992	0.005	0.0107
	500	-	-8726.63	0.0025	0.0012	0.0017	0.0053	1.0000	1.000	0.000	-	0.0023	0.0063	0.0019	0.0051	0.995	1.000	0.012	-
		1.0	-8755.22	0.0024	0.0011	0.0019	0.0140	1.0000	1.000	0.000	0.0018	0.0022	0.0064	0.0014	0.0079	0.995	1.000	0.013	0.0011
		2.0	-8755.37	0.0024	0.0013	0.0020	0.0166	1.0000	1.000	0.000	0.0033	0.0022	0.0063	0.0015	0.0093	0.994	1.000	0.016	0.0015
		3.0	-8755.66	0.0024	0.0010	0.0019	0.0154	1.0000	1.000	0.000	0.0051	0.0022	0.0065	0.0015	0.0103	0.996	1.000	0.011	0.0023
		4.0	-8756.02	0.0024	0.0010	0.0019	0.0146	1.0000	1.000	0.000	0.0068	0.0022	0.0065	0.0015	0.0104	0.998	1.000	0.005	0.0031
		5.0	-8756.26	0.0024	0.0010	0.0019	0.0152	1.0000	1.000	0.000	0.0085	0.0022	0.0065	0.0015	0.0109	0.999	1.000	0.002	0.0040
	1000	-	-17630.13	0.0012	0.0015	0.0009	0.0035	1.0000	1.000	0.000	-	0.0010	0.0039	0.0009	0.0037	1.000	1.000	0.000	-
		1.0	-17662.45	0.0012	0.0017	0.0009	0.0076	1.0000	1.000	0.000	0.0009	0.0010	0.0040	0.0007	0.0053	0.994	1.000	0.014	0.0006
		2.0	-17662.73	0.0012	0.0018	0.0010	0.0089	1.0000	1.000	0.000	0.0017	0.0010	0.0040	0.0007	0.0059	0.998	1.000	0.006	0.0008
		3.0	-17662.92	0.0012	0.0016	0.0010	0.0079	1.0000	1.000	0.000	0.0025	0.0010	0.0041	0.0007	0.0063	0.999	1.000	0.003	0.0011
		4.0	-17663.14	0.0012	0.0015	0.0009	0.0077	1.0000	1.000	0.000	0.0034	0.0010	0.0041	0.0007	0.0063	0.999	1.000	0.003	0.0015
		5.0	-17663.29	0.0012	0.0015	0.0009	0.0080	1.0000	1.000	0.000	0.0043	0.0010	0.0041	0.0007	0.0064	1.000	1.000	0.000	0.0019
20	200	-	-8372.39	0.0084	0.0042	0.0044	0.0082	1.0000	1.000	0.000	-	0.0050	0.0133	0.0044	0.0038	0.952	1.000	0.107	-
		1.0	-8398.24	0.0084	0.0037	0.0056	0.0279	1.0000	1.000	0.000	0.0043	0.0050	0.0119	0.0035	0.0103	0.961	1.000	0.087	0.0020
		2.0	-8400.53	0.0082	0.0028	0.0065	0.0373	1.0000	1.000	0.000	0.0085	0.0049	0.0115	0.0036	0.0148	0.981	0.999	0.041	0.0032
		3.0	-8400.41	0.0082	0.0033	0.0076	0.0410	1.0000	1.000	0.000	0.0128	0.0049	0.0112	0.0042	0.0198	0.994	0.999	0.012	0.0052
		4.0	-8402.15	0.0085	0.0044	0.0057	0.0294	1.0000	1.000	0.000	0.0169	0.0049	0.0115	0.0041	0.0188	0.995	0.999	0.010	0.0074
		5.0	-8401.62	0.0084	0.0046	0.0072	0.0412	1.0000	1.000	0.000	0.0213	0.0049	0.0111	0.0047	0.0232	0.995	0.995	0.006	0.0099
	500	-	-21380.25	0.0038	0.0076	0.0022	0.0083	1.0000	1.000	0.000	-	0.0020	0.0064	0.0017	0.0037	0.998	1.000	0.005	-
		1.0	-21420.12	0.0042	0.0051	0.0025	0.0192	1.0000	1.000	0.000	0.0018	0.0020	0.0057	0.0012	0.0074	0.991	1.000	0.021	0.0010
		2.0	-21420.95	0.0041	0.0056	0.0027	0.0223	1.0000	1.000	0.000	0.0034	0.0020	0.0056	0.0012	0.0089	0.995	1.000	0.012	0.0014
		3.0	-21421.37	0.0041	0.0052	0.0028	0.0226	1.0000	1.000	0.000	0.0051	0.0020	0.0056	0.0013	0.0103	0.997	1.000	0.006	0.0021
		4.0	-21422.06	0.0041	0.0051	0.0026	0.0200	1.0000	1.000	0.000	0.0068	0.0020	0.0056	0.0013	0.0100	0.999	1.000	0.003	0.0029
		5.0	-21422.43	0.0041	0.0051	0.0027	0.0215	1.0000	1.000	0.000	0.0086	0.0020	0.0056	0.0013	0.0106	0.999	1.000	0.002	0.0037
	1000	-	-43182.16	0.0022	0.0017	0.0012	0.0020	1.0000	1.000	0.000	-	0.0010	0.0027	0.0008	0.0026	1.000	1.000	0.001	-
		1.0	-43138.76	0.0026	0.0010	0.0015	0.0181	1.0000	1.000	0.000	0.0009	0.0011	0.0023	0.0006	0.0063	0.993	1.000	0.016	0.0005
		2.0	-43140.25	0.0026	0.0012	0.0015	0.0199	1.0000	1.000	0.000	0.0017	0.0011	0.0023	0.0006	0.0071	0.998	1.000	0.004	0.0007
		3.0	-43140.55	0.0026	0.0011	0.0016	0.0195	1.0000	1.000	0.000	0.0026	0.0011	0.0024	0.0007	0.0077	0.999	1.000	0.002	0.0011
		4.0	-43140.92	0.0026	0.0011	0.0015	0.0188	1.0000	1.000	0.000	0.0034	0.0011	0.0024	0.0007	0.0076	0.999	1.000	0.003	0.0014
		5.0	-43141.15	0.0026	0.0012	0.0015	0.0192	1.0000	1.000	0.000	0.0043	0.0011	0.0024	0.0007	0.0078	1.000	1.000	0.001	0.0018

Table 3.2: Simulation Study II: Performance measures for the MML (first row for each combination of number of items and sample size) and the PMML estimates of a Beta factor model with unknown heteroscedastic items. Results by number of items (p), sample size (n), and influence factor (γ). AvMSE stands for the average Mean Squared Error across simulations, AvAB for the average Absolute Bias across simulations, AvCER for average Correct Estimation Rate across simulations, TPR for True Positive Rate, and FPR for False Positive Rate. Results for the Alasso penalty with automatic selection of the tuning parameter for the scale parameter λ_σ, with additional parameter $a = 2$.

3.7. Empirical Applications

3.7.1 PISA 2018: A semi-confirmatory joint model for item response and response times

In Section 2.5.1, we presented a confirmatory GLVM-LSS model for item responses (IR) and response times (RT), using data from the 2018 PISA computer-based mathematics exam. Our study focused on a sample of Brazilian students, consisting of $n = 1280$ individuals. For our analysis, we selected nine binary items from the first testlet, along with their corresponding response times. The purpose was to explore the relationship between IR and RT, which has received significant attention in educational research literature because it provides valuable insights into students' abilities, test-taking strategies, and also helps in item calibration and test design (van der Linden, 2007, 2008; van der Linden and Guo, 2008; van der Linden et al., 2010). The relationship between IR and RT is often associated with the concept of the 'speed-accuracy trade-off' (Zimmerman, 2011). This trade-off suggests that individuals who take more time and respond slowly tend to achieve higher scores on exams compared to their peers who respond quickly but make more errors.

The model consists of, on one hand, IR that follow a Bernoulli distribution conditional on the latent ability (z_1), $y_i \mid \mathbf{z} \sim \text{Bernoulli}(\pi_i(z_1))$. The measurement equations for the location parameter of the IR (the probability of responding correctly) are:

$$\text{logit}(\pi_i) = \alpha_{i0,\pi} + \alpha_{i1,\pi} z_1 \tag{3.21}$$

Here, $\alpha_{i0,\pi}$ and $\alpha_{i1,\pi}$ represent item i's difficulty and discrimination parameters, respectively. On the other hand, the logarithm of RT (log-RT) are assumed to follow a Skew-Normal distribution conditional on the latent speed trait (z_2), $\log(t_i) \mid \mathbf{z} \sim \text{SN}(\mu_i(z_2), \sigma_i(z_2), \nu_i(z_2))$. The location ($\mu_i \in \mathbb{R}$), scale ($\sigma_i \in \mathbb{R}^+$), and shape ($\nu_i \in (0, 1)$) parameters characterising this re-parameterised version of the SN distribution (see Appendix A1.1) are modelled as linear functions of z_2. After choosing appropriate link functions for the distributional parameters, the measurement equations for the distributional parameters of the log-RT are given by:

$$\mu_i = \alpha_{i0,\mu} + \alpha_{i1,\mu} z_2 \tag{3.22}$$

$$\log(\sigma_i) = \alpha_{i0,\sigma} + \alpha_{i1,\sigma} z_2 \tag{3.23}$$

$$\text{logit}(\nu_i) = \alpha_{i0,\nu} + \alpha_{i1,\nu} z_2 \tag{3.24}$$

Furthermore, the latent ability and the latent speed trait are correlated and assumed to follow a multivariate Normal distribution, $(z_1, z_2)^\intercal \sim \mathbb{N}(\mathbf{0}, \mathbf{\Phi})$. For identifiability purposes, we assume $\mathbf{\Phi}$ is a correlation matrix (i.e., $\text{diag}(\mathbf{\Phi}) = 1$), with diagonal elements denoted by $\phi_{12} = \phi_{21} = \phi_{\mathbf{z}}$.

We fit the GLVM-LSS model described above using the penalised estimation framework outlined in Section 3.3. We refer to this model as a 'semi-confirmatory' GLVM-LSS since i) we inform the model structure by assuming that IRs only depend on the latent ability and the log-RTs only depend on the latent speed trait, but ii) we also allow all the distributional parameters characterising the log-RTs to be functions of z_2. In reality, some items may not exhibit heteroscedasticity and/or varying skewness across the latent speed trait. The penalised estimation deals with this by effectively shrinking unnecessary factor loadings in the scale and shape measurement equations for log-RT towards zero. The proposed penalised estimation framework strikes a balance between a confirmatory model structure, where specific relationships between latent variables and the IR and RT are pre-specified, and an exploratory model that incorporates data-driven insights regarding the presence of heteroscedasticity and varying skewness.

We use the Alasso penalty due to its oracle property, and the additional parameter in the penalty term is set to $a = 2$, which is a common practice in the literature for model selection (Zou, 2006). In this context, only the factor loadings are penalised, $\Theta_p^\intercal = \text{vec(A)}$, while the intercepts and factor correlations remain unpenalised, $\Theta_u^\intercal = (\boldsymbol{\alpha}_0^\intercal, \text{vech}(\mathbf{\Phi})^\intercal)$. Six different models are estimated, each corresponding to a different value of the influence factor, $\gamma = \{1, \ldots, 6\}$. These varying values allow for different weights on the model complexity term of the approximate UBRE in the automatic model selection procedure. The initial parameter values and the weights in the Alasso penalty are based on the unpenalised model parameter estimates, specifically the MLEs for Model 7 in Section 2.5.1. The results of this analysis are summarised in Table 3.3.

The model with the best fit, as indicated by the GBIC, is achieved by setting the influence factor to $\gamma = 5$. The effective degrees of freedom (edf) measure the complexity of the selected model. In this case, the edf is calculated to be 61.31, which is substantially lower than the number of estimated parameters in the full model (73). The estimated parameters for the selected model are presented in Table 3.4. The factor loadings in the measurement equations for the location parameters characterising the IRs ($\hat{\alpha}_{i1,\pi}$'s) and the log-RTs ($\hat{\alpha}_{i1,\mu}$'s) are similar to those obtained from the unpenalised solution (Table 2.7).

However, some of the factor loadings in the measurement equations for the scale ($\hat{\alpha}_{i1,\sigma}$'s) and shape ($\hat{\alpha}_{i1,\nu}$'s) parameters have been shrunk towards zero due to the penalisation. These findings highlight the presence of heteroscedasticity and varying skewness in the log-RT of certain items,

Model	GBIC	EDF	$\hat{\lambda}_\mu$	$\hat{\lambda}_\sigma$	$\hat{\lambda}_\nu$
Unpenalised	25548.18	73.00	-	-	-
Penalised					
$\gamma = 1$	25509.28	67.14	2.55e-04	5.34e-05	3.80e-04
$\gamma = 2$	25499.03	65.05	5.14e-04	1.40e-04	9.65e-04
$\gamma = 3$	25491.53	63.26	8.00e-04	2.63e-04	1.91e-03
$\gamma = 4$	25492.64	61.89	1.05e-03	2.86e-04	2.94e-03
$\gamma = 5$	**25491.32**	61.31	1.38e-03	4.03e-04	3.99e-03
$\gamma = 6$	25492.27	60.60	1.65e-03	5.35e-04	5.36e-03

Table 3.3: PISA 2018: Model fit and model complexity results for the MML and PMML estimation of the joint model for item responses and response times. The influence factor γ controls the relative importance of the model complexity term, given by the effective degrees of freedom (EDF). $\hat{\lambda}_\mu$ is the estimated tuning parameter for the location parameter, $\hat{\lambda}_\sigma$ the estimated tuning parameter for the scale parameter, and $\hat{\lambda}_\nu$ the estimated tuning parameter for the shape parameter.

indicating that the whole conditional distribution of the log-RTs, and not only their conditional means, are influenced by the latent speed trait. In particular, from the penalised solution it can be observed that the log-RT for Items 2, 3, 5, and 8 display some degree of heteroscedasticity, although not to a significant extent. The factor loadings in the scale measurement equations for the rest of the items have been penalised towards zero, indicating homoscedasticity. In terms of the shape parameter, we observe varying skewness along the latent speed trait for the log-RT of Items 2, 3, 6, 7, and 8.

As an example, Figure 3.1 shows the response times (in log scale) and the associated fitted SN distributions parameterised by the factor loading in Table 3.4 for items 8 and 9. Item 8 has full effects of the latent speed trait on the location (μ_8), scale (σ_8) and scale (ν_8) parameters. Figure 3.1a shows how the conditional distribution of $\log(t_8)$ changes for different levels of the the latent speed trait. On the contrary, item 9 (Figure 3.1b) has constant variance and skewness. Indeed, the implied value of the skeweness parameter ($\text{logit}(0.39) = 0.6$) suggests that the marginal distribution of $\log(t_9)$ is almost symmetrical (as $\nu \approx \text{logit}(0) = 0.5$, the reparameterised SN distribution discussed in Appendix A1.1 tends to the Normal distribution).

The estimated correlation between the latent ability and the latent speed trait is $\hat{\phi}_{\mathbf{z}} = -0.29$ for the selected model[3]. This suggest that, for this test and sample, the speed-accuracy trade-off hypothesis holds. Previous studies have found correlations between the latent ability and the latent speed trait of similar magnitude in the context of large scale educational testing of quantitative

[3]The estimated correlation between the latent ability and the latent speed trait was -0.28 for the penalised models with influence factor $\gamma = \{1, 2, 3\}$ and -0.29 for $\gamma = \{4, 5, 6\}$.

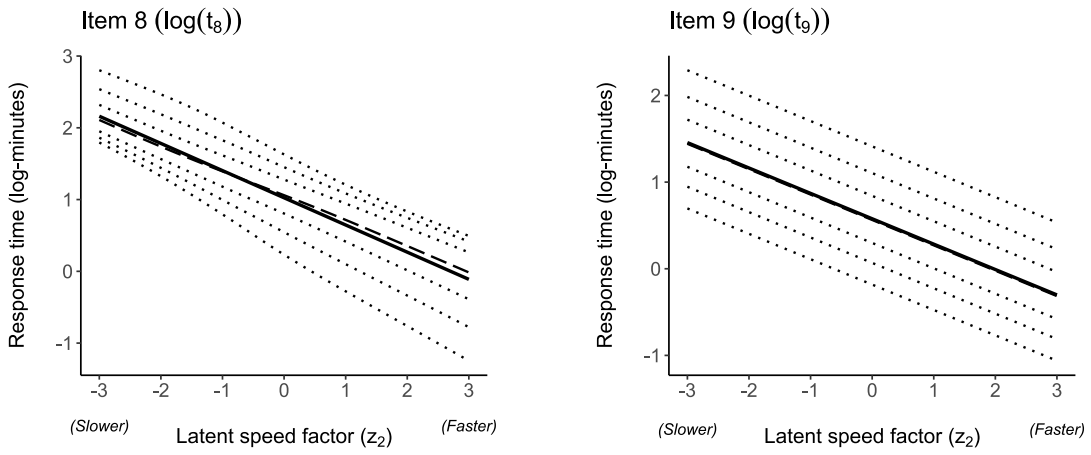
	Parameters for item responses (IR)		Parameters for response times (log-RT)					
Item	$\hat{\alpha}_{i0,\pi}$	$\hat{\alpha}_{i1,\pi}$	$\hat{\alpha}_{i0,\mu}$	$\hat{\alpha}_{i1,\mu}$	$\hat{\alpha}_{i0,\sigma}$	$\hat{\alpha}_{i1,\sigma}$	$\hat{\alpha}_{i0,\nu}$	$\hat{\alpha}_{i1,\nu}$
Item 1	0.64	0.77	0.19	-0.16	-0.95		0.65	
Item 2	-0.46	1.02	0.30	-0.22	-0.89	-0.06	1.59	-0.68
Item 3	-0.04	1.94	0.43	-0.24	-0.60	-0.04	-1.04	0.73
Item 4	-0.69	0.96	0.45	-0.34	-0.86		-1.27	
Item 5	-2.84	2.28	1.00	-0.35	-0.68	0.11	-1.01	
Item 6	-0.90	0.27	0.16	-0.35	-0.97		0.11	-0.78
Item 7	-4.79	2.48	0.65	-0.32	-1.15		0.43	-0.39
Item 8	-3.67	2.39	1.02	-0.38	-1.03	0.09	-1.20	-1.59
Item 9	-2.73	1.45	0.58	-0.29	-0.90		0.39	
Estimated latent correlation (z_1, z_2): $\hat{\phi}_{\mathbf{z}} = -0.29$								

Table 3.4: PISA 2018: Estimated coefficients of the penalised model for joint model of item responses and response times (full model). Alasso penalty with additional parameter $a = 2$, influence factor $\gamma = 5$. Blank spaces correspond to factor loadings that have been shrunk to zero in the estimation process.

subjects (see, e.g., van der Linden and Guo, 2008).

Further investigation by researchers with substantive knowledge in the field is necessary to explore potential explanations for why certain items display heteroscedasticity and/or varying skewness along the latent speed trait dimension in the log-RT. It would be valuable to consider factors such as the wording or content of the exam questions, as they may influence how students with different latent abilities process information, leading to heterogeneous response times across different levels of the latent speed trait. Future research should also aim to explore the implications of items displaying heteroscedasticity or varying skewness in relation to important psychometric concepts such as measurement invariance or differential item functioning. Additionally, investigating whether these findings reflect poor item quality would be beneficial.

Overall, the main advantage of the proposed penalised estimation framework is its ability to efficiently and rapidly identify items that exhibit higher order moments depending on the latent variables. This allows researchers to gain insights into the underlying structure of the data and potentially identify areas for further investigation.



(a) (b)

Figure 3.1: PISA 2018: Fitted conditional expected values (solid line, ——), median (dashed line, - - -), and percentiles (dotted lines, ·····) for log-RTs of items 8 and 9.

3.7.2 The Holzinger and Swineford (1939) dataset

The Holzinger and Swineford dataset (Holzinger and Swineford, 1939) is a classical example in Psychometrics containing measures of mental ability test scores of a sample of $n = 301$ seventh- and eighth-grade children from two different schools. This dataset (or subsets of it) has been used in applications of CFA (Jöreskog, 1969), EFA (Browne, 2001), and several applications of penalised Normal linear factor model (Trendafilov et al., 2017; Jacobucci et al., 2016; Huang et al., 2017; Jin et al., 2018). We focus on nine mental ability tests measuring three correlated latent traits: spatial visualisation (Items 1, 2, 3), verbal intelligence (Items 4, 5, 6), and speed (Items 7, 8, 9). We estimate the model parameters for a series of GLVM-LSS models with Normally distributed items, including homoscedastic and heteroscedastic models, using the proposed penalised marginal maximum likelihood procedure with automatic selection of the tuning parameters. The model structure is similar to that in Simulation Study I.

Due to its oracle property, we choose the Alasso penalty (with additional parameter $a = 2$). Moreover, we tried a sequence of values for the influence factor $\gamma = \{1, ..., 6\}$. The data set was mean-centred, so no intercept is included in the measurement equations for the location parameter[4]. We penalise the factor loadings in the location and scale measurement equations, $\Theta_p^\intercal = \text{vec(A)}$, while the intercepts and factor correlations are unpenalised, $\Theta_u^\intercal = (\boldsymbol{\alpha}_0^\intercal, \text{vech}(\boldsymbol{\Phi})^\intercal)$. For comparison purposes, we also fit unpenalised exploratory models (EFA), where we impose the minimum number

[4]Previous papers using this dataset (e.g., Jin et al., 2018) also divide each column by its corresponding standard deviation. We avoid this practice to better show how the scale parameter varies along the latent variable scale. The standard deviations of the 9 items range between 1.01 and 1.29, thus we do not expect scaling to affect the results significantly.

of restrictions on the model parameters for identification purposes allowing the latent variables to be correlated; and confirmatory models (CFA) with correlated latent variables, where we impose restrictions on the model parameters such that the factor loading matrices for the location and scale parameters, A_μ and A_σ, have simple structures. Results are summarised in Table 3.5.

	Homoscedastic Model			Heteroscedastic Model			
Model	GBIC	EDF	$\hat{\lambda}_\mu$	GBIC	EDF	$\hat{\lambda}_\mu$	$\hat{\lambda}_\sigma$
Unpenalised							
EFA	7601.12	33.00	-	7621.80	60.00	-	-
CFA	7595.20	21.00	-	7559.00	30.00	-	-
Penalised							
$\gamma = 1$	7570.24	26.95	6.13e-04	7522.71	41.01	6.13e-04	1.84e-12
$\gamma = 2$	7562.53	24.96	1.75e-03	7502.75	32.11	1.75e-03	2.21e-06
$\gamma = 3$	7562.77	22.98	6.07e-03	**7498.59**	28.96	6.07e-03	2.73e-10
$\gamma = 4$	7564.17	22.31	8.89e-03	7500.60	26.97	8.89e-03	6.17e-16
$\gamma = 5$	7565.60	21.75	1.04e-02	7507.30	29.41	1.04e-02	6.59e-03
$\gamma = 6$	7568.25	21.29	1.49e-02	7541.99	40.41	1.49e-02	2.21e-06

Table 3.5: Holzinger and Swineford dataset: Model fit and model complexity results for the MML and PMML estimation of homoscedastic and heteroscedastic Normal linear factor models. The influence factor γ controls the relative importance of the model complexity term, given by the effective degrees of freedom (EDF). $\hat{\lambda}_\mu$ is the estimated tuning parameter for the location parameter and $\hat{\lambda}_\sigma$ the estimated tuning parameter for the scale parameter.

Several important conclusions are derived from the results in Table 3.5. Firstly, notice that for the homoscedastic and heteroscedastic models, both the unpenalised EFA and the CFA models had worse fit than the penalised models as judged by their corresponding GBICs. This is probably due to the unnecessary complexity of the former, and the strict assumption of no cross-loadings of the latter. This result shows that the introduction of sparsity via the penalised estimation benefited the analysis, and, as argued by Huang et al. (2017), that complex models do not necessarily outperform simpler ones when complexity is also taken into account in the model selection criterion.

Secondly, in all cases but the EFA, the heteroscedastic model provides a better fit than the homoscedastic model. This means that some items in this data set display heteroscedasticity along the latent variables scale. In particular, the penalised heteroscedastic model with influence factor $\gamma = 3$ yields the best fit as suggested by its GBIC value (7498.59). Model parameter estimates are reported in Table 3.6. Thirdly, a closer inspection of the effective degrees of freedom (edf) shows that, for $\gamma = 3$, the heteroscedastic model (edf = 28.96) is only slightly more complex than its homoscedastic counterpart (edf = 22.98). However, for the unpenalised EFA, the edf (i.e., the total number of estimated parameters) go from 33 to 60 in the homoscedastic and heteroscedastic

Measurement model							
	Location parameter (μ_i)			Scale parameter (σ_i)			
	$\hat{\alpha}_{ij,\mu}$			$\hat{\alpha}_{i0,\sigma}$	$\hat{\alpha}_{ij,\sigma}$		
Item	Spatial	Verbal	Speed	Intercept	Spatial	Verbal	Speed
Item 1	0.808	0	0	-0.207			
Item 2	0.486			0.021	0.141		-0.048
Item 3	0.783	-0.182		-0.215	0.215		
Item 4	0	0.943	0	-0.449			
Item 5	-0.003	1.050		-0.431	-0.019	-0.082	0.038
Item 6		0.808		-0.593		0.288	
Item 7	0	0	0.721	-0.243			0.119
Item 8	0.062		0.705	-0.467		-0.120	0.242
Item 9	0.395		0.427	-0.314			0.150
Structural model							
Spatial	1						
Verbal	0.472	1					
Speed	0.215	0.183	1				

Table 3.6: Holzinger and Swineford dataset: Estimated coefficients for the penalised heteroscedastic Normal linear factor model. Alasso penalty with additional parameter $a = 2$, influence factor $\gamma = 3$. Underlined parameters are fixed to their respective values for identification purposes. Blank spaces correspond to factor loadings that were shrunk to zero in the estimation process.

models, respectively. This suggests that including latent variable effects on the scale parameter, but also fitting the model via the proposed PMML estimation framework, improves model fit while keeping model complexity at levels only supported by the data.

From the parameter estimates in Table 3.6, we see that the penalised solution recovered a nearly perfect structure in the factor loading matrix for the location parameter, except for four cross-loadings ($\hat{\alpha}_{51,\mu}$, $\hat{\alpha}_{81,\mu}$, $\hat{\alpha}_{91,\mu}$, $\hat{\alpha}_{32,\mu}$) which were identified as non-zero. This result is consistent to with previous results in the penalised LVM literature using this dataset (e.g., Jin et al., 2018; Geminiani et al., 2021). The factor loading matrix for the scale parameter tells a similar story. Most items display some degree of heteroscedasticity along the latent dimension that they were originally designed to measure. This suggests heterogeneity in the responses of the tests across different levels of the latent trait. However, we see that $\hat{\alpha}_{ij,\sigma}$'s for Items 1 and 4 were shrunk to zero, meaning that these items are homoscedastic.

3.8. Discussion

In this Chapter, we present a penalised marginal maximum likelihood estimation with automatic selection of tuning parameters for the class of Generalised Latent Variable Models for Location, Scale, and Shape parameters (GLVM-LSS) introduced in Chapter 2. The GLVM-LSS framework extends the traditional Generalised Linear Latent Variable Model (GLLVM, Skrondal and Rabe-Hesketh, 2004; Bartholomew et al., 2011) by explicitly modelling the location, scale, and shape parameters parameters characterising the items' conditional distributions as linear functions of the latent variables. This also allows for exploring the relationships between the latent variables and the items' higher order moments.

In most applications, it is of interest to obtain sparse factor loading matrices where most items load on a few latent variables and the cross loadings are close to zero. While rotation techniques help to produce factor loading matrices with simple structures (Mulaik, 2009, Chapters 10-12), in practice, the rotated solutions often end up being overly dense. Moreover, most rotation methods resort to hard-thresholding of estimated factor loadings when deciding which parameters are sufficiently close to zero (Hair et al., 2010). This approach is not only subjective in nature, but also affects the effective degrees of freedom of the model, which are used to evaluate model fit.

An alternative is to fit the model via penalised estimation. By introducing a sparsity-inducing term in the objective function of the estimation problem, we simultaneously estimate the model parameters and obtain a sparse solution. Unlike rotation methods, 'simple structures' are only obtained if supported by the data. Penalised maximum likelihood estimation has recently gained traction in the LVM literature (see, e.g. Hirose and Yamamoto, 2014; Chen et al., 2015; Sun et al., 2017; Trendafilov et al., 2017; Huang et al., 2017; Jin et al., 2018; Battauz, 2020, to name but a few). The level of sparsity is determined by a non-negative tuning parameter, which is typically selected by fitting a sequence of candidate models with different values of the tuning parameter defined over a (usually one-dimensional) grid, and then picking the optimal value that yields the lowest information criteria. However, for models with multiple tuning parameters defined over a multi-dimensional grid, this process can be computationally expensive and time-consuming, which limits the application of these methods in applied research. Recently, Geminiani et al. (2021) proposed a penalised estimation for the Normal linear factor model that uses an automatic procedure for selecting the optimal value of the tuning parameter. The selection of the tuning parameter is is based on minimising an approximate unbiased risk estimate, which is proportional to the AIC. This minimisation problem requires smooth approximations of the L_1-penalty functions. Their framework provides a unified estimation and inferential framework for penalised estimation of

Normal linear factor models.

Our proposed penalised framework is an extension of the approach in Geminiani et al. (2021) to LVMs that do not have closed-form solutions. By using local approximations of the L_1 penalties, we are able to implement a two-step penalised marginal maximum likelihood estimation strategy that combines the advantages of the EM-algorithm and (quasi-)Newton algorithms. In the GLVM-LSS framework, the level of sparsity in the factor loadings for the location, scale, and shape parameters are determined independently by a vector tuning parameters. As such, the automatic procedure provides an efficient and computationally convenient way of selecting the optimal value of the tuning parameter. We demonstrate the properties of the proposed estimation method through simulation studies and empirical applications in educational testing. The code and replication files are available online.

While the penalised estimation with automatic selection of the tuning parameter provides a computationally efficient and theoretically grounded way of obtaining sparse factor loading matrices in the GLVM-LSS context, there are still some aspects that can be improved in future research. One important challenge of the proposed framework is to obtain standard errors for the penalised parameter estimates using the locally approximated penalty term. The asymptotic properties of the penalised estimator with L_1-norm based penalties have been extensively studied in the regression context (see, e.g., Fan and Li, 2001; Zou, 2006). However, to the best of our knowledge, no work has addressed how (and/or if) the asymptotic properties of the penalised estimator are influenced by the local approximation of the L_1-norm penalty term. We conjecture that the oracle property of the Alasso is not affected by the local approximation, and that potentially depends on the constant $\bar{c}$ that determines the closeness between the local approximation and the L_1-norm. Formally, this would imply that $\lim_{n\to\infty} \lim_{\bar{c}\to 0} \mathbb{P}(||\hat{\Theta}_{\mathcal{A}}(\boldsymbol{\lambda}) - \mathbf{0}||_1 < \epsilon) = 1$, for any $\epsilon > 0$ and appropriate tuning parameter vector $\boldsymbol{\lambda}$. This claim requires further research and thus we leave this as an open problem to explore in the future.

The proposed penalised framework can also be extended to other types of penalties, such as the elastic net (Zou and Hastie, 2005), the grouped Lasso (Yuan and Lin, 2006), or the fused Lasso (Tibshirani et al., 2006), to accommodate for different interests in the estimation of the GLVM-LSS. For example, the elastic net or the grouped Lasso can be used for inducing cross-group equality of loadings and intercepts when assessing measurement invariance in multiple group settings or differential item functioning (DIF) (Huang, 2018; Geminiani et al., 2021; Bauer et al., 2020). Likewise, the fused Lasso can be used for collapsing redundant categories in factor models with categorical data Battauz (2020). Furthermore, the ridge penalty (a L_2-norm based penalty) can be used to avoid over-fitting in LVMs with non-linear effects of the latent variables on the

distributional parameters of interest (Falk and Cai, 2016a,b; Rizopoulos and Moustaki, 2008).

Chapter 4

Conclusions and Future Research

In this concluding chapter, we summarise the main contributions of this dissertation and discuss potential avenues for future extensions of the GLVM-LSS framework introduced in Chapter 2 and the penalised marginal maximum likelihood estimation framework with automatic tuning parameter selection presented in Chapter 3.

In Chapter 2, we present a class of Generalised Latent Variable Models for Location, Scale, and Shape parameters (GLVM-LSS). This framework expands upon traditional LVMs by modelling the distributional parameters characterising the observed variables' conditional distributions as linear functions of the latent variables. This also allows for exploring the relationships between the latent variables and the items' higher order moments, which are often expressed in terms of the distributional parameters. By modelling the whole conditional distribution in terms of the latent variables, rather than just the conditional mean, the GLVM-LSS offers a more comprehensive understanding of the data. This approach proves valuable in real-world applications where the observed variables do not necessarily satisfy the exponential family distributional assumption, or where there is substantive interest in studying the relationship between the manifest variables and the latent variables beyond the mean. The model parameters are estimated via a full information maximum likelihood. We use a computationally efficient two-stage optimisation procedure that combines the EM-algorithm with direct maximisation of the marginal log-likelihood using (quasi-)Newton methods. Simulation studies and empirical analyses of real-world data are also presented to illustrate the effectiveness of our proposed method. The GLVM-LSS opens up exciting possibilities for applied researchers to gain deeper insights into the relationships between manifest variables and latent variables of interest.

The GLVM-LSS framework is applicable to both confirmatory and exploratory settings. In

confirmatory settings, researchers impose restrictions on the measurement and structural models based on substantive theory to guide their hypotheses. The main focus is on the estimation and inference of the model parameters. On the other hand, in exploratory settings, the model parameters are estimated without imposing any restrictions, except for those necessary for model identification.

In many applications, obtaining sparse factor loading matrices is of interest, where most items are expected to load heavily on a few latent variables, while cross-loadings are close to zero. Traditional rotation methods often yield overly dense solutions, requiring subjective thresholding procedures to determine which parameters are 'substantively' different from zero. This can impact the effective degrees of freedom of the model, which are important for evaluating model fit. An alternative approach is to employ penalised maximum likelihood (PML) estimation. By introducing a sparsity-inducing penalty term in the objective function, the resulting estimation procedure yields sparse solutions for the model parameters.

