Title
Learning Factor Analysis Structures: A Clique Search Method on Correlation Thresholded Graphs and a Piecewise Linear Spline Approach

Author
Kim, Dale

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Learning Factor Analysis Structures: A Clique Search Method on Correlation Thresholded Graphs and a Piecewise Linear Spline Approach

A dissertation submitted in partial satisfaction
of the requirements for the degree Doctor of Philosophy
in Statistics

by

Dale S. Kim

2022

ABSTRACT OF THE DISSERTATION

Learning Factor Analysis Structures:
A Clique Search Method on Correlation Thresholded Graphs
and a Piecewise Linear Spline Approach

by

Dale S. Kim
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2022
Professor Qing Zhou, Chair

Factor analysis is a widely used method for modeling a set of observed variables by a set of unobserved latent factors. Despite their widespread application, existing methods for factor analysis suffer from some or all of the following weaknesses: requiring the number of factors to be known, lack of theoretical guarantees for learning the model structure, and nonidentifiability of the parameters due to rotation invariance properties of the likelihood. To address these concerns, this dissertation proposes two main methods. First, we propose a fast correlation thresholding (CT) algorithm that simultaneously learns the number of latent factors and a model structure that leads to identifiable parameters. This approach translates this structure learning problem into the search for so-called independent maximal cliques in a thresholded correlation graph that can be easily constructed from the observed data. Moreover, we present a routine to find all independent maximal cliques very efficiently by checking the neighborhood of each node in the graph. Finite-sample error bound and high-dimensional consistency for the structure learning of this method is also presented. Second, we consider the problem of non-linear factor analysis, and propose a piecewise linear spline method under an EM-algorithm framework. In many practical settings, learning a non-linear model may obviate the

need for multiple latent factors, and also allow the model to avoid rotational invariance nonidentifiability. This method is explored by simulation, and a preliminary study into the non-linear multidimensional extension is also presented.

The dissertation of Dale S. Kim is approved.

Ying Nian Wu

Jingyi Li

Oscar Hernan Madrid Padilla

Qing Zhou, Committee Chair

University of California, Los Angeles

2022

For Moose.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would like to express my gratitude to several of the faculty in the Department of Statistics at UCLA. First and foremost, I would like to thank my advisor, Dr. Qing Zhou, for being an academic exemplar in both teaching and research. Thank you for taking me as a student for these past five years. I am constantly astounded at the wealth of knowledge you have accumulated, and your sharp intuition on how problems should be approached. I would also like to thank Dr. Jessica Li and Dr. Ying Nian Wu, who served on my committee and have taught some of the courses in the program that I consider to be the most valuable. My training was greatly enriched by your classes, and I appreciate the time and patience that was spent on ensuring that practical subject-matter knowledge was imparted onto all your students.

Additionally, I would like to thank several of my graduate student comrades: Dr. Hans Li, Jiayi Li, Stephen Smith, and Gabriel Ruiz. I greatly appreciate the companionship all of you provided throughout my time in the program. My graduate student experience would not be the same without our fervent discussions about statistics, careers, movies, and as well as playing badminton together.

And finally, to my friends, who have been there since before the start of my journey into academia: Colin Chang, Jonathan Tsai, Ann Tseng, Jeanette Yang, and Samson Wong (you're listed alphabetically, so don't read into it). I appreciate all of your support and confidence throughout all of my graduate studies. I will note however, there were some among you who didn't come to my graduation. You know who you are. You owe me dinner.

VITA

Education

Ph.D., Statistics — September 2019 - June 2022 (Expected)
University of California, Los Angeles, CA

Ph.D., Quantitative Psychology — September 2015 - December 2021
Minor: Computational Cognition
University of California, Los Angeles, CA

M.S., Statistics — September 2019 - June 2020
University of California, Los Angeles, CA

M.A., Quantitative Psychology — September 2015 - June 2019
University of California, Los Angeles, CA

B.S., Psychology — January 2011 - June 2014
University of Washington, Seattle, WA

Publications

Kim, D. S. & Zhou, Q. (2022). A Correlation Thresholding Algorithm for Factor Analysis Models. *arXiv*. https://arxiv.org/abs/2203.01471

Kim, D. S. & McCabe, C. M. (2022). The Partial Derivative Framework for Substantive Regression Effects. *Psychological Methods*, 27(1), 121-141.

Kim, D. S. (2021). A Hybrid EM Algorithm for Linear Two-Way Interactions with Missing Data. *arXiv*. https://arxiv.org/abs/2111.06998

McCabe, C. M., Halvorson, M. A., King, K. M., Cao, X., & Kim, D. S., (2021). Interpreting interaction effects in generalized linear models of nonlinear probabilities and counts. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2020.1868966

Halvorson, M. A., McCabe, C. M., Kim, D. S., Cao, X. & King, K. M. (2021). Making sense of some odd ratios: A tutorial and improvements to present practices in reporting and visualizing quantities of interest for binary and count outcome models. *Psychology of Addictive Behaviors.* Advance online publication. https://doi.org/10.1037/adb0000669

McCabe, C. J., Kim, D. S. & King, K. M. (2018) Tools and Recommendations for the Visual Display of Interactions. *Advances in Methods and Practices in Psychological Science*, 1(2), 147-165. https://doi.org/10.1177/2515245917746792

King, K. M., Kim, D. S. & McCabe, C. J. (2018) Random responses inflate statistical estimates in heavily skewed addictions data. *Drug and Alcohol Dependence*, 183, 102-110. https://doi.org/10.1016/j.drugalcdep.2017.10.033

Kim, D. S., Reise, S. P. & Bentler, P. M. (2018) Identifying Aberrant Data in Structural Equation Models with IRLS-ADF. *Structural Equation Modeling*, 24(3), 343-358. https://doi.org/10.1080/10705511.2017.1379881

Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A. & King, K. M. (2018) Detecting Careless Responders with Infrequency Scales Using an Error Balancing Threshold. *Behavior Research Methods*, 50(5), 1960-1970. https://doi.org/10.3758/s13428-017-0964-9

Reise, S. P., Kim, D. S., Mansolf, M. & Widaman, K. F. (2016) Is the Bifactor Model a Better Model or is it Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, 51(6), 818-838. https://doi.org/10.1080/00273171.2016.1243461

Chapter 1

Introduction

Factor analysis is a commonly used multivariate technique which conceptualizes observed variables as a function of unobserved latent factors. Methods and discussions have appeared in a variety of fields, particularly the social sciences, such as psychology (Reise et al., 2000), sociology (Werts et al., 1973), education (Schreiber et al., 2006), and epidemiology (Martínez et al., 1998). It is generally assumed that the number of latent factors is less than the number of observed variables, hence serving as a dimension reduction procedure in this sense. Many applications use factor analysis to relate observed variables to hypothetical constructs that cannot be directly observed. These may include personality (McCrae and Costa, 1987), emotional states (Lovibond and Lovibond, 1995), symptomatology (Nisenbaum et al., 1998), or political ideology (Bollen, 1980).

1.1 The Factor Analysis Model

Let $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be a (column) vector of observed variables. The factor analysis model specifies the joint distribution of X in the form of a structural equation model:

$$X = \Lambda L + \epsilon, \tag{1.1}$$

where $L = (L_1, \ldots, L_d) \sim \mathcal{N}_d(0, \Phi)$ is a vector of latent variables or factors, $\epsilon = (\epsilon_1, \ldots, \epsilon_p) \sim \mathcal{N}_p(0, \Omega)$ is a vector of independent errors with a diagonal Ω, and $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{p \times d}$ is a matrix of coefficients, or factor loadings. For convenience, an additive mean vector μ is omitted from the model without loss of generality. The functional

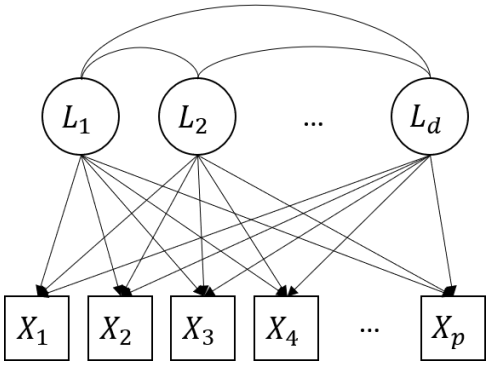

Figure 1.1: An illustration of a general factor analysis model. Directed edges denote a functional relation. Undirected edges denote a correlation. The dependence of X on ϵ and the independence among the ϵ variables are considered to be implicitly understood and thus omitted.

relations of the model are depicted in Figure 1.1, which illustrates that all observed variables X are a function of the latent variables L. We assume that $d < p$, since factor analysis is generally used as a dimension simplification technique. In the context of Λ, X_i is a function of L_j if and only if $\lambda_{ij} \neq 0$, in which case we may say that L_j is a *parent* of X_i and X_i a *child* of L_j. We assume that every X_i has at least one parent, and every L_j has at least one child, i.e., there are no rows or columns of full zeroes in Λ. No functional relations are assumed among the L variables, and they are only assumed to be correlated (oblique factor analysis) or uncorrelated (orthogonal factor analysis). We are considering the more general case of oblique factor analysis models in this study.

Subsequently, factor analysis models can readily accommodate causal assumptions and are often used as causal models. In such models, causal relations are generally assumed to be from L to X, such that a causal relation exists from L_j to X_i if and only if $\lambda_{ij} \neq 0$. Although factor analysis models make useful causal representations, we note that the methods presented herein will not be using such assumptions.

The model stated in Equation (1.1) implies a covariance structure Σ for X as follows:

$$\Sigma(\theta) \coloneqq \mathrm{Var}(X) = \mathrm{Var}(\Lambda L + \epsilon) = \Lambda \Phi \Lambda^T + \Omega, \tag{1.2}$$

letting $\theta = \{\Lambda, \Phi, \Omega\}$. We write $\Sigma(\theta)$ to make explicit that we are referring to Σ as a function of the parameters Λ, Φ, and Ω. At times, it will be easier to deal with observed variables which are unit variance scaled. Let $D_\sigma = \mathrm{diag}(\Sigma)^{1/2}$, i.e. a diagonal matrix with

entries $\Sigma_{ii}^{1/2}$. Then we define a unit variance scaled X as $\widetilde{X}$ in the following manner:

$$\widetilde{X} \coloneqq D_\sigma^{-1} X = D_\sigma^{-1}(\Lambda L + \epsilon) = \widetilde{\Lambda} L + \widetilde{\epsilon}, \tag{1.3}$$

where $\widetilde{\Lambda} = D_\sigma^{-1}\Lambda$ and $\widetilde{\epsilon} = D_\sigma^{-1}\epsilon$. Similarly, it follows that a correlation matrix $\widetilde{\Sigma}$ can be expressed as:

$$\widetilde{\Sigma}(\theta) \coloneqq D_\sigma^{-1}\Sigma D_\sigma^{-1} = \widetilde{\Lambda}\Phi\widetilde{\Lambda}^T + \widetilde{\Omega}, \tag{1.4}$$

where $\widetilde{\Omega} = D_\sigma^{-1}\Omega D_\sigma^{-1}$. Note that the factor analysis model for Σ and $\widetilde{\Sigma}$ are often used interchangeably, and the elements of $\widetilde{\Sigma}(\theta)$ may be referred to as ρ_{ij}. Finally, notice that the structure of a factor analysis model is entailed by the number of factors d and the support of Λ, denoted $\mathcal{A}(\Lambda)$. Therefore we will define the *structure* of a factor analysis model as the pair $(d, \mathcal{A}(\Lambda))$.

Given the structure of a factor analysis model $(d, \mathcal{A}(\Lambda))$, maximum likelihood is most widely used for estimating the parameters, based on the Gaussian log-likelihood for $X \sim \mathcal{N}_p(0, \Sigma(\theta))$:

$$\ell(\theta) = \frac{n}{2}\log|\Sigma(\theta)^{-1}| - \frac{n}{2}\mathrm{tr}(\Sigma(\theta)^{-1}S), \tag{1.5}$$

where S is the sample covariance matrix. However, there is no closed-form solution for the MLE (Jöreskog, 1967). Therefore, iterative algorithms, such as Newton-Raphson (Jennrich and Robinson, 1969) or Expectation-Maximization (Rubin and Thayer, 1982), are employed, which can be computationally intensive when the number of observed variables p is large. Furthermore, the parameters Λ and Φ as in Equation (1.2) are in general not identifiable, often referred to as rotational nonidentifiability in the literature (Anderson and Rubin, 1956). This issue must be taken care of with additional criteria for parameter estimation or restrictions on the model structure.

In summary, all methods of learning factor analysis must address three fundamental issues: (1) determine the number of factors, (2) learn the support of Λ, (3) resolve the rotational nonidentifiability issue. As we will review, an overabundance of literature has been dedicated to addressing these issues *separately*, all with varying degrees of success. From a practitioner's perspective, this has led to a combinatoric medley of methods for

determining the number of factors, determining the structure of the model, and choosing a solution from the rotationally equivalent set. In contrast, we seek to address all three issues simultaneously from a unified framework.

1.2 Review of Prior Work

Structure learning in the context of factor analysis typically refers to restrictions imposed on Λ. We are interested in sparse structures, where many entries of Λ are zero. Sparse structures are favorable in that they allow a clean interpretation of the model so it is clear as to which latent variables relate to each observed variable. If causal relations can be assumed, then learning the structure of Λ can be seen as a problem of learning causal relations. In such applications, selecting the most relevant relations through sparsity can greatly improve interpretability.

Prior work on learning factor analysis may be constraint-based or score-based. Constraint-based methods involve the analyzing permutations of correlations and partial correlations among the observed variables for constraints that would be implied by potential models (Scheines et al., 1998; Silva et al., 2006). However, we note that the focus of these algorithms is to construct equivalence classes of possible models and can be computationally demanding. In contrast, our goal is to develop efficient methods for learning and estimating a single model output in this work.

Single model output methods of sparse factor analysis usually involve score-based techniques of search or estimation. To seek sparse solutions in particular, restrictions are usually imposed on the model in Equation 1.1. A typical restriction would be to set $\Phi = I_d$, where variants such as ℓ_1 penalties (Choi et al., 2010; Ning and Georgiou, 2011), non-concave penalties (Hirose and Yamamoto, 2014b), and alternative parameterizations (Trendafilov et al., 2017) have been applied. Other work has investigated this topic with prior restrictions on Λ, e.g., assuming one entry per row (Adachi and Trendafilov, 2018). In contrast we focus on estimating these parameters without these restrictions. We review other relevant work on this problem in what follows.

1.2.1 Exploratory Factor Analysis

Currently, the main methods of learning a sparse structure on Λ fall under the umbrella of Exploratory Factor Analysis (EFA). EFA itself may or may not include a sparsity learning step depending on the purpose of the model. In practice, it is an algorithm that works generally as follows (Ford et al., 1986; Howard, 2016):

1. Given d as an input, set $\Phi = I_d$ and estimate an unconstrained Λ and diagonal Ω (e.g., via MLE by optimizing Equation 1.5).
2. Use a rotation criterion to find Φ.
3. (Optional) For sparsity, small entries of Λ may be set to zero if they are less than some threshold τ.
4. (Optional) Use a model selection procedure to choose among several choices of d.