Penalised maximum likelihood estimation has gained popularity in the LVM literature. However, determining the optimal value of the tuning parameters that control the level of sparsity in the PML estimates can be computationally expensive and time-consuming. The choice of these tuning parameters is crucial in obtaining an appropriate balance between model complexity and parsimony. In Chapter 3, we introduce a penalised marginal maximum likelihood estimation approach with automatic selection of tuning parameters for the GLVM-LSS framework. This methodology builds upon the previous work of Geminiani et al. (2021), adapting it to LVMs that lack closed-form solutions. The selection of the tuning parameter is accomplished by minimising an approximate unbiased risk estimate, which is proportional to the AIC. To solve this minimisation problem, we employ smooth approximations of the L_1-penalty functions. Our estimation strategy combines the benefits of the EM-algorithm and (quasi-)Newton algorithms, resulting in a two-step penalised marginal maximum likelihood estimation procedure. In the GLVM-LSS framework, the level of sparsity in the factor loadings for the location, scale, and shape parameters is determined independently by a vector of tuning parameters. The automatic procedure allows for a flexible and efficient selection of the optimal tuning parameter vector. We validate the properties of our proposed estimation method through extensive simulation studies and empirical applications in the field of educational testing.

Moving forward, we envision several potential extensions of the research in this dissertation. In Section 4.1, we outline a framework for the GLVM-LSS model with non-linear measurement equations. This extension would broaden the applicability of the GLVM-LSS framework to capture more complex data patterns. In Section 4.2, we discuss the potential application of the penalised

estimation with automatic selection of tuning parameters to LVMs with ordinal items.

4.1. Extension 1: Generalised Additive Latent Variable Model for Location, Scale, and Shape parameters

In many real-world applications, assuming a linear structure in the measurement equations may not adequately capture the complex relationships between the distributional parameters of the observed variables and the latent variables. For example, in educational testing, the discrimination power of items can vary for different levels of latent ability: easier (harder) items might have poor discrimination among high-ability (low-ability) subjects. This suggests the presence of nonlinear relationships between the items and the respondents' ability (see, e.g., McDonald, 1965, 1967; Molenaar et al., 2010).

In addition to the location parameter measurement equation, which is typically associated with the conditional mean of the items, non-linearities can also be present in the measurement equations for the scale and shape parameters. Higher-order moments of the observed variables can be influenced by nonlinear effects originating from the latent variables. This is an open research question in the LVM literature.

Using the same notation as in previous chapters, let $\mathbf{y} = (y_1, ..., y_p)^\intercal \in \mathbb{R}^p$ be a vector of observed variables and $\mathbf{z} = (z_1, ..., z_q)^\intercal \in \mathbb{R}^q$ a vector of latent variables, with $q \ll p$. Let $g_i(\mathbf{z}; \boldsymbol{\alpha}_{i,\varphi})$ denote a general function of the latent variables parameterised by $\boldsymbol{\alpha}_{i,\varphi}$. This notation allows us to express the measurement equations in a more general form. Equation 2.3 can be expressed as

$$\upsilon_{i,\varphi}(\varphi_i) = g_{i,\varphi}(\mathbf{z}; \boldsymbol{\alpha}_{i,\varphi}) \tag{4.1}$$

where $\upsilon_{i,\varphi}$ is a parameter-specific link function (e.g., identity, log, logit, etc.), and $g_{i,\varphi} : \mathbb{R}^q \to \mathbb{R}$ is a (non-linear) multivariate smooth function of the latent variables for a general location, scale, or shape parameter $\varphi_i \in \boldsymbol{\theta}_i$. Assume the mapping g_i belongs to a simple class of functions that can be decomposed as the sum of q independent functions:

$$g_{i,\varphi}(\mathbf{z}; \boldsymbol{\alpha}_{i,\varphi}) := \alpha_{i0,\varphi} + \sum_{j=1}^{q} h_{ij,\varphi}(z_j) \tag{4.2}$$

where, for each $j = 1, ..., q$, $h_{ij,\varphi}$ is a smooth unidimensional function of the latent variable z_j. This formulation corresponds to a Generalised Additive Model (GAM, Hastie and Tibshirani, 1990; Wood, 2017) structure on the location, scale, and shape measurement equations of the GLVM-LSS.

We refer to this model as the Generalised Additive Latent Variable Model for Location, Scale, and Shape parameters (GaLVM-LSS) framework. The functions $h_{ij,\varphi}$ can be approximated through basis splines (B-splines) functions (Eilers and Marx, 1996; de Boor, 2001). The measurement models can be expressed as:

$$\upsilon_{i,\varphi}(\varphi_i) = \alpha_{i0,\varphi} + \sum_{j=1}^{q} \boldsymbol{\alpha}_{ij,\varphi}^{\intercal} B_{j,\varphi}(z_j) \tag{4.3}$$

In the above, $B_{j,\varphi}(z_j)$ is a design vector of size d_j that results from evaluating B-splines on a set of d_j knots (fixed points) along the domain of z_j, denoted by $\tilde{z}_1, \ldots, \tilde{z}_{d_j}$; and $\boldsymbol{\alpha}_{ij,\varphi}$ is its corresponding d_j-dimensional vector of 'factor loadings' defining the shape of the nonlinear function. For simplicity, assume the same number of knots $d = d_1 = \cdots = d_q \ll n$ for all $j = 1, ..., q$. Lastly, denote $B_{j,\varphi}^{(i)}(z_j) = \boldsymbol{\alpha}_{ij,\varphi}^{\intercal} B_{j,\varphi}(z_j)$ as the linear combination of the B-splines for item i.

The measurement equations of the type in (4.3) are a general case of the non-linear factor model in Yalcin and Amemiya (2001); Rizopoulos and Moustaki (2008) and, to some extent, Sardy and Victoria-Feser (2012). While GAM-type formulations have been extensively used to model non-linear relationships between latent variables in the structural part of the LVM (see, e.g. Song and Lu, 2010; Song et al., 2013; Finch, 2015), their application in the measurement part has been more limited. Some earlier exceptions include the works of Ramsay (1991); Ramsay and Winsberg (1991), where monotonic splines were used to model the conditional probability of a correct response (location parameter) in the context of Item Response Theory (IRT).

However, to the best of our knowledge, there have been few instances where splines or similar techniques have been applied to model other distributional parameters or higher order moments of observed variables as non-linear functions of the latent variables. This highlights the novelty and potential of the GaLVM-LSS framework in extending the application of non-linear models to the measurement part of LVMs.

For more compact notation, we introduce the following matrix formulation:

$$\upsilon_{\varphi}(\boldsymbol{\varphi}) = \boldsymbol{\alpha}_{0,\varphi} + \mathrm{A}_{\varphi} \mathbf{B}_{\varphi}(\mathbf{z}), \tag{4.4}$$

where $\upsilon_{\varphi}(\cdot)$ is a vector valued link function, $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p)^{\intercal}$ is a p-dimensional vector of the same location, scale, or shape distributional parameters for all items, $\boldsymbol{\alpha}_0 = (\alpha_{10,\varphi}, \ldots, \alpha_{p0,\varphi})^{\intercal}$ is a p-dimensional vector of intercept terms in the measurement equations for $\varphi \in \boldsymbol{\theta}$. The matrix A_{φ} is a $p \times qd$ matrix of 'factor loadings' with rows given by $\boldsymbol{\alpha}_{i,\varphi} = (\boldsymbol{\alpha}_{i1,\varphi}, \ldots, \boldsymbol{\alpha}_{iq,\varphi})^{\intercal}$ for $i = 1, \ldots, p$; and $\mathbf{B}_{\varphi}(\mathbf{z})$ is a qd-dimensional vector defined as $\mathbf{B}_{\varphi}(\mathbf{z}) = (B_{1,\varphi}(z_1), \ldots, B_{q,\varphi}(z_q))^{\intercal}$.

To further simplify notation, we can compactly write the system of all measurement equations above, following equation (2.4):

$$\upsilon_{\varphi}(\boldsymbol{\theta}) = \boldsymbol{\alpha}_0 + \mathrm{A}\mathbf{B}(\mathbf{z}), \tag{4.5}$$

where $\boldsymbol{\theta}^\intercal = (\boldsymbol{\mu}^\intercal, \boldsymbol{\sigma}^\intercal, \boldsymbol{\nu}^\intercal, \boldsymbol{\tau}^\intercal)$ is a vector of distributional parameters; $\boldsymbol{\alpha}_0^\intercal = (\boldsymbol{\alpha}_{0,\mu}^\intercal, \boldsymbol{\alpha}_{0,\sigma}^\intercal, \boldsymbol{\alpha}_{0,\nu}^\intercal, \boldsymbol{\alpha}_{0,\tau}^\intercal)$ is a vector of intercepts; $\mathrm{A}^\intercal = [\mathrm{A}_\mu^\intercal, \mathrm{A}_\sigma^\intercal, \mathrm{A}_\nu^\intercal, \mathrm{A}_\tau^\intercal]$ is matrix of 'factor loadings' that determine the shape of the B-spline approximation; and $\mathbf{B}^\intercal(\mathbf{z}) = (\mathbf{B}_\mu^\intercal(\mathbf{z}), \mathbf{B}_\sigma^\intercal(\mathbf{z}), \mathbf{B}_\nu^\intercal(\mathbf{z}), \mathbf{B}_\tau^\intercal(\mathbf{z}))^\intercal$ a stacked vector of B-spline vectors. For simplicity, the latent variables are assumed to be independent and distributed standard Normal, $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \mathbb{I}_q)$.

Estimation and parameter computation

To estimate the model parameters in the GaLVM-LSS framework, $\Theta = (\boldsymbol{\alpha}_0, \mathrm{vec}(\mathrm{A}))$, we can adapt the penalised marginal maximum likelihood estimation framework presented in Chapter 3. The combination of B-splines with difference penalties on the model parameters is called the P-splines framework (Eilers and Marx, 1996; Wood, 2017). The penalty term is introduced to avoid over-fitting and to control the smoothness of the fitted function $\hat{g}_i$. The penalty is applied to the second derivative of the B-spline functions to ensure that the *rate of change* of the fitted function between knots is smooth. A penalty is imposed for each item $i = 1, \ldots, p$. The vector of penalised parameters is $\Theta_p = \mathrm{vec}(A)$. The intercept terms are not penalised.

For computational convenience, cubic B-splines are commonly used, as their second-order derivatives result in a penalty function that penalises the squared difference between factor loadings of adjacent basis expansions. Therefore, the objective function of the estimation problem for the GaLVM-LSS model is the penalised marginal log-likelihood, similar to the equation (3.5):

$$\begin{aligned}
\ell_p(\Theta; \mathbf{y}, \boldsymbol{\lambda}) &= \ell(\Theta; \mathbf{y}) - \mathcal{P}_{\boldsymbol{\lambda}}(\Theta) \\
&= \ell(\Theta; \mathbf{y}) - \sum_{\varphi \in \boldsymbol{\theta}} \lambda_\varphi \sum_{i=1}^{p} \sum_{j=1}^{q} \int_{\tilde{z}_1}^{\tilde{z}_d} B_{j,\varphi}^{(i)''}(z_j)\, \mathrm{d}z_j \\
&= \ell(\Theta; \mathbf{y}) - \sum_{\varphi \in \boldsymbol{\theta}} \lambda_\varphi \sum_{i=1}^{p} \sum_{j=1}^{q} \boldsymbol{\alpha}_{ij,\varphi}^\intercal \mathrm{P}^\intercal \mathrm{P} \boldsymbol{\alpha}_{ij,\varphi}
\end{aligned} \tag{4.6}$$

where $\ell(\Theta; \mathbf{y})$ is the log-likelihood function in (2.5) with distributional parameters modelled through

the non-linear measurement equations of the type in (4.3). The matrix P is

$$\mathrm{P} = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & \cdots & & 0 & -1 & 1 \end{bmatrix}, \qquad \text{such that} \qquad \mathrm{P}\boldsymbol{\alpha}_{ij,\varphi} = \begin{bmatrix} \alpha_{ij(2),\varphi} - \alpha_{ij(1),\varphi} \\ \alpha_{ij(3),\varphi} - \alpha_{ij(2),\varphi} \\ \vdots \\ \alpha_{ij(d),\varphi} - \alpha_{ij(d-1),\varphi} \end{bmatrix} \tag{4.7}$$

and hence

$$\boldsymbol{\alpha}_{ij,\varphi}^{\intercal}\mathrm{P}^{\intercal}\,\mathrm{P}\boldsymbol{\alpha}_{ij,\varphi} = \boldsymbol{\alpha}_{ij,\varphi}^{\intercal} \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & \cdots & & & 1 & -1 \end{bmatrix} \boldsymbol{\alpha}_{ij,\varphi} \tag{4.8}$$

The penalty term in equation (4.6) is twice-differentiable and does not involve the L_1-norm. Therefore, there is no need to approximate the penalty term using the techniques discussed in Chapter 3. This implies that the two-step penalised marginal maximum likelihood estimation strategy, which combines the advantages of the EM-algorithm and (quasi-)Newton algorithms, can be directly implemented. However, careful attention should be given to the computational implementation of the estimation procedure, as the use of B-splines can introduce additional computational complexity compared to linear models. Efficient algorithms and numerical techniques specific to B-splines estimation should be used to ensure computational feasibility.

Furthermore, due to the separability between the penalty terms and the tuning parameters, it is possible to adapt the automatic selection procedure of the tuning parameters vector. This aspect is discussed in Wood (2017) for GAMs with observed covariates. The automatic selection of tuning parameters can help determine the appropriate level of penalisation and achieve a balance between model complexity and goodness of fit.

It is important to note that while the proposed framework is theoretically feasible, further considerations regarding model identifiability should be explored on a case-by-case basis. Non-linear LVMs with a large number of parameters, such as those involving B-splines, may require careful model specification and regularisation techniques to ensure identifiability and prevent over-fitting.

4.2. Extension 2: Single- and Multiple-Index IRT models

Item response theory (IRT) models are used in educational and psychological testing to measure latent constructs of interest and to explore psychometric properties of test items.

Consider a test with dichotomous items $\mathbf{y} = (y_1, \ldots, y_p)^\intercal \in \{0,1\}^p$, measuring latent variables $\mathbf{z} = (z_1, \ldots, z_q)^\intercal \in \mathbb{R}^q$, with $q \ll p$. Latent variables are often assumed to follow a multivariate Normal distribution, $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Phi})$. For simplicity, very often it is assumed that $\boldsymbol{\Phi} = \mathbb{I}_q$, so that latent variables are uncorrelated and have unit variance. However, in many applications in educational testing, tests are assumed to measure a single, unidimensional construct (i.e., $q = 1$), and thus $\mathbf{z} \equiv z \sim \mathbb{N}(0,1)$.

'Simple' unidimensional IRT models[1], such as the Rasch model (Rasch, 1960) (also know as the one-parameter logistic (1PL) model) or the two-parameter logistic model (2PL, Lord and Novick, 1968; Reckase, 2009), can be expressed in the GLVM-LSS framework by assuming $y_i \,|\, \mathbf{z} \sim$ Bernoulli$(\pi_i(z))$, along with logit link functions in the measurement equations for the location parameter, i.e., $\upsilon_{i,\pi}(\pi(z)) = \text{logit}(\pi(z))$, for all $i = 1, \ldots, p$. The normal-ogive IRT model (Lord and Novick, 1968) is equivalent to using probit link functions, i.e., $\upsilon_{i,\pi}(\pi(z)) = \Phi^{-1}(\pi(z))$, where $\Phi^{-1}(\cdot)$ is the inverse of the CDF for the standard Normal distribution (see Table A1 in Appendix A2).

In the context of analysing test data, psychometricians have special interest in the item response function (IRF), also called item characteristic curve (ICC) in the educational testing literature. This function gives the probability that an individual with a latent ability level z' answers correctly to item i: $\mathbb{P}(y_i = 1 \,|\, z') = \pi_i(z')$. Let $P_i(z)$ denote the IRF for item i. For the 'simple' IRT models mentioned above, the IRFs for items $i = 1, \ldots, p$, are given by:

$$P_i(z) = \upsilon_{i,\pi}^{-1}(\alpha_{i0,\pi} + \alpha_{i1,\pi} \cdot z)\,, \tag{4.9}$$

where $\upsilon_{i,\pi}^{-1}(\cdot)$ is the inverse link function (i.e., the inverse of the logit or probit functions), and the vector $\boldsymbol{\alpha}_{i,\pi} = (\alpha_{i0,\pi}, \alpha_{i1,\pi})^\intercal$ contains the difficulty and discrimination parameters, respectively. In this context, $P_i(z)$ is said to be parametric, as it is fully characterised by the functional form of the link function, $\upsilon_{i,\pi}$, and the model parameters in $\boldsymbol{\alpha}_{i,\pi}$.

However, IRFs cannot always be modelled well with 'simple' parametric IRT models, and the

[1] Other IRT models, such as the three parameter (3PL, Birnbaum, 1968) and the four parameter logistic IRT models (4PL, Barton and Lord, 1981), are not covered by the GLVM-LSS framework without introducing new parameters and assuming different functional forms for the link functions.

resulting parametric IRFs are sometimes too restrictive to capture the shape of the true underlying IRFs. Some works on asymmetric IRFs have been proposed in Samejima (1997, 2000); Bazán et al. (2006, 2014); Bolfarine and Bazán (2010), but are still parametric in the sense that depend on the functional forms of the link functions and the model parameters.

A vast body of literature on non-parametric (N-) and semi-parametric (Sp-) IRT models has been developed over the past few decades as an alternative to parametric IRT models. NIRT and SpIRT models are more flexible and are particularly useful when: i) there is limited knowledge about the functional form of the items' IRFs; ii) the researcher wants to explore whether a given parametric IRT model fits the data correctly; and iii) some of the fundamental assumptions in IRT models (i.e., unidimensionality, local independence, monotonicity, etc.) need to be tested for individual or groups of items (see, e.g., Douglas and Cohen, 2001; Junker and Sijtsma, 2001; Stout, 2001; Lee, 2007b).

Despite sharing common methodological and theoretical background, NIRT and SpIRT models are different in a number of ways. On one hand, NIRT models estimate IRFs using either descriptive statistics (e.g., Mokken (1971); Mokken and Lewis (1982); see Sijtsma (2005), Sijtsma and van der Ark (2020, Chapter 3) for an overview) or kernel smoothing techniques (e.g., Ramsay, 1991; Douglas, 1997; Douglas and Cohen, 2001). In the latter, for each item $i = 1, ..., p$, $P_i(z)$ is estimated via local averaging of a surrogate ability value, $\tilde{z}$, which usually corresponds to the total/sum test score. It is common practice that IRFs in unidimensional NIRT models are restricted to be monotone non-decreasing functions (i.e., $P_i(z_a) < P_i(z_b)$ whenever $z_a < z_b$ and for all $i = 1, ..., p$), but it is not always the case (see, e.g., Lee, 2007b).

On the other hand, SpIRT models use splines to approximate $P_i(z)$. Some works include Winsberg et al. (1984); Ramsay (1988); Ramsay and Abrahamowicz (1989), where IRFs are modelled using splines with (non-decreasing) monotonicity constraints (I-splines,Ramsay, 1988). Similar to the kernel smoothing technique in the NIRT framework, the n-dimensional vector of values for the latent variable is not observed, and thus the smooth regressions of y_i are fitted on a (monotonic) transformation of the total/sum score. Alternatively, Ramsay and Winsberg (1991) proposed a MML estimation where the latent variable is treated as a nuisance parameter and is integrated out in the estimation process. SpIRT models are also related to non-linear IRT models that use high order polynomials of the latent variable to estimate the IRFs (e.g., Liang and Browne, 2015; Falk and Cai, 2016a,b). However, in these cases, the link function is known.

Despite their flexibility, NIRT and SpIRT models face some limitations. The most important is, perhaps, that are limited to the unidimensionality assumption of the latent space (i.e., $q = 1$).

Empirical applications and substantive theory often suggest that unidimensionality is unrealistic in psychological, social, behavioural, and health sciences. Additionally, NIRT and SpIRT models often require large sample sizes and test lengths to produce accurate estimates of the IRFs.

In this section, we propose a more general model for SpIRT models with multiple latent variables. We draw inspiration from the projection pursuit regression (PPR) model (Friedman and Stuetzle, 1981), a popular technique used for non-linear dimensionality reduction. In the PPR framework, we are interested in modelling a (univariate) regression of the type:

$$\mathbb{E}(y \,|\, \mathbf{x}) := \mu_y(\mathbf{x}) = \beta_0 + \sum_{k=1}^{K} g_k(\boldsymbol{\beta}_k^\intercal \mathbf{x}) \tag{4.10}$$

where y is a continuous outcome variable, $\mathbf{x}$ is a (large) p-dimensional vector of *observed* covariates, $K \leq p$ is an unknown integer, $\boldsymbol{\beta}_k$ are regression parameters satisfying $||\boldsymbol{\beta}_k||_1 = 1$ for all k, and $g_k : \mathbb{R} \to \mathbb{R}$ are arbitrary (unknown) non-linear functions satisfying $\mathbb{E}(g_k(\boldsymbol{\beta}_k^\intercal \mathbf{x})) = 0$ for all k. Some additional technical constraints are needed for identifiability (Friedman and Stuetzle, 1981; Chen, 1991). Note that, when $\boldsymbol{\beta}_k^\intercal \mathbf{x} = x_k$ and $K = p$, the PPR regression in (4.10) reduces to an additive model (Hastie and Tibshirani, 1990; Wood, 2017). A generalised PPR framework for outcome variables with restricted domains (binary, count, positive, etc.) was proposed in Lingjærde and Liestøl (1998). The PPR framework is often labelled as an *universal approximator*, because, if K is sufficiently large, for an appropriate choice of g_k the PPR model can approximate any continuous function in $\mathbb{R}^p$ at an arbitrary level of accuracy (Hastie et al., 2009). This generality comes with challenges in model interpretation. A special case of the PPR is when $K = 1$, which is known as the single-index model (SIM) in the econometrics literature (Ichimura, 1993; Härdle et al., 1993). The SIM framework still offers good prediction power, similar to the PPR model, but with a lower toll on interpretability.

In the LVM context, our proposal includes modelling the IRFs following a SIM framework. We call this model a single-index IRT model (SI-IRT). Formally, for each item $i = 1, ..., p$, let the probability of answering correctly (conditional on the latent variables) be given by:

$$\mathbb{P}(y_i = 1 \,|\, \mathbf{z}) := P_i(\mathbf{z}) = g_{i,\pi}(\boldsymbol{\alpha}_{i,\pi}^\intercal \mathbf{z}) \,, \tag{4.11}$$

where $g_{i,\pi} : \mathbb{R} \to [0, 1]$ is an *unknown* monotone (non-decreasing) function, and $\boldsymbol{\alpha}_{i,\pi} \in \mathbb{R}^q$ is a vector of parameters that can be interpreted as factor loadings, with the identification constraint $||\boldsymbol{\alpha}_{i,\pi}||_1 = 1$. Indeed, the individual entries in $\boldsymbol{\alpha}_{i,\pi}$ give the direction (i.e., the relative importance) of the corresponding latent variable in the projection $\boldsymbol{\alpha}_{i,\pi}^\intercal \mathbf{z} \in \mathbb{R}$. If an intercept term $\alpha_{i0,\pi}$ is included in (4.11), it can be interpreted as a guessing parameter, but further exploration of

the identifiability of this IRT model is required. The SI-IRT framework overcomes some of the limitations of the SpIRT models when $q > 1$. Moreover, if $1 < K \ll q$, then the IRFs of the type in (4.11) can be expressed as:

$$\mathbb{P}(y_i = 1 \,|\, \mathbf{z}) := P_i(\mathbf{z}) = \sum_{k=1}^{K} \beta_{ik} \cdot g_{ik,\pi}(\boldsymbol{\alpha}_{ik,\pi}^{\intercal}\mathbf{z}) \,, \tag{4.12}$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iK})^{\intercal}$ is the vector of parameters that serve as 'weights' for the corresponding $g_{ik,\pi}$'s. In this case, we impose the additional restriction $\boldsymbol{\beta}_i^{\intercal}\mathbf{1}_K = 1$, where $\mathbf{1}_K$ is a K-dimensional vector of ones, for $i = 1, \ldots, p$. This model, which we call a multiple-index IRT model (MI-IRT), would be useful when complex interactions between the latent variables are of substantive interest (see, e.g., Kenny and Judd (1984); Jöreskog and Yang (1996); Klein and Moosbrugger (2000); Marsh et al. (2004); Kelava et al. (2011) for LVMs with continuous items that include interactions between latent variables).

To account for the monotonicity of the IRFs, the functions $g_{i,\pi} : \mathbb{R} \to [0, 1]$ can be approximated using I-splines (Ramsay, 1988). Similar to the measurement equations in (4.3) involving B-splines, the $g_{ik,\pi}$'s in (4.12) (and $g_{i,\pi}$ in 4.11) can be written as:

$$g_{ik,\pi}(\boldsymbol{\alpha}_{ik,\pi}^{\intercal}\mathbf{z}) = \breve{\boldsymbol{\alpha}}_{ik,\pi}^{\intercal} I_{ik,\pi}(\boldsymbol{\alpha}_{ik,\pi}^{\intercal}\mathbf{z}) \,, \qquad k = 1, \ldots, K \,, \; i = 1, \ldots, p \tag{4.13}$$

where $I_{ik,\pi}(\boldsymbol{\alpha}_{ik,\pi}^{\intercal}\mathbf{z})$ is a d-dimensional design vector resulting from evaluating the I-splines basis functions on a set of d knots (fixed points) along the domain of $\boldsymbol{\alpha}_{ik,\pi}^{\intercal}\mathbf{z} \in \mathbb{R}$; and $\breve{\boldsymbol{\alpha}}_{ik,\pi}$ is the corresponding vector of coefficients defining the shape of the approximation. The entries of $\breve{\boldsymbol{\alpha}}_{ik,\pi}$ satisfy the restrictions $\breve{\boldsymbol{\alpha}}_{ik,\pi}^{\intercal}\mathbf{1}_d \leq 1$ and $\min(\breve{\boldsymbol{\alpha}}_{ik,\pi}) \geq 0$.

Estimation and parameter computation

We propose a penalised marginal maximum likelihood estimation of the parameters in the SI-IRT and MI-IRT models. The vector of model parameters include the 'weight' parameters $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iK})^{\intercal}$, the factor loadings $\boldsymbol{\alpha}_i^{\intercal} = (\boldsymbol{\alpha}_{i1,\pi}^{\intercal}, \ldots, \boldsymbol{\alpha}_{iK,\pi}^{\intercal})$, and the coefficients associated with the monotone spline approximations to the functions $g_{i1,\pi}, \ldots, g_{iK,\pi}$, $\breve{\boldsymbol{\alpha}}_i^{\intercal} = (\breve{\boldsymbol{\alpha}}_{i1,\pi}^{\intercal}, \ldots, \breve{\boldsymbol{\alpha}}_{iK,\pi}^{\intercal})$, for all items $i = 1, \ldots, p$. Define $\omega_i^{\intercal} = (\boldsymbol{\beta}_i^{\intercal}, \boldsymbol{\alpha}_i^{\intercal}, \breve{\boldsymbol{\alpha}}_i^{\intercal})$ as the vector of model parameters for item i. Moreover, assume the latent variables follow a multivariate standard Normal distribution, $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \mathbb{I}_q)$. Then, the model parameters in the SI/MI-IRT are $\Theta = (\omega_1, \ldots, \omega_p)^{\intercal}$. The objective function is similar to that in equation (4.6), with the smoothness penalty applied on the I-splines coefficients, i.e., $\Theta_p^{\intercal} = (\breve{\boldsymbol{\alpha}}_1^{\intercal}, \ldots, \breve{\boldsymbol{\alpha}}_p^{\intercal})$.

For a sample of n independent and identically distributed observations, the penalised marginal log-likelihood for the MI-IRT model is given by

$$\begin{aligned}\ell_p(\Theta; \mathbf{y}, \lambda_\pi) &= \ell(\Theta; \mathbf{y}, \lambda_\pi) - \mathcal{P}_{\lambda_\pi}(\Theta) \\ &= \sum_{m=1}^{n} \log \left[\int_{\mathbb{R}^q} \prod_{i=1}^{p} [P_i(\mathbf{z})^{y_i} \; Q_i(\mathbf{z})^{1-y_i}] \; p(\mathbf{z}; \Theta_z) \; \mathrm{d}\mathbf{z} \right] - \lambda_\pi \sum_{i=1}^{p} \sum_{k=1}^{K} \breve{\boldsymbol{\alpha}}_{ik,\pi}^{\intercal} \mathrm{P}^{\intercal} \, \mathrm{P} \breve{\boldsymbol{\alpha}}_{ik,\pi} \quad (4.14)\end{aligned}$$

where, for simplicity, $Q_i(\mathbf{z}) = 1 - P_i(\mathbf{z})$, the IRFs denoted by $P_i(\mathbf{z})$ are given in (4.12) (or 4.11 for the SI-IRT model) with approximated functions parameterised as in (4.13); P is the difference penalty matrix in (4.7), and thus the penalty term is similar to that in (4.8). The PMML estimate is then given by $\hat{\Theta} = \arg\max \ell_p(\Theta; \mathbf{y}, \lambda_\pi)$, subject to the additional parameter constraints described earlier for model identifiability.

The solution for $\hat{\Theta}$ is not available in a closed form. Therefore, we compute the estimates of the model parameters in the SI-IRT and MI-IRT models using a procedure that combines high-dimensional numerical integration techniques and iterative optimisation algorithms, similar to the that introduced in Chapters 2 and 3.

Compared to the estimation framework for the GLVM-LSS model, the estimation of SI-IRT and MI-IRT models can be significantly more computationally intensive. This increased complexity arises from the need to iteratively evaluate numerical integrals involving latent variables, while simultaneously exploring a relatively high-dimensional space of model parameters. To address these challenges, stochastic approximation methods have been proposed to efficiently handle high-dimensional latent variables and reduce computation time (see, e.g., Gu and Kong, 1998; Cai, 2010). However, the estimation of SI-IRT and MI-IRT models also involves considerations of the penalty term and constraints on the model parameters within the optimisation problem. In the context LVMs, Zhang and Chen (2022) proposed a stochastic proximal algorithm that handles constraints on the model parameters, including penalty terms. In this case, the automatic selection procedure for the smoothing parameter described in Chapter 3 can be used to compute the optimal value of the smoothness parameter, denoted as $\hat{\lambda}_\pi$.

Given the complexity of the proposed model, we believe that adapting the estimation framework developed by Zhang and Chen (2022) to the current context, while incorporating the automatic selection of the smoothing parameter procedure from Geminiani et al. (2021), would provide a promising avenue for further research in the field of complex IRT models.

In light of the strong connection between the PPR framework and neural networks (NN), an interesting question arises regarding whether the SI-IRT and MI-IRT models presented here can be

considered as 'vanilla' or preliminary versions of a more comprehensive NN-IRT model. This raises the possibility of exploring the power and flexibility of neural networks to enhance the modelling of item response data. Lastly, it is worth noting that the PPR structure assumed in (4.3) for the IRF can be potentially extended to the measurement models for the location, scale, and shape parameters in the GLVM-LSS.

We raise a cautionary note, though, similar to the one for the GaLVM-LSS in Section 4.1. The proposed SI-IRT and MI-IRT models are theoretically feasible, but require further exploration of (both necessary and sufficient) conditions to ensure model identifiability.

Concluding remarks

Although several methodological developments are possible under the GLVM-LSS framework, these need to be motivated by applications in real-world problems that require more complex models with latent variables. The proposed methodology offers opportunities to explore how the conditional distribution of observed variables relate to latent variables that often play key roles when formulating and testing hypothesis in substantive research of related disciplines. We believe that new methodologies should have practical applications and be developed with applied users in mind.

Likewise, the development of complex statistical models often reveals gaps in the literature that require new estimation algorithms, new sampling/imputing techniques, new approaches of dealing with latent variables, etc. We anticipate that the proposed GLVM-LSS, and the extensions discussed in this Chapter, will continue to foster research in these areas.

Chapter A

Appendix for Chapter 2

A1. Parametric Distributions and related quantities

In this Appendix we present analytical expressions for the log-likelihood of the distributions implemented in this work, along with the first- and second-order derivatives.

A1.1 Continuous distributions

Normal Distribution: We denote a random variable following a Normal (or Gaussian) distribution as $Y \sim \mathbb{N}(\mu, \sigma^2)$. The Normal distribution is parameterised by $\boldsymbol{\theta} = (\mu, \sigma)^\intercal$, where $\mu \in \mathbb{R}$ is the location parameter, and $\sigma \in \mathbb{R}^+$ is the scale parameter. If y is a value sampled from Y, its contribution to the log-likelihood function is given by:

$$\log f(\boldsymbol{\theta}; y) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$$

The mean and variance of a Normally distributed random variable can be expressed directly in terms of the location and scale parameters as $\mathbb{E}(Y) = \mu$ and $\text{Var}(Y) = \sigma^2$, respectively. The default link function for μ is the identity link and for σ the log link. The first-order derivatives are:

$$\frac{\partial \log f(\boldsymbol{\theta}; y)}{\partial \mu} = \frac{y-\mu}{\sigma^2}$$
$$\frac{\partial \log f(\boldsymbol{\theta}; y)}{\partial \sigma} = \frac{(y-\mu)^2}{\sigma^3} - \frac{1}{\sigma}$$

The second-order derivatives in the observed information matrix are:

$$\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial\mu)^2} = -\frac{1}{\sigma^2}$$
$$\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{\partial\mu\ \partial\sigma} = -\frac{2(y-\mu)}{\sigma^3}$$
$$\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial\sigma)^2} = -\frac{3(y-\mu)^2}{\sigma^4} + \frac{1}{\sigma^2}$$

By taking the expectation of the expressions above with respect to Y, the second-order derivatives in the expected information matrix are:

$$\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial\mu)^2}\right] = -\frac{1}{\sigma^2}$$
$$\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{\partial\mu\ \partial\sigma}\right] = 0$$
$$\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial\sigma)^2}\right] = -\frac{2}{\sigma^2}$$

Implementation: An object of class `dist_glvmlss`, given by an element in the list `family` defined for item i as `family[[i]] = Normal(mu.link = "identity", sg.link = "log")`. The Normal distribution is the default distribution of the `family` list argument of the `glvmlss()` function.

Skew-Normal Distribution: Consider the Skew-Normal distribution introduced in Azzalini (1985, 2005). Here we follow the material and notation in Azzalini (2013, Chapters 1-3). In its original parameterisation (also called direct parameterisation, DP), a random variable following a Skew-Normal distribution (SN) is denoted as $Y \sim \text{SN}_{\text{DP}}(\xi, \omega^2, \alpha)$. The SN distribution is parameterised by $\boldsymbol{\theta}_{\text{DP}} = (\xi, \omega, \alpha)^\intercal$, where $\xi \in \mathbb{R}$ is a location parameter, $\omega \in \mathbb{R}^+$ is a scale parameter, and $\alpha \in \mathbb{R}$ is a shape (or slant) parameter. If y is a valued sampled from Y, its contribution to the log-likelihood function is given by

$$\log f(\boldsymbol{\theta}_{\text{DP}}; y) = -\frac{1}{2}\log(2\pi) - \log(\omega) - \frac{(y-\xi)^2}{2\omega^2} + \zeta_0\left(\alpha\frac{y-\xi}{\omega}\right)$$

where $\zeta_0(\cdot)$ is defined as $\zeta_0(x) = \log\{2\Phi(x)\}$, and $\Phi(\cdot)$ is the standard Normal distribution function. The mean, variance, and skewness of a SN random variable are given by

$$\begin{aligned} \mathbb{E}(Y) := \mu &= \xi + b\omega\delta \\ \text{Var}(Y) := \sigma^2 &= \omega^2(1 - b^2\delta^2) \end{aligned} \tag{A1.1}$$

$$\text{Skewness}(Y) := \gamma_1 = \frac{4-\pi}{2}\frac{b^3\alpha^3}{(1+(1-b^2)\alpha^2)^{3/2}}$$

where $b = \sqrt{2/\pi}$ and $\delta = \alpha(1+\alpha^2)^{-1/2}$. From the above, we see that parameters in $\boldsymbol{\theta}_{\text{DP}}$ do not have a direct interpretation in terms of the random variable's moments. Let $z = (y-\xi)/\omega$ be a realisation of a SN random variable $Z \sim \text{SN}(0,1,\alpha)$. Moreover, define $\zeta_1(x) = \text{d}\zeta_0(x)/\text{d}x = \phi(x)/\Phi(x)$, where $\phi(\cdot)$ is the standard Normal density function. The first-order derivatives of the SN log-density function are:

$$\begin{aligned}
\frac{\partial \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{\partial \xi} &= \frac{z}{\omega} - \frac{\alpha}{\omega}\zeta_1(\alpha z) \\
\frac{\partial \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{\partial \omega} &= -\frac{1}{\omega} + \frac{z^2}{\omega} - \frac{\alpha}{\omega}\zeta_1(\alpha z)\, z \\
\frac{\partial \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{\partial \alpha} &= \zeta_1(\alpha z)\, z
\end{aligned} \tag{A1.2}$$

To simplify notation, let $\mathbb{S}_{\boldsymbol{\theta}_{\text{DP}}} = \nabla_{\boldsymbol{\theta}_{\text{DP}}} \log f(\boldsymbol{\theta}_{\text{DP}}; y)$ be the vector of first-order derivatives with entries in (A1.2). Moreover, define $\zeta_2(\cdot)$ as $\zeta_2(x) = \text{d}^2\zeta_0(x)/(\text{d}x)^2 = -\zeta_1(x)^2 - x\zeta_1(x)$. The second-order derivatives in the observed information matrix are:

$$\begin{aligned}
\frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{(\partial \xi)^2} &= -\frac{1}{\omega^2} + \left(\frac{\alpha}{\omega}\right)^2 \zeta_2(\alpha z) \\
\frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{\partial \xi\, \partial \omega} &= \frac{1}{\omega^2}\left(-2z + \alpha\zeta_1(\alpha z) + \alpha^2\zeta_2(\alpha z)\, z\right) \\
\frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{\partial \xi\, \partial \alpha} &= -\frac{1}{\omega}\left(\zeta_1(\alpha z) + \alpha\zeta_2(\alpha z)\, z\right) \\
\frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{(\partial \omega)^2} &= \frac{1}{\omega^2}\left(1 - 3z^2 + 2\alpha\zeta_1(\alpha z)\, z + \alpha^2\zeta_2(\alpha z)\, z^2\right) \\
\frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{\partial \omega\, \partial \alpha} &= -\frac{1}{\omega}\left(\zeta_1(\alpha z)\, z + \alpha\zeta_2(\alpha z)\, z^2\right) \\
\frac{\partial^2 \log f(\boldsymbol{\theta}_{\text{DP}}; y)}{(\partial \alpha)^2} &= z^2\, \zeta_2(\alpha z)
\end{aligned} \tag{A1.3}$$

In matrix notation, let $\mathbb{H}_{\boldsymbol{\theta}_{\text{DP}}} = \nabla_{\boldsymbol{\theta}_{\text{DP}}} \nabla_{\boldsymbol{\theta}_{\text{DP}}^\intercal} \log f(\boldsymbol{\theta}_{\text{DP}}; y)$ be the symmetric matrix of second-order derivatives with entries given by (A1.3). Taking the expectation of the expressions above involves expectations of some non-linear functions of $Z \sim \text{SN}(0,1,\alpha)$. Particularly,

$$\mathbb{E}(Z^k \zeta_1(\alpha Z)) = \begin{cases} \dfrac{b}{(1+\alpha^2)^{(k+1)/2}} \times \{1 \times 3 \times \cdots \times (k-1)\}, & \text{for } k = 0, 2, 4, \ldots \\ 0, & \text{for } k = 1, 3, 5, \ldots \end{cases}$$

and $\mathbb{E}(Z^k\zeta_1(\alpha Z)^2)$, which are evaluated numerically:

$$\hat{\mathbb{E}}(Z^k\zeta_1(\alpha Z)^2) = \frac{1}{n}\sum_{m=1}^{n}\left(z_m^k\zeta_1(\hat{\alpha}z_m)^2\right), \quad \text{for } k = 0, 1, ...$$

With the above, the second-order derivatives in the expected information matrix are:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}_{\mathrm{DP}}; y)}{(\partial\xi)^2}\right] &= -\frac{1}{\omega^2} - \left(\frac{\alpha}{\omega}\right)^2 \mathbb{E}\left(\zeta_1(\alpha Z)^2\right) \\
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}_{\mathrm{DP}}; y)}{\partial\xi\ \partial\omega}\right] &= \frac{1}{\omega^2}\left(-\frac{b\alpha(1+2\alpha^2)}{(1+\alpha^2)^{3/2}} - \alpha^2\mathbb{E}\left(Z\zeta_1(\alpha Z)^2\right)\right) \\
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}_{\mathrm{DP}}; y)}{\partial\xi\ \partial\alpha}\right] &= -\frac{1}{\omega}\left(\frac{b}{(1+\alpha^2)^{3/2}} - \alpha\mathbb{E}\left(Z\zeta_1(\alpha Z)^2\right)\right) \\
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}_{\mathrm{DP}}; y)}{(\partial\omega)^2}\right] &= \frac{1}{\omega^2}\left(-2 - \alpha^2\mathbb{E}\left(Z^2\,\zeta_1(\alpha Z)^2\right)\right) \\
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}_{\mathrm{DP}}; y)}{\partial\omega\ \partial\alpha}\right] &= \alpha\mathbb{E}\left(Z^2\zeta_1(\alpha Z)^2\right) \\
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}_{\mathrm{DP}}; y)}{(\partial\alpha)^2}\right] &= \mathbb{E}\left(Z^2\,\zeta_1(\alpha Z)^2\right)
\end{aligned} \tag{A1.4}$$

In matrix notation, let $\mathcal{I}_{\boldsymbol{\theta}_{\mathrm{DP}}} = \mathbb{E}[\nabla_{\boldsymbol{\theta}_{\mathrm{DP}}}\nabla_{\boldsymbol{\theta}_{\mathrm{DP}}^\intercal}\log f(\boldsymbol{\theta}_{\mathrm{DP}}; y)]$ be the symmetric matrix of expected second-order derivatives with entries given by (A1.4). One important limitation of the DP is that $\mathcal{I}_{\boldsymbol{\theta}_{\mathrm{DP}}}$ becomes singular for values of, and within a neighbourhood of, $\alpha = 0$. This issue has been discussed in previous works (e.g., Pewsey, 2000; Chiogna, 2005; Di Ciccio and Monti, 2011, and Chapter 3 of Azzalini, 2013). The singularity prevents the application of standard asymptotic theory of ML estimation and yields non-normal asymptotic distribution of the MLE. Moreover, it can lead to computational issues during the estimation process. While a comprehensive treatment of the theoretical properties of the ML estimates of the SN distribution is beyond the scope of this paper, interested readers can refer to the aforementioned citations.

To overcome this limitation, we adopt the centred parameterisation (CP) of the SN distribution. The CP enables us to express the SN distribution directly in terms of the parameters $\boldsymbol{\theta}_{\mathrm{CP}} = (\mu, \sigma, \gamma_1)^\intercal$. The relationship between the CP and DP parametrisations is given by (A1.1) and the inverse mappings:

$$\begin{aligned}
\xi &= \mu - b\omega\delta \\
\omega &= \frac{\sigma}{(1-b^2\delta^2)^{1/2}} \\
\alpha &= \frac{R}{(b^2-(1-b^2)R^2)^{1/2}}, \quad \text{with } R = \sqrt[3]{\frac{2|\gamma_1|}{4-\pi}} \times \mathrm{sign}(\gamma_1)
\end{aligned} \tag{A1.5}$$

The admissible set for the parameters $\boldsymbol{\theta}_{\text{CP}}$ is $\mathbb{R} \times \mathbb{R}^+ \times (-\gamma_1^{\max}, \gamma_1^{\max})$, where $\gamma_1^{\max} = \sqrt{2}(4-\pi) \cdot (\pi-2)^{-3/2} \approx 0.9953$ represents the upper bound for the index of skewness. Since the components of $\boldsymbol{\theta}_{\text{CP}}$ are smooth functions of $\boldsymbol{\theta}_{\text{DP}}$, we can expect that the CP MLE exhibit regular asymptotic properties. Moreover, the CP offers the advantage of interpreting μ directly as a location parameter, which is not the case for ξ. Similarly, $(\sigma, \gamma_1)^\intercal$ are preferred over $(\omega, \alpha)^\intercal$, as σ and γ_1 are scale and shape parameters, respectively.

To model γ_1 as a function of the latent factors, we apply an additional monotone transformation to the index of skewness parameter. Since there is no natural link function $\upsilon_{\gamma_1} : \mathbb{R} \to (-\gamma_1^{\max}, \gamma_1^{\max})$, we instead model a scaled skewness parameter $\nu \in (0,1)$, such that:

$$\nu = \frac{\gamma_1 + \gamma_1^{\max}}{2\gamma_1^{\max}}$$

The admissible set for $\boldsymbol{\theta} = (\mu, \sigma, \nu)^\intercal$ is $\mathbb{R} \times \mathbb{R}^+ \times (0,1)$. As this is a one-to-one transformation, the invariance property of the MLE implies that $\hat{\boldsymbol{\theta}}$ will also be the MLE under the DP and CP. The default link function for μ is the identity link, for σ the log link, and for ν the logit link.

Let $\mathbb{D}$ be the Jacobian matrix with entries $\mathbb{D}_{[r,s]} = \partial \boldsymbol{\theta}_{\text{DP}[r]} / \partial \boldsymbol{\theta}_{[s]}$:

$$\mathbb{D} = \begin{bmatrix} \dfrac{\partial \xi}{\partial \mu} & \dfrac{\partial \xi}{\partial \sigma} & \dfrac{\partial \xi}{\partial \gamma_1}\dfrac{\mathrm{d}\gamma_1}{\mathrm{d}\nu} \\ 0 & \dfrac{\partial \omega}{\partial \sigma} & \dfrac{\partial \omega}{\partial \gamma_1}\dfrac{\mathrm{d}\gamma_1}{\mathrm{d}\nu} \\ 0 & 0 & \dfrac{\mathrm{d}\alpha}{\mathrm{d}\gamma_1}\dfrac{\mathrm{d}\gamma_1}{\mathrm{d}\nu} \end{bmatrix}$$

The terms in $\mathbb{D}$ are:

$$\begin{aligned}
\frac{\partial \xi}{\partial \mu} &= 1 \\
\frac{\partial \xi}{\partial \sigma} &= -b\,\delta(1 - b^2\delta^2)^{-1/2} \\
\frac{\partial \xi}{\partial \gamma_1} &= \frac{-\sigma\, b\, \delta}{3(1 - b^2\delta^2)^{1/2}\,\gamma_1} \\
\frac{\partial \omega}{\partial \sigma} &= (1 - b^2\delta^2)^{-1/2} \\
\frac{\partial \omega}{\partial \gamma_1} &= \frac{\sigma b^2 \delta}{(1 - b^2\delta^2)^{3/2} \cdot (1 + \alpha^2)^{3/2}} \frac{\mathrm{d}\alpha}{\mathrm{d}\gamma_1} \\
\frac{\mathrm{d}\alpha}{\mathrm{d}\gamma_1} &= \frac{2}{3(4-\pi)} \left(\frac{1}{TR^2} + \frac{1 - b^2}{T^3} \right) \\
\frac{\mathrm{d}\gamma_1}{\mathrm{d}\nu} &= 2\gamma_1^{\max}
\end{aligned}$$

where $T = \left[b^2 - (1-b^2)R^2\right]^{1/2}$ and R is defined in (A1.5).

The first-order derivatives of the SN distribution parameterised by $\boldsymbol{\theta} = (\mu, \sigma, \nu)^\intercal$ are given by $\mathbb{S}_{\boldsymbol{\theta}} = \mathbb{D}^\intercal \mathbb{S}_{\boldsymbol{\theta}_{\text{DP}}}$, the matrix of second-order derivatives is given by $\mathbb{H}_{\boldsymbol{\theta}} = \mathbb{D}^\intercal \mathbb{H}_{\boldsymbol{\theta}_{\text{DP}}} \mathbb{D}$, and the matrix of expected second-order derivatives by $\mathcal{I}_{\boldsymbol{\theta}} = \mathbb{D}^\intercal \mathcal{I}_{\boldsymbol{\theta}_{\text{DP}}} \mathbb{D}$. The individual entries in $\mathbb{S}_{\boldsymbol{\theta}}$, $\mathbb{H}_{\boldsymbol{\theta}}$, $\mathcal{I}_{\boldsymbol{\theta}}$ are used in the score vectors and observed- and expected-information matrices in the estimation procedure.

A note on inference: If $\gamma_1 \to 0$, then $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{CP}} - \boldsymbol{\theta}_{\text{CP}}) \xrightarrow{d} \mathbb{N}(0, \text{diag}(\sigma^2, \frac{\sigma^2}{2}, 6))$ (Chiogna, 2005), where the third term is the asymptotic variance of the sample coefficient of skewness (see, e.g., DasGupta, 2008, Theorem 3.8). The scaled skewness parameter ν results from the continuous and differentiable mapping $h : (-\gamma_1^{\max}, \gamma_1^{\max}) \to (0, 1)$, defined above. Application of the Delta theorem yields:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathbb{N}\left(0, \text{diag}(\sigma^2, \frac{\sigma^2}{2}, [h'(\gamma_1)]^2 6)\right) \approx \mathbb{N}\left(0, \text{diag}(\sigma^2, \frac{\sigma^2}{2}, 23.77)\right)$$

Implementation: An object of class `dist_glvmlss`, given by an element in the list `family` defined for item i as `family[[i]] = SkewNormal(mu.link = "identity", sg.link = "log", nu.link = "logit")`.

Beta Distribution: We denote a continuous random variable in the $(0, 1)$ interval following a Beta distribution as $Y \sim \text{Beta}(\alpha, \beta)$. In its original form, the Beta distribution is parameterised by $\boldsymbol{\theta}_o = (\alpha, \beta)^\intercal$, with shape parameters $(\alpha, \beta) > 0$. If y is a value sampled from Y, its contribution to the log-likelihood function is given by:

$$\log f(\boldsymbol{\theta}_o; y) = \log \Gamma(\alpha + \beta) - \log \Gamma(\alpha) - \log \Gamma(\beta) + (\alpha - 1)\log(y) + (\beta - 1)\log(1 - y)$$

where $\Gamma(\cdot)$ is the gamma function. The first-order derivatives of the Beta log-density function are:

$$\begin{aligned}
\frac{\partial \log f(\boldsymbol{\theta}_o; y)}{\partial \alpha} &= \psi_0(\alpha + \beta) - \psi_0(\alpha) + \log(y) \\
\frac{\partial \log f(\boldsymbol{\theta}_o; y)}{\partial \beta} &= \psi_0(\alpha + \beta) - \psi_0(\beta) + \log(1 - y)
\end{aligned}$$

where $\psi_0(x) = \text{d} \log \Gamma(x)/\text{d}x$ is the digamma function. To simplify notation, let $\mathbb{S}_{\boldsymbol{\theta}_o} = \nabla_{\boldsymbol{\theta}_o} \log f(\boldsymbol{\theta}_o; y)$ be the vector of first-order derivatives with entries described above. The second-order derivatives in the observed information matrix are:

$$\begin{aligned}
\frac{\partial^2 \log f(\boldsymbol{\theta}_o; y)}{(\partial \alpha)^2} &= \psi_1(\alpha + \beta) - \psi_1(\alpha) \\
\frac{\partial^2 \log f(\boldsymbol{\theta}_o; y)}{\partial \alpha \, \partial \beta} &= \psi_1(\alpha + \beta)
\end{aligned}$$

$$\frac{\partial^2 \log f(\boldsymbol{\theta}_o; y)}{(\partial\beta)^2} = \psi_1(\alpha+\beta) - \psi_1(\beta)$$

where $\psi_1(x) = \mathrm{d}^2 \log\Gamma(x)/(\mathrm{d}x)^2$ is the trigamma function. Let $\mathbb{H}_{\boldsymbol{\theta}_o} = \nabla_{\boldsymbol{\theta}_o}\nabla_{\boldsymbol{\theta}_o^\intercal} \log f(\boldsymbol{\theta}_o; y)$ be the symmetric matrix of second-order derivatives with entries given by the expressions above. Note how the entries for the expected information matrix are the same, i.e., $\mathcal{I}_{\boldsymbol{\theta}_o} = \mathbb{H}_{\boldsymbol{\theta}_o}$. The mean and variance of a Beta distributed random variable are

$$\mathbb{E}(Y) = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \text{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Note how the parameters $\boldsymbol{\theta}_o$ cannot be interpreted directly as location or scale parameters. The location-scale parameterisation in Rigby et al. (2020) suggests using the parameter vector $\boldsymbol{\theta} = (\mu, \sigma)^\intercal$, where $\mu \in (0,1)$ is a location parameter and $\sigma \in (0,1)$ is a scale parameter. The relationship between $\boldsymbol{\theta}_o = (\alpha, \beta)^\intercal$ and $\boldsymbol{\theta} = (\mu, \sigma)^\intercal$ is given by

$$\mu = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \sigma = (\alpha+\beta+1)^{-1/2};$$

and the inverse mappings

$$\alpha = \frac{\mu(1-\sigma^2)}{\sigma^2} \qquad \text{and} \qquad \beta = \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}.$$

Under this parametrisation, $\mathbb{E}(Y) = \mu$ and $\text{Var}(Y) = \sigma^2\mu(1-\mu)$. The default link function for μ and σ is the logit link.

Let $\mathbb{D}$ be the Jacobian matrix with entries $\mathbb{D}_{[r,s]} = \partial\boldsymbol{\theta}_{o[r]}/\partial\boldsymbol{\theta}_{[s]}$:

$$\mathbb{D} = \begin{bmatrix} \dfrac{\partial\alpha}{\partial\mu} & \dfrac{\partial\alpha}{\partial\sigma} \\ \dfrac{\partial\beta}{\partial\mu} & \dfrac{\partial\beta}{\partial\sigma} \end{bmatrix}$$

The terms in $\mathbb{D}$ are:

$$\begin{aligned} \frac{\partial\alpha}{\partial\mu} &= \frac{1}{\sigma^2} - 1 \\ \frac{\partial\alpha}{\partial\sigma} &= \frac{-2\mu}{\sigma^3} \\ \frac{\partial\beta}{\partial\mu} &= 1 - \frac{1}{\sigma^2} \\ \frac{\partial\beta}{\partial\sigma} &= \frac{2\mu-2}{\sigma^3} \end{aligned}$$

The first-order derivatives of the Beta distribution parameterised by $\boldsymbol{\theta} = (\mu, \sigma)^\intercal$ are given by $\mathbb{S}_{\boldsymbol{\theta}} = \mathbb{D}^\intercal \mathbb{S}_{\boldsymbol{\theta}_o}$, and the matrix of second-order derivatives is given by $\mathbb{H}_{\boldsymbol{\theta}} = \mathbb{D}^\intercal \mathbb{H}_{\boldsymbol{\theta}_o} \mathbb{D}$. The matrix of expected second-order derivatives is $\mathcal{I}_{\boldsymbol{\theta}} = \mathbb{H}_{\boldsymbol{\theta}}$. The individual entries in $\mathbb{S}_{\boldsymbol{\theta}}$, $\mathbb{H}_{\boldsymbol{\theta}}$, $\mathcal{I}_{\boldsymbol{\theta}}$ are used in the score vectors and observed- and expected-information matrices in the estimation procedure.

Implementation: An object of class `dist_glvmlss`, given by an element in the list `family` defined for item i as `family[[i]] = Beta(mu.link = "logit", sg.link = "logit")`.

Gamma Distribution: We denote a continuous random variable following a Gamma distribution as $Y \sim \text{Gamma}(\alpha, \beta)$. The Gamma distribution is parameterised by $\boldsymbol{\theta} = (\alpha, \beta)^\intercal$, where $\alpha \in \mathbb{R}^+$ is a shape parameter and $\beta \in \mathbb{R}^+$ is a scale parameter. If y is drawn from a Gamma distribution, its contribution to the log-likelihood function is

$$\log f(\boldsymbol{\theta}; y) = (\alpha - 1) \cdot \log(y) - \frac{y}{\beta} - \log \Gamma(\alpha) - \alpha \log(\beta)$$

where, $\Gamma(\cdot)$ is the Gamma function. For a Gamma distributed random variable, we have $\mathbb{E}(Y) = \alpha\beta$ and $\text{Var}(Y) = \alpha\beta^2$. The first-order derivatives of the log-density function with respect to $\boldsymbol{\theta}$ are

$$\begin{aligned}
\frac{\partial \log f(\boldsymbol{\theta}; y)}{\partial \alpha} &= \log(y) - \psi_0(\alpha) - \log(\beta) \\
\frac{\partial \log f(\boldsymbol{\theta}; y)}{\partial \beta} &= \frac{y}{\beta^2} - \frac{\alpha}{\beta}
\end{aligned}$$

The second-order derivatives of the log-density function with respect to $\boldsymbol{\theta}$ are

$$\begin{aligned}
\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial \alpha)^2} &= -\psi_1(\alpha) \\
\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{\partial \alpha \partial \beta} &= -\frac{1}{\beta} \\
\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial \beta)^2} &= \frac{\alpha\beta - 2y}{\beta^3}
\end{aligned}$$

and taking the expectation with respect to Y yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial \alpha)^2}\right] &= -\psi_1(\alpha) \\
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{\partial \alpha \partial \beta}\right] &= -\frac{1}{\beta} \\
\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial \beta)^2}\right] &= -\frac{\alpha}{\beta^2}
\end{aligned}$$

where $\psi_0(x) = \mathrm{d}\log\Gamma(x)/\mathrm{d}x$ is the digamma function and $\psi_1(x) = \mathrm{d}^2\log\Gamma(x)/(\mathrm{d}x)^2$ is the trigamma function.

Implementation: An object of class `dist_glvmlss`, given by an element in the list `family` defined for item i as `family[[i]] = Gamma(mu.link = "log", sg.link = "log")`.

A1.2 Discrete distributions

Bernoulli Distribution: We denote a binary random variable following a Bernoulli distribution as $Y \sim \text{Bernoulli}(\pi)$. The Bernoulli distribution is parameterised by $\boldsymbol{\theta} = \pi \in (0, 1)$, which represents the probability of 'success', i.e., $\mathbb{P}(Y = 1) = \pi$. If y is a realisation of Y, its contribution to the log-likelihood function is given by

$$\log f(\boldsymbol{\theta}; y) = y \cdot \log(\pi) + (1 - y) \cdot \log(1 - \pi)$$

For a Bernoulli distributed random variable, we have $\mathbb{E}(Y) = \pi$ and $\text{Var}(Y) = \pi\,(1 - \pi)$. The first-order derivative of the log-density function with respect to π is

$$\frac{\partial \log f(\boldsymbol{\theta}; y)}{\partial \pi} = \frac{y - \pi}{\pi \cdot (1 - \pi)}$$

The second-order derivative of the log-density function with respect to π is

$$\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial \pi)^2} = -\frac{\pi^2 - 2y\pi + y}{(\pi - 1)^2\pi^2}$$

and taking the expectation with respect to Y yields

$$\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial \pi)^2}\right] = -\frac{1}{\pi(1 - \pi)}$$

Implementation: An object of class `dist_glvmlss`, given by an element in the list `family` defined for item i as `family[[i]] = Binomial(n = 1, mu.link = "logit")`.

Poisson Distribution: We denote a discrete random variable following a Poisson distribution as $Y \sim \text{Poisson}(\lambda)$. The Poisson distribution is parameterised by $\boldsymbol{\theta} = \lambda \in \mathbb{R}^+$, representing the rate of occurrence. If y is a realisation of Y, its contribution to the log-likelihood function is given

by

$$\log f(\boldsymbol{\theta}; y) = y \cdot \log(\lambda) - \lambda - \log(y!)$$

For a Poisson distributed random variable, we have $\mathbb{E}(Y) = \text{Var}(Y) = \lambda$. The first-order derivative of the log-density function with respect to λ is

$$\frac{\partial \log f(\boldsymbol{\theta}; y)}{\partial \lambda} = \frac{y}{\lambda} - 1$$

The second-order derivative of the log-density function with respect to π is

$$\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial \lambda)^2} = -\frac{y}{\lambda^2}$$

and taking the expectation with respect to Y yields

$$\mathbb{E}\left[\frac{\partial^2 \log f(\boldsymbol{\theta}; y)}{(\partial \lambda)^2}\right] = -\frac{1}{\lambda}$$

Implementation: An object of class `dist_glvmlss`, given by an element in the list `family` defined for item i as `family[[i]] = Poisson(mu.link = "log")`.

Zero-Inflated Poisson Distribution: We denote a discrete random variable following a Zero-Inflated Poisson (ZIP) distribution as $Y \sim \text{ZIP}(\pi, \lambda)$. The ZIP distribution is parameterised by $\boldsymbol{\theta} = (\pi, \lambda)^\intercal$, which define the two-component mixture:

$$Y \sim \begin{cases} 0, & \text{with probability } \pi \\ \text{Poisson}(\lambda), & \text{with probability } 1 - \pi \end{cases}$$

If y is a value sampled from Y, its contribution to the likelihood function is given by:

$$f(\boldsymbol{\theta}; y) = \begin{cases} \pi + (1 - \pi) \cdot e^{-\lambda}, & \text{if } y = 0 \\ (1 - \pi) \cdot \dfrac{\lambda^y \cdot e^{-\lambda}}{y!}, & \text{if } y > 0 \end{cases}$$

which can be expressed more succinctly as

$$f(\boldsymbol{\theta}; y) = \left(\pi + (1 - \pi) \cdot e^{-\lambda}\right)^{\mathbb{1}_0(y)} \times \left((1 - \pi) \cdot \frac{\lambda^y \cdot e^{-\lambda}}{y!}\right)^{(1 - \mathbb{1}_0(y))},$$

where $\mathbb{1}_0(y)$ is an indicator function that takes the value of 1 if $y = 0$ and 0 if $y > 0$. The

contribution to the log-likelihood function is thus:

$$\log f(\boldsymbol{\theta}; y) = \mathbb{1}_0(y) \log\left(\pi + (1-\pi)\cdot e^{-\lambda}\right) + (1-\mathbb{1}_0(y)) \log\left((1-\pi)\cdot \frac{\lambda^y \cdot e^{-\lambda}}{y!}\right)$$

For ZIP random variables, $\mathbb{E}(Y) = (1-\pi)\lambda$ and $\mathrm{Var}(Y) = (1-\pi)\lambda + \pi(1-\pi)\lambda^2$. For simplicity we will refer to λ as the location parameter and π as the scale parameter.

An important limitation, however, is that we do not know whether an observation $y = 0$ was sampled from the 'perfect zero' state or from the Poisson process. Following Hall (2000); Wang (2010), to account for this uncertainty we introduce an auxiliary binary latent variable b, such that $b = 1$ if $y = 0$ is from the 'perfect zero' state, and $b = 0$ otherwise. The joint distribution of (y, b) can be factorised as $f(y, b) = f(y \mid b) f(b)$, with

$$f(y \mid b; \boldsymbol{\theta}) = \left(\frac{\lambda^y \cdot e^{-\lambda}}{y!}\right)^{(1-b)}$$

$$f(b; \boldsymbol{\theta}) = \pi^b \cdot (1-\pi)^{(1-b)}$$

and thus, the complete-data log-density is $\log f(\boldsymbol{\theta}; y, b) = \log f(y \mid b; \boldsymbol{\theta}) + \log f(b; \boldsymbol{\theta})$. In the estimation of $\boldsymbol{\theta}$, we treat b as missing data, as in the EM-algorithm. Let $\log \tilde{f}(\boldsymbol{\theta}; y) = \mathbb{E}_{b \mid y}[\log f(\boldsymbol{\theta}; y, b)]$ be the expectation of $\log f(\boldsymbol{\theta}; y, b)$ taken over the conditional distribution $f(b \mid y)$. Given the linearity of $\log \tilde{f}(\boldsymbol{\theta}; y)$ on b, we have that:

$$\begin{aligned}
\log \tilde{f}(\boldsymbol{\theta}; y) &= \mathbb{E}_{b \mid y}[\log f(y \mid b; \boldsymbol{\theta})) + \log f(b; \boldsymbol{\theta})] \\
&= (1-\hat{b}) \cdot (y \cdot \log(\lambda) - \lambda - \log(y!)) + \hat{b} \log(\pi) + (1-\hat{b}) \log(1-\pi)
\end{aligned}$$

where $\hat{b}$ is the expected value of b, conditional on y:

$$\begin{aligned}
\hat{b} = \mathbb{E}_b(b \mid y; \boldsymbol{\theta}) &= \mathbb{P}(b = 1 \mid y; \boldsymbol{\theta}) \\
&= \frac{f(y \mid b = 1, \boldsymbol{\theta}) \cdot f(b = 1; \boldsymbol{\theta})}{f(y \mid b = 1; \boldsymbol{\theta}) \cdot f(b = 1; \boldsymbol{\theta}) + f(y \mid b = 0; \boldsymbol{\theta}) \cdot f(b = 0; \boldsymbol{\theta})} \\
&= \frac{\pi}{\pi + e^{-\lambda}(1-\pi)} \\
&= \begin{cases} \left[1 + \exp(-\mathrm{logit}(\pi) - \lambda)\right]^{-1}, & \text{if } \; y = 0 \\ 0, & \text{if } \; y > 0 \end{cases}
\end{aligned}$$

The first-order derivatives of $\log \tilde{f}(\boldsymbol{\theta}; y)$ with respect to the parameters in $\boldsymbol{\theta}$ are:

$$\frac{\partial \log \tilde{f}(\boldsymbol{\theta}; y)}{\partial \lambda} = (1 - \hat{b}) \left(\frac{y}{\lambda} - 1 \right)$$
$$\frac{\partial \log \tilde{f}(\boldsymbol{\theta}; y)}{\partial \pi} = \frac{\hat{b} - \pi}{\pi \cdot (1 - \pi)}$$

The second-order derivatives in the observed information matrix are:

$$\frac{\partial^2 \log \tilde{f}(\boldsymbol{\theta}; y)}{(\partial \lambda)^2} = -(1 - \hat{b}) \frac{y}{\lambda^2}$$
$$\frac{\partial^2 \log \tilde{f}(\boldsymbol{\theta}; y)}{\partial \lambda \, \partial \pi} = 0$$
$$\frac{\partial^2 \log \tilde{f}(\boldsymbol{\theta}; y)}{(\partial \pi)^2} = -\frac{\pi^2 - 2\hat{b}\pi + \hat{b}}{(\pi - 1)^2 \pi^2}$$

By taking the expectation of the expressions above with respect to $Y \,|\, b \sim \text{Poisson}(\lambda)$, the second-order derivatives in the expected information matrix are:

$$\mathbb{E}\left[\frac{\partial^2 \log \tilde{f}(\boldsymbol{\theta}; y)}{(\partial \lambda)^2} \right] = \frac{-(1 - \hat{b})}{\lambda}$$
$$\mathbb{E}\left[\frac{\partial^2 \log \tilde{f}(\boldsymbol{\theta}; y)}{\partial \lambda \, \partial \pi} \right] = 0$$
$$\mathbb{E}\left[\frac{\partial^2 \log \tilde{f}(\boldsymbol{\theta}; y)}{(\partial \pi)^2} \right] = -\frac{\pi^2 - 2\hat{b}\pi + \hat{b}}{(\pi - 1)^2 \pi^2}$$

Implementation: An object of class `dist_glvmlss`, given by an element in the list `family` defined for item i as `family[[i]] = ZIpoisson(mu.link = "log", sg.link = "logit")`.

A2. Derivations for the score vectors, information matrices, and link functions

In the following, we refer the readers to the the corresponding sections in Chapter 2 and Chapter 3 for the corresponding notation.