To explain Step 2 , note that if Λ is unconstrained, the factor analysis model is not identifiable. That is, there may be many (Λ, Φ) pairs that exist such that $\Sigma(\theta) = \Lambda\Phi\Lambda^T + \Omega$. This issue is referred to as a lack of *rotational uniqueness*. To see this, let M be a $d \times d$ invertible matrix. Then, beginning with an orthogonal factors model that would be obtained by Step 1 of EFA ($\Phi = I_d$), we have

$$\begin{aligned} \Sigma(\theta) &= \Lambda\Lambda^T + \Omega \\ &= \Lambda_M \Phi_M \Lambda_M^T + \Omega, \end{aligned} \tag{1.6}$$

letting $\Lambda_M = \Lambda M$ and $\Phi_M = M^{-1}M^{-T}$. To ensure that Φ_M remains a valid correlation matrix, we impose the constraint that $\text{diag}(M^{-1}M^{-T}) = I_d$. Hence, estimating Φ (Step 2) amounts to finding a suitable M.

To find M, additional constraints called "rotation criteria" can be imposed (Browne, 2001). One common example of such a criteria is to choose M such that the following is minimized:

$$f(\Lambda_M) = (1-\kappa)\sum_{i=1}^{p}\sum_{j=1}^{d}\sum_{l\neq j}^{d}\lambda_{ij}^2\lambda_{il}^2 + \kappa\sum_{j=1}^{d}\sum_{i=1}^{p}\sum_{k\neq i}^{p}\lambda_{ij}^2\lambda_{kj}^2, \quad \kappa \in [0,1], \tag{1.7}$$

subject to the aforementioned constraints of M. This is known as the Crawford-Ferguson family of rotation criteria (Crawford and Ferguson, 1970). We can see that the term $\sum_{j=1}^{d}\sum_{l\neq j}^{d}\lambda_{ij}^2\lambda_{il}^2 \geq 0$, where equality holds if and only if there is at most one non-zero element in the ith row of Λ_M. The term $\sum_{i=1}^{p}\sum_{k\neq i}^{p}\lambda_{ij}^2\lambda_{kj}^2$ behaves the same way except it acts upon the jth column of Λ_M. Thus, Equation 1.7 is a weighted penalty on the row and column magnitudes of Λ_M, which is parameterized by κ. The most common parameterization choice is $\kappa = 1/p$, which is also known as varimax rotation (Kaiser, 1958).

Arguably, the biggest criticism of EFA is this lack of rotational uniqueness and the *ad hoc* nature of rotation criteria: Different criteria may yield very different solutions. Further, Step 3 of the EFA algorithm is another source of subjectivity, due to the lack of a principled way to enforce sparsity in Λ. Even though a rotation criterion may minimize certain magnitudes of the entries of Λ_M, rotation alone is insufficient to produce entries that are exactly zero. Hence, one must choose an arbitrary threshold τ by which to set low magnitude entries of Λ to zero.

1.2.2 Penalized Exploratory Factor Analysis

As a potential solution to the subjectivity problems in EFA, penalized methods also have been developed. Instead of rotating factor coefficients, penalized EFA can achieve sparse solutions directly in estimation. While penalized estimation additionally requires tuning parameters, these can be selected in an objective manner, for example by using the Bayesian Information Criterion (BIC; Schwarz, 1978) or cross-validation (CV; Scharf and Nestler, 2019).

These methods maximize a penalized likelihood (or minimize other loss functions) of the form:

$$\ell_p(\theta) = \ell(\theta) - p(\Lambda), \tag{1.8}$$

where $p(\cdot)$ is some penalty function. One example is the LASSO penalty (Tibshirani, 1996), which has been adapted to EFA (Choi et al., 2010; Ning and Georgiou, 2011) as

follows:

$$p_\kappa(\Lambda) = \kappa\|\Lambda\|_1 = \kappa \sum_{j=1}^{p} \sum_{k=1}^{d} |\lambda_{jk}|, \tag{1.9}$$

for a regularization parameter κ. Another common example is the minimax-concave penalty (MCP; Zhang, 2010), which has been utilized in penalized EFA as well (Hirose and Yamamoto, 2014a,b):

$$p_{\kappa,\gamma}(\Lambda) = \kappa \sum_{j=1}^{p} \sum_{k=1}^{d} \int_0^{|\lambda_{jk}|} \left(1 - \frac{x}{\kappa\gamma}\right)_+ dx, \tag{1.10}$$

where κ, γ are regularization parameters. In both cases, the regularization parameters are chosen by some model selection procedure (BIC, CV), while the number of latent factors d is generally regarded as given.

1.2.3 Choosing d

One of the early well-known methods is to set d equal to the number of eigenvalues greater than 1, also known as the Kaiser-Guttman criterion (Guttman, 1954; Kaiser, 1960). This number was shown to be a lower bound on the number of factors in the population (Guttman, 1954). It also has been argued that the minimum variance a latent variable should be able to explain is at least equal to that of a single observed variable under null correlation (Kaiser, 1960). However, subsequent authors have criticized the lower bound argument as not useful since the goal is to determine the actual number of factors (Cliff, 1988; Lance et al., 2006; Velicer and Jackson, 1990). Furthermore, empirical studies have shown inaccurate results in a variety of situations (Browne, 1968; Patil et al., 2008).

Another well-known method is to examine the successive differences between the ordered eigenvalues. Then the greatest of these differences is argued to be a border at which the eigenvalues that represent the latent variables' variance ends, and the eigenvalues that represent the error variances begins. This is often done in an informal manner by checking for an "elbow" in the plot of ordered eigenvalues (known as the Scree Test; Cattell, 1966), or more formally by examining the actual differences between eigenvalues (Raîche et al., 2013). Some criticisms of this approach are that the differences between

successive eigenvalues may not be clear or substantial and are vulnerable to sampling fluctuations (Hayton et al., 2004; Linn, 1968). While this method has been shown to perform better than the Kaiser-Guttman criterion, it still suffers from poor reliability (Hakstian et al., 1982; Streiner, 1998; Zwick and Velicer, 1986).

A potential problem that interferes with the efficacy of these methods is that sampling variability is not accounted for. Thus, other methods use a parametric bootstrap to generate data from a hypothetical null correlation model, to obtain a reference distribution of ordered eigenvalues. This null model provides a comparison to the hypothesis that all eigenvalues represent error variance (zero latent variables). Thus the number of eigenvalues higher than this reference distribution is taken to be d. This is known as parallel analysis (Horn, 1965) and can be done with respect to the average bootstrapped eigenvalue or some empirical quantile (Glorfeld, 1995). This method is essentially the same as the Kaiser-Guttman criterion, except that it takes sampling variation into account. For this reason, authors have criticized parallel analysis for the same shortcomings as the Kaiser-Guttman method (Crawford and Koopman, 1973; Wayne et al., 2000).

1.3 Motivation and Contributions

As reviewed above, the problems with current methods can be categorized into three main issues: determining the number of latent variables d, learning the support of Λ, and determining a rotationally unique solution. To address this, we propose a correlation thresholding algorithm to learn the support of Λ and the number of latent variables simultaneously, whose solution is guaranteed to be rotationally unique. We also establish high-dimensional consistency for learning the structure $(d, \mathcal{A}(\Lambda))$. Our method first constructs an undirected graph on p nodes, corresponding to $X_1, \ldots, X_p$, by thresholding sample correlations among these observed variables. We find that, under certain assumptions, there is a perfect correspondence between the latent factors and a class of maximal cliques (which we call independent maximal cliques defined in Section 2.2) in the correlation thresholded graph. Therefore, the structure learning problem can be converted to a search for all independent maximal cliques in the graph. We prove that

all the independent maximal cliques can be found by checking the neighborhood of each node, of which the computational complexity is no more than and usually well below $O(k^2 p)$, where k is the maximum neighborhood size. Another enormous advantage of our algorithm is that it can provide a set of candidate model structures without maximizing the likelihood (Equation 1.5) or a penalized likelihood, and is well applicable to problems with thousands of variables, as demonstrated in the numerical results.

The rest of this dissertation is organized as follows. We describe our correlation thresholding algorithm in Chapter 2 and develop theoretical justifications for its use in Section 2.3. We then test our correlation thresholding algorithm along with other methods with a series of simulation studies in Section 2.4 and a real data example in Section 2.5. In Chapter 3 we examine non-linear generalizations of the factor analysis model, and present a simple piecewise linear spline method for estimating it. Finally, we conclude with some comments on future work and extensions in Chapter 4.

Notation throughout this article will be as follows. Define $[n] \coloneqq \{1, \ldots, n\}$. Let $A \subseteq [n]$ and $B \subseteq [p]$ be index sets. The complement of A is denoted as A^c. For a matrix $M = (m_{ij}) \in \mathbb{R}^{n \times p}$, we define M_{AB} as the submatrix of M consisting of the rows indexed by A and columns indexed by B. Similarly for a vector $V \in \mathbb{R}^n$, we define V_A as the subvector of V consisting of the entries indexed by A. We denote the support of M as $\mathcal{A}(M) \coloneqq \{(i,j) : m_{ij} \neq 0\}$. We use $\mathbf{0}$ to represent a matrix or vector of zeroes, whose dimension can be inferred from context and I_n denotes the $n \times n$ identity matrix.

Chapter 2

The Correlation Thresholding Algorithm

2.1 Overview

To begin, we review several terms and definitions from graph theory. We define a graph $\mathcal{G}$ as an ordered pair (V, E), explicitly denoted as $\mathcal{G}(V, E)$, where V is a set of vertices and $E \subseteq V \times V$ is a set of edges. For convenience, we will use $V = X$ to mean that the elements of the vertex set V represent the index set of the random vector X. We also restrict our attention to *undirected* graphs, where we only consider edges $(i, j) \in E$ such that $i < j$. A *clique* of $\mathcal{G}(V, E)$ is a subset of vertices $C \subseteq V$ such that all pairs of distinct vertices in C are connected by an edge. Finally, a *maximal clique* is a clique that cannot be extended by including more vertices from V.

We now give a simple example to demonstrate how we will use graphical models for the factor analysis model as in Equation (1.1). Consider the following parameters:

$$\widetilde{\Lambda} = \begin{bmatrix} \widetilde{\lambda}_{11} & \\ \widetilde{\lambda}_{21} & \\ \widetilde{\lambda}_{31} & \widetilde{\lambda}_{32} \\ & \widetilde{\lambda}_{42} \\ & \widetilde{\lambda}_{52} \end{bmatrix}, \Phi = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}, \widetilde{\Omega} = \begin{bmatrix} \widetilde{\omega}_1 & & & & \\ & \widetilde{\omega}_2 & & & \\ & & \widetilde{\omega}_3 & & \\ & & & \widetilde{\omega}_4 & \\ & & & & \widetilde{\omega}_5 \end{bmatrix}. \tag{2.1}$$

This model is illustrated in Figure 2.1 (left). Note that these parameters imply the

following correlation matrix:

$$\widetilde{\Sigma}(\theta) = \widetilde{\Lambda}\Phi\widetilde{\Lambda}^T + \widetilde{\Omega} = \begin{bmatrix} \widetilde{\lambda}_{11}^2 + \widetilde{\omega}_1 & \widetilde{\lambda}_{11}\widetilde{\lambda}_{21} & \widetilde{\lambda}_{11}\widetilde{\lambda}_{31} & & \\ \widetilde{\lambda}_{11}\widetilde{\lambda}_{21} & \widetilde{\lambda}_{21}^2 + \widetilde{\omega}_2 & \widetilde{\lambda}_{21}\widetilde{\lambda}_{31} & & \\ \widetilde{\lambda}_{11}\widetilde{\lambda}_{31} & \widetilde{\lambda}_{21}\widetilde{\lambda}_{31} & \widetilde{\lambda}_{31}^2 + \widetilde{\lambda}_{32}^2 + \widetilde{\omega}_3 & \widetilde{\lambda}_{32}\widetilde{\lambda}_{42} & \widetilde{\lambda}_{32}\widetilde{\lambda}_{52} \\ & & \widetilde{\lambda}_{32}\widetilde{\lambda}_{42} & \widetilde{\lambda}_{42}^2 + \widetilde{\omega}_4 & \widetilde{\lambda}_{42}\widetilde{\lambda}_{52} \\ & & \widetilde{\lambda}_{32}\widetilde{\lambda}_{52} & \widetilde{\lambda}_{42}\widetilde{\lambda}_{52} & \widetilde{\lambda}_{52}^2 + \widetilde{\omega}_5 \end{bmatrix}. \quad (2.2)$$

From here, let us convert $\widetilde{\Sigma}(\theta) = (\rho_{ij})_{p\times p}$ to a graph that simply encodes the non-zero correlations. To do this, we define a *thresholded correlation graph* $\mathcal{G}(X, E(\tau))$, where the edge set is determined by thresholding the values of $|\rho_{ij}|$ at $\tau \geq 0$, i.e.,

$$E(\tau) \coloneqq \{(i,j) : |\rho_{ij}| > \tau, i < j\}. \quad (2.3)$$

For this example, $\mathcal{G}(X, E(0))$ is depicted in Figure 2.1 (right). There are two key observations to make regarding the relation between this factor analysis model and its thresholded correlation graph. First, the number of latent variables ($d = 2$) equals the number of maximal cliques in $\mathcal{G}(X, E(0))$. Second, these maximal cliques have a one-to-one correspondence to the children sets of the latent variables in the factor analysis model. In Figure 2.1, the maximal cliques $\{X_1, X_2, X_3\}$ and $\{X_3, X_4, X_5\}$ (right panel) are the respective children sets of L_1 and L_2 (left panel). The primary motivation of our algorithm is to leverage this correspondence by using such a thresholded correlation graph to gain insight into the support of Λ. In later sections we will develop the theoretical details for these relations to hold formally in more general settings.

For now, let us generalize this model to allow a correlation between L_1 and L_2, i.e. $\phi_{12} \neq 0$ in (2.1). In this case, the edge detection procedure is not as simple as thresholding for non-zero correlations. Generally, we begin with a saturated correlation matrix (no sparsities), since variables that do not share parents will be correlated by virtue of their parents being correlated. However, in most practical settings, the correlation of variables that do not share parents, e.g. X_1 and X_4, will have a lower magnitude than those

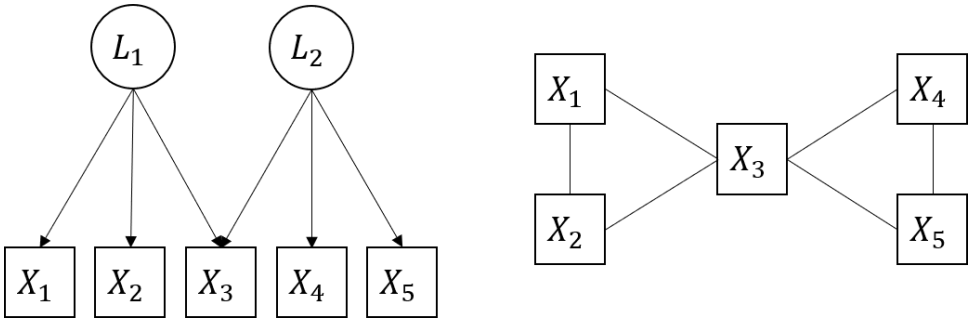

Figure 2.1: (Left) graphical representation of the factor analysis model described in Equation (2.1); (right) the corresponding thresholded correlation graph $\mathcal{G}(X, E(0))$.

variables whose parents are shared, e.g. X_1 and X_2. To see why this could be the case, consider $\widetilde{\Sigma}(\theta)$ in Equation (2.2) again if the correlation $\phi_{12} \neq 0$:

$$\begin{bmatrix} \widetilde{\lambda}_{11}^2 + \widetilde{\omega}_1^2 & \widetilde{\lambda}_{11}\widetilde{\lambda}_{21} & \widetilde{\lambda}_{11}\widetilde{\lambda}_{31} + \widetilde{\lambda}_{11}\widetilde{\lambda}_{32}\phi_{12} & \boldsymbol{\widetilde{\lambda}_{11}\widetilde{\lambda}_{42}\phi_{12}} & \boldsymbol{\widetilde{\lambda}_{11}\widetilde{\lambda}_{52}\phi_{12}} \\ \widetilde{\lambda}_{11}\widetilde{\lambda}_{21} & \widetilde{\lambda}_{21}^2 + \widetilde{\omega}_2^2 & \widetilde{\lambda}_{21}\widetilde{\lambda}_{31} + \widetilde{\lambda}_{21}\widetilde{\lambda}_{32}\phi_{12} & \boldsymbol{\widetilde{\lambda}_{21}\widetilde{\lambda}_{42}\phi_{12}} & \boldsymbol{\widetilde{\lambda}_{21}\widetilde{\lambda}_{45}\phi_{12}} \\ \widetilde{\lambda}_{11}\widetilde{\lambda}_{31} + \widetilde{\lambda}_{11}\widetilde{\lambda}_{32}\phi_{12} & \widetilde{\lambda}_{21}\widetilde{\lambda}_{31} + \widetilde{\lambda}_{21}\widetilde{\lambda}_{32}\phi_{12} & \widetilde{\lambda}_{31}^2 + \widetilde{\lambda}_{32}^2 + \widetilde{\omega}_3^2 & \widetilde{\lambda}_{31}\widetilde{\lambda}_{42}\phi_{12} + \widetilde{\lambda}_{32}\widetilde{\lambda}_{42} & \widetilde{\lambda}_{31}\widetilde{\lambda}_{52}\phi_{12} + \widetilde{\lambda}_{32}\widetilde{\lambda}_{52} \\ \boldsymbol{\widetilde{\lambda}_{11}\widetilde{\lambda}_{42}\phi_{12}} & \boldsymbol{\widetilde{\lambda}_{21}\widetilde{\lambda}_{42}\phi_{12}} & \widetilde{\lambda}_{31}\widetilde{\lambda}_{42}\phi_{12} + \widetilde{\lambda}_{32}\widetilde{\lambda}_{42} & \widetilde{\lambda}_{42}^2 + \widetilde{\omega}_4^2 & \widetilde{\lambda}_{42}\widetilde{\lambda}_{52} \\ \boldsymbol{\widetilde{\lambda}_{11}\widetilde{\lambda}_{52}\phi_{12}} & \boldsymbol{\widetilde{\lambda}_{21}\widetilde{\lambda}_{45}\phi_{12}} & \widetilde{\lambda}_{31}\widetilde{\lambda}_{52}\phi_{12} + \widetilde{\lambda}_{32}\widetilde{\lambda}_{52} & \widetilde{\lambda}_{42}\widetilde{\lambda}_{52} & \widetilde{\lambda}_{52}^2 + \widetilde{\omega}_5^2 \end{bmatrix}.$$

Since $|\phi_{12}| < 1$, we see that it has a shrinking effect on the correlations between variables that do not share parents (bolded for emphasis). That is, if $|\phi_{12}|$ is small enough, there would exist some threshold by which these correlations (bold) were below and correlations among variables with shared parents (not bold) were above. Then this threshold could identify and eliminate edges between pairs of variables that did not share parents, yielding a structurally informative graph as in Figure 2.1 (right). Finding such a threshold and using the thresholded correlation graph to learn the factor analysis structure is the key intuition behind our algorithm.

A main assumption here is that there exists some threshold τ_0 such that $\mathcal{G}(X, E(\tau_0))$ can differentiate between pairs of variables that share latent factor parents and pairs that do not. Let the *parent set* of X_i be $\Pi_i \coloneqq \{j : \lambda_{ij} \neq 0, j \in [d]\}$. Then we can formalize this notion by denoting the set of pairs that share parents as

$$E_0 \coloneqq \{(i, j) \in [p] \times [p] : \Pi_i \cap \Pi_j \neq \emptyset, i < j\}. \tag{2.4}$$

Subsequently, we denote the set of pairs that do not share parents (the complement of E_0) as:

$$E_0^c = \{(i,j) \in [p] \times [p : \Pi_i \cap \Pi_j = \emptyset, i < j\}. \tag{2.5}$$

Essentially, we would like to find some threshold τ_0 that is able to separate the E_0 and E_0^c sets by the magnitude of the correlations. We will define this notion as *thresholdable*. Recall that ρ_{ij} is the correlation between X_i and X_j given by $\widetilde{\Sigma}(\theta)$ in (1.4).

Definition 1 (Thresholdable)**.** A set of parameters θ is called *thresholdable* if there exists a threshold τ_0 such that

$$\max\{|\rho_{kl}| : (k,l) \in E_0^c\} < \tau_0 < \min\{|\rho_{ij}| : (i,j) \in E_0\}. \tag{2.6}$$

That is, if θ is thresholdable, then we can correctly sort the index pairs of X into the E_0 and E_0^c sets using τ_0, i.e, $E(\tau_0) = E_0$. This allows us to move forward with the graphical logic as shown in the previous example with orthogonal factors (i.e., Figure 2.1).

In practice, given a sample of X, we can define an estimate of E_0 for a candidate τ_k as

$$\hat{E}(\tau_k) \coloneqq \{(i,j) : |r_{ij}| > \tau_k, i < j\}, \tag{2.7}$$

where r_{ij} denotes the sample correlation. Putting these ideas together, the core task of our algorithm is to search for a suitable τ_0. This can be done by searching over a set of candidate set $\tau_k \in [0,1]$ and analyzing their respective thresholded correlation graphs $\mathcal{G}(X, \hat{E}(\tau_k))$. The aforementioned graphical concept of maximal cliques can then be leveraged to learn the number of latent variables and the support of Λ. This essentially yields a set of candidate models for which we can utilize model selection procedures (e.g., BIC) to select a final model.

Remark 1. Note that the thresholded graphs defined in Equation (2.3) are acting upon the marginal correlations of the observed variables. This is in contrast to the well-known conditional independence graphs in Gaussian graphical models, which correspond to the support of the inverse correlation matrix, or $\widetilde{\Sigma}^{-1}$ (Lauritzen, 1996). In the case of factor

analysis models of Equation (1.1), conditional independence graphs over the observed variables are generally not sparse if the latent factors are correlated. That is, for any pair (i, j), we have

$$X_i \not\perp X_j | \{X_k : k \in V - \{i, j\}\}$$

because their correlation is always confounded by the latent variables which are unaccounted for. In this case, the conditional independence graph will be a complete graph, and thus is not informative for estimating the structure of the factor analysis model.

Remark 2. Ancestral graphs are another way of representing multivariate relations with latent variables (Richardson and Spirtes, 2002). Algorithms that determine these relations (FCI and related variants; Spirtes et al., 2000; Spirtes, 2001; Colombo et al., 2012) typically analyze the conditional independence among the observed variables. In the case of the factor analysis models of Equation (1.1), the result would be a complete graph without any edge orientations, since the relation $X_i \not\perp X_j | A$ would hold for all (i, j) and any $A \subseteq V - \{i, j\}$ (for a list of orientation rules see Zhang, 2008). Again, this graph is not useful for structure estimation of the factor analysis model.

2.2 The Algorithm

We now apply the framework from the previous section to construct our correlation thresholding (CT) algorithm. Denote by $R = (r_{ij}) \in \mathbb{R}^{p \times p}$ the sample correlation matrix among $X_1, \ldots, X_p$, and let a set of chosen thresholds be $\tau = \{\tau_k : \tau_k \in [0, 1], k \in [m]\}$. Then the CT algorithm is described in Algorithm 1.

A couple of notes are in order. First, of particular interest to our algorithm, we will define a specific kind of maximal clique, which we term as *independent maximal clique*:

Definition 2 (Independent Maximal Clique)**.** Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the set of all maximal cliques in a graph $\mathcal{G}$. Then, C_i is an independent maximal clique if

$$C_i \nsubseteq \bigcup_{j \neq i} C_j. \tag{2.8}$$

Algorithm 1: The Correlation Thresholding Algorithm

input : The sample correlation matrix R and a set of thresholds τ.
output : Parameter estimates $\hat{\theta}$.

for $k \in [m]$ **do**
  Calculate $\mathcal{G}(X, \hat{E}(\tau_k))$ and extract the set of independent maximal cliques: $\mathcal{C}_k = \{C_1, \dots C_{|\mathcal{C}_k|}\}$;
  Set $\hat{d}_k = |\mathcal{C}_k|$;
  Initialize $\hat{A}_k = \emptyset$;
  for $(i, j) \in [p] \times [\hat{d}_k]$ **do**
    if $i \in C_j$ **then** add (i, j) to $\hat{A}_k$;
  end
  Estimate $\hat{\theta}_k$ given $(\hat{d}_k, \hat{A}_k)$, i.e. subject to $\lambda_{ij} = 0$ for all $(i, j) \notin \hat{A}_k$;
end
Select one of the m estimates from $\{\hat{\theta}_k : k \in [m]\}$ via a model selection procedure.

Essentially, an independent maximal clique is a maximal clique that contains a vertex that is not a member of any other maximal clique. We call such a vertex a *unique member* of the independent maximal clique. We use the word "independent" as an analog to the notion of linear independence in a vector space. That is, an independent maximal clique cannot be covered by the union of any of the other maximal cliques. This condition is needed for structural identifiability, which will be described in a later section.

To quickly find all independent maximal cliques in a graph, we can employ the following Lemma.

Lemma 1. *Given a graph $\mathcal{G}(X, E)$, let $\mathrm{ne}(X_i)$ be the set of vertices that contains X_i and every node that shares an edge with X_i (the neighbors of X_i).*

1. *If $\mathrm{ne}(X_i)$ is a clique, then $\mathrm{ne}(X_i)$ is also an independent maximal clique and X_i is a unique member of this clique.*

2. *If C is an independent maximal clique, then $C = \mathrm{ne}(X_i)$ for any unique member $X_i \in C$.*

Proof. First, we prove that $ne(X_i)$ must be a maximal clique by contradiction. Suppose $ne(X_i)$ is a clique, but not maximal. Then $ne(X_i)$ can be extended by another node $X_j \notin ne(X_i)$, such that the union $X_j \cup ne(X_i)$ is a clique. This implies that there is an edge between X_i and X_j and thus $X_j \in ne(X_i)$. This leads to a contradiction, and

therefore, $ne(X_i)$ must be maximal. Second, we prove that X_i is not a part of any other maximal clique, once again by contradiction. Suppose that $X_i \in A$, where A is a maximal clique and $A \neq ne(X_i)$. By the definition of $ne(X_i)$, we must have $A \subset \text{ne}(X_i)$, i.e., a proper subset of $\text{ne}(X_i)$, which contradicts the hypothesis that A is maximal. Therefore, X_i is not a part of any other maximal clique, making $ne(X_i)$ an independent maximal clique. This completes the proof of the first statement.

Now we prove the second statement. Let X_i be any unique member of an independent maximal clique C. Suppose $ne(X_i)$ is not a subset of C, which means there is a vertex $X_j \notin C$ but is a neighbor of X_i. Then $\{X_i, X_j\}$ either is a maximal clique or can be grown to a maximal clique $C' \neq C$. This contradicts the fact that X_i is a unique member of C. Therefore, $ne(X_i)$ must be a subset of C and thus is a clique. By the first statement of this lemma, $ne(X_i)$ is also an independent maximal clique and thus we must have $ne(X_i) = C$. □

We can use Lemma 1 to find all independent maximal cliques in a graph, in the worst-case scenario by checking whether $ne(X_i)$ is a clique for every node X_i. The computational cost for checking if $ne(X_i)$ is a clique has a brute force complexity of $O(k^2)$, assuming a maximum neighbor size of k. Thus, the total computational cost on all p nodes can be no greater than and usually well below $O(k^2 p)$, which is very efficient even for large graphs.

Returning to the CT algorithm, these independent maximal cliques are extracted in Step 2, and Steps 3 through 7 use these cliques to learn the number of latent factors d and the support of Λ. The number of independent maximal cliques is set as the estimate of d, which is also the number of columns in Λ. Then, the nodes in each C_j determine if λ_{ij} is zero or non-zero for each $i \in [p]$, allowing us to construct a candidate support $\hat{A}_k$. These are the steps that apply the logic set forth in Section 2.1 for structural learning.

Steps 8 and 10 are general in that they can utilize any estimation and model selection methods. In our implementation, we will prefer to use maximum likelihood estimation (1.5) and BIC for model selection. That is, after using maximum likelihood estimation to

obtain $\hat{\theta}_k$, we calculate BIC with

$$\text{BIC}(\hat{\theta}_k) = q_k \log(n) - 2\ell(\hat{\theta}_k), \tag{2.9}$$

where q_k is the number of free parameters in $\hat{\theta}_k$. Then the output of the CT algorithm would be the estimated parameters using

$$\hat{\theta} = \text{argmin BIC}(\hat{\theta}_k). \tag{2.10}$$

We note that a preferred pairing of methods would be one where both parameter estimation and model selection are consistent. This leads to the final output model having consistent parameters and model structure, as we will show in Section 2.3.4.

Finally, note that the generation of candidate structures does not depend on the parameter estimates $\hat{\theta}_k$. Hence, it is possible to forego the parameter estimation and model selection parts (Steps 8 and 10) of the algorithm. The output would then be a set of candidate structures $\{(\hat{d}_k, \hat{A}_k)) : k \in [m]\}$ without any $\hat{\theta}_k$ estimates. This is useful in high-dimensional settings, when parameter estimation can be computationally slow, but the structure learning portion remains fast. In this case, the structures may serve as a set of candidate models to be further evaluated.

We provide a broad summary of the procedure in Figure 2.2. The main assertion of the algorithm is that under a good choice of $\{\tau_k : k \in [m]\}$ among other identifiability conditions (described below), the correct structure of Λ will be recovered by one of the $\hat{E}(\tau_k)$ (Steps 3 - 7). Then, given that the correct model is among the final set of candidate models, a consistent model selection criterion will be able to recover it (Step 10). In the following sections, we describe the precise conditions under which this can be achieved, as well as establishing statistical consistency for the algorithm.

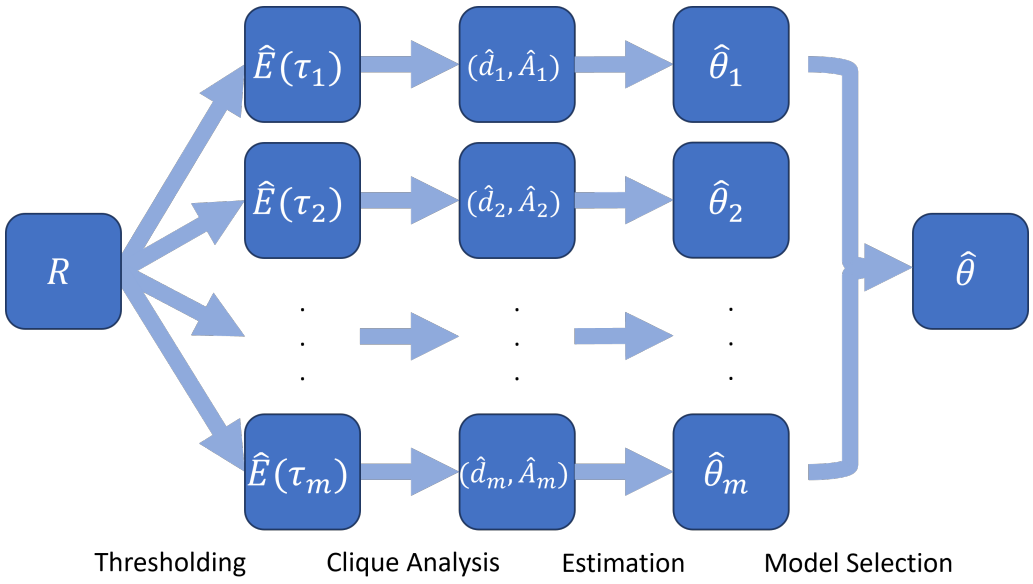

Figure 2.2: Overview of the CT algorithm.