Score vectors for intercepts and factor loadings

The score vectors used in t^{th} iteration of the EM-algorithm are given by:

$$\begin{aligned}
\mathbb{S}^{[t]}_{[\bar{k}_{i,\varphi}]} &= \frac{\partial \mathcal{Q}(\Theta; \Theta^{[t]})}{\partial \boldsymbol{\alpha}_{i,\varphi}} \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}_{i,\varphi}} \left[\sum_{m=1}^{n} \int_{\mathbb{R}^q} \sum_{i=1}^{p} \log f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i) \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \right] \\
&= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \frac{\partial \log f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i)}{\partial \boldsymbol{\alpha}_{i,\varphi}} \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \varphi_i} \cdot \frac{\partial \varphi_i}{\partial \eta_{i,\varphi}} \cdot \frac{\partial \eta_{i,\varphi}}{\partial \boldsymbol{\alpha}_{i,\varphi}} \right] \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z}
\end{aligned}$$

The score vectors used in the direct optimisation algorithm are given by:

$$\begin{aligned}
\mathbb{S}_{[\bar{k}_{i,\varphi}]}(\Theta) &= \frac{\partial \ell(\Theta; \mathbf{y})}{\partial \boldsymbol{\alpha}_{i,\varphi}} \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}_{i,\varphi}} \left[\sum_{m=1}^{n} \log \left(\int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i) \right] \; p(\mathbf{z}; \Phi) \; \mathrm{d}\mathbf{z} \right) \right] \\
&= \sum_{m=1}^{n} \frac{1}{f(\mathbf{y}_m)} \int_{\mathbb{R}^q} \left[\prod_{i' \neq i} f_{i'}(y_{i'm} \,|\, \mathbf{z}; \boldsymbol{\theta}_{i'}) \frac{\partial f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i)}{\partial \boldsymbol{\alpha}_{i,\varphi}} \right] \; p(\mathbf{z}; \Phi) \; \mathrm{d}\mathbf{z} \\
&= \sum_{m=1}^{n} \frac{\prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i)}{f(\mathbf{y}_m)} \int_{\mathbb{R}^q} \frac{\partial \log f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i)}{\partial \boldsymbol{\alpha}_{i,\varphi}} \; p(\mathbf{z}; \Phi) \; \mathrm{d}\mathbf{z} \\
&= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \frac{\partial \log f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i)}{\partial \boldsymbol{\alpha}_{i,\varphi}} \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta) \; \mathrm{d}\mathbf{z}
\end{aligned}$$

evaluated at the value of the parameter estimates at iteration t, i.e, $\mathbb{S}^{[t]}_{[\bar{k}_{i,\varphi}]} := \mathbb{S}_{[\bar{k}_{i,\varphi}]}(\Theta^{[t]})$. Note the equivalence between the score vector under the complete data specification in the EM-algorithm and the score vector for the marginal log-likelihood, i.e., $\nabla_\Theta \ell(\Theta; \mathbf{y}) \equiv \nabla_\Theta \mathcal{Q}(\Theta; \Theta^{[t]})$ (Louis, 1982).

Score vector for factor correlations

Recall that the factor correlation matrix is reparameterised through a Cholesky decomposition as $\boldsymbol{\Phi} = \mathrm{L}\mathrm{L}^\intercal$, where L is a lower triangular matrix, and that we denote L_j as the j^{th} row of L, with $\mathrm{L}_{j,[k]}$ being the k^{th} element of L_j.

The score vectors for the factor correlations used in t^{th} iteration of the EM-algorithm have

entries of the form:

$$
\begin{aligned}
\mathbb{S}^{[t]}_{\mathrm{L}_{j,[k]}} &= \frac{\partial \mathcal{Q}(\Theta; \Theta^{[t]})}{\partial \mathrm{L}_{j,[k]}} \\
&= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial \log p(\mathbf{z}; \mathrm{L})}{\partial \mathrm{L}_{j,[k]}} \right] \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&= -\frac{1}{2} \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial \log \det(\mathrm{LL}^\intercal)}{\partial \mathrm{L}} \frac{\partial \mathrm{L}}{\partial \mathrm{L}_{j,[k]}} + \frac{\partial \, \mathbf{z}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathbf{z}}{\partial \mathrm{L}} \frac{\partial \mathrm{L}}{\partial \mathrm{L}_{j,[k]}} \right] \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&= -\frac{1}{2} \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[2 \operatorname{tr}\left(\mathrm{L}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathrm{D}_{jk} \right) - 2 \operatorname{tr}\left(\mathrm{L}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathbf{z}\mathbf{z}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathrm{D}_{jk} \right) \right] \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&= -n \operatorname{tr}\left(\mathrm{L}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathrm{D}_{jk} \right) + \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\mathbf{z}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathrm{D}_{jk} \mathrm{L}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathbf{z} \right] \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&= -n \operatorname{tr}\left(\mathrm{L}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathrm{D}_{jk} \right) + \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\mathbf{z}^\intercal \mathrm{G}_{jk} \mathbf{z} \right] \; p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]}) \; \mathrm{d}\mathbf{z} \\
&= -n \operatorname{tr}\left(\mathrm{L}^\intercal (\mathrm{LL}^\intercal)^{-1} \mathrm{D}_{jk} \right) + \sum_{m=1}^{n} \left[\operatorname{tr}\left(\mathrm{G}_{jk} \mathbb{V}^{[t]}_m \right) + \breve{\mathbf{z}}^{[t]\intercal}_m \mathrm{G}_{jk} \breve{\mathbf{z}}^{[t]}_m \right]
\end{aligned}
$$

where $\mathrm{D}_{jk} = \partial \mathrm{L} / \partial \mathrm{L}_{j,[k]}$ is a square matrix of dimension q, with a value of 1 in the $[j,k]$ position and zero elsewhere; $\mathrm{G}_{jk} = (\mathrm{LL}^\intercal)^{-1} \mathrm{D}_{jk} \mathrm{L}^\intercal (\mathrm{LL}^\intercal)^{-1}$; and the conditional mean $\breve{\mathbf{z}}^{[t]}_m = \mathbb{E}(\mathbf{z} \,|\, \mathbf{y}_m; \Theta^{[t]})$ and conditional variance $\mathbb{V}^{[t]}_m = \mathbb{E}((\mathbf{z} - \breve{\mathbf{z}}^{[t]}_m)(\mathbf{z} - \breve{\mathbf{z}}^{[t]}_m)^\intercal \,|\, \mathbf{y}_m; \Theta^{[t]})$ are computed using the properties of the trace operator and the linearity of the conditional expectation.

As before, the score vectors for the factor correlations from the marginal log-likelihood are equivalent to those from the complete data log-likelihood. For simplicity, we derive this equivalence using the derivatives with respect to the covariance/correlation matrix $\boldsymbol{\Phi}$, but the results can be easily extended to the Cholesky parameterisation $\boldsymbol{\Phi} = \mathrm{LL}^\intercal$. Indeed:

$$
\begin{aligned}
\mathbb{S}_{\boldsymbol{\Phi}}(\Theta) &= \frac{\partial \ell(\Theta; \mathbf{y})}{\partial \boldsymbol{\Phi}} \\
&= \frac{\partial}{\partial \boldsymbol{\Phi}} \left[\sum_{m=1}^{n} \log \left(\int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i) \right] \; p(\mathbf{z}; \boldsymbol{\Phi}) \; \mathrm{d}\mathbf{z} \right) \right] \\
&= \sum_{m=1}^{n} \frac{1}{f(\mathbf{y}_m)} \int_{\mathbb{R}^q} \left[\prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i) \right] \frac{\partial p(\mathbf{z}; \boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} \; \mathrm{d}\mathbf{z} \\
&= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial p(\mathbf{z}; \boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} \right] \frac{\prod_{i=1}^{p} f_i(y_{im} \,|\, \mathbf{z}; \boldsymbol{\theta}_i)}{f(\mathbf{y}_m)} \; \mathrm{d}\mathbf{z} \\
&= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial \log p(\mathbf{z}; \boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} \right] p(\mathbf{z} \,|\, \mathbf{y}_m; \Theta) \; \mathrm{d}\mathbf{z}
\end{aligned}
$$

The above is equivalent to the second line in the derivation of the score vector for the factor correlations for the complete-data log-likelihood presented above.

Observed information matrix (marginal log-likelihood)

The marginal log-likelihood observed information matrix has matrix entries given by:

$$\mathcal{H}_{[\bar{k}_{i,\varphi},\bar{k}_{i',\tilde{\varphi}}]} = \frac{\partial^2 \ell(\Theta;\mathbf{y})}{\partial \boldsymbol{\alpha}_{i,\varphi} \partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} = \frac{\partial \mathbb{S}_{i,\varphi}(\Theta;\mathbf{y})}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}}$$

$$= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \left[\frac{\partial^2 \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi} \partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} \cdot p(\mathbf{z} \,|\, \mathbf{y}_m) + \frac{\partial \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi}} \cdot \underbrace{\frac{\partial p(\mathbf{z} \,|\, \mathbf{y}_m)}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}}}_{\mathcal{H}_0^{\intercal}} \right] \mathrm{d}\mathbf{z}; \qquad \text{(A2.1)}$$

The vector $\mathcal{H}_0$ in the integrand above is

$$\begin{aligned}
\mathcal{H}_0 &= \frac{\partial p(\mathbf{z} \,|\, \mathbf{y}_m)}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} = \frac{\partial}{\partial \boldsymbol{\alpha}_{j,\tilde{\varphi}}} \left[\frac{f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})}{f(\mathbf{y}_m)} \right] \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \left[\frac{f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})}{\int_{\mathbb{R}^q} f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \right] \\
&= \frac{\left[\frac{\partial f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \cdot \int_{\mathbb{R}^q} f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z} \right] - \left[\frac{\partial \left(\int_{\mathbb{R}^q} f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z} \right)}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \cdot f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z}) \right]}{\left[\int_{\mathbb{R}^q} f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z} \right]^2} \\
&= \frac{f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})}{\int_{\mathbb{R}^q} f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \cdot \left[\frac{\partial f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \frac{1}{f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})} \right] \\
&\quad - \frac{f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})}{\int_{\mathbb{R}^q} f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \cdot \left[\int_{\mathbb{R}^q} \frac{\partial f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \frac{1}{f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})} \, \mathrm{d}\mathbf{z} \cdot \frac{f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z})}{\int_{\mathbb{R}^q} f(\mathbf{y}_m \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \right] \\
&= p(\mathbf{z} \,|\, \mathbf{y}_m) \cdot \frac{\partial}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \log f(\mathbf{y}_m \,|\, \mathbf{z}) - \; p(\mathbf{z} \,|\, \mathbf{y}_m) \cdot \left[\int_{\mathbb{R}^q} p(\mathbf{z} \,|\, \mathbf{y}_m) \frac{\partial}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \log f(\mathbf{y}_m \,|\, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right] \\
&= p(\mathbf{z} \,|\, \mathbf{y}_m) \cdot \frac{\partial}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \log f_{i'}(y_{i'm} \,|\, \mathbf{z}) - \; p(\mathbf{z} \,|\, \mathbf{y}_m) \cdot \left[\int_{\mathbb{R}^q} p(\mathbf{z} \,|\, \mathbf{y}_m) \frac{\partial}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}} \log f_{i'}(y_{i'm} \,|\, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right]
\end{aligned}$$

With the above, equation (A2.1) becomes

$$\mathcal{H}_{[\bar{k}_{i,\varphi},\bar{k}_{i',\tilde{\varphi}}]} = \sum_{m=1}^{n} \int_{\mathbb{R}^q} p(\mathbf{z} \,|\, \mathbf{y}_m) \cdot \frac{\partial^2 \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi} \partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} \, \mathrm{d}\mathbf{z}$$

$$+\sum_{m=1}^{n}\int_{\mathbb{R}^q} p(\mathbf{z}\,|\,\mathbf{y}_m)\cdot\frac{\partial\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\boldsymbol{\alpha}_{i,\varphi}}\cdot\frac{\partial\log f_{i'}(y_{i'm}\,|\,\mathbf{z})}{\partial\boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}}\;\mathrm{d}\mathbf{z} \tag{A2.2}$$
$$-\sum_{m=1}^{n}\int_{\mathbb{R}^q} p(\mathbf{z}\,|\,\mathbf{y}_m)\cdot\frac{\partial\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\boldsymbol{\alpha}_{i,\varphi}}\;\mathrm{d}\mathbf{z}\cdot\int_{\mathbb{R}^q} p(\mathbf{z}\,|\,\mathbf{y}_m)\frac{\partial\log f_{i'}(y_{i'm}\,|\,\mathbf{z})}{\partial\boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}}\;\mathrm{d}\mathbf{z}$$

Note that, when $i \neq i'$, the first summand in (A2.2) matrix of second derivatives in the first summation is a null matrix, i.e.,

$$\frac{\partial^2\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\boldsymbol{\alpha}_{i,\varphi}\partial\boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} = \frac{\partial\eta_{i,\varphi}}{\partial\boldsymbol{\alpha}_{i,\varphi}}\cdot\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\cdot\frac{\partial^2\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i\partial\tilde{\varphi}_{i'}}\cdot\frac{\partial\tilde{\varphi}_{i'}}{\partial\eta_{i',\tilde{\varphi}}}\cdot\frac{\partial\eta_{i',\tilde{\varphi}}}{\partial\boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} = \mathbf{0}\,,$$

but when $i = i'$, it becomes

$$\begin{aligned}
&= \frac{\partial}{\partial\boldsymbol{\alpha}_{i,\tilde{\varphi}}^{\intercal}}\left[\frac{\partial\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\boldsymbol{\alpha}_{i,\varphi}}\right]\\
&= \frac{\partial}{\partial\boldsymbol{\alpha}_{i,\tilde{\varphi}}^{\intercal}}\left[\frac{\partial\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i}\cdot\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\cdot\frac{\partial\eta_{i,\varphi}}{\partial\boldsymbol{\alpha}_{i,\varphi}}\right]\\
&= \frac{\partial}{\partial\tilde{\varphi}_i}\left[\frac{\partial\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i}\cdot\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\cdot\frac{\partial\eta_{i,\varphi}}{\partial\boldsymbol{\alpha}_{i,\varphi}}\right]\cdot\frac{\partial\tilde{\varphi}_i}{\partial\eta_{i,\tilde{\varphi}}}\cdot\frac{\partial\eta_{i,\tilde{\varphi}}}{\partial\boldsymbol{\alpha}_{i,\tilde{\varphi}}^{\intercal}}\\
&= \frac{\partial\eta_{i,\varphi}}{\partial\boldsymbol{\alpha}_{i,\varphi}}\cdot\frac{\partial}{\partial\tilde{\varphi}_i}\left[\frac{\partial\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i}\cdot\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\right]\cdot\frac{\partial\tilde{\varphi}_i}{\partial\eta_{i,\tilde{\varphi}}}\cdot\frac{\partial\eta_{i,\tilde{\varphi}}}{\partial\boldsymbol{\alpha}_{i,\tilde{\varphi}}^{\intercal}}\\
&= \frac{\partial\eta_{i,\varphi}}{\partial\boldsymbol{\alpha}_{i,\varphi}}\cdot\left[\frac{\partial^2\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i\partial\tilde{\varphi}_i}\cdot\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}+\frac{\partial}{\partial\tilde{\varphi}_i}\left(\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\right)\cdot\frac{\partial\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i}\right]\cdot\frac{\partial\tilde{\varphi}_i}{\partial\eta_{i,\tilde{\varphi}}}\cdot\frac{\partial\eta_{i,\tilde{\varphi}}}{\partial\boldsymbol{\alpha}_{i,\tilde{\varphi}}^{\intercal}}
\end{aligned}$$

which for $\varphi_i = \tilde{\varphi}_i$ is:

$$= \frac{\partial\eta_{i,\varphi}}{\partial\boldsymbol{\alpha}_{i,\varphi}}\cdot\left[\frac{\partial^2\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i^2}\cdot\left(\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\right)^2+\frac{\partial}{\partial\varphi_i}\left(\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\right)\cdot\frac{\partial\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i}\cdot\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\right]\cdot\frac{\partial\eta_{i,\varphi}}{\partial\boldsymbol{\alpha}_{i,\varphi}^{\intercal}}\,, \tag{A2.3}$$

but for $\varphi_i \neq \tilde{\varphi}_i$ simplifies to:

$$= \frac{\partial\eta_{i,\varphi}}{\partial\boldsymbol{\alpha}_{i,\varphi}}\cdot\left[\frac{\partial^2\log f_i(y_{im}\,|\,\mathbf{z})}{\partial\varphi_i\partial\tilde{\varphi}_i}\cdot\frac{\partial\varphi_i}{\partial\eta_{i,\varphi}}\cdot\frac{\partial\tilde{\varphi}_i}{\partial\eta_{i,\tilde{\varphi}}}\right]\cdot\frac{\partial\eta_{i,\tilde{\varphi}}}{\partial\boldsymbol{\alpha}_{i,\tilde{\varphi}}^{\intercal}} \tag{A2.4}$$

Expected information matrix (marginal log-likelihood)

The matrix entries in the expected information matrix are given by

$$\mathcal{I}_{[\bar{k}_{i,\varphi},\bar{k}_{i',\tilde{\varphi}}]} = -\mathbb{E}_{\mathbf{y}}\left[\mathcal{H}_{[\bar{k}_{i,\varphi},\bar{k}_{i',\tilde{\varphi}}]}\right]$$

Taking the expectation of (A2.2) with respect to $\mathbf{y}$ yields:

$$\mathcal{I}_{[\bar{k}_{i,\varphi},\bar{k}_{i',\tilde{\varphi}}]} = \sum_{m=1}^{n} \int_{\mathbb{R}^q} \mathbb{E}_{\mathbf{y}} \left[\frac{\partial^2 \log f_i(y_{im} \,|\, \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi} \partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} \right] p(\mathbf{z} \,|\, \mathbf{y}_m) \, \mathrm{d}\mathbf{z}$$

as the second and third summand in (A2.2) become null matrices (the expected value of the score vector is zero). Moreover, note that $\mathcal{I}_{[\bar{k}_{i,\varphi},\bar{k}_{i',\tilde{\varphi}}]} = \mathbf{0}$ for $i \neq i'$, and as such $\mathcal{I}$ is block diagonal.

Link functions

We present some analytical expressions for the partial derivatives involved in the score and Hessian functions. For entries in the parameter vector $\theta_i = (\mu_i, \sigma_i, \tau_i, \nu_i)^{\intercal}$, $i = 1, ..., p$, the link functions are monotonic, measurable, and differentiable mappings $\upsilon_i : \mathbb{R} \to \mathbb{R}$, that relate the systematic component (or predictor) to the corresponding location, shape, or scale parameters $\varphi_i \in \theta_i$, $\upsilon_i(\varphi_i) = \eta_{i,\varphi}$. Note that

$$\frac{\partial \varphi_i}{\partial \eta_{i,\varphi}} = \left(\frac{\partial \eta_{i,\varphi}}{\partial \varphi_i} \right)^{-1}$$

The link functions currently implemented are:

Link function	Parameter range	$\upsilon_i(\varphi_i) = \eta_{i,\varphi}$	$\left(\frac{\partial}{\partial \varphi} \eta_{i,\varphi} \right)^{-1}$
Identity link	$\varphi_i \in (-\infty, \infty)$	$\varphi_i = \eta_{i,\varphi}$	1
Log link	$\varphi_i \in (0, \infty)$	$\log(\varphi_i) = \eta_{i,\varphi}$	φ_i
Logit link	$\varphi_i \in (0, 1)$	$\log\left(\frac{\varphi_i}{1-\varphi_i} \right) = \eta_{i,\varphi}$	$\varphi_i \cdot (1 - \varphi_i)$
Probit link	$\varphi_i \in (0, 1)$	$\Phi^{-1}(\varphi_i) = \eta_{i,\varphi}$	$1/\phi^{-1}(\varphi_i)$

Table A1: Link functions and their derivatives

A3. A note on trust-region algorithms

Let the negative of the marginal log-likelihood in (2.1), $-\ell(\Theta; \mathbf{y})$, be the objective function to be minimised in the direct optimisation step of the estimation procedure discussed in Chapter 2. At iteration t of the trust-region algorithm, we construct a model function $\breve{\ell}^{[t]}$ that acts as a local approximation of $-\ell$ when evaluated at the current point $\Theta^{[t]}$, i.e., $-\ell(\Theta^{[t]}; \mathbf{y}) \approx \breve{\ell}^{[t]}$. To protect against bad approximations of $-\ell$, the algorithm restricts the search for a solution within a region

around $\Theta^{[t]}$, bounded by the trial step in the search process, $\mathbf{e}$. The model function is usually a quadratic approximation of the objective function about $\Theta^{[t]}$,

$$\breve{\ell}(\mathbf{e}; \Theta^{[t]}) = -\left\{ \ell(\Theta^{[t]}; \mathbf{y}) + \mathbf{e}^{\intercal}\mathbb{S}^{[t]} + \frac{1}{2}\mathbf{e}^{\intercal}\mathcal{H}^{[t]}\mathbf{e} \right\}$$

The trust-region algorithm restricts the search for the minimiser of $\breve{\ell}^{[t]}(\mathbf{e}; \Theta^{[t]})$ to a region defined by K-dimensional ball around $\Theta^{[t]}$ with radius $||\mathbf{e}||_2 \leq \Delta^{[t]}$, where $||\cdot||_2$ is the Euclidean norm, and the scalar $\Delta^{[t]} > 0$ denotes the trust-region radius at iteration t. The trust-region algorithm iteratively solves the following optimisation sub-problem: First, it chooses the trial step size $\mathbf{e}$ such that $\breve{\ell}^{[t]}(\mathbf{e}; \Theta^{[t]})$ becomes a good local approximation of $\ell(\Theta^{[t]})$, such that

$$\mathbf{e}^{[t]} = \underset{\mathbf{e} \in \mathbb{R}^K}{\arg\min}\, \breve{\ell}^{[t]}(\mathbf{e}; \Theta^{[t]}) \quad \text{subject to} \quad ||\mathbf{e}||_2 \leq \Delta^{[t]} \tag{A3.1}$$

and, secondly, updates the parameter vector as

$$\Theta^{[t+1]} = \Theta^{[t]} + \mathbf{e}^{[t]}$$

The size of the trust-region is critical, as the update happens only if the trial step produces an improvement over the objective function. This is measured by the agreement between the model function $\breve{\ell}^{[t]}(\mathbf{e}; \Theta^{[t]})$ and the objective function $\ell(\Theta^{[t]})$, as:

$$r^{[t]} = \frac{-\left\{ \ell(\Theta^{[t]}) - \ell(\Theta^{[t]} + \mathbf{e}^{[t]}) \right\}}{\breve{\ell}^{[t]}(\mathbf{0}; \Theta^{[t]}) - \breve{\ell}^{[t]}(\mathbf{e}^{[t]}; \Theta^{[t]})}$$

The numerator measures the actual reduction (agreement) in the objective function, and the denominator the predicted reduction. If $r^{[t]}$ is negative, the model function is not a good approximation of the objective function, thus the trial step $\mathbf{e}^{[t]}$ is rejected and we proceed to solve the sub-problem in equation (A3.1) for a smaller trust-region (i.e., smaller $\Delta^{[t]}$ and repeat search). If $r^{[t]}$ is close to 1, the model function is an adequate approximation of the objective function, thus the parameter vector is update and the trust-region is enlarged for the next iteration (i.e., $\Delta^{[t+1]} > \Delta^{[t]}$). If $r^{[t]}$ is positive but not close to 1, the parameters are updated but the trust region is unaltered (i.e., $\Delta^{[t+1]} = \Delta^{[t]}$). If $r^{[t]}$ is positive and close to zero, the parameters are updated and the trust-region is shrunken (i.e., $\Delta^{[t+1]} < \Delta^{[t]}$). We iterate until convergence, which is assessed by the stopping criteria $|\,\ell(\Theta^{[t+1]}) - \ell(\Theta^{[t]})\,| < \epsilon$ for a sufficiently small $\epsilon > 0$. We use the trust-region algorithm implementation in the `R` package `trust` (Geyer, 2020). Further details on the trust-region algorithm are found in Radice et al. (2016); Marra et al. (2017), and Nocedal

and Wright (2006, Chapter 4).

A4. Numerical Integration: The Gaussian-Hermite quadrature

The Gaussian-Hermite quadrature (GHQ) is a popular numerical integration technique in statistics, known for its numerical and computational efficiency. GHQ approximates a univariate integral of the form

$$I(f) = \int_{-\infty}^{\infty} \exp(-z^2) f(z) \mathrm{d}z$$

where $f(z)$ is an integrable function on $(-\infty, \infty)$. The integral is approximated as a discrete sum of R components:

$$I(f) \approx I_{GH}(f; \{z_r, w_r\}_{r=1}^{R}) = \sum_{r=1}^{R} w_r \cdot f(z_r) + \mathcal{R}_R$$

where the quadrature points z_r (also known as nodes) correspond to the r^{th} zero of the Hermite polynomial of degree R, $H_R(z)$, and the weights w_r are given by:

$$w_i = \frac{2^{R-1} R! \sqrt{\pi}}{R^2 (H_{R-1}(z_r))^2} \qquad (r = 1, 2, ..., R)$$

The remainder term $\mathcal{R}_R$ takes the form

$$\mathcal{R}_R = \frac{R! \sqrt{\pi}}{2^R (2R)!} f^{(2R)}(z)$$

for some z, where $f^{(2R)}(z)$ denotes the $2R^{\text{th}}$-order derivative of $f(z)$. The nodes $\{z_r\} r = 1^R$ are symmetrical around zero. Under standard regularity conditions, the accuracy of GHQ improves as the number of quadrature points increases, as the remainder term $\mathcal{R}_R$ implies that the approximation is exact if $f(z)$ is a polynomial of degree at most $R-1$. Further details on Hermite polynomials can be found in Davis and Rabinowitz (1975, Chapter 2), and tables of nodes and weights can be found in Stroud and Secrest (1966).

The GHQ can be extended to multivariate integrals in q dimensions. The approximation takes the form:

$$\int_{\mathbb{R}^q} \exp(-\mathbf{z}'\mathbf{z}) f(\mathbf{z}) \, \mathrm{d}\mathbf{z} \approx \sum_{r_1=1}^{R_1} \cdots \sum_{r_q=1}^{R_q} w_{r_1} \cdots w_{r_q} f(z_{r_1}, ..., z_{r_q}) = \sum_{r=1}^{R} w_r f(\mathbf{z}_r)$$

where $\mathbf{z}r = (zr_1, ..., zr_q)' \in \mathbb{R}^q$ is the r^{th} (out of $R = \prod_{l=1}^{q} R_l$) quadrature point, and $w_r = \prod_{l=1}^{q} w_{r_l}$ is its corresponding weight. If the integrand does not include the Gauss function $\exp(-z^2)$ as

a factor, we can still use GHQ by multiplying and dividing the integrand by this factor. The GHQ approximation works well when the function $\exp(-z^2)g(z)$ is smooth enough. This is the case when computing the multivariate integrals of the score vectors and observed and expected information matrices in Chapters 2 and 3. In this case, the weights are adjusted by a factor of $(2\pi)^{-1/2} \cdot \exp(z_{r_l}^2/2)$ to obtain the posterior distribution $p(\mathbf{z} \mid \mathbf{y})$ in the integrand. Therefore, we can compute the score vectors as:

$$\begin{aligned}
\mathbb{S}_{[\bar{k}_{i,\varphi}]} &= \sum_{m=1}^{n} \int_{\mathbb{R}^q} p(\mathbf{z} \mid \mathbf{y}_m) \cdot \frac{\partial \log f_i(y_{im} \mid \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi}} \, \mathrm{d}\mathbf{z} \\
&= \sum_{m=1}^{n} \int_{\mathbb{R}^q} \frac{p(\mathbf{z}) \cdot f(\mathbf{y}_m \mid \mathbf{z})}{f(\mathbf{y}_m)} \cdot \frac{\partial \log f_i(y_{im} \mid \mathbf{z})}{\partial \boldsymbol{\alpha}_{i,\varphi}} \, \mathrm{d}\mathbf{z} \\
&\approx \sum_{m=1}^{n} \sum_{r=1}^{R} (2\pi)^{-q/2} \cdot \exp(\mathbf{z}_r^2/2) \cdot w_r \cdot \frac{f(\mathbf{y}_m \mid \mathbf{z}_r)}{f(\mathbf{y}_m)} \cdot \frac{\partial \log f_i(y_{im} \mid \mathbf{z}_r)}{\partial \boldsymbol{\alpha}_{i,\varphi}} \\
&\approx \sum_{m=1}^{n} \sum_{r=1}^{R} \tilde{w}_r \cdot \frac{\partial \log f_i(y_{im} \mid \mathbf{z}_r)}{\partial \boldsymbol{\alpha}_{i,\varphi}} \, ,
\end{aligned}$$

where, $\tilde{w}_r = (2\pi)^{-q/2} \cdot \exp(\mathbf{z}_r^2/2) \cdot w_r \cdot f(\mathbf{y}_m \mid \mathbf{z}_r)/f(\mathbf{y}_m)$, and, similarly, the observed information matrices as

$$\mathbb{H}_{[\bar{k}_{i,\varphi}, \bar{k}_{i,\tilde{\varphi}}]} \approx \sum_{m=1}^{n} \sum_{r=1}^{R} \tilde{w}_r \cdot \frac{\partial^2 \log f_i(y_{im} \mid \mathbf{z}_r)}{\partial \boldsymbol{\alpha}_{i,\varphi} \partial \boldsymbol{\alpha}_{i,\tilde{\varphi}}^{\intercal}}$$

and the expected information matrices as

$$\begin{aligned}
\mathcal{H}_{[\bar{k}_{i,\varphi}, \bar{k}_{i',\tilde{\varphi}}]} \approx & \sum_{m=1}^{n} \sum_{r=1}^{R} \tilde{w}_r \cdot \frac{\partial^2 \log f_i(y_{im} \mid \mathbf{z}_r)}{\partial \boldsymbol{\alpha}_{i,\varphi} \partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} \\
& + \sum_{m=1}^{n} \sum_{r=1}^{R} \tilde{w}_r \cdot \frac{\partial \log f_i(y_{im} \mid \mathbf{z}_r)}{\partial \boldsymbol{\alpha}_{i,\varphi}} \cdot \frac{\partial \log f_{i'}(y_{i'm} \mid \mathbf{z}_r)}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} \\
& - \sum_{m=1}^{n} \left[\sum_{r=1}^{R} \tilde{w}_r \cdot \frac{\partial \log f_i(y_{im} \mid \mathbf{z}_r)}{\partial \boldsymbol{\alpha}_{i,\varphi}} \cdot \sum_{r=1}^{R} \tilde{w}_r \cdot \frac{\partial \log f_{i'}(y_{i'm} \mid \mathbf{z}_r)}{\partial \boldsymbol{\alpha}_{i',\tilde{\varphi}}^{\intercal}} \right]
\end{aligned}$$

A5. Asymptotic properties of the MML estimator

We derive the asymptotic properties of the MML estimator introduced in Section 2.3. Let $\Theta^* \in \Xi$ be the true population parameters, with $\Xi \subseteq \mathbb{R}^K$ the parameter space. Consider the assumptions:

(A1) Θ^* is an interior point of Ξ.

(A2) Within a neighbourhood of Θ^*, the first three derivatives of the marginal log-likelihood exist and are bounded.

(A3) $\Theta^* \in \Xi$ is identified. That is, the marginal distribution generated by Θ^*, $f(\mathbf{y}; \Theta^*)$, is unique. Consequently, the marginal log-likelihood evaluated at Θ^*, $\ell(\Theta^*; \mathbf{y})$, is unique too.

Assumptions (A1)-(A3) are standard regularity conditions (e.g., Barndorff-Nielsen and Cox, 1994). Specifically, assumptions (A1) and (A2) are necessary to approximate the marginal log-likelihood by a Taylor expansion about Θ^*. Assumption (A3) is a requirement for the consistency of the MLE. From assumptions (A1)-(A3) we can further derive the following assumptions:

(A4) $\mathbb{E}[\mathbb{S}(\Theta^*)] = \mathbf{0}$, where $\mathbb{S}(\Theta^*) = \nabla_\Theta \ell(\mathbf{y}; \Theta)|_{\Theta=\Theta^*}$. Moreover, assume $\mathbb{S}(\Theta^*) = \mathcal{O}_P(n)$, where the big-O notation in probability means that for any $\epsilon > 0$, exists a finite $M(\epsilon) > 0$ such that $\mathbb{P}(\, ||\mathbb{S}(\Theta^*)/n||_1 \geq M(\epsilon)) < \epsilon$, for sufficiently large n.

(A5) $\text{Var}[\mathbb{S}(\Theta^*)] = \mathbb{E}[\mathbb{S}(\Theta^*)\mathbb{S}(\Theta^*)^\intercal] = \mathcal{I}(\Theta^*) = -\mathbb{E}[\mathbb{H}(\Theta^*)]$, where $\mathbb{H}(\Theta^*) = \nabla_\Theta \nabla_{\Theta^\intercal} \ell(\mathbf{y}; \Theta)|_{\Theta=\Theta^*}$.

(A6) The observed information matrix the sum of n individual contributions from the independent and identically distributed observations in the sample, $\mathbb{H}(\Theta^*) = \sum_{m=1}^n \mathbb{H}_m(\Theta^*)$, with $\mathbb{H}_m(\Theta^*) = \nabla_\Theta \nabla_{\Theta^\intercal} \log f(\mathbf{y}_m; \Theta)|_{\Theta=\Theta^*}$. Likewise, the expected information matrix is $\mathcal{I}(\Theta^*) = \sum_{m=1}^n \mathcal{I}_m(\Theta^*)$, with $\mathcal{I}_m(\Theta^*) = -\mathbb{E}[\mathbb{H}_m(\Theta^*)]$. Assume $\mathcal{I}_m(\Theta^*)$ is constant, i.e., $-\mathcal{I}_m(\Theta^*) = \mathcal{O}(1)$ for all $m = 1, \ldots, n$. Thus, $\mathcal{I}(\Theta^*) = n\mathcal{I}_m(\Theta^*)$. It follows that $-\mathcal{I}(\Theta^*) = \mathcal{O}(n)$, that is, $\mathcal{I}(\Theta^*)$ is constant in the limit $n \to \infty$.

(A7) Similarly, $\mathbb{S}(\Theta^*) = \sum_{m=1}^n \mathbb{S}_m(\Theta^*)$, with $\nabla_\Theta \log f(\mathbf{y}_m; \Theta)|_{\Theta=\Theta^*}$. Moreover, $\mathbb{E}[\mathbb{S}_m(\Theta^*)] = \mathbf{0}$ and $\text{Var}[\mathbb{S}_m(\Theta^*)] = -\mathbb{E}[\mathbb{H}_m(\Theta^*)] = \mathcal{I}_m(\Theta^*)$.