2.3 Theoretical Analysis

In this section, we establish theoretical guarantees for the CT algorithm and discuss the key underlying assumptions. We assume throughout that the factor analysis model in Equation (1.1) holds, i.e., $X \sim \mathcal{N}_p(0, \Sigma(\theta))$ and $\epsilon \sim \mathcal{N}_p(0, \Omega)$. Proofs of these results can be found in Supplementary Materials **??**.

2.3.1 On the Thresholdability of θ

One of the more fundamental assumptions of the CT algorithm is the thresholdability of θ. In this section, we examine this assumption in more detail. Specifically, a necessary and sufficient condition for thresholdability is as follows:

Lemma 2. *Recall the definitions of E_0 and E_0^c in Equations 2.4 and 2.5, respectively. A set of parameters θ is thresholdable if and only if:*

$$\max_{(k,l)\in E_0^c} |\widetilde{\Lambda}_{kE}\Phi_{EF}\widetilde{\Lambda}_{lF}^T| < \min_{(i,j)\in E_0} |\widetilde{\Lambda}_{iA}\Phi_{AB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iC}\Phi_{CB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iA}\Phi_{AC}\widetilde{\Lambda}_{jC}^T + \widetilde{\Lambda}_{iC}\Phi_{CC}\widetilde{\Lambda}_{jC}^T|, \tag{2.11}$$

where $A = A(i,j) = \Pi_i - \Pi_j$, $B = B(i,j) = \Pi_j - \Pi_i$, $C = C(i,j) = \Pi_i \cap \Pi_j$, $E = \Pi_k$, and $F = \Pi_l$.

First it will be convenient to partition the parent variables of any pair (X_i, X_j) as $\Pi_i \cup \Pi_j = \{L_A, L_B, L_C\}$, where:

$$\begin{aligned} A &= \Pi_i - \Pi_j \\ B &= \Pi_j - \Pi_i \\ C &= \Pi_i \cap \Pi_j. \end{aligned} \tag{2.12}$$

Then we may re-cast Equation (1.1) for any pair $(\widetilde{X}_i, \widetilde{X}_j)$ as follows:

$$\begin{bmatrix} \widetilde{X}_i \\ \widetilde{X}_j \end{bmatrix} = \begin{bmatrix} \widetilde{\Lambda}_{iA} & \mathbf{0} & \widetilde{\Lambda}_{iC} \\ \mathbf{0} & \widetilde{\Lambda}_{jB} & \widetilde{\Lambda}_{jC} \end{bmatrix} \begin{bmatrix} L_A \\ L_B \\ L_C \end{bmatrix} + \begin{bmatrix} \widetilde{\epsilon}_i \\ \widetilde{\epsilon}_j \end{bmatrix}. \tag{2.13}$$

We then obtain the correlation of between X_i and X_j from this form as follows:

$$\text{Var}\left(\begin{bmatrix} \widetilde{X}_i \\ \widetilde{X}_j \end{bmatrix}\right) = \begin{bmatrix} \widetilde{\Lambda}_{iA} & \mathbf{0} & \widetilde{\Lambda}_{iC} \\ \mathbf{0} & \widetilde{\Lambda}_{jB} & \widetilde{\Lambda}_{jC} \end{bmatrix} \begin{bmatrix} \Phi_{AA} & \Phi_{AB} & \Phi_{AC} \\ \Phi_{BA} & \Phi_{BB} & \Phi_{BC} \\ \Phi_{CA} & \Phi_{CB} & \Phi_{CC} \end{bmatrix} \begin{bmatrix} \widetilde{\Lambda}_{iA}^T & \mathbf{0} \\ \mathbf{0} & \widetilde{\Lambda}_{jB}^T \\ \widetilde{\Lambda}_{iC}^T & \widetilde{\Lambda}_{jC}^T \end{bmatrix} + \begin{bmatrix} \widetilde{\omega}_i & 0 \\ 0 & \widetilde{\omega}_j \end{bmatrix}, \tag{2.14}$$

for which we multiply through and take the off-diagonal to be:

$$\rho_{ij} = \widetilde{\Lambda}_{iA}\Phi_{AB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iC}\Phi_{CB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iA}\Phi_{AC}\widetilde{\Lambda}_{jC}^T + \widetilde{\Lambda}_{iC}\Phi_{CC}\widetilde{\Lambda}_{jC}^T. \tag{2.15}$$

Writing ρ_{ij} in this way yields a useful decomposition with respect to the structure of the factor analysis model. Specifically, this can be thought of as the correlation between X_i and X_j due to their non-shared parents being correlated (Φ_{AB}), their non-shared parents being correlated with their shared parents (Φ_{AC}, Φ_{CB}) and simply having shared parents (Φ_{CC}). Thus, if X_i and X_j have no shared parents, then the index set C is empty. This reduces Equation (2.15) to:

$$\rho_{ij} = \widetilde{\Lambda}_{iA}\Phi_{AB}\widetilde{\Lambda}_{jB}^T. \tag{2.16}$$

The result of Lemma 2 follows by characterizing the definition of thresholdability (2.6)

directly in terms of θ. That is, if for all (X_i, X_j) that share parents and for all (X_k, X_l) that do not share parents, θ is thresholdable if and only if:

$$\max_{(k,l)\in E_0^c} |\widetilde{\Lambda}_{kE}\Phi_{EF}\widetilde{\Lambda}_{lF}^T| < \min_{(i,j)\in E_0} |\widetilde{\Lambda}_{iA}\Phi_{AB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iC}\Phi_{CB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iA}\Phi_{AC}\widetilde{\Lambda}_{jC}^T + \widetilde{\Lambda}_{iC}\Phi_{CC}\widetilde{\Lambda}_{jC}^T|. \tag{2.17}$$

□

We first note that the Gaussian assumption is not needed for Lemma 2. Rather, it depends only on the correlation structure $\widetilde{\Sigma}(\theta)$ and is agnostic to the underlying distribution. Second, to illustrate the application of Lemma 2 we will consider several scenarios of interest. Let us begin with the case where the latent variables are orthogonal.

Corollary 1. *If $\Phi = I_d$, then θ is thresholdable.*

Proof. From Equation (2.11), we can see that if $\Phi = I_d$, then the Φ_{AB}, Φ_{CB}, Φ_{AC}, and Φ_{EF} matrices are all zero matrices, and Φ_{CC} is an identity matrix. Thus Equation 2.11 reduces to

$$0 < \min_{(i,j)\in E_0} |\widetilde{\Lambda}_{iC}\widetilde{\Lambda}_{jC}^T|, \tag{2.18}$$

which trivially holds. □

This can be seen from a straightforward substitution of $\Phi = I_d$ into Inequality 2.11, where the left side is zero and the right side positive. Another common scenario is when Λ has exactly one non-zero entry per row, i.e., $|\Pi_i| = 1$, for all $i \in [p]$. This is called "independent cluster structure" (Harris and Kaiser, 1964) or "perfect simple structure" (Jennrich, 2006). Under this structure, the children sets of latent variables are mutually exclusive, which is a common design for factor analysis models due to its ease of causal interpretability. Such structures lead to a simplification of the thresholdability condition as shown in the following corollary.

Corollary 2. *If Λ has exactly one non-zero entry per row, then θ is thresholdable if*

$$\max_{(k,l)\in E_0^c} |\widetilde{\lambda}_{ke}\widetilde{\lambda}_{lf}\phi_{ef}| < \min_{(i,j)\in E_0} |\widetilde{\lambda}_{ic}\widetilde{\lambda}_{jc}|, \tag{2.19}$$

where $\Pi_i = \Pi_j = \{c\}$, $\Pi_k = \{e\}$, *and* $\Pi_l = \{f\}$.

Proof. The defining characteristic of the independent cluster structure is that Λ has exactly one non-zero entry. This implies that each observed variable has only one latent variable parent. Thus, the relevant parent sets will reduce to $\Pi_i = \Pi_j = \{c\}$, $\Pi_k = \{e\}$, and $\Pi_l = \{f\}$. That is, each pair of observed variables will either have one shared parent, or no shared parents, but not both. Hence for each pair of variables that share parents, the Φ_{AB}, Φ_{CB}, and Φ_{AC} matrices will not exist and $\Phi_{CC} = 1$. Corollary 2 follows by simplifying Equation 2.11 with these reductions. □

Following Corollary 2, we can further observe that if Λ follows an independent cluster structure and $\widetilde{\lambda}_{ij}$'s are homogeneous in magnitude, then θ is thresholdable.

Corollary 3. *If* Λ *has exactly one non-zero entry per row and* $|\widetilde{\lambda}_{ij}| = \lambda$ *for all* $(i, j) \in \mathcal{A}(\Lambda)$, *then* θ *is thresholdable.*

Proof. The result follows from Corollary 2 and setting all $|\widetilde{\lambda}_{ij}| = \lambda$. The condition for thresholdability then becomes

$$\max_{(k,l)\in E_0^c} |\lambda^2 \phi_{ef}| < \min_{(i,j)\in E_0} |\lambda^2|, \tag{2.20}$$

which holds since $\phi_{ef} \in [0, 1)$. Note we exclude the pathological case of $\phi_{ef} = 1$ as this is not distinguishable from considering L_e and L_f as the same latent variable. □

Corollaries 1, 2 and 3 involve desirable properties of factor analytic designs. First, it has been suggested that latent variable models should be designed such that the latent factors be distinguishable from one another, or that they are not too highly correlated (Whitely, 1983). If the latent factors are too highly correlated, then a factor solution with less dimensions may be better suited. Second, an independent cluster structure yields mutually exclusive subsets of children for each latent variable. In other words, each observed variable provides a "measurement" of a single latent variable alone. This design is very common in educational and psychological test construction (Hattie, 1985;

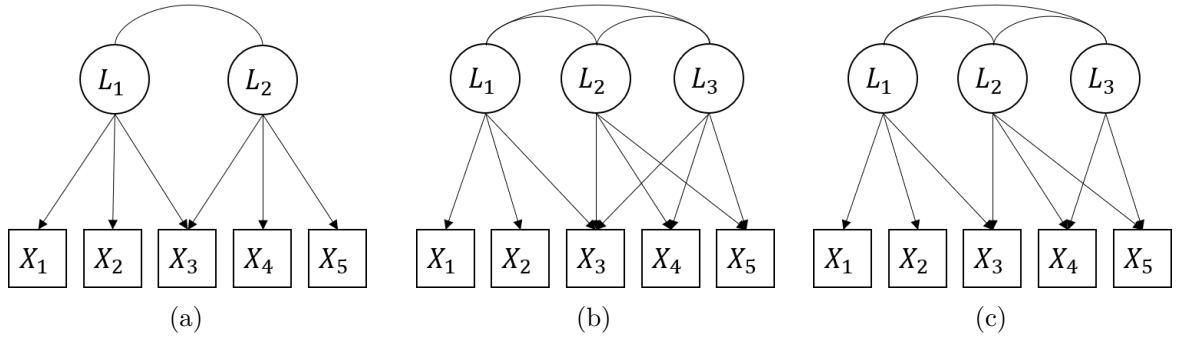

Figure 2.3: Three structures that yield the same graph $\mathcal{G}(X, E_0)$.

Anderson and Gerbing, 1988), and much methodology has been devoted to seeking these types of factor solutions (Scheines et al., 1998; Jennrich, 2001, 2006; Silva et al., 2006).

2.3.2 Structural Identifiability

In this section we study the conditions under which the structure for Λ can be recovered from the thresholded correlation graph. To demonstrate the problem of structural identifiability, consider some counter examples in Figure 2.3. All these structures will yield the same graph with edge set E_0 (Equation 2.4). Specifically, the independent maximal cliques that are yielded by them are $\{1, 2, 3\}$ and $\{3, 4, 5\}$, despite all having different structures. This can be seen by noting that some latent variables do not yield maximal cliques in $\mathcal{G}(X, E_0)$, or yield the same independent maximal clique as another latent variable. For example, in Figure 2.3b, both L_2 and L_3 yield the clique $\{3, 4, 5\}$. Thus, L_2 and L_3 cannot be distinguished from each other through independent maximal cliques alone. Similarly, in Figure 2.3c, L_3 yields the clique $\{4, 5\}$, but it is not maximal since L_2 yields $\{3, 4, 5\}$. In this case, L_3 cannot be identified as a latent variable, since its clique is subsumed by the one yielded by L_2. Hence, we must consider the problem of multiple structures corresponding to the same edge set E_0.

It would be ideal to find a one-to-one correspondence between the latent variables and the independent maximal cliques (Definition 2). If such a correspondence holds for a given Λ, we will call Λ (or θ) *independent maximal clique identifiable*. It turns out that the following simple condition is sufficient for this correspondence to hold:

Condition 1 (Unique Child Condition)**.** Let the *child set* of a latent variable be denoted

$ch(L_k) = \{i \in [p] : \lambda_{ik} \neq 0\}$. If

$$U_k \coloneqq ch(L_k) - \bigcup_{j \neq k} ch(L_j) \neq \emptyset, \qquad \forall\, k \in [d], \tag{2.21}$$

i.e., if each latent variable L_k has a non-empty set of unique children U_k, then we say that the *unique child condition* holds. It essentially means that all latent parents have at least one unique child variable.

Given this condition, we can obtain a bijection between the latent variables and the independent maximal cliques in $\mathcal{G}(X, E_0)$. We state this in the following lemma.

Lemma 3. *If the unique child condition holds in Λ (Condition 1), then the set $\{\mathrm{ch}(L_k) : k \in [d]\}$ is identical to the set of independent maximal cliques in $\mathcal{G}(X, E_0)$.*

Proof. Recall the definition of E_0, which we re-state for convenience:

$$E_0 \coloneqq \{(i, j) : \Pi_i \cap \Pi_j \neq \emptyset\}.$$

Pick any $k \in [d]$. By definition, every $X_j \in ch(L_k)$ shares a common parent L_k and thus $ch(L_k)$ forms a clique in $\mathcal{G} = \mathcal{G}(X, E_0)$. Let U_k be the set of unique children of L_k. Under the unique child condition, $U_k \neq \emptyset$, so we can pick an $X_i \in U_k$. Then X_i does not have an edge connected to any node other than $ch(L_k)$ by the definition of E_0. This implies every clique that includes X_i must be a subset of $ch(L_k)$. Thus, $ch(L_k)$ is the only maximal clique that includes X_i, making it an independent maximal clique. The above argument shows that each $ch(L_k), k \in [d]$ is an independent maximal clique. Since $\cup_k ch(L_k) = X$, any other maximal clique, if it exists, cannot be independent, and thus, $\{ch(L_k) : k \in [d]\}$ is the set of independent maximal cliques in $\mathcal{G}$. □

Recall the key observation that the dimension of L and support of Λ, i.e. $(d, \mathcal{A}(\Lambda))$, completely encodes the structure of the model. The CT algorithm leverages Lemma 3 to recover the structure of a factor analysis model $(d, \mathcal{A}(\Lambda))$ by finding independent maximal cliques in an estimated graph $\mathcal{G}(X, \hat{E}(\tau_k))$.

Remark 3. We note our use of $\mathcal{A}(\Lambda)$ defines a model structure up to a column permutation of Λ. That is, we consider different ordering or labeling of the factors to be equivalent, since they define the same $\Sigma(\theta)$ in Equation 1.2.

We note that the unique child condition is practical for a wide range of factor analysis designs. In Section 2.3.1 and Corollary 2, we discussed the independent cluster structure (one non-zero per row of Λ) as a commonly used design. The unique child condition is much more general in comparison, since it only requires a single unique child as opposed to all children being unique to their parents. Further, observed variables can be designed to fulfill the unique child condition a priori, which is typical for psychometric settings (Hattie, 1985; Anderson and Gerbing, 1988).

2.3.3 Rotational Uniqueness

An important consideration with a factor analysis model is the identifiability of the parameters $\theta = \{\Lambda, \Phi, \Omega\}$. The lack of rotational uniqueness implies that there may be many (Λ, Φ) pairs that exist such that $\Sigma(\theta) = \Lambda\Phi\Lambda^T + \Omega$. To see this, let M be a $d \times d$ invertible matrix. Then we have

$$\begin{aligned} \Sigma(\theta) &= \Lambda\Phi\Lambda^T + \Omega \\ &= \Lambda M M^{-1}\Phi M^{-T} M^T \Lambda^T + \Omega \\ &= \Lambda_M \Phi_M \Lambda_M^T + \Omega, \end{aligned} \tag{2.22}$$

letting $\Lambda_M = \Lambda M$ and $\Phi_M = M^{-1}\Phi M^{-T}$. To ensure that Φ_M remains a valid correlation matrix, we impose the constraint that $\text{diag}(M^{-1}\Phi M^{-T}) = I_d$. However, we will show that the solutions learned by the CT algorithm resolves this nonidentifiability issue given that the zero constraints implied by $\mathcal{A}(\hat{\Lambda})$ are preserved. Formally, we define the notion of rotational uniqueness as follows.

Definition 3 (Rotational Uniqueness)**.** For a set of parameters $\theta = \{\Lambda, \Phi, \Omega\}$, denote a rotated set of parameters as $\theta_M = \{\Lambda M, M^{-1}\Phi M^{-T}, \Omega\}$, where M is an invertible $d \times d$

matrix. Let us define a set of *constraint preserving rotations* as

$$\mathcal{M}_{CP} = \mathcal{M}_{CP}(\theta) \coloneqq \{M : \Sigma(\theta_M) = \Sigma(\theta), \mathcal{A}(\Lambda M) \subseteq \mathcal{A}(\Lambda), \text{diag}(M^{-1}\Phi M^{-T}) = I_d\}. \tag{2.23}$$

Then:

1. If $\mathcal{M}_{CP} = \{I_d\}$, then θ is said to be *globally rotationally unique*.

2. If $\mathcal{M}_{CP}$ is a set of signature matrices, then θ is said to be *locally rotationally unique*, where signature matrices are diagonal matrices whose diagonal elements are ± 1.

First, note that the condition $\mathcal{A}(\Lambda M) \subseteq \mathcal{A}(\Lambda)$ ensures that the zero constraints implied by $\mathcal{A}(\Lambda)$ are persevered. Second, all matrix factorizations will have the signature matrix rotation as a source of non-uniqueness unless the signs of the main diagonal (or a permutation thereof) are fixed and non-zero. Since the model in Equation 1.1 makes no assumptions regarding the signs in Λ, local rotational uniqueness is the best type of rotational uniqueness we can establish. Two local rotational uniqueness properties relevant to the CT algorithm are described in Corollaries 4 and 5.

Corollary 4. *If the unique child condition holds in Λ, then θ is locally rotationally unique.*

Proof. Define an index set for the rows of $\Lambda \in \mathbb{R}^{p \times d}$ which have zeroes in the jth column as

$$Z_j \coloneqq \{i : \lambda_{ij} = 0\} \subseteq [p],$$

and define

$$\Lambda^{[j]} \coloneqq \Lambda_{Z_j, -j},$$

which is a submatrix of size $|Z_j| \times (d-1)$. Adapted from Peeters (2012), two sufficient conditions for Λ that yield local rotational uniqueness for our model are:

Condition 1: Λ has at least $d-1$ zeroes in each column.

Condition 2: $\text{rank}(\Lambda^{[j]}) = d-1$ for all $j \in [d]$.

An example of $\Lambda^{[j]}$ is as follows:

$$\Lambda = \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & 0 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{52} & \lambda_{53} \\ 0 & \lambda_{62} & 0 \\ 0 & 0 & \lambda_{73} \\ 0 & 0 & \lambda_{83} \\ \lambda_{91} & 0 & \lambda_{93} \end{bmatrix}, \quad \Lambda^{[1]} = \begin{bmatrix} \lambda_{42} & 0 \\ \lambda_{52} & \lambda_{53} \\ \lambda_{62} & 0 \\ 0 & \lambda_{73} \\ 0 & \lambda_{83} \end{bmatrix}, \quad \Lambda^{[2]} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{73} \\ 0 & \lambda_{83} \\ \lambda_{91} & \lambda_{93} \end{bmatrix}, \quad \Lambda^{[3]} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{62} \end{bmatrix}. \tag{2.24}$$

These conditions can be seen to be satisfied by the unique child condition as follows. Let U_j be the set of unique children for L_j as defined in Equation (2.21). For all $j, k \in [d]$, and $i \in [p]$ we can re-cast U_j as:

$$U_j = \{i : \lambda_{ij} \neq 0, \lambda_{ik} = 0, k \neq j\}, \tag{2.25}$$

and let the index of non-unique variables be:

$$\overline{U} = \{i : i \notin \cup_{j=1}^{d} U_j\}. \tag{2.26}$$

Let us permute the rows of Λ according to an order that satisfies $(U_1, \ldots, U_d, \overline{U})$. Denoting a permutation matrix that yields such a row ordering as P, we have:

$$P\Lambda = \begin{bmatrix} \Lambda_{U_1 1} & & \\ & \ddots & \\ & & \Lambda_{U_d d} \\ \Lambda_{\overline{U}1} & \cdots & \Lambda_{\overline{U}d} \end{bmatrix}. \tag{2.27}$$

That is, we can permute the rows of Λ such that its upper part is block-diagonal with d

blocks. Then there must be at least $d-1$ zeroes in each column, satisfying Condition 1. It is easily seen that $P\Lambda$ also satisfies Condition 2, as any $(P\Lambda)^{[j]}$ will also have its upper part be block-diagonal, and thus full rank $(d-1)$. □

Corollary 5. *Any $\hat{\theta}_k$ for $k \in [m]$, produced by Step 8 of the CT algorithm, is locally rotationally unique.*

Proof. As described in Section 2.2, Steps 5 through 7 of the CT algorithm construct the support $\hat{A}_k$ deterministically based on a set of independent maximal cliques $\mathcal{C}_k$ (from Step 2). Since by Definition 2 independent maximal cliques always have a unique node, the sparsity pattern in $\hat{A}_k$ is guaranteed to follow the unique child condition (Condition 1). By Corollary 4, $\hat{\theta}_k$ will be locally rotationally unique due to this pattern. □

Note that Corollary 5 holds regardless if Condition 1 is true in the population structure. Thus the CT algorithm can be used as a model approximation tool for finding locally rotationally unique structures.

2.3.4 Error Bounds and Consistency

In this section, we establish the consistency of the CT algorithm. The crucial part of the argument depends on the structure learning consistency of the algorithm (Steps 3 through 7, described in Section 2.2). We will call a structural estimate $(\hat{d}, \mathcal{A}(\hat{\Lambda}))$ consistent if

$$\lim_{n\to\infty} \mathbb{P}\Big[(\hat{d}, \mathcal{A}(\hat{\Lambda})) = (d, \mathcal{A}(\Lambda))\Big] = 1, \tag{2.28}$$

given an i.i.d. sample of size n from the model in Equation (1.1). By Lemma 3, the model structure $(\mathcal{A}(\Lambda), d)$ can be recovered exactly from the set of independent maximal cliques in $\mathcal{G}(X, E_0)$ when the unique child condition holds. Therefore, structural consistency holds when $\lim_{n\to\infty} \mathbb{P}(\hat{E}(\tau_0) = E_0) = 1$ under the unique child condition for a suitable τ_0, where $\hat{E}(\tau)$ is the edge set of an estimated correlation thresholded graph as defined in Equation (2.7). In what follows, it will be useful to define a gap of separation for a

thresholdable θ as

$$\gamma \coloneqq \frac{1}{2}\left[\min\{|\rho_{ij}| : (i,j) \in E_0\} - \max\{|\rho_{ij}| : (i,j) \in E_0^c\}\right]. \tag{2.29}$$

Theorem 1. *Assume the model described in Equation* (1.1) *holds for* X *and that the correlations between all pairs* (X_i, X_j) *are bounded such that* $\max_{i\neq j}|\rho_{ij}| \leq M < 1$*. If* θ *is thresholdable with a gap* $\gamma > 0$*, then*

$$\mathbb{P}(\hat{E}(\tau_0) \neq E_0) \leq Cp(p-1)(n-2)\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4} := \eta, \tag{2.30}$$

where $0 < C < \infty$ *only depends on* M*. If additionally the unique child condition holds (Condition 1), then we have*

$$\mathbb{P}((\hat{d}, \mathcal{A}(\hat{\Lambda})) = (d, \mathcal{A}(\Lambda))) \geq 1 - \eta, \tag{2.31}$$

where $(\hat{d}, \mathcal{A}(\hat{\Lambda}))$ *is the estimated model structure by the CT algorithm with cutoff* τ_0*.*

To obtain our result, we will leverage existing estimation error bounds on the event $|r_{ij} - \rho_{ij}| \geq \epsilon$ for some $\epsilon > 0$. To do this it will be convenient to re-cast our event of interest to $\hat{E}(\tau_0) \neq E_0$. For clarity, let us first consider the event $\hat{E}(\tau_0) = E_0$, which by definition, holds if and only if:

$$\Big(\bigcap_{(i,j)\in E_0} |r_{ij}| > \tau_0\Big) \cap \Big(\bigcap_{(i,j)\in E_0^c} |r_{ij}| < \tau_0\Big). \tag{2.32}$$

Then by De Morgan's laws, we can say $\hat{E}(\tau_0) \neq E$ if and only if:

$$\Big(\bigcup_{(i,j)\in E_0} |r_{ij}| \leq \tau_0\Big) \cup \Big(\bigcup_{(i,j)\in E_0^c} |r_{ij}| \geq \tau_0\Big), \tag{2.33}$$

which is to say that $\hat{E}(\tau_0) \neq E_0$ holds if and only if any r_{ij} is on the opposite side of τ_0 as their population analog ρ_{ij}. From here, the strategy is to derive bounds for $\mathbb{P}(|r_{ij}| \leq \tau_0)$ if $(i,j) \in E_0$, and $\mathbb{P}(|r_{ij}| \geq \tau_0)$ if $(i,j) \in E_0^c$, for all (i,j). To determine these bounds, we

make use of a concentration inequality for $\mathbb{P}(|r_{ij} - \rho_{ij}| \geq \epsilon)$ from Lemma 1 of Kalisch and Bühlmann (2007). We re-state this as follows:

Lemma 4. *Assuming X_i and X_j are Gaussian random variables with correlation $|\rho_{ij}| \leq M < 1$. Let r_{ij} be the sample correlation calculated from an i.i.d. sample of size n. Then for any $0 < \epsilon \leq 2$,*

$$\mathbb{P}(|r_{ij} - \rho_{ij}| \geq \epsilon) \leq C_0(n-2)\left(\frac{4-\epsilon^2}{4+\epsilon^2}\right)^{n-4}, \tag{2.34}$$

where $0 < C_0 < \infty$ only depends on M.

For our purposes, we set $\epsilon = \gamma$ and select as τ_0 the mid-point of $\min_{E_0}(|\rho_{ij}|)$ and $\max_{E_0^c}(|\rho_{ij}|)$, which will be the best choice to uniformly bound all $\mathbb{P}(|r_{ij}| \leq \tau_0)$ if $(i,j) \in E_0$ and $\mathbb{P}(|r_{ij}| \geq \tau_0)$ if $(i,j) \in E_0^c$. The uniformity of the bound follows by seeing that $\gamma \leq \left||\rho_{ij}| - \tau_0\right|$ for all (i,j). That is, there is no ρ_{ij} that is closer to τ_0 than the length of γ.

We begin with the scenario where $(i,j) \in E_0^c$. Given the left-hand side of Equation (2.34) and setting $\epsilon = \gamma$, we have:

$$\begin{aligned}
\mathbb{P}(|r_{ij} - \rho_{ij}| \geq \gamma) &\geq \mathbb{P}(|r_{ij}| - |\rho_{ij}| \geq \gamma) \\
&\geq \mathbb{P}(|r_{ij}| - |\rho_{ij}| \geq \tau_0 - |\rho_{ij}|) \\
&= \mathbb{P}(|r_{ij}| \geq \tau_0).
\end{aligned} \tag{2.35}$$

Hence, $\mathbb{P}(|r_{ij}| \geq \tau_0)$ is bounded from above by the right-hand side of Equation (2.34) if $(i,j) \in E_0^c$. We can use the same strategy to conclude that, for $(i,j) \in E_0$,

$$\mathbb{P}(|r_{ij} - \rho_{ij}| \geq \gamma) \geq \mathbb{P}(|r_{ij}| \leq \tau_0). \tag{2.36}$$

Since these two events have the same upper bound, let us combine them by defining:

$$B_{ij} = B(r_{ij}, \tau_0) := \begin{cases} |r_{ij}| \leq \tau_0 & \text{if } (i,j) \in E_0 \\ |r_{ij}| \geq \tau_0 & \text{if } (i,j) \in E_0^c \end{cases}. \tag{2.37}$$

Noting that $\hat{E}(\tau_0) \neq E(\tau_0)$ holds if and only if $\bigcup_{(i,j)} B_{ij}$ holds, what remains is to find a bound of the latter event. This can be done with the union bound:

$$\begin{aligned}\mathbb{P}(\hat{E}(\tau_0) \neq E(\tau_0)) = \mathbb{P}\Big(\bigcup_{(i,j)} B_{ij}\Big) &\leq \sum_{(i,j)} \mathbb{P}\left(B_{ij}\right) \\ &\leq \frac{p(p-1)}{2} \max_{(i,j)}\{\mathbb{P}(B_{ij})\} \\ &\leq Cp(p-1)(n-2)\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4},\end{aligned} \tag{2.38}$$

where $0 < C < \infty$ only depends on M. This result follows by recognizing that all $\mathbb{P}(B_{ij})$ are uniformly bounded as in Lemma 4. Finally, this implies

$$\mathbb{P}(\hat{E}(\tau_0) = E_0) \geq 1 - Cp(p-1)(n-2)\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4} \tag{2.39}$$

and thus, (2.31) follows immediately under the unique child condition by Lemma 3. □

Due to the exponential decay of the term $[(4-\gamma^2)/(4+\gamma^2)]^{n-4}$, consistency is trivially implied under a fixed p regime. More generally speaking, for any joint distribution of X under which the central limit theorem holds for the sample correlations $\{r_{ij}\}$, structural consistency would also follow. By the classical central limit theorem and the delta method, this would include the class of distributions with finite fourth-order moments (Ferguson, 1996). Furthermore, we will use the bound described in Inequality 2.30 to develop a consistency result with high-dimensional accommodations where the dimension $p = p_n \gg n$.