(A8) The observed information matrix can be decomposed into its mean and a stochastic part, $\mathbb{H}(\Theta^*) = \mathbb{E}[\mathbb{H}(\Theta^*)] + \epsilon_\mathbb{H}$, where $\epsilon_\mathbb{H} = \mathcal{O}_P(\sqrt{n})$ is a negligible error term (Kauermann, 2005).

We now present the asymptotic distribution of the MLE:

Theorem A5.1. *(Asymptotic distribution of the MLE) Under assumptions (A1)-(A6), the maximum likelihood estimator (MLE), denoted by $\hat{\Theta}$, has the following asymptotic distribution:*

$$\sqrt{n}(\hat{\Theta} - \Theta^*) \xrightarrow{d} \mathbb{N}(\mathbf{0}, n\mathcal{I}(\Theta^*)^{-1})$$

Proof: Start by performing a Taylor expansion of the score vector $\mathbb{S}(\hat{\Theta})$ about Θ^*. For simplicity, all the orders higher than the first order are omitted. The first-order expansion of $\mathbb{S}(\hat{\Theta})$ about Θ^* is approximately

$$\mathbb{S}(\hat{\Theta}) \approx \mathbb{S}(\Theta^*) + \mathbb{H}(\Theta^*)(\hat{\Theta} - \Theta^*)$$

By definition, the left-hand-side (LHS) of the above approximation is $\mathbb{S}(\hat{\Theta}) = \mathbf{0}$. Multiplying on both sides by $\sqrt{n}$ and re-arranging gives

$$\sqrt{n}(\hat{\Theta} - \Theta^*) = -\mathbb{H}(\Theta^*)^{-1}\sqrt{n}\mathbb{S}(\Theta^*)$$

and further dividing and multiplying by n on the right-hand-side (RHS) yields:

$$\sqrt{n}(\hat{\Theta} - \Theta^*) = -\left[\frac{\mathbb{H}(\Theta^*)}{n}\right]^{-1}\sqrt{n}\,\frac{\mathbb{S}(\Theta^*)}{n} \tag{A5.1}$$

The term $\mathbb{S}(\Theta^*)/n$ in (A5.1) is the sample mean of $\{\mathbb{S}_m(\Theta^*) : m = 1, \dots n\}$. Thus, by the central limit theorem and assumptions (A5)-(A7):

$$\sqrt{n}\,\frac{\mathbb{S}(\Theta^*)}{n} \xrightarrow{d} \mathbb{N}\left(\mathbf{0}, \frac{\mathcal{I}(\Theta^*)}{n}\right)$$

Moreover, by the law of large numbers and assumption (A8):

$$-\left[\frac{\mathbb{H}(\Theta^*)}{n}\right]^{-1} \to -\left[\frac{\mathcal{I}(\Theta^*)}{n}\right]^{-1}$$

Therefore,

$$\sqrt{n}(\hat{\Theta} - \Theta^*) \xrightarrow{d} \mathbb{N}\left(\mathbf{0}, \left[\frac{\mathcal{I}(\Theta^*)}{n}\right]^{-1}\frac{\mathcal{I}(\Theta^*)}{n}\left[\frac{\mathcal{I}(\Theta^*)}{n}\right]^{-1}\right) \overset{d}{=} \mathbb{N}\left(\mathbf{0}, \left[\frac{\mathcal{I}(\Theta^*)}{n}\right]^{-1}\right)$$

□

A6. Software Implementation

In this Appendix, we present the implementation of the GLVM-LSS in the statistical software `R` to contribute to reproducible research practices and transparent dissemination of results. We discuss the estimation of the GLVM-LSS in the PISA 2018 empirical application (Section 2.5.1).

Data preparation

Here, we present the code for data preparation using the original PISA 2018 files. We select the 9 items from the first testlet (see

```r
rm(list = ls())
set.seed(1234)

source("R/prep.R")      # Code for pre-fitting prepration
source("R/fams.R")      # Code with distributions
source("R/glvmlss.R")   # Code for fitting and simulating GLVM-LSS models
source("R/misc.R")      # Code with miscellaneous functions

library(dplyr)

# Original datafiles from PISA
itemfile <- haven::read_sav("Data/PISA/CY6_MS_CMB_STU_COG.sav")
timefile <- haven::read_sas("Data/PISA/cy6_ms_cmb_stu_ttm.sas7bdat")

# Data cleaning
BR_item <- itemfile %>%
  filter(CNTRYID == 76) %>%
  select(CNTSTUID, starts_with("CM")) %>%
  select(CNTSTUID, ends_with("S")) %>%
  rename_with(~ stringr::str_remove(., 'S'), .cols = starts_with("CM")) %>%
  select(CNTSTUID, CM033Q01, CM474Q01, CM155Q01, CM155Q04,
  CM411Q01, CM411Q02, CM803Q01, CM442Q02, CM034Q01) %>%
  arrange(CNTSTUID) %>%
  rename(ID = CNTSTUID)

BR_time <- itemfile %>%
    filter(CNTRYID == 76) %>%
    select(CNTSTUID, starts_with("CM")) %>% select(CNTSTUID, ends_with("T")) %>%
    rename_with(~ stringr::str_remove(., 'T'), .cols = starts_with("CM")) %>%
    arrange(CNTSTUID) %>%
    rename(ID = CNTSTUID) %>%
    select(colnames(UK_item))

BR_item <- BR_item %>% rename_with(~ paste0("Y", 1:9), .cols = starts_with("CM"))
BR_time <- BR_time %>% rename_with(~ paste0("T", 1:9), .cols = starts_with("CM"))

data <- full_join(BR_time, BR_item, "ID") %>% select(!ID)

# log-RT in minutes
data[,paste0("T",1:9)] <- log(data[,paste0("T",1:9)]/(1000*60))
# Complete cases
data <- data[complete.cases(data),]
```

Confirmatory restrictions

Restrictions on the model parameters should be defined in a list that is included as an additional control option in the main function.

```
iRes <- list(c("mu","Y1","Z2",0), c("mu","Y2","Z2",0), c("mu","Y3","Z2",0),
        c("mu","Y4","Z2",0), c("mu","Y5","Z2",0), c("mu","Y6","Z2",0),
        c("mu","Y7","Z2",0), c("mu","Y8","Z2",0), c("mu","Y9","Z2",0),
        c("mu","T1","Z1",0), c("mu","T2","Z1",0), c("mu","T3","Z1",0),
        c("mu","T4","Z1",0), c("mu","T5","Z1",0), c("mu","T6","Z1",0),
        c("mu","T7","Z1",0), c("mu","T8","Z1",0), c("mu","T9","Z1",0),
        c("sigma","Y1","Z2",0), c("sigma","Y2","Z2",0), c("sigma","Y3","Z2",0),
        c("sigma","Y4","Z2",0), c("sigma","Y5","Z2",0), c("sigma","Y6","Z2",0),
        c("sigma","Y7","Z2",0), c("sigma","Y8","Z2",0), c("sigma","Y9","Z2",0),
        c("sigma","T1","Z1",0), c("sigma","T2","Z1",0), c("sigma","T3","Z1",0),
        c("sigma","T4","Z1",0), c("sigma","T5","Z1",0), c("sigma","T6","Z1",0),
        c("sigma","T7","Z1",0), c("sigma","T8","Z1",0), c("sigma","T9","Z1",0),
        c("nu","Y1","Z2",0), c("nu","Y2","Z2",0), c("nu","Y3","Z2",0),
        c("nu","Y4","Z2",0), c("nu","Y5","Z2",0), c("nu","Y6","Z2",0),
        c("nu","Y7","Z2",0), c("nu","Y8","Z2",0), c("nu","Y9","Z2",0),
        c("nu","T1","Z1",0), c("nu","T2","Z1",0), c("nu","T3","Z1",0),
        c("nu","T4","Z1",0), c("nu","T5","Z1",0), c("nu","T6","Z1",0),
        c("nu","T7","Z1",0), c("nu","T8","Z1",0), c("nu","T9","Z1",0))
```

In confirmatory models, the users must specify each restriction in the form `c("parameter", "item", "latent variable", "value")`. In the example above, the restriction of the form `c("mu","Y1","Z2",0)` means that the factor loading of the item Y1 on the latent speed trait is set in the measurement equation for the location parameter is set to $\alpha_{12,\mu} = 0$.

Distributions

The distributions for each item (in this case, item responses and response times) should be included in a list with elements of the class `dist_glvmlss`. In this case, the first 9 observed variables, the log-RT, are assumed to follow a Skew-Normal distribution, and items 10 to 18, the observed IR, are assumed to follow a Bernoulli distribution.

```
famSN <- vector("list", ncol(data))
for(i in 1:9){ famSN[[i]] <- SkewNormal()}
for(i in 10:18){ famSN[[i]] <- Binomial()}
```

Estimation

To estimate the GLVM-LSS, we use the function `glvmlss`. Model 7 in Section 2.5.1 results from:

```
mod_SN_comp <- glvmlss(data = data, family = famSN,
                       mu.eq = ~ Z1+Z2, sg.eq = ~ Z1+Z2, nu.eq = ~ Z1+Z2,
                       f.scores = T,
                       verbose = T, iden.res = iRes, corr.lv = T,
                       solver = "nlminb", EM_use2d = F, EM_iter = 5000,
                       iter.lim = 1000,
                       est.ci = "Approximate")

names(mod_SN_comp)
```

The `glvmlss` has three main components: i) `data`, which corresponds to the $n \times p$ matrix of observed variables, ii) `family`, which corresponds to list with elements of the class `dist_glvmlss` created earlier, and iii) the formulas corresponding to the location (`mu.eq`), scale (`sg.eq`) and shape (`nu.eq`/`ta.eq`) parameters. The formula object follows the conventional notation in `R`, with `Z1, ..., Zp` denoting the number of latent variables in the model (for computational simplicity, we do not recommend using more than 3).

Control variables can be included directly, or in a list with the name `control`. The following options, along with their default values, are available for the user:

```
control <- list(
    EM_iter = 30,                          # Number of user-defined EM iterations
    EM_use2d = T,                            # Use GD (FALSE) or NR (TRUE) update
    iter.lim = 300,              # Limit for iterations (quasi-Newton algorithm)
    DirectMaxFlag = T,  # Skip (FALSE) or Keep (TRUE) quasi-Newton maximisation
    EM_appHess = F,     # Use score product to approximate Hessian in EM (TRUE)
    EM_lrate = 0.001,           # Learning rate when using GD update rule in EM
    est.ci = "Approximate",                  # Options: "Standard, "Approximate"
    solver = "nlminb",   # Options: "trust", "L-BGFS-B", (any other in "optim")
    start.val = NULL,              # Starting values, options a list or "random"
    mat.info = "Hessian",       # Options: "Hessian", "Fisher" for info. matrix
    iden.res = NULL,                                 # Identification restrictions
    tol = sqrt(.Machine$double.eps)                  # Stopping criteria tolerance
    corr.lv = FALSE,                         # Correlated latent variables (LVs)?
    Rz = NULL,    # Covariance for LVs (if corr.lv == T, estimated, else fixed)
    var.lv = rep(1,q),      # Estimate variances of LVs? (if 1, fixed, else NA)
    nQP = if(q == 1) 40 else { if(q == 2) 25 else 10 },     # Quadrature points
    verbose = FALSE,                           # Display steps in console (TRUE)?
```

```
    f.scores = F,                                  # Compute factor scores (TRUE)?
    seed = 1234)                                 # Seed when start.val == "random"
```

Chapter B

Appendix for Chapter 3

B1. Non-convex Penalty Functions

The Lasso and Alasso are known for their variable selection capabilities, but they are consistent for model selection only under certain restricted conditions (Zhao and Yu, 2006; Zou, 2006). Additionally, the Lasso lacks the *oracle property*. To address these limitations, non-convex alternatives have been proposed. Two popular non-convex L_1-type penalties are the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and the Minimax Concave Penalty (MCP, Zhang, 2010). Following the general formulation of the penalty term in Section 3.3, the parameter-specific penalty contributions to $\mathcal{P}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}; \boldsymbol{w})$ for the SCAD are:

$$\mathcal{P}_{\lambda,k}(\alpha_k; \lambda_{\varphi,k}, w_k) = \begin{cases} \lambda_{\varphi,k}|\alpha_k| & \text{if } |\alpha_k| \leq \lambda_{\varphi,k} \\ \dfrac{2a\lambda_{\varphi,k}|\alpha_k| - \alpha_k^2 - (\lambda_{\varphi,k})^2}{2(a-1)} & \text{if } \lambda_{\varphi,k} < |\alpha_k| \leq a\lambda_{\varphi,k} \\ \dfrac{(\lambda_{\varphi,k})^2(a+1)}{2} & \text{if } |\alpha_k| > \lambda_{\varphi,k} \end{cases}$$

with implicit weights w_k depending on the value of α_k relative to $\lambda_{\varphi,k}$. The SCAD penalty coincides with the Lasso until $|\alpha_k| = \lambda_{\varphi,k}$, then becomes a quadratic function until $|\alpha_k| = a\lambda_{\varphi,k}$, and then remains constant for $|\alpha_k| > \lambda_{\varphi,k}$. In essence, the SCAD penalty can be seen as a quadratic spline function with knots at $\lambda_{\varphi,k}$ and $a\lambda_{\varphi,k}$. The choice of the additional tuning parameter a typically falls between 2.5 and 4.5 (Huang et al., 2017), with 3.7 being a commonly used value in the literature (Fan and Li, 2001).

For the MCP, the the parameter-specific penalty contributions are:

$$\mathcal{P}_{\lambda,k}(\alpha_k;\lambda_{\varphi,k},w_k)=\begin{cases}\lambda_{\varphi,k}|\alpha_k|-\dfrac{\alpha_k^2}{2a} & \text{if } |\alpha_k|\leq a\lambda_{\varphi,k}\\ \dfrac{a(\lambda_{\varphi,k})^2}{2} & \text{if } |\alpha_k|>a\lambda_{\varphi,k}\end{cases}$$

The MCP, like the SCAD, incorporates a penalisation *rate* similar to that of the Lasso. However, the MCP relaxes this rate towards zero as the absolute value of the coefficient approaches $a\lambda_{\varphi,k}$. The choice of the additional tuning parameter a typically falls between 1.5 and 3.5 (Huang, 2018). Increasing the value of a results in stronger penalisation for small values of α_k and weaker penalisation for large values. Notably, both the SCAD and the MCP approach the Lasso when a tends towards infinity.

B2. Local Approximations to Penalty Functions

Define the function $e_0(x)=|x|$. The idea is to construct a family of convex C^2-functions (i.e., functions that have both a continuous first derivative and a continuous second derivative), denoted by $e_c(x)$, such that $e_c\to e_0$ as $c\to 0$. Koch (1996) proposed the family of functions $e_c:\mathbb{R}\to\mathbb{R}$ defined by

$$e_c(x)=(x^2+c^2)^{1/2}\,,\qquad c>0\,,\ x\in\mathbb{R} \tag{B2.1}$$

From (B2.1), it follows that $\lim_{c\to 0}e_c(x)\to e_0(x)$.

In the current context, the approximation above is $||\mathrm{R}_k\Theta||_1\approx((\mathrm{R}_k\Theta)^\intercal(\mathrm{R}_k\Theta)+\bar{c})^{1/2}$. Let $\tilde{\Theta}$ be a vector in the neighbourhood of Θ. In the following, we drop the notational dependence on the tuning parameters and weights $(\boldsymbol{\lambda},\boldsymbol{w})$ from the penalty function. A first-order Taylor expansion of $\mathcal{P}_{\boldsymbol{\lambda}}(\Theta)$ about $\tilde{\Theta}$ is

$$\begin{aligned}\mathcal{P}_{\boldsymbol{\lambda}}(\Theta)&\approx\mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta})+\nabla_\Theta\mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta})^\intercal(\Theta-\tilde{\Theta})\\&=\mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta})+\left[\sum_{k=1}^{k^*}\nabla_\Theta\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)^\intercal\right](\Theta-\tilde{\Theta})\\&=\mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta})+\left[\sum_{k=1}^{k^*}\left(\frac{\partial\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\right)\left(\frac{\partial||\mathrm{R}_k\tilde{\Theta}||_1}{\partial\mathrm{R}_k\tilde{\Theta}}\right)^\intercal\left(\frac{\partial\mathrm{R}_k\tilde{\Theta}}{\partial\tilde{\Theta}}\right)^\intercal\right](\Theta-\tilde{\Theta})\end{aligned} \tag{B2.2}$$

The first term in the summation indexed by k, the scalar $\partial\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)/\partial||\mathrm{R}_k\tilde{\Theta}||_1$, depends on the specific functional form of the penalty term $\mathcal{P}_{\lambda,k}(\cdot)$. Approximating the second term in the

summation gives:

$$\frac{\partial||\mathrm{R}_k\tilde{\Theta}||_1}{\partial \mathrm{R}_k\tilde{\Theta}} \approx \frac{\partial}{\partial \mathrm{R}_k\tilde{\Theta}}\left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta})+\bar{c}\right]^{1/2}$$
$$= \left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta})+\bar{c}\right]^{-1/2}\mathrm{R}_k\tilde{\Theta};$$

The third term in the summation equals $\partial \mathrm{R}_k\tilde{\Theta}/\partial\tilde{\Theta} = \mathrm{R}_k$. Including the vector $(\tilde{\Theta}-\Theta)$ inside the square brackets results in:

$$\mathcal{P}_{\boldsymbol{\lambda}}(\Theta) \approx \mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta}) + \sum_{k=1}^{k^*}\left[\frac{\partial \mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta})+\bar{c}\right]^{-1/2}(\mathrm{R}_k\tilde{\Theta})^\intercal\mathrm{R}_k(\Theta-\tilde{\Theta})\right]$$
$$= \mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta}) + \sum_{k=1}^{k^*}\mathcal{D}_k$$

with $\mathcal{D}_k$ denoting the product inside the squared brackets. Letting $\mathrm{R}_k\tilde{\Theta} \approx \mathrm{R}_k\Theta$ (Fan and Li, 2001), the last term in $\mathcal{D}_k$ can be approximated as (Ulbricht, 2010):

$$(\mathrm{R}_k\tilde{\Theta})^\intercal\mathrm{R}_k(\Theta-\tilde{\Theta}) \approx (\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k(\Theta-\tilde{\Theta})$$
$$= (\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k\Theta - (\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k\tilde{\Theta}$$
$$= \frac{1}{2}\left[(\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k\Theta - 2(\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k\tilde{\Theta} + (\mathrm{R}_k\tilde{\Theta})^\intercal\mathrm{R}_k\tilde{\Theta}\right]$$
$$+ \frac{1}{2}\left[(\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k\Theta - (\mathrm{R}_k\tilde{\Theta})^\intercal\mathrm{R}_k\tilde{\Theta}\right]$$
$$= \frac{1}{2}\left[(\Theta-\tilde{\Theta})^\intercal\mathrm{R}_k^\intercal\mathrm{R}_k(\Theta-\tilde{\Theta})\right] + \frac{1}{2}\left[(\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k\Theta - (\mathrm{R}_k\tilde{\Theta})^\intercal\mathrm{R}_k\tilde{\Theta}\right]$$
$$\approx \frac{1}{2}\left[(\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k\Theta - (\mathrm{R}_k\tilde{\Theta})^\intercal\mathrm{R}_k\tilde{\Theta}\right],$$

where the last step is justified because the quadratic form of $(\tilde{\Theta}-\Theta)$ can be neglected due to the proximity between the two vectors. With the above, $\mathcal{D}_k$ becomes

$$\mathcal{D}_k = \frac{\partial \mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta})+\bar{c}\right]^{-1/2}(\mathrm{R}_k\tilde{\Theta})^\intercal\mathrm{R}_k(\Theta-\tilde{\Theta})$$
$$\approx \frac{\partial \mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta})+\bar{c}\right]^{-1/2}\cdot\frac{1}{2}\left[(\mathrm{R}_k\Theta)^\intercal\mathrm{R}_k\Theta - (\mathrm{R}_k\tilde{\Theta})^\intercal\mathrm{R}_k\tilde{\Theta}\right]$$
$$= \frac{1}{2}\Theta^\intercal\left[\frac{\partial \mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta})+\bar{c}\right]^{-1/2}\mathrm{R}_k^\intercal\mathrm{R}_k\right]\Theta$$
$$- \frac{1}{2}\tilde{\Theta}^\intercal\left[\frac{\partial \mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\left[(\mathrm{R}_k\tilde{\Theta})^\intercal(\mathrm{R}_k\tilde{\Theta})+\bar{c}\right]^{-1/2}\mathrm{R}_k^\intercal\mathrm{R}_k\right]\tilde{\Theta}$$

$$= \frac{1}{2}\left[\Theta^{\intercal}\mathcal{S}_k(\tilde{\Theta})\Theta - \tilde{\Theta}^{\intercal}\mathcal{S}_k(\tilde{\Theta})\tilde{\Theta}\right]$$

with $\mathcal{S}_k(\tilde{\Theta}) = \left(\partial\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)/\partial||\mathrm{R}_k\tilde{\Theta}||_1\right)\left[(\mathrm{R}_k\tilde{\Theta})^{\intercal}(\mathrm{R}_k\tilde{\Theta}) + \bar{c}\right]^{-1/2}\mathrm{R}_k$, and because R_k is an idempotent matrix (i.e., $\mathrm{R}_k^{\intercal}\mathrm{R}_k = \mathrm{R}_k\mathrm{R}_k = \mathrm{R}_k$). Define $\mathcal{S}(\tilde{\Theta}) = \sum_{k=1}^{k^*}\mathcal{S}_k(\tilde{\Theta})$. Also, given that the penalty term only depends on Θ and all the terms depending only on $\tilde{\Theta}$ are constant, the penalty term can be expressed as:

$$\mathcal{P}_{\boldsymbol{\lambda}}(\Theta) \approx \mathcal{P}_{\boldsymbol{\lambda}}(\tilde{\Theta}) + \frac{1}{2}\left[\Theta^{\intercal}\mathcal{S}(\tilde{\Theta})\Theta\right] - \frac{1}{2}\left[\tilde{\Theta}^{\intercal}\mathcal{S}(\tilde{\Theta})\tilde{\Theta}\right] \approx \frac{1}{2}\left[\Theta^{\intercal}\mathcal{S}(\tilde{\Theta})\Theta\right] \tag{B2.3}$$

The resulting approximated penalty matrix $\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})$ is a $K \times K$ block diagonal matrix of the form

$$\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta}) = \begin{bmatrix} S_{\boldsymbol{\lambda}}(\tilde{\Theta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{B2.4}$$

where $S_{\boldsymbol{\lambda}}(\tilde{\Theta})$ is a diagonal $k^* \times k^*$ matrix with entries

$$S_{\boldsymbol{\lambda}}(\tilde{\Theta})_{[k,k]} = \frac{\partial\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\left[(\mathrm{R}_k\tilde{\Theta})^{\intercal}(\mathrm{R}_k\tilde{\Theta}) + \bar{c}\right]^{-1/2} \qquad \text{for} \quad k = 1, ..., k^*$$

and the sub-matrices $\mathbf{0}$ are the null matrices of dimension $(K - k^*) \times (K - k^*)$. The local approximations of the Lasso, Alasso, SCAD, and MCP penalties are:

Lasso penalty: The derivative of the Lasso penalty with respect to the L_1-norm of its argument is:

$$\left[\frac{\partial\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\right]_{[k]} = \frac{\partial\left(\lambda_{\varphi,k}|\tilde{\alpha}_k|\right)}{\partial|\tilde{\alpha}_k|} = \lambda_{\varphi,k}$$

and thus:

$$S_{\boldsymbol{\lambda}}(\tilde{\Theta})_{[k,k]} = \frac{\partial\left(\lambda_{\varphi,k}|\tilde{\alpha}_k|\right)}{\partial|\tilde{\alpha}_k|}\left[\tilde{\alpha}_k^2 + \bar{c}\right]^{-1/2} = \lambda_{\varphi,k}\left[\tilde{\alpha}_k^2 + \bar{c}\right]^{-1/2} \qquad \text{for} \quad k = 1, ..., k^*$$

Adaptive Lasso (Alasso): The derivative of the Alasso with respect to the L_1-norm of its argument is:

$$\left[\frac{\partial\mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial||\mathrm{R}_k\tilde{\Theta}||_1}\right]_{[k]} = \frac{\partial\left(\lambda_{\varphi,k}|\tilde{\alpha}_k| \div |\hat{\alpha}_k|^a\right)}{\partial|\tilde{\alpha}_k|} = \frac{\lambda_{\varphi,k}}{|\hat{\alpha}_k|^a}$$

and therefore,

$$S_{\boldsymbol{\lambda}}(\tilde{\Theta})_{[k,k]} = \frac{\partial\left(\lambda_{\varphi,k}|\tilde{\alpha}_k| \div |\hat{\alpha}_k|^a\right)}{\partial|\tilde{\alpha}_k|}\left[\tilde{\alpha}_k^2 + \bar{c}\right]^{-1/2} = \frac{\lambda_{\varphi,k}}{|\hat{\alpha}_k|^a}\left[\tilde{\alpha}_k^2 + \bar{c}\right]^{-1/2} \qquad \text{for} \quad k = 1, ..., k^*$$

SCAD: For the Smoothly Clipped Absolute Deviation (SCAD) penalty, the derivative with respect to the L_1-norm of its argument is (Fan and Li, 2001):

$$\left[\frac{\partial \mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial ||\mathrm{R}_k\tilde{\Theta}||_1}\right]_{[k]} = \lambda_{\varphi,k}\left[\mathbb{1}(|\tilde{\alpha}_k| \leq \lambda_{\varphi,k}) + \frac{\max(a\lambda_{\varphi,k} - |\tilde{\alpha}_k|, 0)}{(a-1)\lambda_{\varphi,k}} \cdot \mathbb{1}(|\tilde{\alpha}_k| > \lambda_{\varphi,k})\right]$$

$$= \begin{cases} \lambda_{\varphi,k} & \text{if } |\tilde{\alpha}_k| \leq \lambda_{\varphi,k} \\ \dfrac{a\lambda_{\varphi,k} - |\tilde{\alpha}_k|}{a-1} & \text{if } \lambda_{\varphi,k} < |\tilde{\alpha}_k| \leq a\lambda_{\varphi,k} \\ 0 & \text{if } |\tilde{\alpha}_k| > a\lambda_{\varphi,k} \end{cases}$$

and therefore, for $k = 1, ..., k^*$:

$$S_{\boldsymbol{\lambda}}(\tilde{\Theta})_{[k,k]} = \lambda_{\varphi,k}\left[\mathbb{1}(|\tilde{\alpha}_k| \leq \lambda_{\varphi,k}) + \frac{\max(a\lambda_{\varphi,k} - |\tilde{\alpha}_k|, 0)}{(a-1)\lambda_{\varphi,k}} \cdot \mathbb{1}(|\tilde{\alpha}_k| > \lambda_{\varphi,k})\right] \cdot \left[\tilde{\alpha}_k^2 + \bar{c}\right]^{-1/2}$$

MCP: For the Minimax Concave Penalty (MCP), the derivative with respect to the L_1-norm of its argument is:

$$\left[\frac{\partial \mathcal{P}_{\lambda,k}(||\mathrm{R}_k\tilde{\Theta}||_1)}{\partial ||\mathrm{R}_k\tilde{\Theta}||_1}\right]_{[k]} = \left[\lambda_{\varphi,k} - \frac{|\tilde{\alpha}_k|}{a}\right] \cdot \mathbb{1}(|\tilde{\alpha}_k| < a\lambda_{\varphi,k})$$

$$= \begin{cases} \lambda_{\varphi,k} - \dfrac{|\tilde{\alpha}_k|}{a} & \text{if } |\tilde{\alpha}_k| \leq a\lambda_{\varphi,k} \\ 0 & \text{if } |\tilde{\alpha}_k| > a\lambda_{\varphi,k} \end{cases}$$

and therefore,

$$S_{\boldsymbol{\lambda}}(\tilde{\Theta})_{[k,k]} = \left(\lambda_{\varphi,k} - \frac{|\tilde{\alpha}_k|}{a}\right) \cdot \mathbb{1}(|\tilde{\alpha}_k| < a\lambda_{\varphi,k}) \cdot \left[\tilde{\alpha}_k^2 + \bar{c}\right]^{-1/2} \qquad \text{for} \quad k = 1, ..., k^*$$

B3. Generalised Information Criterion (GIC)

In this Appendix, we present the main derivations from Konishi and Kitagawa (2008) adapted to the GLVM-LSS model. To facilitate the notation, we use Lebesgue integral notation for expected values. Consider a set of n observations $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$, where each observation is generated from the same unknown true p-variate distribution function $G(\mathbf{y})$ with the associated multivariate density function $g(\mathbf{y})$. These observations can be seen as realisations of the random vector $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_n)^\intercal$ consisting of independent and identically distributed random variables. Next, define a collection of parametric multivariate distributions $\{f(\mathbf{y}; \Theta) : \Theta \in \Xi \subseteq \mathbb{R}^K\}$, where Θ is a K-

dimensional vector of parameters and Ξ is the parameter space. Furthermore, we assume that the true density $g(\mathbf{y})$ belongs to this collection, which means that there exists a $\Theta_0 \in \Xi$ such that $g(\mathbf{y}) = f(\mathbf{y}; \Theta_0)$. A statistical model is obtained by evaluating the parametric distribution at a vector of estimated parameters, $f(\mathbf{y}; \hat{\Theta})$.

We assume that each parameter α_k in Θ can be expressed as a real-valued function of the true distribution G, denoted as $\alpha_k = T_k(G)$. Here, $T_k(G)$ is defined over the set of all distributions on the sample space and does not depend on the sample size n. In practice, we do not have access to the true distribution G, but we can estimate it using the empirical distribution based on the observed sample, denoted by $\hat{G}$. As a result, the estimator $\hat{\alpha}_k$ can be written as:

$$\hat{\alpha}_k = \hat{\alpha}_k(\mathbf{y}_1, ..., \mathbf{y}_n) = T_k(\hat{G}), \qquad \text{for} \quad k = 1, ..., K = \dim(\Theta)$$

where $\hat{G}$ represents the empirical distribution function with a probability mass function $\hat{g}(\mathbf{y}_m) = 1/n$ for any $m = 1, ..., n$. Since $\hat{\alpha}_k = T_k(\hat{G})$ depends on the observed data through the empirical distribution, it is commonly referred to as a *statistical functional*. We define the K-dimensional statistical functional vector as $\mathbf{T}(G) = (T_1(G), ..., T_K(G))^\intercal$, where $\mathbf{T}(G)$ represents the solution to the equations:

$$\mathbb{E}_{g(\mathbf{y})}(\boldsymbol{\omega}(\mathbf{y}, \mathbf{T}(G))) = \int \boldsymbol{\omega}(\mathbf{y}, \mathbf{T}(G)) \, \mathrm{d}G(\mathbf{y}) = \mathbf{0}, \tag{B3.1}$$

where the vector-valued function $\boldsymbol{\omega} = (\omega_1, ..., \omega_K)^\intercal$ collect the real-valued mappings $\omega_k(\mathbf{y}, \mathbf{T}(G))$ defined over $(\mathbb{R}^n \times \mathbb{R}^p) \times \Xi$. For example, for the MML estimation problem described in Chapter 2, the functional takes the form:

$$\boldsymbol{\omega}(\mathbf{y}, \mathbf{T}(G)) = \frac{\partial \ell(\Theta; \mathbf{y})}{\partial \Theta}\Big|_{\Theta = \mathbf{T}(G)} = \mathbb{S}(\Theta; \mathbf{y})\big|_{\Theta = \mathbf{T}(G)},$$

and for the penalised MML estimation problem described in Section 3.3.2, $\boldsymbol{\omega}(\mathbf{y}, \mathbf{T}(G))$ is

$$\boldsymbol{\omega}(\mathbf{y}, \mathbf{T}(G)) = \frac{\partial \ell_p(\Theta; \mathbf{y})}{\partial \Theta}\Big|_{\Theta = \mathbf{T}(G)} = \Big\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\Big\}\Big|_{\Theta = \mathbf{T}(G)} \tag{B3.2}$$

Thus, the penalised MML estimate can be expressed as $\hat{\Theta} = \mathbf{T}(\hat{G}) = (T_1(\hat{G}), ..., T_K(\hat{G}))^\intercal$, where $\mathbf{T}(\hat{G})$ is the solution to the system of penalised marginal log-likelihood equations:

$$\boldsymbol{\omega}(\mathbf{y}, \mathbf{T}(\hat{G})) = \sum_{m=1}^{n} \boldsymbol{\omega}(\mathbf{y}_m, \mathbf{T}(\hat{G})) = \sum_{m=1}^{n} \boldsymbol{\omega}(\mathbf{y}_m, \hat{\Theta}) = \left\{\sum_{m=1}^{n} \frac{\partial \log f(\mathbf{y}_m; \hat{\Theta})}{\partial \Theta} - n\mathcal{S}_{\boldsymbol{\lambda}}(\hat{\Theta})\hat{\Theta}\right\} = \mathbf{0}$$

It is common for the model selection problem to prioritise accurate predictions on independent

new data that were not used for estimating $\hat{\Theta}$. Let $\breve{\mathbf{y}}$ denote such data, generated from the unknown true distribution $g(\breve{\mathbf{y}})$. In this context, the performance of a statistical model $f(\breve{\mathbf{y}}; \hat{\Theta})$ can be assessed by measuring the Kullback-Leibler distance between the model and the true distribution, evaluated at the new data points. This distance is given by:

$$\begin{aligned} D_{\mathrm{KL}}(g(\breve{\mathbf{y}}); f(\breve{\mathbf{y}}; \hat{\Theta})) &:= \mathbb{E}_{g(\breve{\mathbf{y}})}\left[\log\left(\frac{g(\breve{\mathbf{y}})}{f(\breve{\mathbf{y}}; \hat{\Theta})}\right)\right] = \int \log\left(\frac{g(\breve{\mathbf{y}})}{f(\breve{\mathbf{y}}; \hat{\Theta})}\right)\, \mathrm{d}G(\breve{\mathbf{y}}) \\ &= \int \log g(\breve{\mathbf{y}})\, \mathrm{d}G(\breve{\mathbf{y}}) - \int \log f(\breve{\mathbf{y}}; \hat{\Theta})\, \mathrm{d}G(\breve{\mathbf{y}}) \end{aligned} \tag{B3.3}$$

In equation (B3.3), the first term involves the true but unknown model, which is not available for direct comparison. Therefore, for the purpose of assessing goodness-of-fit and prediction, we focus on the second term, referred to as the expected (marginal) log-likelihood: $\mathbb{E}_{g(\breve{\mathbf{y}})}[\log f(\breve{\mathbf{y}}; \hat{\Theta})] = \int \log f(\breve{\mathbf{y}}; \hat{\Theta})\, \mathrm{d}G(\breve{\mathbf{y}})$. A larger value of the expected (marginal) log-likelihood indicates a closer similarity between $g(\breve{\mathbf{y}})$ and $f(\breve{\mathbf{y}}; \hat{\Theta})$, suggesting a better fit of the proposed model to the true model in terms of information. However, it is important to note that the expected (marginal) log-likelihood still depends on $g(\breve{\mathbf{y}})$, and the key challenge lies in obtaining a good estimator. To address this, we replace the unknown probability distribution G with the empirical distribution function $\hat{G}$, yielding:

$$\begin{aligned} \mathbb{E}_{\hat{g}(\breve{\mathbf{y}})}\left[\log f(\breve{\mathbf{y}}; \hat{\Theta})\right] = \int \log f(\breve{\mathbf{y}}; \hat{\Theta})\, \mathrm{d}\hat{G}(\breve{\mathbf{y}}) &= \sum_{m=1}^{n} \log f(y_m; \hat{\Theta}) \cdot \hat{g}(\breve{\mathbf{y}}) \\ &= \frac{1}{n}\sum_{m=1}^{n} \log f(\mathbf{y}_m; \hat{\Theta}) = \frac{1}{n}\ell(\hat{\Theta}; \mathbf{y}) \end{aligned}$$

It is worth noting the relationship between $\mathbb{E}_{\hat{g}(\breve{\mathbf{y}})}[\log f(\breve{\mathbf{y}}; \hat{\Theta})]$ and the maximum marginal log-likelihood, with a scaling factor proportional to the sample size, $n^{-1}\ell(\hat{\Theta}; \mathbf{y})$. This relationship supports the use of the maximum marginal log-likelihood as a natural estimator of $n \cdot \mathbb{E}_{g(\breve{\mathbf{y}})}[\log f(\breve{\mathbf{y}}; \hat{\Theta})]$. According to the law of large numbers, as the sample size n approaches infinity, the estimated expected marginal log-likelihood converges to its true expectation:

$$\mathbb{E}_{\hat{g}(\breve{\mathbf{y}})}\left[\log f(\breve{\mathbf{y}}; \hat{\Theta})\right] = \frac{1}{n}\sum_{m=1}^{n} \log f(\mathbf{y}_m; \hat{\Theta}) \xrightarrow{n\to\infty} \mathbb{E}_{g(\breve{\mathbf{y}})}\left[\log f(\breve{\mathbf{y}}; \hat{\Theta})\right]$$

Hence, one can argue that $\ell(\hat{\Theta}; \mathbf{y}) \xrightarrow{n\to\infty} n \cdot \mathbb{E}_{g(\breve{\mathbf{y}})}[\log f(\breve{\mathbf{y}}; \hat{\Theta})]$. This implies that the goodness-of-fit or predictive accuracy of a set of competing models can be determined by comparing the values of their maximum (marginal) log-likelihood, $\ell(\hat{\Theta}; \mathbf{y})$. However, using the same data points $\mathbf{y}$ to

estimate both $\mathbb{E}_{g(\check{\mathbf{y}})}[\log f(\check{\mathbf{y}};\hat{\Theta})]$ and $\hat{\Theta}$ introduces bias into $\ell(\hat{\Theta};\mathbf{y})$. Furthermore, the magnitude of this bias varies with the dimension of the parameter vector. Evaluating and correcting this bias allows for a fair comparison between competing models. This is precisely why Konishi and Kitagawa (2008) define information criteria as bias-corrected log-likelihood-based measures of the goodness-of-fit of a statistical model.

The bias of the marginal log-likelihood as an estimator of the expected marginal log-likelihood can be defined as:

$$b(G) := \mathbb{E}_{g(\mathbf{y})}\left[\sum_{m=1}^{n} \log f(\mathbf{y}_m;\hat{\Theta}) - n \cdot \mathbb{E}_{g(\check{\mathbf{y}})}[\log f(\check{\mathbf{y}};\hat{\Theta})]\right]$$

where the outer expectation is evaluated with respect to the joint distribution of the sample, $g(\mathbf{y}) = \prod_{m=1}^{n} g(\mathbf{y}_m)$, and the inner expectation is taken with respect to the true distribution, $g(\check{\mathbf{y}})$. After some derivations (refer to Konishi and Kitagawa, 2008, Chapter 3), it can be shown that the bias term for a model estimated through MML is asymptotically equivalent to

$$\begin{aligned} b(G) &= \operatorname{tr}\left(\left[-\int \frac{\partial^2 \ell(\Theta;\mathbf{y})}{\partial\Theta\partial\Theta^\intercal}\Big|_{\Theta=\mathbf{T}(G)} \mathrm{d}G(\mathbf{y})\right]^{-1}\left[\int \frac{\partial \ell(\Theta;\mathbf{y})}{\partial\Theta}\frac{\partial \ell(\Theta;\mathbf{y})}{\partial\Theta^\intercal}\Big|_{\Theta=\mathbf{T}(G)} \mathrm{d}G(\mathbf{y})\right]\right) + o(1) \\ &= \operatorname{tr}\left(\left[-\int \frac{\partial \boldsymbol{\omega}(\mathbf{y},\mathbf{T}(G))}{\partial\Theta^\intercal} \mathrm{d}G(\mathbf{y})\right]^{-1}\left[\int \boldsymbol{\omega}(\mathbf{y},\mathbf{T}(G))\,\boldsymbol{\omega}(\mathbf{y},\mathbf{T}(G))^\intercal \mathrm{d}G(\mathbf{y})\right]\right) + o(1) \\ &= \operatorname{tr}\left(\mathbb{E}_{g(\mathbf{y})}\left[-\mathcal{H}(\Theta;\mathbf{y})|_{\Theta=\mathbf{T}(G)}\right]^{-1}\ \mathbb{E}_{g(\mathbf{y})}\left[\mathbb{S}(\Theta;\mathbf{y})\,\mathbb{S}(\Theta;\mathbf{y})^\intercal|_{\Theta=\mathbf{T}(G)}\right]\right) + o(1) \end{aligned} \tag{B3.4}$$

where $\operatorname{tr}(\cdot)$ is the matrix trace operator, and $o(1)$ describes an error term of the approximation that goes to zero asymptotically. Define the $K \times K$ positive semi-definite matrices $J(G) = \mathbb{E}_{g(\mathbf{y})}\left[-\mathcal{H}(\Theta;\mathbf{y})|_{\Theta=\mathbf{T}(G)}\right]$ and $I(G) = \mathbb{E}_{g(\mathbf{y})}\left[\mathbb{S}(\Theta;\mathbf{y})\,\mathbb{S}(\Theta;\mathbf{y})^\intercal|_{\Theta=\mathbf{T}(G)}\right]$, corresponding to the expected information matrix and the product of the scores, respectively, evaluated with respect to the true distribution $g(\mathbf{y})$. According to asymptotic maximum likelihood theory, if the true distribution is included in the class of parametric models being compared, we have $J(G) = I(G)$ for the MML estimator. Therefore, the bias term can be expressed as $\operatorname{tr}(J(G)^{-1}I(G)) = \operatorname{tr}(\mathbb{I}_K) = K$, which represents the degrees of freedom of the model estimated by MML..

Since the bias term depends on the unknown true distribution $g(\mathbf{y})$, it must be estimated using observed data and the empirical distribution $\hat{g}(\mathbf{y})$. Let $\hat{J}(\hat{G})$ and $\hat{I}(\hat{G})$ be consistent estimators of $J(G)$ and $I(G)$. An estimator for the bias term is $\hat{b}(\hat{G}) = \operatorname{tr}(\hat{J}(\hat{G})^{-1}\hat{I}(\hat{G}))$. The Generalised Information Criterion (GIC, Konishi and Kitagawa, 1996, 2008) is formulated by evaluating and

correcting for the bias of the marginal log-likelihood:

$$\text{GIC}(\hat{\Theta}) = -2\left(\sum_{m=1}^{n} \log f(\mathbf{y}_m; \hat{\Theta}) - \hat{b}(\hat{G})\right) = -2\ell(\hat{\Theta}; \mathbf{y}) + 2 \cdot \hat{b}(\hat{G}) \tag{B3.5}$$

The relationship between the Generalised Information Criterion (GIC) and other model selection criteria is immediate. The GIC serves as an extension of the Akaike Information Criterion (AIC) when the estimated bias term reflects the degrees of freedom of the model, corresponding to the number of estimated parameters. Similarly, it extends the Bayesian Information Criterion (BIC) by replacing the weight assigned to the penalty term of 2 with $\log(n)$. However, the specific form of the estimated bias term, denoted as $\hat{b}(\hat{G})$, can vary depending not only on the relationship between the true distribution that generates the data and the proposed model but also on the method used to estimate the parameters.

In a penalised maximum likelihood framework, the bias term differs from the total number of parameters utilised in the AIC and BIC. Notably, when employing the BIC to choose the tuning parameter in a PMML setting, as observed in much of the literature on penalised factor models (excluding Geminiani et al., 2021), the BIC tends to over-correct for the bias in the log-likelihood. Consequently, there is a risk of selecting a tuning parameter that leads to a model that does not adequately approximate the true data generating process.

In Konishi and Kitagawa (1996, 2008, Chapters 3, 5, and 7), it is demonstrated that the bias term can be expressed in terms of statistical functionals, as defined earlier in this section. This formulation enables the evaluation, estimation, and correction of the bias in the marginal log-likelihood for various estimators, including the penalised maximum marginal likelihood approach. To begin, consider a functional $\mathbf{T}(G)$ and define the directional derivative with respect to the distribution G as the vector-valued function $\mathbf{T}^{(1)}(\breve{\mathbf{y}}; G) = (T_1^{(1)}(\breve{\mathbf{y}}; G), \ldots, T_K^{(1)}(\breve{\mathbf{y}}; G))^\intercal$, which satisfies the following equation for any distribution function $H(\breve{\mathbf{y}})$:

$$\begin{aligned}\lim_{\epsilon \to 0} \frac{\mathbf{T}((1-\epsilon)G + \epsilon H) - \mathbf{T}(G)}{\epsilon} &= \frac{\partial}{\partial \epsilon} \left\{\mathbf{T}((1-\epsilon)G + \epsilon H) - \mathbf{T}(G)\right\} \Big|_{\epsilon=0} \\ &= \int \mathbf{T}^{(1)}(\breve{\mathbf{y}}; G) \, \mathrm{d}\left\{H(\breve{\mathbf{y}}) - G(\breve{\mathbf{y}})\right\}\end{aligned}$$

To ensure uniqueness, the following condition must hold: $\int \mathbf{T}^{(1)}(\breve{\mathbf{y}}; G) \, \mathrm{d}G(\breve{\mathbf{y}}) = 0$. Additionally, consider $H(\breve{\mathbf{y}}) = \delta_{\breve{\mathbf{y}}}$, where $\delta_{\breve{\mathbf{y}}}$ represents a probability mass of 1 at $\breve{\mathbf{y}}$ and zero elsewhere. Under

these conditions, the aforementioned equality required for defining $\mathbf{T}^{(1)}(\breve{\mathbf{y}}; G)$ becomes:

$$\frac{\partial}{\partial \epsilon}\left\{\mathbf{T}((1-\epsilon)G + \epsilon\delta_{\breve{\mathbf{y}}}) - \mathbf{T}(G)\right\}\Big|_{\epsilon=0} = \int \mathbf{T}^{(1)}(\breve{\mathbf{y}}; G)\ \mathrm{d}\delta_{\breve{\mathbf{y}}} = \mathbf{T}^{(1)}(\breve{\mathbf{y}}; G)$$

This expression, called the *influence function*, describes the effect of an infinitesimal contamination at the point $\breve{\mathbf{y}}$ on the true distribution G. To derive the influence function, we need to compute the derivative of the functional. We begin by substituting $(1-\epsilon)G + \epsilon\delta_{\breve{\mathbf{y}}}$ for G and, for the PMML estimator, the functional in (B3.2) for $\boldsymbol{\omega}(\mathbf{y}, \mathbf{T}(G))$ in (B3.1). This yields the following expression:

$$\begin{aligned}
\mathbf{0} &= \int \boldsymbol{\omega}(\mathbf{y}, \mathbf{T}(G))\ \mathrm{d}G(\mathbf{y}) \\
&= \int \left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}((1-\epsilon)G+\epsilon\delta_{\breve{\mathbf{y}}})} \mathrm{d}\left\{(1-\epsilon)G(\mathbf{y}) + \epsilon\delta_{\breve{\mathbf{y}}}(\mathbf{y})\right\}
\end{aligned}$$

Differentiating on both sides with respect to ϵ yields:

$$\begin{aligned}
\mathbf{0} &= \int \frac{\partial}{\partial\epsilon}\left[\left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}((1-\epsilon)G+\epsilon\delta_{\breve{\mathbf{y}}})} \mathrm{d}\left\{(1-\epsilon)G(\mathbf{y}) + \epsilon\delta_{\breve{\mathbf{y}}}(\mathbf{y})\right\}\right] \\
&= \int \left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}((1-\epsilon)G+\epsilon\delta_{\breve{\mathbf{y}}})} \frac{\partial}{\partial\epsilon}\mathrm{d}\left\{(1-\epsilon)G(\mathbf{y}) + \epsilon\delta_{\breve{\mathbf{y}}}(\mathbf{y})\right\} \\
&\quad + \int \frac{\partial}{\partial\epsilon}\left[\left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}((1-\epsilon)G+\epsilon\delta_{\breve{\mathbf{y}}})}\right] \mathrm{d}\left\{(1-\epsilon)G(\mathbf{y}) + \epsilon\delta_{\breve{\mathbf{y}}}(\mathbf{y})\right\} \\
&= \int \left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}((1-\epsilon)G+\epsilon\delta_{\breve{\mathbf{y}}})} \mathrm{d}\left\{-G(\mathbf{y}) + \delta_{\breve{\mathbf{y}}}(\mathbf{y})\right\} \\
&\quad + \int \frac{\partial}{\partial\Theta^{\intercal}}\left[\left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}((1-\epsilon)G+\epsilon\delta_{\breve{\mathbf{y}}})}\right] \\
&\qquad \times \frac{\partial}{\partial\epsilon}\left[\mathbf{T}((1-\epsilon)G + \epsilon\delta_{\breve{\mathbf{y}}})\right]\ \mathrm{d}\left\{(1-\epsilon)G(\mathbf{y}) + \epsilon\delta_{\breve{\mathbf{y}}}(\mathbf{y})\right\}
\end{aligned} \tag{B3.6}$$

Evaluating the expression in (B3.6) at $\epsilon = 0$, and considering the result in equation (B3.1) gives:

$$\begin{aligned}
\mathbf{0} &= \int \left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}(G)} \mathrm{d}\left\{-G(\mathbf{y}) + \delta_{\breve{\mathbf{y}}}(\mathbf{y})\right\} \\
&\quad + \int \frac{\partial}{\partial\Theta^{\intercal}}\left[\left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}(G)}\right] \times \frac{\partial}{\partial\epsilon}\left[\mathbf{T}((1-\epsilon)G + \epsilon\delta_{\breve{\mathbf{y}}})\right]\Big|_{\epsilon=0} \mathrm{d}G(\mathbf{y}) \\
&= \int \left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}(G)} \mathrm{d}\delta_{\breve{\mathbf{y}}}(\mathbf{y}) - \cancel{\int \left\{\mathbb{S}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\Theta\right\}\Big|_{\Theta=\mathbf{T}(G)} \mathrm{d}G(\mathbf{y})} \\
&\quad + \int \left[\mathcal{H}(\Theta; \mathbf{y}) - n\mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta})\right]\Big|_{\Theta=\mathbf{T}(G)} \mathrm{d}G(\mathbf{y}) \times \frac{\partial}{\partial\epsilon}\left[\mathbf{T}((1-\epsilon)G + \epsilon\delta_{\breve{\mathbf{y}}})\right]\Big|_{\epsilon=0}
\end{aligned}$$

With the above, the influence function can be written as:

$$\begin{aligned}
\mathbf{T}^{(1)}(\breve{\mathbf{y}}; G) &:= \frac{\partial}{\partial \epsilon} \left[\mathbf{T}((1-\epsilon)G + \epsilon \delta_{\breve{\mathbf{y}}}) \right] \Big|_{\epsilon=0} \\
&= \left[\int -\mathcal{H}_p(\Theta; \mathbf{y}) \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\mathbf{y}) \right]^{-1} \left[\int \mathbb{S}_p(\Theta; \mathbf{y}) \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}\delta_{\breve{\mathbf{y}}}(\mathbf{y}) \right] \\
&= \left[\int -\mathcal{H}_p(\Theta; \mathbf{y}) \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\mathbf{y}) \right]^{-1} \left[\mathbb{S}_p(\Theta; \breve{\mathbf{y}}) \Big|_{\Theta=\mathbf{T}(G)} \right] \\
&= \mathbb{E}_{g(\mathbf{y})} \left[-\mathcal{H}_p(\Theta; \mathbf{y}) |_{\Theta=\mathbf{T}(G)} \right]^{-1} \left[\mathbb{S}_p(\Theta; \breve{\mathbf{y}}) |_{\Theta=\mathbf{T}(G)} \right] \\
&= \mathcal{R}(\boldsymbol{\omega}, G)^{-1} \boldsymbol{\omega}(\breve{\mathbf{y}}, \mathbf{T}(G))
\end{aligned} \tag{B3.7}$$

With the influence function $\mathbf{T}^{(1)}(\breve{\mathbf{y}}; G)$, Konishi and Kitagawa (2008, Chapter 7) show that the asymptotic bias term for the PMML estimator can be expressed in terms of the influence function as:

$$\begin{aligned}
b(G) &= \operatorname{tr} \left(\int \mathbf{T}^{(1)}(\breve{\mathbf{y}}; G) \, \frac{\partial \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta^\intercal} \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\breve{\mathbf{y}}) \right) + o(1) \\
&= \operatorname{tr} \left(\int \mathcal{R}(\boldsymbol{\omega}, G)^{-1} \boldsymbol{\omega}(\breve{\mathbf{y}}, \mathbf{T}(G)) \, \frac{\partial \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta^\intercal} \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\breve{\mathbf{y}}) \right) + o(1) \\
&= \operatorname{tr} \left(\mathcal{R}(\boldsymbol{\omega}, G)^{-1} \int \boldsymbol{\omega}(\breve{\mathbf{y}}, \mathbf{T}(G)) \, \frac{\partial \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta^\intercal} \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\breve{\mathbf{y}}) \right) + o(1) \\
&= \operatorname{tr} \left(\mathcal{R}(\boldsymbol{\omega}, G)^{-1} \, \mathcal{W}(\boldsymbol{\omega}, G) \right) + o(1)
\end{aligned} \tag{B3.8}$$

which is similar to the asymptotic MML bias in (B3.4). The $K \times K$ matrix $\mathcal{W}(\boldsymbol{\omega}, G)$ is defined as:

$$\begin{aligned}
\mathcal{W}(\boldsymbol{\omega}, G) &= \int \boldsymbol{\omega}(\breve{\mathbf{y}}, \mathbf{T}(G)) \, \frac{\partial \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta^\intercal} \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\breve{\mathbf{y}}) \\
&= \int \frac{\partial \ell_p(\Theta; \breve{\mathbf{y}})}{\partial \Theta} \, \frac{\partial \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta^\intercal} \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\breve{\mathbf{y}}) \\
&= \int \frac{\partial \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta} \, \frac{\partial \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta^\intercal} \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\breve{\mathbf{y}}) - \int n \mathcal{S}_{\boldsymbol{\lambda}}(\tilde{\Theta}) \Theta \, \frac{\partial \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta^\intercal} \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\breve{\mathbf{y}}) \\
&= - \int \frac{\partial^2 \ell(\Theta; \breve{\mathbf{y}})}{\partial \Theta \partial \Theta^\intercal} \Big|_{\Theta=\mathbf{T}(G)} \, \mathrm{d}G(\breve{\mathbf{y}}) \equiv \mathcal{W}(G)
\end{aligned}$$

with the last line following from standard asymptotic likelihood theory and the fact that the penalty function is independent of the true distribution G. In addition, the dependence on $\boldsymbol{\omega}$ is dropped because the functional is no longer part of the expression. The estimated bias is obtained by replacing the unknown true distribution G with the empirical distribution $\hat{G}$, resulting in

$\hat{b}(\hat{G}) = \text{tr}(\mathcal{R}(\boldsymbol{\omega}, \hat{G})^{-1}\, \mathcal{W}(\hat{G}))$. In more detail, this can be expressed as follows:

$$\begin{aligned}
\mathcal{R}(\boldsymbol{\omega}, \hat{G}) &= -\int \left(\sum_{m=1}^{n} \frac{\partial^2 \log f(\mathbf{y}_m; \Theta)}{\partial \Theta \partial \Theta^\intercal} - n\mathcal{S}_{\boldsymbol{\lambda}}(\Theta) \right) \Big|_{\Theta = \mathbf{T}(\hat{G})} \mathrm{d}\hat{G}(\mathbf{y}) \\
&= -\frac{1}{n} \sum_{m=1}^{n} \left(\frac{\partial^2 \log f(\mathbf{y}_m; \Theta)}{\partial \Theta \partial \Theta^\intercal} - n\mathcal{S}_{\boldsymbol{\lambda}}(\Theta) \right) \Big|_{\Theta = \mathbf{T}(\hat{G})} \\
&= -\frac{1}{n} \left(\frac{\partial^2 \ell(\hat{\Theta}; \mathbf{y})}{\partial \Theta \partial \Theta^\intercal} - n\mathcal{S}_{\boldsymbol{\lambda}}(\hat{\Theta}) \right) = -\frac{1}{n} \mathcal{H}_p(\hat{\Theta})
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{W}(\hat{G}) &= -\int \left(\sum_{m=1}^{n} \frac{\partial^2 \log f(\mathbf{y}_m; \Theta)}{\partial \Theta \partial \Theta^\intercal} \right) \Big|_{\Theta = \mathbf{T}(\hat{G})} \mathrm{d}\hat{G}(\mathbf{y}) \\
&= -\frac{1}{n} \sum_{m=1}^{n} \left(\frac{\partial^2 \log f(\mathbf{y}_m; \Theta)}{\partial \Theta \partial \Theta^\intercal} \right) \Big|_{\Theta = \mathbf{T}(\hat{G})} = -\frac{1}{n} \left(\frac{\partial^2 \ell(\hat{\Theta}; \mathbf{y})}{\partial \Theta \partial \Theta^\intercal} \right) = -\frac{1}{n} \mathcal{H}(\hat{\Theta})
\end{aligned}$$

With the above, and following (B3.8), the estimated bias term for the PMML estimator is $\hat{b}(\hat{G}) = \text{tr}\left(\mathcal{H}_p(\hat{\Theta})^{-1} \mathcal{H}(\hat{\Theta})\right)$. Thus, for a given GLVM-LSS model with parameters $\hat{\Theta}$ estimated via PMML, the GIC is defined as

$$\text{GIC}(\hat{\Theta}, \lambda) = -2\ell(\hat{\Theta}) + 2 \cdot \text{tr}\left(\mathcal{H}_p(\hat{\Theta})^{-1} \mathcal{H}(\hat{\Theta})\right) \tag{B3.9}$$

and an approximate Generalised Bayesian Information Criterion (GBIC, Konishi et al., 2004) as:

$$\text{GBIC}(\hat{\Theta}, \lambda) = -2\ell(\hat{\Theta}) + \log(n) \cdot \text{tr}\left(\mathcal{H}_p(\hat{\Theta})^{-1} \mathcal{H}(\hat{\Theta})\right) \tag{B3.10}$$

B4. Automatic Selection of the Tuning Parameter Vector

B4.1 Estimation and Computation

The optimal value for $\boldsymbol{\lambda} = (\lambda_\mu, \lambda_\sigma, \lambda_\nu, \lambda_\tau)^\intercal$ is determined by minimising the GBIC value (due to the consistency in model selection property of the BIC):

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} \in [0, \infty)^S}{\arg\min}\, \text{GBIC}(\hat{\Theta}, \boldsymbol{\lambda})$$

where S is the dimension of the vector $\boldsymbol{\lambda}$, indicating the number of different tuning parameters associated with the location, scale, or shape parameters indexing the items' conditional distributions within the measurement part of the GLVM-LSS model. In this section of the Appendix, we outline

an automatic and data-driven procedure for estimating $\hat{\boldsymbol{\lambda}}$, based on the statistical properties of the MLEs and the relationship between an unbiased risk estimate (UBRE) and the AIC.

To simplify the notation, we omit the specific point at which the score vector is evaluated, denoting it as $\mathbb{S} = \mathbb{S}(\Theta)$. We introduce special notation to indicate the evaluation of the score vector at a particular point during iteration t, denoted as $\mathbb{S}^{[t]} = \mathbb{S}(\Theta^{[t]})$. Similarly, $\hat{\mathbb{S}} = \mathbb{S}(\hat{\Theta})$ represents the score vector evaluated at the PMLE. The same convention is applied to the information matrices: $\mathcal{H}$, $\mathcal{H}^{[t]}$, and $\hat{\mathcal{H}}$ denote the observed information matrix, while $\mathcal{I}$, $\mathcal{I}^{[t]}$, and $\hat{\mathcal{I}}$ denote the expected information matrix, respectively. We adopt a similar notation for the approximate penalty term, where we omit the dependence on the parameter vector, tuning parameters, and weights. Thus, we have $\mathcal{S}_{\boldsymbol{\lambda}} = \mathcal{S}_{\boldsymbol{\lambda}}(\Theta; \boldsymbol{\lambda}, \boldsymbol{w})$, $\mathcal{S}_{\boldsymbol{\lambda}}^{[t]} = \mathcal{S}_{\boldsymbol{\lambda}}(\Theta^{[t]}; \boldsymbol{\lambda}, \boldsymbol{w})$, and $\hat{\mathcal{S}}_{\boldsymbol{\lambda}} = \mathcal{S}_{\boldsymbol{\lambda}}(\hat{\Theta}; \boldsymbol{\lambda}, \boldsymbol{w})$. Furthermore, we define $\mathbb{J} = -\mathcal{H}$.

Let $\boldsymbol{\lambda}_0$ be the (initial) fixed value for the tuning parameter vector. In a neighbourhood of the (local) mode, the quasi-Newton and/or trust-region algorithm update rules behave like a classic unconstrained Newton-Raphson algorithm (Nocedal and Wright, 2006). Consequently, at iteration $t+1$, we have $\mathbb{S}_p^{[t+1]} \approx 0$, allowing us to express the Newton-Raphson update step as follows:

$$
\begin{aligned}
0 &\approx \mathbb{S}_p^{[t+1]} \\
&\approx \mathbb{S}_p^{[t]} + \mathcal{H}_p^{[t]}(\Theta^{[t+1]} - \Theta^{[t]}) \\
\mathbb{S}^{[t]} - n\mathcal{S}_{\boldsymbol{\lambda}_0}^{[t]}\Theta^{[t]} &= \left[\mathbb{J}^{[t]} + n\mathcal{S}_{\boldsymbol{\lambda}_0}^{[t]}\right]\Theta^{[t+1]} - \mathbb{J}^{[t]}\Theta^{[t]} - n\mathcal{S}_{\boldsymbol{\lambda}_0}^{[t]}\Theta^{[t]} \\
\left[\mathbb{J}^{[t]} + n\mathcal{S}_{\boldsymbol{\lambda}_0}^{[t]}\right]\Theta^{[t+1]} &= \mathbb{J}^{[t]}\Theta^{[t]} + \mathbb{S}^{[t]} \\
\left[\mathbb{J}^{[t]} + n\mathcal{S}_{\boldsymbol{\lambda}_0}^{[t]}\right]\Theta^{[t+1]} &= \sqrt{\mathbb{J}^{[t]}}^{\intercal}\left[\sqrt{\mathbb{J}^{[t]}}\Theta^{[t]} + \sqrt{\mathbb{J}^{[t]}}^{-\intercal}\mathbb{S}^{[t]}\right] \\
\Theta^{[t+1]} &= \left[\mathbb{J}^{[t]} + n\mathcal{S}_{\boldsymbol{\lambda}_0}^{[t]}\right]^{-1}\sqrt{\mathbb{J}^{[t]}}^{\intercal}\,\mathbb{K}^{[t]}
\end{aligned}
\tag{B4.1}
$$

where we use the notation $A^{-\intercal} = (A^{\intercal})^{-1}$ for any matrix A, and $\sqrt{B}$ denotes the unique squared root of a positive semi-definite matrix B, i.e., $B = \sqrt{B}^{\intercal}\sqrt{B}$. This matrix is obtained through an eigenvalue (or Cholesky) decomposition. In (B4.1), we can rewrite $\mathbb{K}^{[t]}$ as $\mathbb{K}^{[t]} = \mu_{\mathbb{K}}^{[t]} + \varepsilon^{[t]}$, where $\mu_{\mathbb{K}}^{[t]} = \sqrt{\mathbb{J}^{[t]}}\Theta^{[t]}$ represents a vector of 'standardised' model parameters, and $\varepsilon^{[t]} = \sqrt{\mathbb{J}^{[t]}}^{-\intercal}\mathbb{S}^{[t]}$ represents a 'standardised' score vector. Note that, according to standard likelihood theory, $\varepsilon \sim \mathbb{N}(0, \mathbb{I}_K)$, with $K = \dim(\Theta)$. Consequently, $\mathbb{K}$ is a random variable that follows a normal distribution $\mathbb{K} \sim \mathbb{N}(\mu_{\mathbb{K}}, \mathbb{I}_K)$, where the expected value $\mathbb{E}(\mathbb{K})$ is given by $\mu_{\mathbb{K}} = \sqrt{\mathbb{J}}\Theta^*$, with Θ^* representing the true parameter vector.

From (B4.1), at convergence we can write the PMLE as:

$$\hat{\Theta}(\boldsymbol{\lambda}_0) = \hat{\Theta} = \left[\hat{\mathbb{J}} + n\hat{\mathcal{S}}_{\boldsymbol{\lambda}_0}\right]^{-1} \sqrt{\hat{\mathbb{J}}}^{\intercal} \; \hat{\mathbb{K}} \tag{B4.2}$$

and thus

$$\mathbb{E}(\hat{\mathbb{K}}) = \hat{\mu}_{\mathbb{K}} = \sqrt{\hat{\mathbb{J}}} \left[\hat{\mathbb{J}} + n\hat{\mathcal{S}}_{\boldsymbol{\lambda}_0}\right]^{-1} \sqrt{\hat{\mathbb{J}}}^{\intercal} \; \hat{\mathbb{K}} = \hat{\mathbb{A}}_{\boldsymbol{\lambda}_0} \hat{\mathbb{K}}$$

where $\hat{\mathbb{A}}_{\boldsymbol{\lambda}_0} = \sqrt{\hat{\mathbb{J}}} \left[\hat{\mathbb{J}} + n\hat{\mathcal{S}}_{\boldsymbol{\lambda}_0}\right]^{-1} \sqrt{\hat{\mathbb{J}}}^{\intercal}$ is the projection (or influence) matrix of the fitting problem, depending on the tuning parameter vector $\boldsymbol{\lambda}_0$ through $\hat{\mathcal{S}}_{\boldsymbol{\lambda}_0}$.