Theorem 2. *Assume the model described in Equation* (1.1) *holds for* X *and that the correlations between all pairs* (X_i, X_j) *are bounded such that* $\max_{i\neq j}|\rho_{ij}| \leq M < 1$ *for some universal constant* M *independent of* n. *If* θ *is thresholdable with a gap* $\gamma = \gamma_n$ *such that* $\gamma_n^2 \geq c_1/(n-4)^b$ *for some* $c_1 > 0$ *and* $b \in [0, 1)$ *when* n *is large, and* $p_n = o(\exp(c(n-4)^{1-b}))$, *where* $0 < c < c_1/8$, *then*

$$\lim_{n\to\infty} \mathbb{P}(\hat{E}(\tau_0) = E_0) = 1. \tag{2.40}$$

If additionally the unique child condition holds (Condition 1), then we also have

$$\lim_{n\to\infty} \mathbb{P}\Big[(\hat{d}, \mathcal{A}(\hat{\Lambda})) = (d, \mathcal{A}(\Lambda))\Big] = 1. \tag{2.41}$$

To begin, we will first examine the growth of a lower bound of $\mathbb{P}(\hat{E}(\tau_0) = E_0)$ as a function of n. Noting from Equation (2.30), an upper bound on the decaying term with n can be derived as follows:

$$\begin{aligned}
\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4} &\leq \left(1 - \frac{\gamma^2}{4}\right)^{n-4} \\
&\leq \left(1 - \frac{c_1}{4(n-4)^b}\right)^{n-4} \\
&= \left(1 - \frac{c_1}{4(n-4)^b}\right)^{(n-4)^b(n-4)^{1-b}} \\
&= \left(\exp\left(-\frac{c_1}{4}\right) + o(1)\right)^{(n-4)^{1-b}} \\
&\leq \exp\left(-\frac{c_2(n-4)^{1-b}}{4}\right),
\end{aligned} \tag{2.42}$$

where we used the limit $\lim_{x\to\infty}(1 + a/x)^x = \exp(a)$ and another constant $c_2 \in (0, c_1)$ such that the $o(1)$ remainder can be dropped. From here, we can form a looser bound on Equation (2.30) as

$$\begin{aligned}
\mathbb{P}(\hat{E}(\tau_0) = E_0) &\geq 1 - Cp(p-1)(n-2)\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4} \\
&\geq 1 - Cp_n(p_n-1)(n-2)\exp\left(-\frac{c_2(n-4)^{1-b}}{4}\right) \\
&= 1 - p(n)f(n),
\end{aligned} \tag{2.43}$$

where $p(n) = p_n(p_n - 1)$ and $f(n) = (n-2)\exp(-c_2(n-4)^{1-b}/4)$. Therefore, we have consistency if $\lim_{n\to\infty} p(n)f(b) = 0$ or if $p(n) = o(1/f(n))$. Comparing the dominating terms of $p(n)$ and $1/f(n)$, consistency is achieved if

$$\begin{aligned}
p_n^2 &= o\left(\exp\left[\frac{c_2(n-4)^{1-b}}{4} - \log n\right]\right) \\
\text{or if } p_n &= o\left(\exp\left[c(n-4)^{1-b}\right]\right),
\end{aligned} \tag{2.44}$$

by choosing a positive constant $c < c_2/8$. □

Note that any fixed value between $\max_{E_0^c}\{|\rho_{ij}|\}$ and $\min_{E_0}\{|\rho_{ij}|\}$ will be a valid choice for τ_0 for structure learning consistency. This result is straightforward to generalize to non-Gaussian forms of X, which could result from non-Gaussian combinations of L and ϵ. All that would be required is to replace our use of Lemma 4 (a Gaussian sample correlation concentration bound) in the proofs of Theorems 1 and 2 with a bound for any non-Gaussian X of interest. So long as this bound is sufficiently well-behaved, the probability bounds in Theorem 1 will hold as will Theorem 2 with different dependencies between p and n.

In the practical context of the CT algorithm, recall that a suitable τ_0 is actually unknown, and the algorithm estimates and selects among a set of models based on a candidate set $\{\tau_k\}$. Assuming that a suitable τ_0 is contained in $\{\tau_k\}$, structural identifiability and consistency implies that the correct model structure is among the set of candidate models, asymptotically. From here, overall parameter consistency follows by simply using a consistent parameter estimation method (Step 8) and a consistent model selection procedure (Step 10) in the algorithm. A straightforward choice would be to use maximum likelihood estimation in conjunction with BIC model selection. Then, asymptotically, the CT algorithm will produce the correct model structure with consistent parameter estimates. One caveat of these results is the reliance on the thresholdability assumption. However, even if thresholdability is violated to a certain degree, the estimated model structure can still be quite accurate as we will show empirically in the simulation studies.

2.4 Simulation Studies

To begin our empirical simulations, we first sought to verify the performance of the CT algorithm under ideal conditions (Section 2.4.1). As such, we designed conditions under which all the assumptions of the CT algorithm were met and compared it against several other existing methods. In Section 2.4.2, we consider situations in which some of these

assumptions are violated, and study the robustness of the CT algorithm. Finally, in Sections 2.4.3 and 2.4.4, we consider high-dimensional ($n < p$) and large p settings, where p and d grow with n. We study how violations to thresholdability and departures from the unique child condition affect model recovery in the high-dimensional setting.

Where possible, we tested the performance of the CT algorithm with a few other methods of factor analysis structure learning. These methods were the MLE with known structure (to serve as a baseline), EFA (described in Section 1.2), EFA-LASSO, and EFA-MCP (both from Hirose and Yamamoto, 2014a). EFA-LASSO and EFA-MCP maximize a penalized likelihood function of the form

$$\ell_p(\theta) = \ell(\theta) - p(\Lambda), \tag{2.45}$$

where $\ell(\theta)$ is the log-likelihood in Equation (1.5) and $p(\Lambda)$ is the LASSO (Tibshirani, 1996) and MCP (Zhang, 2010) penalty functions, respectively.

Note that the three EFA methods all require d as an input. To make a comparison as fair as possible, the CT algorithm was used to give the EFA methods a set of d to work with, which was more informative than ad hoc choices. More specifically, we ran the CT algorithm to Step 3, where d is estimated from the number of independent maximal cliques. Thereafter, we replaced the support learning portion (Steps 5 through 7) of the algorithm with one of the EFA procedures. Then the support of the model was saved from the EFA methods and was used to resume the algorithm from Step 8, where the MLE was estimated from the support.

The simulations were done in the `R` language (4.0.2; R Core Team, 2020). The `lavaan` package (Rosseel, 2012) was used in the estimation phases of the CT algorithm (Step 8), and was used to estimate the baseline MLE solution. For the cutoffs τ_k, 40 equidistant points from 0 to 1 were input for the CT algorithm. For EFA, the `psych` package (Revelle, 2019) was used to obtain MLE solutions for unconstrained Λ. We left the rotation option to the package default oblimin method (Crawford, 1975), however we note that the rotation choice does not affect the results since we will only be examining the likelihood of $\Sigma(\hat{\theta})$. And finally, the LASSO and MCP variants of EFA were estimated

with the `fanc` package (Hirose and Yamamoto, 2014b,a). The tuning parameters were left at the package defaults of 30 values for a single tuning parameter in LASSO and 270 combinations of two tuning parameters in MCP.

2.4.1 Basic Simulation Study

We generated data sets from a Gaussian distribution. The mean vector was set to $\mu = \mathbf{0}$ for all conditions, and the covariance matrix Σ was parameterized by θ which varied by condition. The number of latent variables (d) was set to 2, 3, 4, and 5, with the structure of Λ set up as follows. We began with assigning five children to each latent variable in mutually exclusive sets ($p = 5d$). In the children set of each latent variable, one variable was designated the unique child of that latent variable to enforce structural identifiability. Then each of the non-unique child variables was assigned to have an extra parent with 0.5 probability, with the parent chosen with uniform probability. The non-zero entries of Λ were drawn from a uniform distribution, $\lambda_{ij} \sim \text{Uniform}(0.4, 0.5)$. To generate Φ, we began by setting its diagonals to one. Then for the off-diagonal elements, we drew a $d \times d$ matrix A with entries from $\text{Uniform}(0, 1)$ and rescaled it such that $A^T A$ had off-diagonals in the range of $[0.1, 0.125]$, a quarter of the magnitudes of λ_{ij}. Then the off-diagonals of Φ were set to the off-diagonals of this rescaled $A^T A$, which ensured Φ would be positive definite and θ would be thresholdable. To enforce thresholdability in the sample, we checked if there existed a τ such that $\hat{E}(\tau) = E_0$, and rejected samples for which this did not hold. Finally, we generated two data sets, each with sample size $n = 1000$, per replication, one for training purposes and one for testing purposes, using 100 replications for each value of d.

We collected several metrics to evaluate performance in terms of true model recovery and computational efficiency. For model fit, we calculated the test data log-likelihood differences from that of the known structure MLE solution. To measure the accuracy of the estimated model structure, we collected a Hamming distance (HD) from the true support of Λ and F_1 score of the estimated support (both described below), and the learned dimension ($\hat{d}$) of the latent variable vector. For the penalized EFA methods, the

number of non-zero columns in $\hat{\Lambda}$ was taken as $\hat{d}$ as they would serve as the de facto number of latent variables (Caner and Han, 2014). Finally, for computational efficiency, we calculated the number of models estimated by each method. This metric was chosen to be agnostic toward the numerical idiosyncrasies between the different software packages.

To compare the estimated and true supports ($\mathcal{A}(\hat{\Lambda})$ vs. $\mathcal{A}(\Lambda)$) we computed the minimum HD over all column permutations of $\hat{\Lambda}$. That is, we define an HD as

$$\text{HD} \coloneqq \min_{P} \left[|\mathcal{A}(\hat{\Lambda}P) \triangle \mathcal{A}(\Lambda)| \right], \tag{2.46}$$

where $\triangle$ is the symmetric difference or disjunctive union between two sets. The permutation matrix P reconciles the fact that the column order of $\hat{\Lambda}$ may not be the same as the column order of Λ, and that $\hat{d}$ may not be the same as d. Put another way, HD is the smallest number of element additions and deletions needed to make the sets $\mathcal{A}(\Lambda)$ and $\mathcal{A}(\hat{\Lambda})$ identical, among all column permutations of $\hat{\Lambda}$.

In addition to HD, we also report the F_1 score, a normed measure of classification. This allows for comparability between models with differing dimensions of Λ, that is differing p and d. Note that the F_1 score is simply the harmonic mean between precision and recall. Once again using a permutation matrices to reconcile different orderings of L, we have

$$F_1(\hat{\Lambda}) \coloneqq \max_{P} \left[\frac{2|\mathcal{A}(\hat{\Lambda}P) \cap \mathcal{A}(\Lambda)|}{2|\mathcal{A}(\hat{\Lambda}P) \cap \mathcal{A}(\Lambda)| + |\mathcal{A}(\hat{\Lambda}P) \triangle \mathcal{A}(\Lambda)|} \right] \in [0, 1], \tag{2.47}$$

and the higher the F_1 score, the more accurate the estimated support of $\hat{\Lambda}$.

To measure computational efficiency we simply counted the number of models each method estimated. For the CT algorithm, this is simply the number of unique structures obtained by the sequence of τ_k. For EFA, this translates to the number of unique d obtained by the sequence of τ_k. For EFA-LASSO and EFA-MCP, this is the number of tuning parameter combinations to search over (30 for LASSO, 270 for MCP), per unique d in the sequence of τ_k.

The average results of the evaluation metrics are displayed in Figure 2.4. In terms of model fit and structure estimation, the CT algorithm almost always performed perfectly,

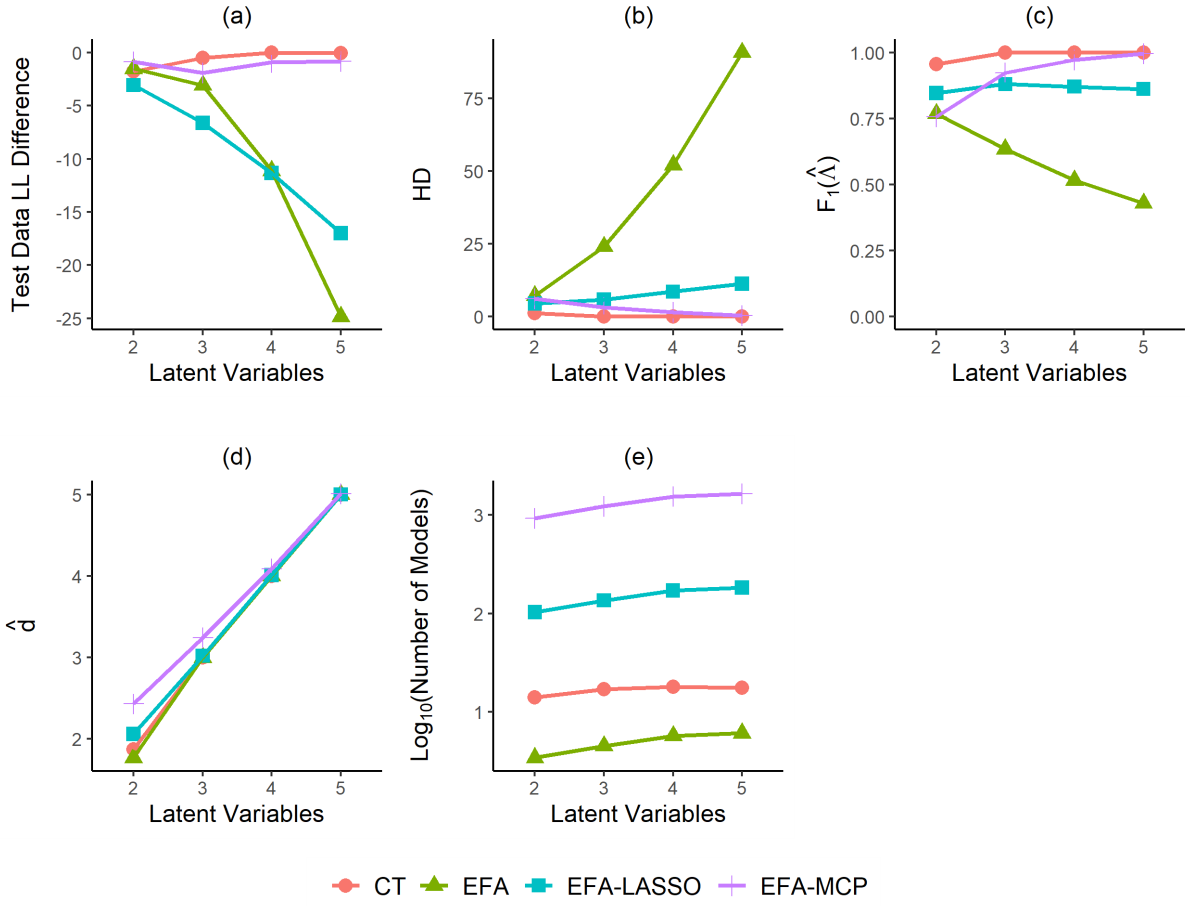

Figure 2.4: Averages for all evaluation metrics of the basic simulation study. The values for the test data log-likelihood had the known structure MLE subtracted for standardization. Hence, a value of zero corresponds to no difference vs. the known structure MLE method.

with test data log-likelihood almost identical to that of the known structure MLE, HD measures approximately zero, and $F_1(\hat{\Lambda})$ at about 1, as shown in Figures 2.4(a) through 2.4(c). This performance was closely followed by EFA-MCP, where EFA-LASSO performed moderately worse and EFA was substantially worse. For test data log-likelihood, HD, and F_1 metrics, the performances of the EFA-LASSO and EFA methods worsened as the number of latent variables grew. For the estimated number of latent variables, all methods had averages close to the true dimension.

The number of models estimated by each method is shown in Figure 2.4(e) in $\log_{10}$ scale. It is seen from the figure that EFA used the least number of solutions, with an overall average of 5.03 models, followed by the CT algorithm which used 16.79. Notice that the CT algorithm may propose multiple different structures with the same d, hence testing more models than EFA. The number of models estimated by EFA-LASSO and EFA-MCP

methods were much higher, using 150.83 and 1357.43 average solutions respectively. To put this into another perspective, we note that the average number of unique d supplied to the penalized EFA algorithms was 5.03. Supposing that we lowered the amount of tuning parameters penalized EFA used in order to match the same number of solutions checked by the CT algorithm, then penalized EFA could only use three to four sets of tuning parameters per d, which is much smaller than the typical number of tuning parameters used by a penalized method. This demonstrates the computational efficiency of the CT algorithm.

To assess variability, we also display box plots of the evaluation metrics of the basic simulation study in Figure 2.5. For the test data log-likelihood differences, the CT algorithm and the EFA-MCP methods showed very little variability, while the EFA and EFA-LASSO methods were much more variable. We see a similar pattern among the HD statistics, where the CT algorithm and EFA-MCP showed much less variability compared to the EFA and EFA-LASSO methods. For $F_1(\hat{\Lambda})$ score, we see that the CT algorithm once again showed very little variability across all numbers of latent variables. For the other EFA methods, there was moderate to large amounts of variability, which decreased as the number of latent variables increased. In particular for the EFA-MCP method, the variability was comparable to the CT algorithm for the 4 and 5 latent variable conditions.

In this basic simulation study we demonstrated that the CT algorithm performs nearly perfectly in all accuracy metrics. In terms of computational efficiency, it was only marginally outperformed by EFA, which conversely performed the worst on all other metrics. EFA-MCP performed almost just as well as the CT algorithm, however it calculated at least 80 times more models than the CT algorithm.

2.4.2 Thresholdability Robustness Study

In the previous simulation study we verified the performance of the CT algorithm when all of the assumptions are met. In practice, the thresholdable θ assumption may be violated and structural consistency will not be guaranteed. Despite this, it may be the case that the CT algorithm still provides a reasonable approximation to the true model.

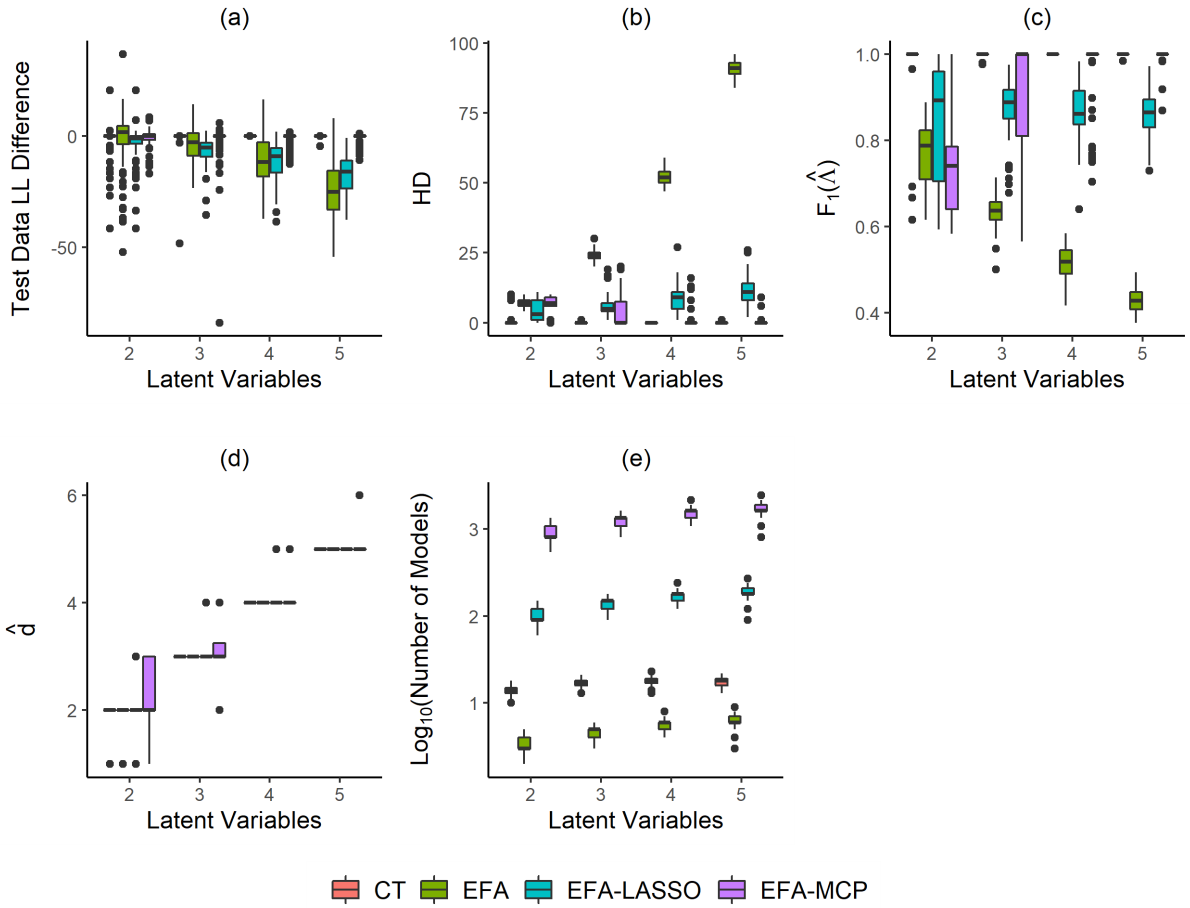

Figure 2.5: Box plots for all evaluation metrics of the basic simulation study are displayed. The values for the test data log-likelihood had the known structure MLE subtracted for standardization. Hence, a value of zero corresponds to no difference vs. the known structure MLE method for these metrics.

Thus in this simulation, we examine conditions under which the thresholdable assumption is violated to varying degrees, and study the robustness of the CT algorithm to these violations.

As in the previous simulation, we generated data sets from a zero-mean Gaussian distribution, with a covariance matrix Σ parameterized by θ which varied by condition. The structure of Λ followed an independent cluster structure (one non-zero entry per row). We focused on this structure since it is the most common factor analysis design and it was the simulation design used in the studies proposing the penalized EFA methods (Hirose and Yamamoto, 2014b,a). We note that more complicated structures were included in the previous simulation study. The number of latent variables (d) was set to 2, 3, 4 and 5, with the number of children per latent variable set to 5. The non-zero entries of Λ

were drawn from a uniform distribution, $\lambda_{ij} \sim \text{Uniform}(0.6, 0.8)$. The latent variable correlation matrix Φ was constructed with the same procedure as the previous simulation, with the exception of setting the range of the off-diagonals to $\alpha[0.6, 0.8]$. The scaling parameter α controlled the frequency of which θ was thresholdable, and was set to 1, 0.75, 0.5, 0.25, and 0. As we empirically show later, $\alpha = 0.5, 0.25, 0$ corresponded to thresholdable conditions, while $\alpha = 1, 0.75$ corresponded to non-thresholdable conditions, generally. Overall, this design resulted in $4 \times 5 = 20$ conditions, for which we conducted 100 replications per condition. As before, we generated two data sets per replication, one for training purposes and one for testing purposes. The sample size of each data set was set to $n = 1000$.

We collected the same basic metrics as the previous simulation. However, since thresholdability was not enforced in this study, we collected the proportion of parameters θ that were thresholdable (Definition 1) and the corresponding proportion for sample thresholdability defined using sample correlations r_{ij} in place of ρ_{ij} in Definition 1. To examine the robustness of these methods to violations of thresholdability, we define an additional metric which conveys the degree to which thresholdability is violated. We use the best possible thresholded correlation graph (in terms of HD to E_0) to represent an upper bound on the best obtainable structure by thresholding. Then for comparability, we represent this structure by its F_1 score. That is, we define a "best" thresholded correlation graph as

$$E_B = E(\tau^*), \qquad \text{where } \tau^* \coloneqq \underset{\tau \in (0,1)}{\text{argmin}} \left[|E(\tau) \triangle E_0| \right] \tag{2.48}$$

is obtained by checking all possible thresholds $\tau \in (0, 1)$ given the population correlations ρ_{ij} between X_i and X_j. Then we calculate its F_1 score as

$$F_1(E_B) \coloneqq \frac{2|E_B \cap E_0|}{2|E_B \cap E_0| + |E_B \triangle E_0|}. \tag{2.49}$$

Note that thresholdability holds if $F_1(E_B) = 1$. We also use a sample version $\hat{E}_B$ of E_B by substituting in $\hat{E}(\tau)$ for $E(\tau)$ in Equation (2.48).

The results of the test data log-likelihood and $\hat{d}$ statistics are displayed in Figure 2.6.

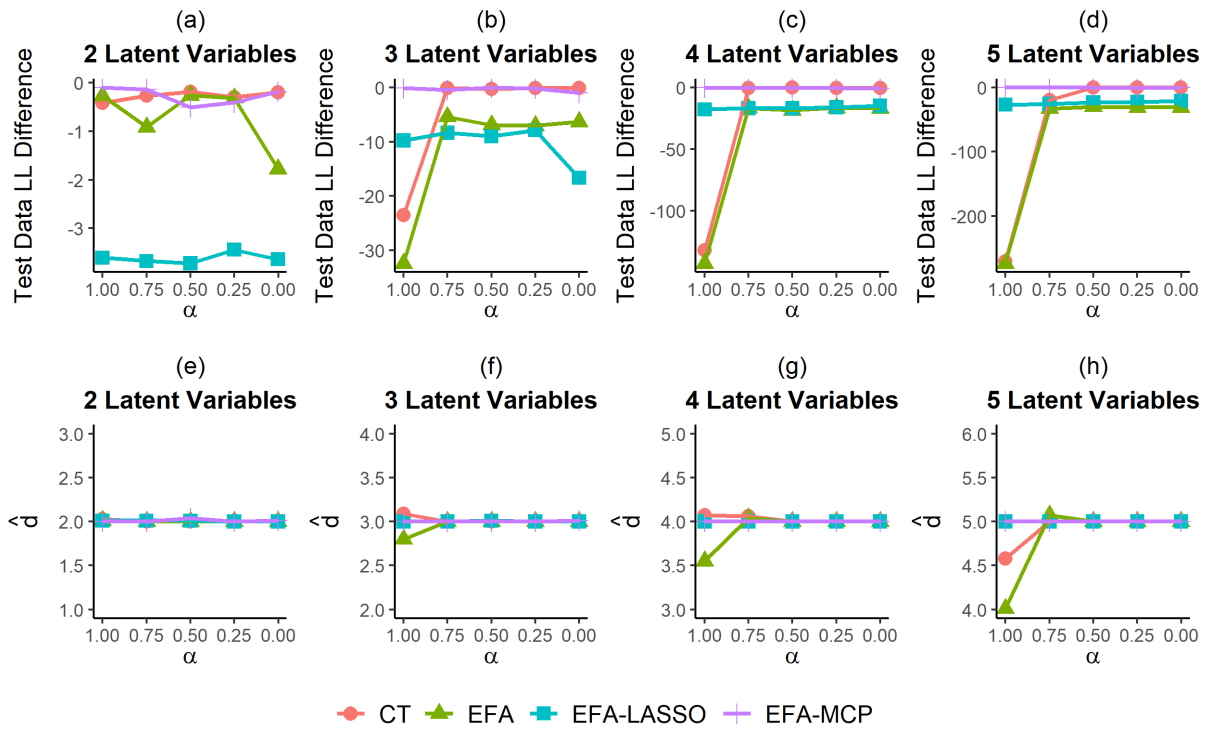

Figure 2.6: Average test data log-likelihood differences and $\hat{d}$ statistics of the robustness simulation study. The test data log-likelihood values had the known structure MLE subtracted for standardization.

For the test data log-likelihood, the performances of the EFA methods were generally ranked from best to worst as EFA-MCP, EFA-LASSO, and EFA, and tended to be stable across α and the number of latent variables. One caveat to this finding was that EFA sometimes displayed a further loss in performance when $\alpha = 1$. Likewise, the CT algorithm only noticeably lost performance when $\alpha = 1$, but performed near perfectly otherwise. As in the prior simulation, EFA-MCP performed near perfectly in all scenarios. For the learned number of latent variables ($\hat{d}$), all methods performed nearly perfectly, except when $\alpha = 1$, where EFA and the CT algorithm suffered some minor to moderate inaccuracies. These inaccuracies were more pronounced as the number of latent variables increased.

The results of HD, $F_1(\hat{\Lambda})$, and number of models are displayed in Figure 2.7. For HD and $F_1(\hat{\Lambda})$, the results were similar to the pattern exhibited by the test data log-likelihood. The performance of EFA methods ordered from best to worst was EFA-MCP, EFA-LASSO, and EFA. This was stable across α and the number of latent variables, with EFA-LASSO

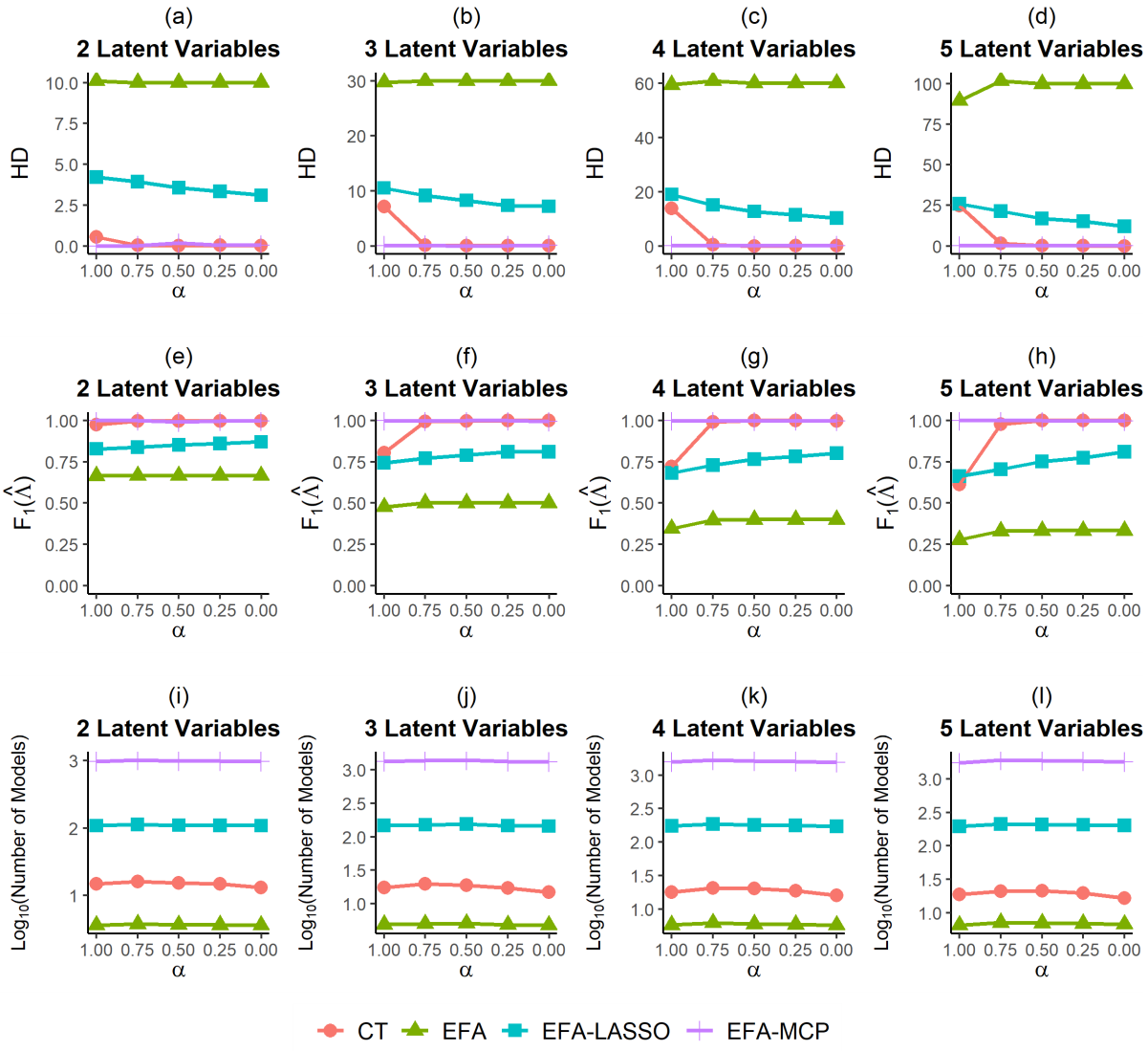

Figure 2.7: Average HD, $F_1(\hat{\Lambda})$, and the number of models (in the $\log_{10}$ scale) from the robustness simulation study.