The goal is to update $\boldsymbol{\lambda}_0$ to a new value such that i) given $\hat{\Theta}(\boldsymbol{\lambda}_0)$, the model complexity, given by the level of sparseness of the factor loading matrices, is guided by the evidence in the observed data, and ii) we ensure that the penalised estimates are as close as possible to the true model parameters. In order to achieve these goals, we seek to minimise the mean squared error (MSE) of the standardised parameters, which is influenced by $\boldsymbol{\lambda}$ through $\hat{\mathbb{A}}_{\boldsymbol{\lambda}}$:

$$\begin{aligned}
\mathbb{E}\left(||\mu_{\mathbb{K}} - \hat{\mu}_{\mathbb{K}}||_2^2\right) &= \mathbb{E}\left(||(\mathbb{K} - \varepsilon) - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2\right) \\
&= \mathbb{E}\left(||(\mathbb{K} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}) - \varepsilon||_2^2\right) \\
&= \mathbb{E}\left(||\mathbb{K} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2 - 2\varepsilon^{\intercal}(\mathbb{K} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}) + \varepsilon^{\intercal}\varepsilon\right) \\
&= \mathbb{E}\left(||\mathbb{K} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2\right) + \mathbb{E}(\varepsilon^{\intercal}\varepsilon) - 2\mathbb{E}\left[\varepsilon^{\intercal}(\mu_{\mathbb{K}} + \varepsilon - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mu}_{\mathbb{K}} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\varepsilon)\right] \\
&= \mathbb{E}\left(||\mathbb{K} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2\right) + 2\mathrm{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}) - K
\end{aligned} \tag{B4.3}$$

where $||\mathbf{x}||_2^2 = \mathbf{x}^{\intercal}\mathbf{x}$ for a column vector $\mathbf{x}$ is the squared Euclidean norm, and the last line on equation (B4.3) results from applying the following identities in Wood (2017, p.53):

$$\begin{aligned}
\mathbb{E}(\varepsilon^{\intercal}\varepsilon) &= \sum_{i=1}^{K} \mathbb{E}(\varepsilon_i^2) = \sum_{i=1}^{K} \mathrm{var}(\varepsilon_i) = K \\
\mathbb{E}(\varepsilon^{\intercal}\mu_{\mathbb{K}}) &= \mathbb{E}(\varepsilon^{\intercal})\mu_{\mathbb{K}} = 0 \\
\mathbb{E}(\varepsilon^{\intercal}\hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mu}_{\mathbb{K}}) &= \mathbb{E}(\varepsilon^{\intercal})\hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mu}_{\mathbb{K}} = 0 \\
\mathbb{E}(\varepsilon^{\intercal}\hat{\mathbb{A}}_{\boldsymbol{\lambda}}\varepsilon) &= \mathbb{E}(\mathrm{tr}(\varepsilon^{\intercal}\hat{\mathbb{A}}_{\boldsymbol{\lambda}}\varepsilon)) = \mathbb{E}(\mathrm{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}\varepsilon\varepsilon^{\intercal})) \\
&= \mathrm{tr}(\mathbb{E}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}\varepsilon\varepsilon^{\intercal})) = \mathrm{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}\mathbb{E}(\varepsilon\varepsilon^{\intercal})) \\
&= \mathrm{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}\mathbb{I}_K) = \mathrm{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}})
\end{aligned}$$

Note that $\mathrm{tr}(\mathbb{A}_{\boldsymbol{\lambda}}) = \mathrm{tr}([\mathbb{J} + n\mathcal{S}_{\boldsymbol{\lambda}}]^{-1}\mathbb{J}) = \mathrm{tr}(\mathcal{H}_p^{-1}\mathcal{H})$, which can be interpreted as the effective

degrees of freedom (edf) of the penalised model. This quantity is also equivalent to the bias term in the GBIC expression given in equation (B3.10). Since Θ^* is unknown, we use an estimate of the mean squared error (MSE) in equation (B4.3):

$$\hat{\mathcal{V}}(\boldsymbol{\lambda}; \hat{\Theta}) = ||\hat{\mathbb{K}} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2 + 2\text{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}) - K \tag{B4.4}$$

which is an (approximate) unbiased risk estimator (UBRE, Wood, 2017, Chapter 6) and an approximate AIC (Appendix B4.2). Thus, for a given PMLE $\hat{\Theta}(\boldsymbol{\lambda}_0) = \hat{\Theta}$, we estimate $\boldsymbol{\lambda}$ as

$$\begin{aligned}\hat{\boldsymbol{\lambda}} &= \underset{\boldsymbol{\lambda}\in[0,\infty)^S}{\arg\min}\, \hat{\mathcal{V}}(\boldsymbol{\lambda}; \hat{\Theta}) \\ &= \underset{\boldsymbol{\lambda}\in[0,\infty)^S}{\arg\min} \left\{ ||\hat{\mathbb{K}} - \hat{\mathbb{A}}_{\boldsymbol{\lambda}}\hat{\mathbb{K}}||_2^2 + 2\text{tr}(\hat{\mathbb{A}}_{\boldsymbol{\lambda}}) - K \right\}\end{aligned} \tag{B4.5}$$

This optimisation problem can be solved by using iterative methods such as (quasi-)Newton or trust-region solvers, as described in Wood (2004, 2017) and Geminiani et al. (2021). Consider a Taylor approximation of $\mathcal{V}(\lambda; \hat{\Theta})$ in (B4.4) about the current value of the tuning parameter vector $\boldsymbol{\lambda}_0$ (we omit the dependence on $\hat{\Theta}$ for notational convenience):

$$\mathcal{V}(\boldsymbol{\lambda}) \approx \mathcal{V}(\boldsymbol{\lambda}_0) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)\dot{\mathcal{V}} + \frac{1}{2}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^\intercal \ddot{\mathcal{V}}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)$$

with $\dot{\mathcal{V}} = \nabla_{\boldsymbol{\lambda}}\mathcal{V}(\boldsymbol{\lambda})$ and $\ddot{\mathcal{V}} = \nabla_{\boldsymbol{\lambda}}\nabla_{\boldsymbol{\lambda}^\intercal}\mathcal{V}(\boldsymbol{\lambda})$. A (local) minimum obtained by solving for $\dot{\mathcal{V}} = \mathbf{0}$. Let $\boldsymbol{\lambda}^{[0]} = \boldsymbol{\lambda}_0$ be the starting value. A NR-type update role for the value of the tuning parameter vector is

$$\boldsymbol{\lambda}^{[r+1]} = \boldsymbol{\lambda}^{[r]} - \ddot{\mathcal{V}}(\boldsymbol{\lambda}^{[r]})^{-1}\dot{\mathcal{V}}(\boldsymbol{\lambda}^{[r]}) \tag{B4.6}$$

which is repeated until convergence. With the updated vector of tuning parameters, say $\hat{\boldsymbol{\lambda}}_1 = \arg\min \mathcal{V}(\boldsymbol{\lambda}; \hat{\Theta}(\boldsymbol{\lambda}_0))$, we continue to update the model parameters to $\hat{\Theta}(\boldsymbol{\lambda}_1)$, and repeat iteratively until convergence of both $\hat{\Theta}(\hat{\boldsymbol{\lambda}}^*)$ and $\hat{\boldsymbol{\lambda}}^*$.

B4.2 Equivalence between the UBRE and the AIC

The Akaike Information Criterion (AIC, Akaike, 1974) is defined as:

$$\text{AIC}(\hat{\Theta}) := -2\ell(\hat{\Theta}; \mathbf{y}) + 2K$$

where $\dim(\hat{\Theta}) = K$ are the number of model parameters. Consider a second-order Taylor expansion of $-2\ell(\hat{\Theta})$ around Θ:

$$\begin{aligned} -2\ell(\hat{\Theta};\mathbf{y}) &\approx -2\ell(\Theta;\mathbf{y}) - 2(\hat{\Theta}-\Theta)^\intercal \nabla_\Theta \ell(\Theta;\mathbf{y}) - (\hat{\Theta}-\Theta)^\intercal \nabla_\Theta \nabla_{\Theta^\intercal} \ell(\Theta;\mathbf{y})(\hat{\Theta}-\Theta) \\ &= -2\ell(\Theta;\mathbf{y}) - 2(\hat{\Theta}-\Theta)^\intercal \mathbb{S} - (\hat{\Theta}-\Theta)^\intercal \mathcal{H}(\hat{\Theta}-\Theta) \end{aligned} \tag{B4.7}$$

As before, let $\mathbb{J} = -\mathcal{H}$. Recall that $\mathbb{K} = \sqrt{\mathbb{J}}\Theta + \sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}$, and note that $\sqrt{\mathbb{J}}$ is symmetric. Thus, we write the second term in equation (B4.7) as:

$$\begin{aligned} (\hat{\Theta}-\Theta)^\intercal \mathbb{S} &= (\hat{\Theta}-\Theta)^\intercal \sqrt{\mathbb{J}}\sqrt{\mathbb{J}}^{-1}\mathbb{S} = \left[\sqrt{\mathbb{J}}(\hat{\Theta}-\Theta)\right]^\intercal \sqrt{\mathbb{J}}^{-1}\mathbb{S} \\ &= (\sqrt{\mathbb{J}}\hat{\Theta} - \sqrt{\mathbb{J}}\Theta)^\intercal \sqrt{\mathbb{J}}^{-1}\mathbb{S} = (\sqrt{\mathbb{J}}\hat{\Theta} - \mathbb{K} + \sqrt{\mathbb{J}}^{-\intercal}\mathbb{S})^\intercal \sqrt{\mathbb{J}}^{-1}\mathbb{S} \\ &= -(\mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta})^\intercal \sqrt{\mathbb{J}}^{-1}\mathbb{S} + \mathbb{S}^\intercal \sqrt{\mathbb{J}}^{-1}\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S} \\ &= -\langle \mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}, \sqrt{\mathbb{J}}^{-1}\mathbb{S}\rangle + ||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 \end{aligned} \tag{B4.8}$$

where $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^\intercal\mathbf{y}$ represents the inner product between two column vectors $\mathbf{x}$ and $\mathbf{y}$. Similarly, after using the identity $||\mathbf{x}||_2^2 \equiv ||-\mathbf{x}||_2^2$ for any column vector $\mathbf{x}$, we can express the last term in (B4.7) as:

$$\begin{aligned} -(\hat{\Theta}-\Theta)^\intercal \mathcal{H}(\hat{\Theta}-\Theta) &= (\hat{\Theta}-\Theta)^\intercal \mathbb{J}(\hat{\Theta}-\Theta) = ||\sqrt{\mathbb{J}}(\hat{\Theta}-\Theta)||_2^2 \\ &= ||\sqrt{\mathbb{J}}\hat{\Theta} - \sqrt{\mathbb{J}}\Theta||_2^2 = ||\sqrt{\mathbb{J}}\hat{\Theta} - \mathbb{K} + \sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 \\ &= ||(\mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}) - \sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 \\ &= ||\mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}||_2^2 + ||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 - 2\langle \mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}, \sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}\rangle \end{aligned} \tag{B4.9}$$

By substituting equations (B4.8) and (B4.9) into equation (B4.7), we obtain:

$$\begin{aligned} -2\ell(\hat{\Theta}) \approx &-2\ell(\Theta) + 2\langle \mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}, \sqrt{\mathbb{J}}^{-1}\mathbb{S}\rangle - 2||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 \\ &+ ||\mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}||_2^2 + ||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 - 2\langle \mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}, \sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}\rangle \\ = &-2\ell(\Theta) - ||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 + ||\mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}||_2^2 \end{aligned}$$

thus, it follows that

$$\begin{aligned} \text{AIC}(\hat{\Theta}) &= -2\ell(\hat{\Theta}) + 2K \approx -2\ell(\Theta) - ||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 + ||\mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}||_2^2 + 2K \\ &= -2\ell(\Theta) - ||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 + ||\mathbb{K} - \sqrt{\mathbb{J}}\hat{\Theta}||_2^2 + 2\text{tr}(\mathbb{A}_{\boldsymbol{\lambda}}) \end{aligned}$$

$$= -2\ell(\Theta) - ||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2 + ||\mathbb{K} - \mathbb{A}_{\boldsymbol{\lambda}}\mathbb{K}||_2^2 + 2\text{tr}(\mathbb{A}_{\boldsymbol{\lambda}}) \tag{B4.10}$$

where $K = \text{tr}(\mathbb{A}_{\boldsymbol{\lambda}})$ denotes the number of estimated parameters in the model (i.e., the effective degrees of freedom). Since we want to minimise the approximate AIC in (B4.10) with respect to the tuning parameter vector $\boldsymbol{\lambda}$, we can ignore the terms that are not affected by it, $-2\ell(\Theta)$ and $-||\sqrt{\mathbb{J}}^{-\intercal}\mathbb{S}||_2^2$, and therefore the approximate AIC in (B4.10) becomes proportional to the (estimated) UBRE criterion in equation (B4.4):

$$\text{AIC}(\hat{\Theta}) \approx ||\mathbb{K} - \mathbb{A}_{\boldsymbol{\lambda}}\mathbb{K}||_2^2 + 2\text{tr}(\mathbb{A}_{\boldsymbol{\lambda}}) \propto \mathcal{V}(\lambda; \hat{\Theta})$$

In this sense, $||\mathbb{K} - \mathbb{A}_{\boldsymbol{\lambda}}\mathbb{K}||_2^2$ is a quadratic approximation of $-2\ell(\hat{\Theta})$ and $\text{tr}(\mathbb{A}_{\boldsymbol{\lambda}})$ represents the effective degrees of freedom of the GLVM-LSS model.

B4.3 Computational details

In order to minimise $\mathcal{V}$, the approximate UBRE, there are some computational considerations to take into account. In this section, we provide the computational details that assist in the automatic estimation of the tuning parameter vector. The notation used in this section is consistent with the notation used in previous sections.

To simplify the evaluation of $\text{tr}(\mathbb{A}_{\boldsymbol{\lambda}}) = \text{tr}(\sqrt{\mathbb{J}}[\mathbb{J} + n\mathcal{S}_{\boldsymbol{\lambda}}]^{-1}\sqrt{\mathbb{J}}^{\intercal})$ in (B4.4), we can perform a QR decomposition of $\sqrt{\mathbb{J}}$, $\sqrt{\mathbb{J}} = \mathbf{QR}$, where $\mathbf{Q}$ is an matrix with orthogonal column vectors and $\mathbf{R}$ is an upper triangular matrix. For numerical stability, a pivoted QR decomposition can be used (Wood, 2017). We define $\mathbf{B}$ as the square root of $n\mathcal{S}\boldsymbol{\lambda}$, such that $\mathbf{B}^{\intercal}\mathbf{B} = n\mathcal{S}_{\boldsymbol{\lambda}}$. The matrix $\mathbf{B}$ can be obtained through an eigenvalue or Cholesky decomposition. To address potential rank deficiency in the fitting problem, we stack the column matrices $\mathbf{R}$ and $\mathbf{B}$ and perform a singular value decomposition (SVD):

$$\begin{bmatrix} \mathbf{R} \\ \mathbf{B} \end{bmatrix} = \mathbf{UDV}^{\intercal},$$

Matrix $\mathbf{U}$ in the SVD has columns that are orthogonal, $\mathbf{D}$ is a diagonal matrix containing the singular values, and $\mathbf{V}$ is an orthogonal matrix. To address rank deficiency, we remove from $\mathbf{D}$ the rows and columns corresponding to singular values that are considered "too small" relative to the largest singular value. Specifically, we remove the rows and columns associated with singular values that are less than the largest singular value multiplied by the square root of the machine precision (approximately 1.5e-8).

Let $\mathbf{U}_1$ and $\mathbf{U}_2$ be sub-matrices of $\mathbf{U}$, i.e., $\mathbf{U}^\intercal = [\mathbf{U}_1^\intercal \ \mathbf{U}_2^\intercal]$, such that $\mathbf{R} = \mathbf{U}_1\mathbf{D}\mathbf{V}^\intercal$ and $\mathbf{B} = \mathbf{U}_2\mathbf{D}\mathbf{V}^\intercal$. This means that $\sqrt{\mathbb{J}} = \mathbf{Q}\mathbf{U}_1\mathbf{D}\mathbf{V}^\intercal$, and thus $\mathbb{J} + n\mathcal{S}_{\boldsymbol{\lambda}} = \mathbf{V}\mathbf{D}\mathbf{U}_1^\intercal\mathbf{Q}^\intercal\mathbf{Q}\mathbf{U}_1\mathbf{D}\mathbf{V}^\intercal + \mathbf{V}\mathbf{D}\mathbf{U}_2^\intercal\mathbf{U}_2\mathbf{D}\mathbf{V}^\intercal = \mathbf{V}\mathbf{D}^2\mathbf{V}^\intercal$, since $\mathbf{U}_\mathbf{1}^\intercal\mathbf{U}_\mathbf{1} + \mathbf{U}_\mathbf{2}^\intercal\mathbf{U}_\mathbf{2} = \mathbf{U}^\intercal\mathbf{U} = \mathbb{I}_\mathbf{K}$. With this result, we have that

$$\begin{aligned}
\mathbb{A}_{\boldsymbol{\lambda}} &= \sqrt{\mathbb{J}}[\mathbb{J} + n\mathcal{S}_{\boldsymbol{\lambda}}]^{-1}\sqrt{\mathbb{J}}^\intercal \\
&= \mathbf{Q}\mathbf{U}_1\mathbf{D}\mathbf{V}^\intercal\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\intercal\mathbf{V}\mathbf{D}\mathbf{U}_1^\intercal\mathbf{Q}^\intercal \\
&= \mathbf{Q}\mathbf{U}_1\mathbf{U}_1^\intercal\mathbf{Q}^\intercal
\end{aligned}$$

and thus, $\text{tr}(\mathbb{A}_{\boldsymbol{\lambda}}) = \text{tr}(\mathbf{Q}\mathbf{U}_1\mathbf{U}_1^\intercal\mathbf{Q}^\intercal) = \text{tr}(\mathbf{U}_1\mathbf{U}_1^\intercal\mathbf{Q}^\intercal\mathbf{Q}) = \text{tr}(\mathbf{U}_1\mathbf{U}_1^\intercal)$, which is relatively easy to compute for new values of $\boldsymbol{\lambda}$.

Second, to compute $\hat{\boldsymbol{\lambda}}$ in equation (B4.6), we need to evaluate the vector of first-order derivatives $\dot{\mathcal{V}}$ and the matrix of second-order derivatives $\ddot{\mathcal{V}}$. We derive expressions that are useful for computing these quantities. Write $\mathbb{A}_{\boldsymbol{\lambda}} = \sqrt{\mathbb{J}}\mathbf{G}^{-1}\sqrt{\mathbb{J}}^\intercal$, where $\mathbf{G}^{-1} = [\mathbb{J} + n\mathcal{S}_{\boldsymbol{\lambda}}]^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\intercal$. Note that $\mathbf{G}^{-1}$ is the only matrix that depends on $\boldsymbol{\lambda}$ through $\mathcal{S}_{\boldsymbol{\lambda}}$. To avoid imposing restrictions on the optimisation problem and to ensure that the tuning parameters remain non-negative, we introduce a transformation by defining $\rho_i = \log(\lambda_i)$, where λ_i is the i-th component of the vector $\boldsymbol{\lambda}$. With this transformation, the optimisation problem in equation (B4.5) becomes:

$$\frac{\partial \mathbf{G}^{-1}}{\partial \rho_i} = -\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_i}\mathbf{G}^{-1} = -\lambda_i \cdot n \cdot \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\intercal\mathcal{S}_{\lambda_i}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\intercal$$

where the matrix $\mathcal{S}_{\lambda_i}$ is a block diagonal matrix similar to the one in (3.3.2), but with zero entries along the diagonal except for the entries involving λ_i. If there is only one tuning parameter, $\mathcal{S}_{\lambda_i}$ is equal to the penalty matrix in equation (3.3.2). Denote $\tilde{\lambda}_i = \lambda_i \cdot n$. The first-order derivatives of the influence matrix with respect to ρ_i are:

$$\begin{aligned}
\frac{\partial \mathbb{A}_{\boldsymbol{\lambda}}}{\partial \rho_i} = \sqrt{\mathbb{J}}\frac{\partial \mathbf{G}^{-1}}{\partial \rho_i}\sqrt{\mathbb{J}}^\intercal &= -\tilde{\lambda}_i \cdot \mathbf{Q}\mathbf{U}_1\mathbf{D}\mathbf{V}^\intercal\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\intercal\mathcal{S}_{\lambda_i}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\intercal\mathbf{V}\mathbf{D}\mathbf{U}_1^\intercal\mathbf{Q}^\intercal \\
&= -\tilde{\lambda}_i \cdot \mathbf{Q}\mathbf{U}_1\mathbf{D}^{-1}\mathbf{V}^\intercal\mathcal{S}_{\lambda_i}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}_1^\intercal\mathbf{Q}^\intercal
\end{aligned} \tag{B4.11}$$

The second-order derivatives are:

$$\begin{aligned}
\frac{\partial^2 \mathbf{G}^{-1}}{\partial \rho_i \partial \rho_j} &= -\left(\frac{\partial \mathbf{G}^{-1}}{\partial \rho_j}\right)\frac{\partial \mathbf{G}}{\partial \rho_i}\mathbf{G}^{-1} - \mathbf{G}^{-1}\frac{\partial^2 \mathbf{G}}{\partial \rho_i \partial \rho_j}\mathbf{G}^{-1} - \mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_i}\left(\frac{\partial \mathbf{G}^{-1}}{\partial \rho_j}\right) \\
&= \mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_j}\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_i}\mathbf{G}^{-1} - \mathbf{G}^{-1}\frac{\partial^2 \mathbf{G}}{\partial \rho_i \partial \rho_j}\mathbf{G}^{-1} + \mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_i}\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_j}\mathbf{G}^{-1}
\end{aligned}$$

$$= \mathbf{G}^{-1}\left(\frac{\partial \mathbf{G}}{\partial \rho_j}\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_i} - \frac{\partial^2 \mathbf{G}}{\partial \rho_i \partial \rho_j} + \frac{\partial \mathbf{G}}{\partial \rho_i}\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_j}\right)\mathbf{G}^{-1}$$

and thus, after some matrix algebra, for $i \neq j$:

$$\begin{aligned}
\frac{\partial^2 \mathbb{A}_{\boldsymbol{\lambda}}}{\partial \rho_i \partial \rho_j} &= \sqrt{\mathbb{J}}\frac{\partial^2 \mathbf{G}^{-1}}{\partial \rho_i \partial \rho_j}\sqrt{\mathbb{J}}^{\intercal} \\
&= \sqrt{\mathbb{J}}\mathbf{G}^{-1}\left(\frac{\partial \mathbf{G}}{\partial \rho_j}\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_i} - \frac{\partial^2 \mathbf{G}}{\partial \rho_i \partial \rho_j} + \frac{\partial \mathbf{G}}{\partial \rho_i}\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \rho_j}\right)\mathbf{G}^{-1}\sqrt{\mathbb{J}}^{\intercal} \\
&= \mathbf{Q}\mathbf{U}_1\mathbf{D}\mathbf{V}^{\intercal}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\intercal}\left(\lambda_j\lambda_i\mathcal{S}_{\lambda_j}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\intercal}\mathcal{S}_{\lambda_i} + \lambda_i\lambda_j\mathcal{S}_{\lambda_i}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\intercal}\mathcal{S}_{\lambda_j}\right)\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\intercal}\mathbf{V}\mathbf{D}\mathbf{U}_1^{\intercal}\mathbf{Q}^{\intercal} \\
&= \tilde{\lambda}_i\tilde{\lambda}_j \cdot \mathbf{Q}\mathbf{U}_1\mathbf{D}^{-1}\mathbf{V}^{\intercal}\left(\mathcal{S}_{\lambda_j}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\intercal}\mathcal{S}_{\lambda_i}\right)^{\ddagger}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}_1^{\intercal}\mathbf{Q}^{\intercal}
\end{aligned}$$

where $(B)^{\ddagger} = B + B^{\intercal}$ for a matrix B and, if $i = j$:

$$\begin{aligned}
\frac{\partial^2 \mathbb{A}_{\boldsymbol{\lambda}}}{(\partial \rho_i)^2} &= \tilde{\lambda}_i^2 \cdot \mathbf{Q}\mathbf{U}_1\mathbf{D}^{-1}\mathbf{V}^{\intercal}\left(\mathcal{S}_{\lambda_i}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\intercal}\mathcal{S}_{\lambda_i}\right)^{\ddagger}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}_1^{\intercal}\mathbf{Q}^{\intercal} - \lambda_i \cdot \mathbf{Q}\mathbf{U}_1\mathbf{D}^{-1}\mathbf{V}^{\intercal}\mathcal{S}_{\lambda_i}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}_1^{\intercal}\mathbf{Q}^{\intercal} \\
&= \tilde{\lambda}_i^2 \cdot \mathbf{Q}\mathbf{U}_1\mathbf{D}^{-1}\mathbf{V}^{\intercal}\left(\mathcal{S}_{\lambda_i}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\intercal}\mathcal{S}_{\lambda_i}\right)^{\ddagger}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}_1^{\intercal}\mathbf{Q}^{\intercal} + \frac{\partial \mathbb{A}_{\boldsymbol{\lambda}}}{\partial \rho_i}
\end{aligned}$$

These results can be further expressed using an indicator function $\mathbb{1}(i = j)$ that takes the value of 1 if $i = j$ and 0 otherwise, as

$$\frac{\partial^2 \mathbb{A}_{\boldsymbol{\lambda}}}{\partial \rho_i \partial \rho_j} = \tilde{\lambda}_i\tilde{\lambda}_j \cdot \mathbf{Q}\mathbf{U}_1\mathbf{D}^{-1}\mathbf{V}^{\intercal}\left(\mathcal{S}_{\lambda_j}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\intercal}\mathcal{S}_{\lambda_i}\right)^{\ddagger}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}_1^{\intercal}\mathbf{Q}^{\intercal} + \mathbb{1}(i = j) \cdot \frac{\partial \mathbb{A}_{\boldsymbol{\lambda}}}{\partial \rho_i} \tag{B4.12}$$

Given the expressions in (B4.11) and (B4.12), we can now compute the expressions for $\dot{\mathcal{V}}$ and $\ddot{\mathcal{V}}$. We simplify the notation by writing $\mathcal{B} = ||\mathbb{K} - \mathbb{A}_{\boldsymbol{\lambda}}\mathbb{K}||_2^2$, and by defining the expressions $\mathcal{K}_1 = \mathbf{U}_1^{\intercal}\mathbf{Q}^{\intercal}\mathbb{K}$, $\mathcal{Z}_{\lambda_i} = \mathbf{D}^{-1}\mathbf{V}^{\intercal}\mathcal{S}_{\lambda_i}\mathbf{V}\mathbf{D}^{-1}$ and $\mathcal{C}_{\lambda_i} = \mathcal{Z}_{\lambda_i}\mathbf{U}_1^{\intercal}\mathbf{U}_1$. Using the trace properties $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, $\mathrm{tr}(A) = \mathrm{tr}(A^{\intercal})$, and $\partial\mathrm{tr}(A) = \mathrm{tr}(\partial A)$, as well as the fact that for a column vector $\mathbf{a}$, $\partial||\mathbf{a}||_2^2 = 2\mathbf{a}$, we have that:

$$\begin{aligned}
\frac{\partial \mathrm{tr}(\mathbb{A}_{\boldsymbol{\lambda}})}{\partial \rho_i} &= \mathrm{tr}\left(\frac{\partial \mathbb{A}_{\boldsymbol{\lambda}}}{\partial \rho_i}\right) = -\tilde{\lambda}_i\mathrm{tr}(\mathcal{C}_{\lambda_i}), \\
\frac{\partial^2 \mathrm{tr}(\mathbb{A}_{\boldsymbol{\lambda}})}{\partial \rho_i \partial \rho_j} &= \mathrm{tr}\left(\frac{\partial^2 \mathbb{A}_{\boldsymbol{\lambda}}}{\partial \rho_i \partial \rho_j}\right) = 2\tilde{\lambda}_i\tilde{\lambda}_j \cdot \mathrm{tr}(\mathcal{Z}_{\lambda_j}\mathcal{C}_{\lambda_i}) - \tilde{\lambda}_i \cdot \mathbb{1}(i = j) \cdot \mathrm{tr}(\mathcal{C}_{\lambda_i}), \\
\frac{\partial \mathcal{B}}{\partial \rho_i} &= 2\tilde{\lambda}_i \cdot \left(\mathcal{K}_1^{\intercal}\mathcal{Z}_{\lambda_i}\mathcal{K}_1 - \mathcal{K}_1^{\intercal}\mathcal{C}_{\lambda_i}\mathcal{K}_1\right), \\
\frac{\partial^2 \mathcal{B}}{\partial \rho_i \partial \rho_j} &= 2\tilde{\lambda}_i\tilde{\lambda}_j \cdot \mathcal{K}_1^{\intercal}\left(\mathcal{Z}_{\lambda_i}\mathcal{C}_{\lambda_j} + \mathcal{Z}_{\lambda_j}\mathcal{C}_{\lambda_i} - \mathcal{Z}_{\lambda_i}\mathcal{Z}_{\lambda_j} - \mathcal{Z}_{\lambda_j}\mathcal{Z}_{\lambda_i} + \mathcal{C}_{\lambda_i}\mathcal{Z}_{\lambda_j}\right)\mathcal{K}_1 + \mathbb{1}(i = j) \cdot \frac{\partial \mathcal{B}}{\partial \rho_i}
\end{aligned}$$

These derivatives are part of the desired quantities $\dot{\mathcal{V}}$ and $\ddot{\mathcal{V}}$, as the i^{th} component of the first-order derivative vector and the $[i,j]^{\text{th}}$ entry of the matrix of second-order derivatives which are used in the NR update step in equation (B4.6) are

$$\dot{\mathcal{V}}_{[i]} = \frac{\partial \mathcal{V}(\lambda)}{\partial \rho_i} = \frac{\partial \mathcal{B}}{\partial \rho_i} + 2 \cdot \frac{\partial \text{tr}(\mathbb{A}_{\boldsymbol{\lambda}})}{\partial \rho_i} \tag{B4.13}$$

$$\ddot{\mathcal{V}}_{[i,j]} = \frac{\partial^2 \mathcal{V}(\lambda)}{\partial \rho_i \partial \rho_j} = \frac{\partial^2 \mathcal{B}}{\partial \rho_i \partial \rho_j} + 2 \cdot \frac{\partial^2 \text{tr}(\mathbb{A}_{\boldsymbol{\lambda}})}{\partial \rho_i \partial \rho_j} \tag{B4.14}$$

B5. Simulation Studies

B5.1 Parameter initialisation

We propose a warm-start strategy for the initial values of the factor loadings in the EM-step of the estimation procedure involves the following steps:

1. Perform a Principal Component Analysis (PCA) on the matrix of observed items. Retain the first q component scores, where q is the number of factors, in a $n \times q$ matrix denoted by $\mathbf{Z}_{\text{PCA}}$.

2. For each item $i = 1, \ldots, p$, use $\mathbf{Z}_{\text{PCA}}$ as a design matrix of observed covariates in a GAMLSS regression (Rigby and Stasinopoulos, 2005). Estimate the intercepts and slopes using the `gamlss` function in the `R` package of the same name. These estimated parameters serve as the initial values for the factor loadings in two-step iterative estimation of the GLVM-LSS model.

For continuous items, the estimated initial values obtained from the GAMLSS regression typically provide good starting points, and the algorithm converges in a reasonable number of steps. However, for count and categorical items, the initial values obtained through PCA and GAMLSS may be further away from the population parameters, resulting in a longer convergence time. Nevertheless, this is not a concern as the EM-algorithm is robust to initial points that are far from the mode. A second alternative in the proposed PMML estimation framework is to set the initial values to the rotated MLEs. In practice we recommend using an L_p-rotation (Liu et al., 2023), but any rotation would do just fine. Lastly, random initial values can be used, particularly when testing for multiple local maxima.

B5.2 Generating Sparse Factor Loading Matrices

We follow the procedure described in each example of Section 3.6 to generate initial dense factor loading matrices for the location, scale, and shape parameters. After generating these matrices, we impose adequate identifiability restrictions on the model parameters $\Theta^\intercal = (\boldsymbol{\alpha}_0^\intercal, \text{vec}(\text{A})^\intercal, \text{vech}(\boldsymbol{\Phi})^\intercal)$.

If we assume independent latent variables, we impose a recursive restriction (type b, described in Section 1.2), and if we assume correlated latent variables, we impose errors-in-variables restrictions on the factor loadings (type d, described in Section 1.2). In both cases, identification restrictions on the stacked factor loading matrix $\text{A}^\intercal = [\text{A}_1^\intercal, \text{A}_2^\intercal]$ are imposed on the $q \times q$ squared matrix A_1. These restrictions ensure that the model parameters are identifiable and that we can estimate them accurately.

We introduce sparsity in the factor loading matrices column-by-column, following the procedure described below:

1. Start with the factor loading matrix corresponding to the location parameter, A_μ, and impose identification restrictions on the first q rows and columns (which form the matrix A_1).
2. From the remaining $a = (p - q)$ rows, select at random $b = \lfloor (p - q)/2 \rfloor$.
3. Set the factor loading for the first latent variable ($\alpha_{i1,\mu}$) to zero for the selected b items. This ensures that these items do not load on the first latent variable.
4. For the second latent variable, select the remaining $a - b$ items that have non-zero loadings in $\alpha_{i1,\mu}$.
5. With a probability of 0.5, assign a value of zero to $\alpha_{i2,\mu}$ for the selected items. . This ensures that these items do not load on the second latent variable (but do load on the first latent variable).
6. Repeat this process for the remaining columns of A_μ if $q > 2$.

By applying this procedure, we ensure that each item loads on at least one latent variable in the measurement equation for the location parameter (μ). The sparsity is introduced column by column, randomly selecting a subset of items and then setting their factor loadings to zero for that specific column (latent factor). Intercepts are not set to zero unless explicitly required.

If the conditional distribution $f_i(y_i|\mathbf{z})$ also has scale or shape parameters, a similar procedure can be applied to introduce sparsity in the corresponding factor loading matrices A_σ, A_ν, and/or A_τ. The procedure is slightly modified and continues as follows:

1. For the factor loading matrix A_σ (or A_ν, A_τ), we do not have to impose identification restrictions and thus we can select from the available p items.

2. To ensure that some items have constant scale (or shape) parameters (i.e., not depending on the latent factors), we randomly select items for which $\alpha_{i1,\sigma} = 0$. These items will have a fixed loading of zero for the first factor.

3. For the remaining factors ($j = 2, ..., q$), we set the factor loadings $\alpha_{ij,\sigma}$ to zero for the selected items. This introduces sparsity by randomly setting the loadings to zero for certain factors and items.

We introduce sparsity in the factor loading matrices for scale or shape parameters while also ensuring that some items have constant scale (or shape) parameters that do not depend on the latent factors. This is useful, for example, when we want to model homoscedastic items in the Gaussian case, where the scale parameter does not vary with the latent factors.

B6. Software Implementation

In this Appendix, we present the implementation of the penalised estimation with automatic selection of tuning parameter vector for the GLVM-LSS framework. The code is written in the statistical software `R` to contribute to reproducible research practices and transparent dissemination of results. We discuss the penalised estimation of the GLVM-LSS in the PISA 2018 empirical application (Section 3.7.1). We refer to Appendix A6 for data preparation and the MLE estimation.

Estimation

The penalised estimation of the GLVM-LSS also makes use of the function `glvmlss`. The selected model model in Section 3.7.1, with additional influence factor $\gamma = 5$, is estimated using the following the code:

```
mod_SN_comp.PENg5 <- glvmlss(data = data, family = famSN,
                        mu.eq = ~ Z1+Z2, sg.eq = ~ Z1+Z2, nu.eq = ~ Z1+Z2,
                        f.scores = T,
                        verbose = T, iden.res = iRes, corr.lv = T,
                        solver = "nlminb", EM_use2d = F, EM_iter = 5000,
                        iter.lim = 1000,
                        est.ci = "Approximate",
                        start.val = mod_SN_comp$b,
```

```
                        Rz = mod_SN_comp$Rz,
                        penalty = "alasso",
                        w.alasso = mod_SN_comp$b,
                        lambda = "auto",
                        gamma = 5)

names(mod_SN_comp)
```

Apart from the main components of the `glvmlss` function, we include starting values for the model parameters (set to be the MLE estimates obtained in Appendix A6), given by the argument `start.val = mod_SN_comp$b`; and starting values for the latent variables covariance matrix, `Rz = mod_SN_comp$Rz`. The functional form of the penalty term is given by the argument `penalty = "alasso"`, with Alasso weights given by `w.alasso = mod_SN_comp$b`, i.e., the MLEs. The argument `lambda` defines the value of the tuning parameter vector. If the user decides to use fixed values (as opposed to the automatic selection procedure described in this dissertation), it should correspond to a vector with as many entries as location, scale, or shape parameters in the model. If a scalar is provided, it is repeated to match the latter. Finally, the argument `gamma` fixes the value of the influence factor. Higher values in this argument imply higher penalty on the model parameter (more sparsity).

Additional control variables in the penalised framework can be included directly, or in a list under the name `control`. The following options, along with their default values, are available for the user:

```
control <- list(
    lamnda = NULL,       # Tuning parameters. Option "auto" or c(v1,v2,...,vD)
    penalty = "none",    # Options: "ridge", "lasso", "alasso", "scad", "mcp"
    w.alasso = NULL,     # A list of factor loadings
    a = NULL,            # Additional parameter a in the Alasso, SCAD, and MCP
    gamma = NULL,        # Influence factor.
    tolb = 1e-4)         # Tolerance level for factor loadings = 0.
```

Bibliography

Abelson, R. P., Kinder, D. R., Peters, M. D., and Fiske, S. T. (1982). Affective and Semantic Components in Political Person Perception. *Journal of Personality and Social Psychology*, 42(4):619–630.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Akaike, H. (1987). Factor Analysis and AIC. *Psychometrika*, 52(3):317–322.

Ali, M. M. (1980). Characterization of the Normal Distribution among the Continuous Symmetric Spherical Class. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):162–164.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In Neyman, J., editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1955*, volume V, pages 111–150. University of California Press.

Asparouhov, T. and Muthén, B. (2016). Structural Equation Models and Mixture Models With Continuous Nonnormal Skewed Distributions. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(4):1–19.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178.

Azzalini, A. (2005). The Skew-normal Distribution and Related Multivariate Families. *Scandinavian Journal of Statistics*, 35(2):159–188.

Azzalini, A. (2013). *The Skew-Normal and Related Families*. Institute of Mathematical Statistics monographs, 3. Cambridge, UK: Cambridge University Press, 1st edition.

Azzalini, A. and Capitanio, A. (1999). Statistical Applications of the Multivariate Skew Normal Distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 61(3):579–602.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and asymptotics*, volume 52 of *Monographs on Statistics and Applied Probability*. Boca Ratón, FL, US: Chapman & Hall / CRC, 1st edition.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(3):293–321.

Bartholomew, D. J. (1984). The foundations of factor analysis. *Biometrika*, 71(2):221–232.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley Series in Probability and Statistics. New York, NY, US: John Wiley & Sons, Ltd, 3rd edition.

Barton, M. A. and Lord, F. M. (1981). An Upper Asymptote for the Three-Parameter Logistic Item Response Model. *ETS Research Report Series*, June(1).

Battauz, M. (2020). Regularized Estimation of the Nominal Response Model. *Multivariate Behavioral Research*, 55(6):811–824.

Bauer, D. J., Belzak, W. C. M., and Cole, V. T. (2020). Simplifying the Assessment of Measurement Invariance over Multiple Background Variables: Using Regularized Moderated Nonlinear Factor Analysis to Detect Differential Item Functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1):43–55.

Bazán, J. L., Branco, M. D., and Bolfarine, H. (2006). A Skew Item Response Model. *Bayesian Analysis*, 1(4):861–892.

Bazán, J. L., Branco, M. D., and Bolfarine, H. (2014). Extensions of the skew-normal ogive item response model. *Bayesian Analysis*, 28(1):1–23.

Bekker, P. A., Merckens, A., and Wansbeek, T. J. (1994). *Identification, Equivalent Models, and Computer Algebra*. Statistical Modeling and Decision Science. San Diego, CA, US: Academic Press.

Bentler, P. M. and Weeks, D. G. (1980). Linear Structural Equations with Latent Variables. *Psychometrika*, 45(3):289–308.

Bianconcini, S. (2014). Asymptotic properties of adaptive maximum likelihood estimators in latent variable models. *Bernoulli*, 20(3):1507–1531.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R., editors, *Statistical Theories of Mental Test Scores*, pages 397–479. Reading, MA, US: Addison-Wesley, 1st edition.

Bock, R. D. and Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459.

Bolfarine, H. and Bazán, J. L. (2010). Bayesian Estimation of the Logistic Positive Exponent IRT Model. *Journal of Educational and Behavioral Statistics*, 35(6):693–713.

Bollen, K. (1989). *Structural Equations with Latent Variables*. New York, NY, US: John Wiley & Sons, Ltd., 1st edition.

Bollen, K. A. (1996). A limited-information estimator for lisrel models with or without heteroscedastic errors. In Marcoulides, G. A. and Schumacker, R. E., editors, *Advanced Structural Equation Modeling: Issues and Techniques*, pages 227–241. Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.

Bolsinova, M. and Molenaar, D. (2018). Modeling Nonlinear Conditional Dependence Between Response Time and Accuracy. *Frontiers in Psychology*, 9(1525):1–12.

Bolsinova, M., Tijmstra, J., and Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 70(2):257–279.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. Methodology in the Social Sciences Series. New York, NY, US: Guilford Press, 2nd edition.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37(1):62–83.

Browne, M. W. (2001). An Overview of Analytic Rotation in Exploratory Factor Analysis. *Multivariate Behavioral Research*, 36(1):111–150.

Broyden, C. G., Dennis Jr, J. E., and Moré, J. J. (1973). On the local and superlinear convergence of quasi-newton methods. *IMA Journal of Applied Mathematics*, 12(3):223–245.

Cagnone, S., Moustaki, I., and Vasdekis, V. (2009). Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology*, 62(2):401–415.

Cai, L. (2010). High-dimensional Exploratory Item Factor Analysis by A Metropolis–Hastings Robbins–Monro Algorithm. *Psychometrika*, 75(1):33–57.

Chang, H.-H. (1996). The Asymptotic Posterior Normality of the Latent Trait for Polytomous IRT Models. *Psychometrika*, 61(3):445–463.

Chang, H.-H. and Stout, W. (1993). The Asymptotic Posterior Normality of the Latent Trait in an IRT Model. *Psychometrika*, 58(1):37–52.

Chen, H. (1991). Estimation of a Projection-Pursuit Type Model. *The Annals of Statistics*, 19(1):142–157.

Chen, L., Dolado, J. J., and Gonzalo, J. (2021). Quantile Factor Models. *Econometrica*, 89(2):875–910.

Chen, Y., Li, X., Liu, J., and Ying, Z. (in press). Item Response Theory: A Statistical Framework for Educational and Psychological Measurement. *Statistical Science*.

Chen, Y., Li, X., and Zhang, S. (2019). Joint Maximum Likelihood Estimation for High-Dimensional Exploratory Item Factor Analysis. *Psychometrika*, 84(1):124–146.

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical Analysis of Q-Matrix Based Diagnostic Classification Models. *Journal of the American Statistical Association*, 110(510):850–866.

Chen, Y., Moustaki, I., and Zhang, H. (2020). A Note on Likelihood Ratio Tests for Models with Latent Variables. *Psychometrika*, 85(4):996–1012.

Chen, Y., Moustaki, I., and Zhang, S. (2023). On the estimation of structural equation models with latent variables. In Hoyle, R. H., editor, *Handbook of Structural Equation Models*, chapter 8, pages 145–162. Mahwah, NJ, US: Erlbaum.

Chen, Y. and Zhang, S. (2021). Estimation Methods for Item Factor Analysis: An Overview. In Zhao, Y. and Chen, D. D.-G., editors, *Modern Statistical Methods for Health Research*, pages 329–350. New York, NY, US: Springer-Verlag.

Chiogna, M. (2005). A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. *Statistical Methods & Applications*, 14:331–341.

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

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. New York, NY, US: Springer-Verlag, 1st edition.

Davis, P. J. and Rabinowitz, P. (1975). *Methods of Numerical Integration*. New York, NY, US: Dover, 1st edition.

Davison, A. C. (2003). *Statistical Models*. Number 11 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press, 1st edition.

De Boeck, P. and Jeon, M. (2019). An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, 10(102):1–11.

de Boor, C. (2001). *A Practical Guide to Splines (Revised Edition)*, volume 27 of *Springer Series in Applied Mathematical Sciences*. New York, NY, US: Springer-Verlag, 2nd edition.

Deary, I. J., Egan, V., Gibson, G., Austin, E. J., Brand, C. R., and Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence*, 23(2):105–132.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38.

Detterman, D. K. and Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, 13(4):349–359.

Di Ciccio, T. J. and Monti, A. C. (2011). Inferential aspects of the skew t-distribution. *Quaderni di Statistica*, 13:1–21.

Dolan, C. V., van der Maas, H. L. J., and Molenaar, P. C. (2002). A framework for ML estimation of parameters of (mixtures of) common reaction time distributions given optional truncation or censoring. *Behavior Research Methods*, 34(3):304.323.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, 62(1):7–28.

Douglas, J. and Cohen, A. (2001). Nonparametric Item Response Function Estimation for Assessing Parametric Model Fit. *Applied Psychological Measurement*, 25(3):234–243.

Eberl, A. and Klar, B. (2020). Asymptotic distributions and performance of empirical skewness measures. *Computational Statistics & Data Analysis*, 146:1–18.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2):407–499.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89–121.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2021). *Regression: Models, methods and applications*. Berlin, Germany: Springer-Verlag, 2nd edition.

Falk, C. F. and Cai, L. (2016a). Maximum Marginal Likelihood Estimation of a Monotonic Polynomial Generalized Partial Credit Model with Applications to Multiple Group Analysis. *Psychometrika*, 81(2):434–460.

Falk, C. F. and Cai, L. (2016b). Semiparametric Item Response Functions in the Context of Guessing. *Journal of Educational Measurement*, 53(2):229–247.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, volume 2. New York, NY, US: John Wiley & Sons, Ltd.

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

Finch, W. H. (2015). Modeling Nonlinear Structural Equation Models: A Comparison of the Two-Stage Generalized Additive Models and the Finite Mixture Structural Equation Model. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1):60–75.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise Coordinate Optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1).

Friedman, J. H. and Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association*, 76(376):817–823.

Fuller, W. A. (1987). *Measurement Error Models*. Wiley Series in Probability and Statistics. New York, NY, US: John Wiley & Sons, Ltd, 1st edition.

Geminiani, E., Marra, G., and Moustaki, I. (2021). Single- and Multiple-Group Penalized Factor Analysis: A Trust-Region Algorithm Approach with Integrated Automatic Multiple Tuning Parameter Selection. *Psychometrika*, 86(1):65–95.

Geyer, C. J. (2020). *`trust`: Trust Region Optimization*. R package version 0.1-8.

Gu, C. (1992). Cross-Validating Non-Gaussian Data. *Journal of Computational and Graphical Statistics*, 1(2):169–179.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Procedures of the National Academy of Sciences*, 95(13):7270–7274.

Guth, J. L. (2019). Are White Evangelicals Populists? The View from the 2016 American National Election Study. *The Review of Faith and International Affairs*, 17(3):20–35.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2010). *Multivariate Data Analysis*. Upper Saddle River, NJ, US: Prentice Hall, 7th edition.

Hall, D. (2000). Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*, 56(4):1030–1039.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, 21(1):157–178.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, US: Springer-Verlag, 2nd edition.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability (43). Boca Ratón, FL, US: Chapman & Hall / CRC.

Hedecker, D., Berbaum, M., and Mermelstein, R. (2006). Location-Scale Models for Multilevel Ordinal Data: Between- and Within-Subjects Variance Modeling. *Journal of Probability and Statistical Science*, 4(1):1–20.

Hedecker, D., Mermelstein, R., and Demirtas, H. (2008). An Application of a Mixed-Effects Location Scale Model for Analysis of Ecological Momentary Assessment (EMA) Data. *Biometrics*, 64(2):627–634.

Hedecker, D., Mermelstein, R., and Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, 31(27):3328–3336.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. New York, NY, US: John Wiley & Sons, Ltd, 1st edition.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8:1–19.

Hessen, D. J. and Dolan, C. V. (2009). Heteroscedastic one-factor models and marginal maximum likelihood estimation. *British Journal of Mathematical and Statistical Psychology*, 62(1):57–77.

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

Hirose, K. and Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, 25:863–875.

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

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

Huang, P.-H. (2020). Penalized Least Squares for Structural Equation Modeling with Ordinal Responses. *Multivariate Behavioral Research*, 55(6):811–824.

Huang, P.-H., Chen, H., and Weng, L.-J. (2017). A Penalized Likelihood Method for Structural Equation Modeling. *Psychometrika*, 82(2):329–354.

Huber, P., Ronchetti, E., and Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 66(4):893–908.

Hui, F. K. C., Tanaka, E., and Warton, D. I. (2018). Order Selection and Sparsity in Latent Variable Models via the Ordered Factor LASSO. *Biometrics*, 74(4):1311–1319.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120.

Irincheeva, I., Cantoni, E., and Genton, M. G. (2012). Generalized Linear Latent Variable Models with Flexible Distribution of Latent Variables. *Scandinavian Journal of Statistics*, 39(4):663–680.

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

Jennrich, R. I. (2001). A Simple General Procedure for Orthogonal Rotation. *Psychometrika*, 66(2):289–306.

Jennrich, R. I. (2002). A Simple General Procedure for Oblique Rotation. *Psychometrika*, 67(1):7–20.

Jennrich, R. I. (2004). Rotation to Simple Loadings Using Component Loss Functions: The Orthogonal Case. *Psychometrika*, 69(2):257–273.

Jennrich, R. I. (2006). Rotation to Simple Loadings Using Component Loss Functions: The Oblique Case. *Psychometrika*, 71(1):173–191.

Jennrich, R. I. (2007). Rotation methods, algorithms, and standard errors. In Cudeck, R. and MacCallum, R. C., editors, *Factor Analysis at 100: Historical Developments and Future Directions*, pages 315–335. Lawrence Erlbaum Associates, Publishers.

Jiang, Y. and Xu, X. (2022). Testing the skewness of skew-normal distribution by bayes factors. *Journal of Statistical Planning and Inference*, 220:24–48.

Jin, S. and Andersson, B. (2020). A note on the accuracy of adaptive Gauss–Hermite quadrature. *Biometrika*, 107(3):737–744.

Jin, S., Moustaki, I., and Yang-Wallentin, F. (2018). Approximated Penalized Maximum Likelihood for Explanatory Factor Analysis: An orthogonal case. *Psychometrika*, 83(3):628–649.

Jöreskog, K. G. and Moustaki, I. (2001). Factor Analysis of Ordinal Variables: A Comparison of Three Approaches. *Multivariate Behavioral Research*, 36(3):347–387.

Jöreskog, K. K. and Golberger, A. S. (1972). Factor Analysis by Generalized Least Squares. *Psychometrika*, 37(3):243–260.

Junker, B. W. and Sijtsma, K. (2001). Nonparametric Item Response Theory in Action: An Overview of the Special Issue. *Applied Psychological Measurement*, 25(3):211–220.

Jöreskog, K. G. (1969). A General Approach to Confirmatory Maximum Likelihood Factor Analysis. *Psychometrika*, 34(2):183–202.

Jöreskog, K. G. (1970a). A General Method for Analysis of Covariance Structures. *Biometrika*, 57(2):239–251.

Jöreskog, K. G. (1970b). A General Method for Estimating a Linear Structural Equation System. *ETS Research Bulletin Series*, 2.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36(2):109–133.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In Goldberger, A. S. and Duncan, O. D., editors, *Structural Equation Models in the Social Sciences*, pages 85–112. New York, NY, US: Seminar Press.

Jöreskog, K. G. and Yang, F. (1996). Nonlinear structural equation models: The kenny-judd model with interaction effects. In Marcoulides, G. A. and Schumacker, R. E., editors, *Advanced*

Structural Equation Modeling: Issues and techniques, chapter 3, pages 57–87. Mahwah, NJ, US: Erlbaum.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., and 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*, 49:169–186.

Kelava, A., Werner, C. S., Schermelleh-Engel, K., Moosbrugger, H., Zapf, D., Ma, Y., Cham, H., and West, L. S. A. . S. G. (2011). Advanced Nonlinear Latent Variable Modeling: Distribution Analytic LMS and QML Estimators of Interaction and Quadratic Effects. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(3):465–491.

Kenny, D. A. and Judd, C. M. (1984). Estimating the Nonlinear and Interaction Effects of Latent Variables. *Psychological Bulletin*, 96(1):201–210.

Kim, Y.-J. and Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 66(2):337–356.

Klein, A. and Moosbrugger, H. (2000). Maximum Likelihood Estimation of Latent Interactions Effects with the LMS Method. *Psychometrika*, 65(4):457–474.

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in germany. *The Annals of Applied Statistics*, 9(2):1024–1052.

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling*, 13(4):275–303.

Kneib, T., Silbersdorff, A., and Säfken, B. (2023). Rage Against the Mean - A Review of Distributional Regression Approaches. *Econometrics and Statistics*, 26:99–123.

Knott, M. and Albanese, M. T. (1993). Conditional Distributions of a Latent Variable and Scoring for Binary Data. *Revista Brasileira de Probabilidade e Estadística (Brazilian Journal of Probability and Statistics)*, 6(2):171–188.

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., Ando, T., and Imoto, S. (2004). Bayesian Information Criteria and Smoothing Parameter Selection in Radial Basis Function Networks. *Biometrika*, 91(1):27–43.

Konishi, S. and Kitagawa, G. (1996). Generalised Information Criteria in Model Selection. *Biometrika*, 83(4):875–890.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer Series in Statistics. New York, NY, US: Springer-Verlag, 1st edition.

Kornely, M. J. K. and Kateri, M. (2022). Asymptotic Posterior Normality of Multivariate Latent Traits in an IRT Model. *Psychometrika*, 87(3):1146–1172.

Krasa, S. and Polborn, M. (2014). Policy Divergence and Voter Polarization in a Structural Model of Elections. *Journal of Law and Economics*, 57(1):31–76.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 12(3):209–229.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*. London : Butterworths, 2nd edition.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. New York, NY, US: Houghton-Mifflin, 1st edition.

Lee, E. R. and Park, S. (2021). Poisson reduced-rank models with sparse loadings. *Journal of the Korean Statistical Society*, 50:1079–1097.

Lee, J. D., Sun, Y., and Saunders, M. A. (2014). Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24.

Lee, S.-Y. (2007a). *Structural Equation Modeling: A Bayesian Approach*. Wiley Series in Probability and Statistics. New York, NY, US: John Wiley & Sons, Ltd, 1st edition.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):619–678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006.

Lee, Y. and Nelder, J. A. (2006). Double hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(2):139–185.

Lee, Y.-S. (2007b). A Comparison of Methods for Nonparametric Estimation of Item Characteristic Curves for Binary Items. *Applied Psychological Measurement*, 31(2):121–134.

Lewin-Koh, S.-C. and Amemiya, Y. (2003). Heteroscedastic factor analysis. *Biometrika*, 90(1):85–97.

Liang, L. and Browne, M. W. (2015). A Quasi-Parametric Method for Fitting Flexible Item Response Function. *Journal of Educational and Behavioral Statistics*, 40(1):5–34.

Lin, T.-I., Wu, P. H., McLachlan, G. J., and Lee, S. X. (2015). A robust factor analysis model using the restricted skew-t distribution. *TEST*, 24(3):510–531.

Lingjærde, O. C. and Liestøl, K. (1998). Generalized projection pursuit regresion. *SIAM Journal on Scientific Computing*, 20.

Liu, M. and Lin, T. I. (2015). Skew-normal factor analysis models with incomplete data. *Journal of Applied Statistics*, 42(4):789–805.

Liu, X., Wallin, G., Chen, Y., and Moustaki, I. (2023). Rotation to Sparse Loadings using L^p Losses and Related Inference Problems. *Psychometrika*, 88(2):527 – 553.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA, US: Addison-Wesley, 1st edition.

Louis, T. A. (1982). Finding the Observed Information Matrix When Using the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233.

Ma, Y. and Genton, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 72(4):1–38.

Magnus, B. E. and Garnier-Villarreal, M. (2021). A Multidimensional Zero-Inflated Graded Response Model for Ordinal Symptom Data. *Psychological Methods*, 27(2):261–279.

Magnus, B. E. and Liu, Y. (2022). Symptom presence and symptom severity as unique indicators of psychopathology: An application of multidimensional zero-inflated and hurdle graded response models. *Educational and Psychological Measurement*, 82(5):938–966.

Magnus, B. E. and Thissen, D. (2017). Item Response Modeling of Multivariate Count Data With Zero Inflation, Maximum Inflation, and Heaping. *Journal of Educational and Behavioral Statistics*, 42(5):531–558.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., and McGovern, M. E. (2017). A Simultaneous Equation Approach to Estimating HIV Prevalence With Nonignorable Missing Responses. *Journal of the American Statistical Association*, 112(518):484–496.

Marsh, H. W., Wen, Z., and Hau, K.-T. (2004). Structural Equation Models of Latent Interactions: Evaluation of Alternative Estimation Strategies and Indicator Construction. *Psychological Methods*, 9(3):275–300.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability (37). Boca Ratón, FL, US: Chapman & Hall / CRC.

McDonald, R. P. (1965). Difficulty factors and non-linear factor analysis. *British Journal of Mathematical and Statistical Psychology*, 18(1):11–23.

McDonald, R. P. (1967). Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika*, 32(1):77–112.

McDonald, R. P. (1982). A Note on the Investigation of Local and Global Identifiability. *Psychometrika*, 47(1):101–103.

McDonald, R. P. and Krane, W. R. (1977). A note on local identifiability and degrees of freedom in the asymptotic likelihood ratio test. *British Journal of Mathematical and Statistical Psychology*, 30(2):198–203.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. New York, NY, US: John Wiley & Sons, Ltd., 2nd edition.

Meijer, E. and Mooijaart, A. (1996). Factor analysis with heteroscedastic errors. *British Journal of Mathematical and Statistical Psychology*, 49(1):189–202.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis*, volume 1 of *Methods and Models in the Social Sciences*. Berlin, DE: De Gruyter Mouton, 1 edition.

Mokken, R. J. and Lewis, C. (1982). A Nonparametric Approach to the Analysis of Dichotomous Item Responses. *Applied Psychological Measurement*, 6(4):417–430.

Molenaar, D. (2015). Heteroscedastic Latent Trait Models for Dichotomous Data. *Psychometrika*, 80(3):625–644.

Molenaar, D. and Bolsinova, M. (2017). A heteroscedastic generalized linear model with a non-normal speed factor for responses and response times. *British Journal of Mathematical and Statistical Psychology*, 70(2):297–316.

Molenaar, D., Cúri, M., and Bazán, J. L. (2022). Zero and One Inflated Item Response Theory Models for Bounded Continuous Data. *Journal of Educational and Behavioral Statistics*, 47(6):693–735.

Molenaar, D., Dolan, C. V., and van der Maas, H. L. (2011). Modeling Ability Differentiation in the Second- Order Factor Model. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(4):578–594.

Molenaar, D., Dolan, C. V., and Verhelst, N. D. (2010). Testing and modelling non-normality within the one-factor model. *British Journal of Mathematical and Statistical Psychology*, 63(2):293–317.

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. J. (2015). A Bivariate Generalized Linear Item Response Theory Modeling Framework to the Analysis of Responses and Response Times. *Multivariate Behavioral Research*, 50(1):56–74.

Montanari, A. and Viroli, C. (2010). A skew-normal factor model for the analysis of student satisfaction towards university courses. *Journal of Applied Statistics*, 37(3):473–487.

Monti, A. C. (2003). A note on the estimation of the skew normal and the skew exponential power distributions. *METRON*, 61(2):205–219.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, 49(2):313–334.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, 56(2):337–357.

Moustaki, I. and Knott, M. (2000). Generalized Latent Trait Models. *Psychometrika*, 65(3):391–411.

Moustaki, I. and Knott, M. (2014). Latent variable models that account for atypical responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(2):343–360.

Moustaki, I. and Steele, F. (2005). Latent variable models for mixed categorical and survival responses, with an application to fertility preferences and family planning in Bangladesh. *Statistical Modelling*, 5(4):327–342.

Moustaki, I. and Victoria-Feser, M. P. (2006). Bounded-Influence Robust Estimation in Generalized Linear Latent Variable Models. *Journal of the American Statistical Association*, 101(474):644–653.

Mulaik, S. A. (2009). *Foundations of Factor Analysis*. Statistics in the Social and Behavioral Sciences Series. Boca Ratón, FL, US: Chapman & Hall / CRC, 2nd edition.

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

Muthén, L. K. and Muthén, B. O. (1998–2017). *Mplus User's Guide*. Los Angeles, CA, US: Muthén & Muthén, 8th edition edition.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 135(2):370–384.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., and Warton, D. I. (2019). Efficient estimation of generalized linear latent variable models. *PLoS ONE*, 14(5):1–20.

Niku, J., Warton, D. I., Hui, F. K. C., and Taskinen, S. (2017). Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, 22(4):498–522.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. New York, NY, US: Springer-Verlag, 2nd edition.

Noel, Y. (2014). A Beta Unfolding Model for Continuous Bounded Responses. *Psychometrika*, 79(4):647–674.

Noel, Y. and Dauvier, B. (2007). A Beta Item Response Model for Continuous Bounded Responses. *Applied Psychological Measurement*, 31(1):47–73.

O'Muircheartaigh, C. and Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(2):177–194.

Pan, Z. and Lin, D. Y. (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61(4):1000–1009.

Parikh, N. and Boyd, S. (2014). Proximal Algorithms. *Foundations and Trends in Optimization*, 3(1):123–231.

Pewsey, A. (2000). Problems of inference for Azzalini's skewnormal distribution. *Journal of Applied Statistics*, 27(7):859–870.

R Core Team (2022). *R: A Language and Environment for Statistical Computing, version 4.2.2*. R Foundation for Statistical Computing, Vienna, Austria.

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

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science*, 3(4):425–461.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56(4):611–630.

Ramsay, J. O. and Abrahamowicz, M. (1989). Binomial Regression With Monotone Splines: A Psychometric Application. *Journal of the American Statistical Association*, 84(408):906–915.

Ramsay, J. O. and Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika*, 56(3):365–379.

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Copenhagen, DK: Danish Institute for Educational Research, 1st edition.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer Series in Statistics for Social and Behavioral Sciences. New York, NY, US: Springer-Verlag, 1 edition.

Reiersøl, O. (1950). On the Identifiability of Parameters in Thurstone's Multiple Factor Analysis. *Psychometrika*, 15(2):121–149.

Revuelta, J., Hidalgo, B., and Alcazar-Córcolesa, M. A. (2022). Bayesian Estimation and Testing of a Beta Factor Model for Bounded Continuous Variables. *Multivariate Behavioral Research*, 57(1):1–22.

Rigby, R. A. and Stasinopoulos, M. D. (2005). Generalized additive models for location, shape and scale. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., and De Bastiani, F. (2020). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. Chapman & Hall/CRC The R Series. Boca Ratón, FL, US: Chapman & Hall / CRC.

Rizopoulos, D. and Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, 61(2):415–438.

Rothenberg, T. J. (1971). Identification in Parametric Models. *Econometrica*, 39(3):577–591.

Sagner, A. G. (2019). *Three essays on quantile factor analysis*. PhD thesis, Boston University.

Samejima, F. (1997). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, 62(4):471–493.

Samejima, F. (2000). Logistic Positive Exponent Family of Models: Virtue of Asymmetric Item Characteristic Curves. *Psychometrika*, 65(3):319–335.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent Variable Models for Mixed Discrete and Continuous Outcomes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 59(3):667–678.

Sánchez, B. N., Houseman, E. A., and Ryan, L. M. (2009). Residual-based diagnostics for structural equation models. *Biometrics*, 65(1):104–115.

Sardy, S. and Victoria-Feser, M. P. (2012). Isotone additive latent variable models. *Statistics and Computing*, 22:647–659.

Schilling, S. and Bock, R. D. (2005). High-dimensional Maximum Marginal Likelihood Item Factor Analysis by Adaptive Quadrature. *Psychometrika*, 70(3):533–555.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464.

Shao, J. (1997). An Asymptotic Theory for Linera Model Selection. *Statistica Sinica*, 7(2):221–242.

Shapiro, A. (1985). Identifiability of Factor Analysis: Some Results and Open Problems. *Linear Algebra and its Applications*, 70:1–7.

Shi, J.-Q. and Lee, S.-Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 62(1):77–87.

Sijtsma, K. (2005). Nonparametric item response theory models. In Kempf-Leonard, K., editor, *Encyclopedia of Social Measurement*, volume 2, pages 875–882. Amsterdam, NL: Elsevier Inc.

Sijtsma, K. and van der Ark, L. A. (2020). *Measurement Models for Psychological Attributes*. Statistics in the Social and Behavioral Sciences Series. Boca Ratón, FL, US: Chapman & Hall / CRC, 1 edition.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: multilevel, longitudinal, and structural equation models*. Interdisciplinary Statistics. Boca Ratón, FL, US: Chapman & Hall, CRC.

Skrondal, A. and Rabe-Hesketh, S. (2007). Latent Variable Modelling: A Survey. *Scandinavian Journal of Statistics*, 34(4):712–745.

Song, X.-Y. and Lu, Z.-H. (2010). Semiparametric Latent Variable Models with Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 19(3):590–608.

Song, X.-Y., Lu, Z.-H., Cai, J.-H., and Ip, E. H.-S. (2013). A Bayesian modeling approach for generalized semiparametric structural equation models. *Psychometrika*, 78(4):624–647.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: Using GAMLSS in R*. Chapman & Hall/CRC The R Series. Boca Ratón, FL, US: Chapman & Hall / CRC.

Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stout, W. (2001). Nonparametric Item Response Theory: A Maturing and Applicable Measurement Modeling Approach. *Applied Psychological Measurement*, 25(3):300–306.

Stroud, A. H. and Secrest, D. (1966). *Gaussian Quadrature Formulas*. New York, NY, US: Prentice Hall, 1st edition.

Sun, J., Chen, Y., Liu, J., Ying, Z., and Xin, T. (2017). Latent Variable Selection for Multidimensional Item Response Theory Models via L_1 Regularization. *Psychometrika*, 81(4):921–939.

Thurstone, L. L. (1947). *Multiple-Factor Analysis: A Development and Expansion of The Vectors of Mind*. Chicago, IL, US: Chicago University Press.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2006). Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67(1):91–108.

Trendafilov, N., Fontanella, S., and Adachi, K. (2017). Sparse Exploratory Factor Analysis. *Psychometrika*, 82(3):778–794.

Tucker-Drob, E. M. (2009). Differentiation of Cognitive Abilities Across the Life Span. *Developmental Psychology*, 45(4):1097–1118.

Ulbricht, J. (2010). *Variable Selection in Generalized Linear Models*. PhD thesis, Ludwig-Maximilians-Universität München.

Umlauf, N., Klein, N., and Zeileis, A. (2018). BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond). *Journal of Computational and Graphical Statistics*, 27(3):612–627.

van der Linden, W. J. (2007). A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, 72(3):287–308.

van der Linden, W. J. (2008). Using Response Times for Item Selection in Adaptive Testing. *Journal of Educational and Behavioral Statistics*, 33(1):5–20.

van der Linden, W. J. (2009). Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, 46(3):247–272.

van der Linden, W. J. and Guo, F. (2008). Bayesian Procedures for Identifying Aberrant Response-Time Patterns in Adaptive Testing. *Psychometrika*, 73(3):365–384.

van der Linden, W. J., Klein Entink, R. H., and Fox, J.-P. (2010). IRT Parameter Estimation With Response Times as Collateral Information. *Applied Psychological Measurement*, 34(5):327–347.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Number 3 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press, 1st edition.

Verkuilen, J. and Smithson, M. (2012). Mixed and Mixture Regression Models for Continuous Bounded Responses Using the Beta Distribution. *Journal of Educational and Behavioral Statistics*, 37(1):82–113.

Wald, A. (1950). A note on the identification of economic relations. In Koopmans, T. C., editor, *Statistical Inference in Dynamic Economic Models*, number 10 in Cowles Comission Monograph, chapter 3, pages 238–244. New York, NY, US: John Wiley & Sons, Ltd.

Wall, M. M., Guo, J., and Amemiya, Y. (2015a). Mixture Factor Analysis for Approximating a Nonnormally Distributed Continuous Latent Factor With Continuous and Dichotomous Observed Variables. *Multivariate Behavioral Research*, 47(2):276–313.

Wall, M. M., Park, J. Y., and Moustaki, I. (2015b). IRT Modeling in the Presence of Zero-Inflation With Application to Psychiatric Disorder Severity. *Applied Psychological Measurement*, 39(8):583–597.

Wang, L. (2010). IRT-ZIP Modeling for Multivariate Zero-Inflated Count Data. *Journal of Educational and Behavioral Statistics*, 35(6):671–692.

Wang, L., Hamaker, E., and Bergeman, C. S. (2012). Investigating Inter-Individual Differences in Short-Term Intra-Individual Variability. *Psychological Methods*, 17(4):567–581.

Winsberg, S., Thissen, D., and Ramsay, J. O. (1984). Fitting Item Characteristic Curves With Spline Functions. *ETS Research Report Series*, December(2).

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*. Texts in Statistical Science. Boca Ratón, FL, US: Chapman & Hall / CRC, 2nd edition.

Woods, C. M. and Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71(2):281–301.

Yalcin, I. and Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statistical Science*, 16(3):275–294.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer Series in Statistics. New York, NY, US: Springer-Verlag, 1st edition.

Yee, T. W. (2020). The VGAM package for negative binomial regression. *Australian & New Zealand Journal of Statistics*, 62(1):116–131.

Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(3):481–493.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 68(1):49–67.

Zeger, S. L. and Karim, M. R. (1991). Generalized Linear Models With Random Effects; A Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86(413):79–86.

Zhang, C.-H. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *Annals of Statistics*, 38(2):894–942.

Zhang, S. and Chen, Y. (2022). Computation for Latent Variable Model Estimation: A Unified Stochastic Proximal Framework. *Psychometrika*, 87(4):1473–1502.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zimmerman, M. E. (2011). Speed-accuracy tradeoff. In Kreutzer, J. S., DeLuca, J., and Caplan, B., editors, *Encyclopedia of Clinical Neuropsychology*, pages 2344–2344. New York, NY, US: Springer-Verlag.

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the LASSO. *The Annals of Statistics*, 35(5):2173–2192.