exhibiting slightly better structure learning as α decreased. The CT algorithm only had noticeably loss in performance when $\alpha = 1$, as with the fit statistics, otherwise was near perfect. The EFA-MCP method, also showed near perfect performance regardless of the condition. For the number of models, there is little variation across α and number of latent variables. The averages ordered from best to worst are EFA (5.42), CT algorithm (17.74), EFA-LASSO (162.45), and EFA-MCP (1462.05).

The thresholdability and $F_1(E_B)$ statistics, reported in Figure 2.8, show that thresholdability generally holds when $\alpha \leq 0.5$. At $\alpha = 0.75$, the average population thresholdability is no less than 97%. However, at $\alpha = 1$ the population thresholdability decreases to

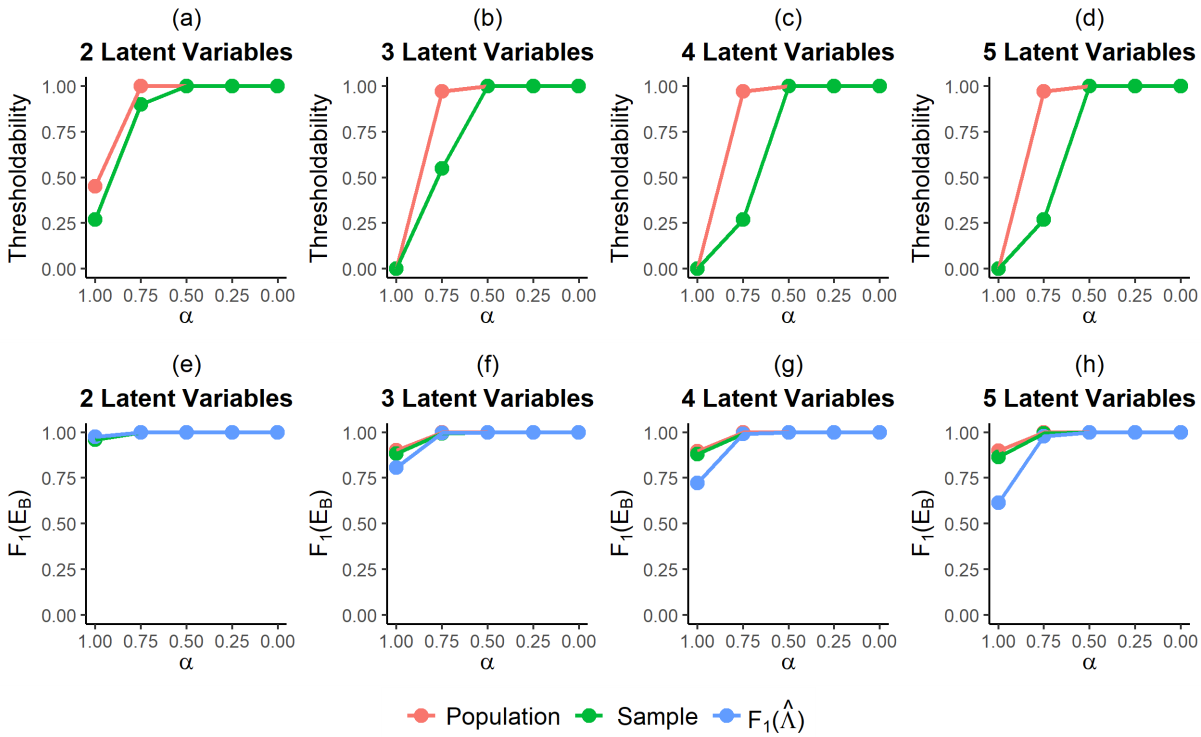

Figure 2.8: Average thresholdability and $F_1(E_B)$ statistics of the robustness simulation study. The $F_1(\hat{\Lambda})$ metric is shown alongside $F_1(E_B)$ for ease of comparison.

45% for 2 latent variables and decreases to 0% for 3 through 5 latent variables. For sample correlations, the pattern is the same, with the most notable difference being more substantial decreases at $\alpha = 0.75$. For the $F_1(E_B)$ score, we see that the average values are all high, being only less than 1 at $\alpha = 1$ and 0.75 as implied by the thresholdability results. In the worst case condition ($\alpha = 1$ and $d = 5$), the population scores were no less than 0.89 and the sample scores were no less than 0.86.

To assess variability, we also display box plots of the evaluation metrics of the thresholdability robustness study simulation study in Figures 2.9 and 2.10. We only show the box plots for the 5 latent variable condition for brevity. For the test data log-likelihood difference, all methods generally showed very little variability, except for the CT algorithm and EFA methods when $\alpha = 1$. The HD statistics showed similar patterns as the test data log-likelihood differences, with the exception that the EFA-LASSO method additionally showed a moderately small amount of variability. For the $F_1(\hat{\Lambda})$ score, all methods showed very little variability except for the CT algorithm at $\alpha = 1$ and the EFA-LASSO showing moderate variability across all α. For $\hat{d}$, we also see very little variability except for the

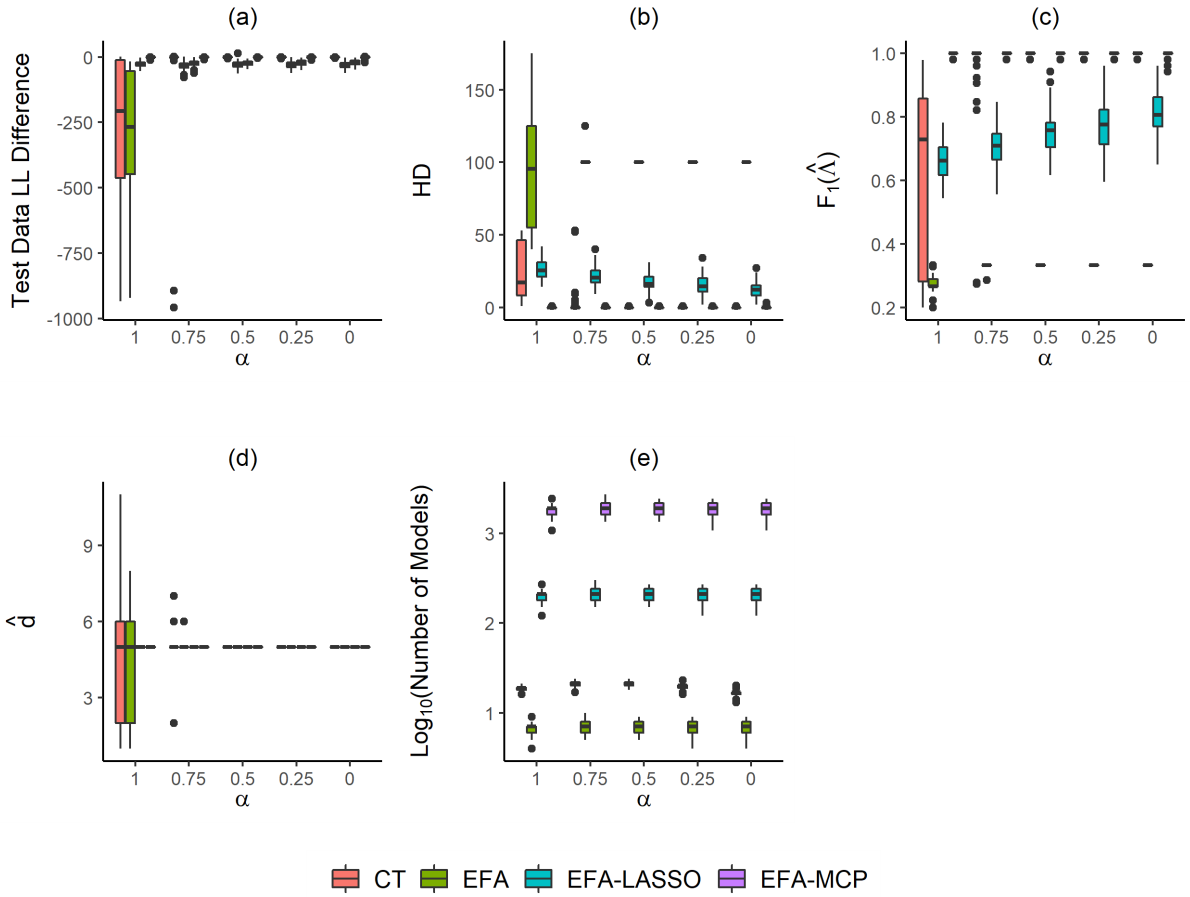

Figure 2.9: Box plots for test data log-likelihood, HD, F_1, and $\hat{d}$ of the thresholdability robustness simulation study are displayed. The values for the test data log-likelihood had the known structure MLE subtracted for standardization. Hence, a value of zero corresponds to no difference vs. the known structure MLE method for these metrics.

CT algorithm and EFA methods in the $\alpha = 1$ condition. The variability in the $\log_{10}$ number of models was uniformly small across all methods and all α. Finally, for the $F_1(E_B)$ statistics (Figure 2.10), we see that there is virtually no variability in both the population and sample except in the $\alpha = 1$ condition.

Overall, this simulation study largely confirms the robustness of the CT algorithm against moderate violations to thresholdability assumption. When thresholdability is met, the CT algorithm performs just as well as the known structure MLE. The results show that the CT algorithm can be robust to large frequencies of thresholdability violations (i.e., up to about 75% when $\alpha = 0.75$), although the performance begins to suffer when thresholdability is always violated (i.e., never thresholdable at $\alpha = 1$): The $F_1(\hat{\Lambda})$ of the CT algorithm only decreased moderately even for non-thresholdable parameters

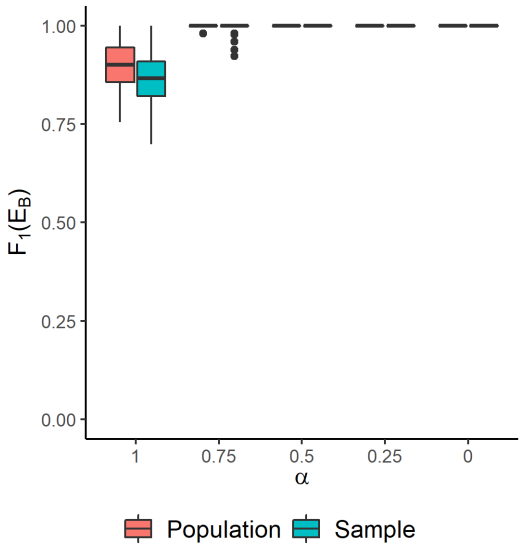

Figure 2.10: Box plots for the thresholdability and $F_1(E_B)$ statistics of the thresholdability robustness simulation study are displayed.

($\alpha = 1, d \geq 3$, lower panels in Figure 2.8), but still outperformed EFA-LASSO and EFA in almost all of these cases (Figure 2.7). Further, we note that the $F_1(E_B)$ scores were robust, and always close to 1, across all cases including non-thresholdable cases. The higher value of $F_1(E_B)$ than $F_1(\hat{\Lambda})$ indicates the possibility of improving structure learning accuracy of the CT algorithm by increasing the number of candidate thresholds, τ_k. On the other hand, the simulation results also show that α need not be too small in order to make θ thresholdable: even at at $\alpha = 0.5$ thresholdability is not violated much at all. Lastly, it is worth reiterating that, in terms of the number of models estimated, the CT algorithm is orders of magnitude more efficient than the competing method of EFA-MCP.

2.4.3 High-Dimensional Thresholdability Study

The previous simulations were designed under the classical low-dimensional setting, where p and d were fixed to small values relative to the sample size n. In this simulation study, we examine the high-dimensional setting, where both p and d grow proportionally with n, and $n < p$. This allows us to study the degree to which the thresholdability assumption is violated by varying (n, p, d), and how robust the CT algorithm is to violations of this nature.

We generated data identically to the previous simulation (Section 2.4.2), except

we varied $n \in \{250, 500, 1000\}$, and set $p = 1.5n$ and $d = 0.1n$. Under these high p settings, both the MLE and penalized EFA methods are prohibitively slow, and thus are not included in this study. For computational considerations, we omitted the estimation step of the CT algorithm (Step 8), since finding the MLE is too time-consuming, and only examined the learned structures, i.e., $(\hat{d}_k, \mathcal{A}(\hat{\Lambda}_k))$ for the set of input thresholds $\{\tau_k,\ k \in [m]\}$. Then we chose the model structure with the minimum HD among all $\{(\hat{d}_k, \mathcal{A}(\hat{\Lambda}_k)), k \in [m]\}$. This would allow us to examine whether the structures estimated by the CT algorithm contained at least one accurate model. Accordingly, we collected HD, $F_1(\hat{\Lambda})$, $\hat{d}$, elapsed time, thresholdability, and $F_1(E_B)$ statistics as our study outcomes.

The results are displayed in Figure 2.11. From the plots for HD and $F_1(\hat{\Lambda})$, we see that the learned structure is more accurate as n grows despite a proportional growth in p. This is predicted by Theorem 1, since $\mathbb{P}(\hat{E}(\tau_0) \neq E_0)$ decays at an exponential rate with n, but grows only at a polynomial rate with p. The estimated number of latent variables $(\hat{d})$ was also fairly accurate on average across all conditions, confirming the CT algorithm is capable of determining the number of latent factors automatically even in such challenging high-dimensional settings. As expected, the structure learning accuracy is also affected by α: The structure learning becomes more accurate as α decreases in terms of both HD and $F_1(\hat{\Lambda})$. This is in good agreement with how thresholdability and $F_1(E_B)$ were affected by α (Figures 2.11(e) through 2.11(h)) in this high-dimensional setting. In contrast to the previous simulation where sample correlations are generally all thresholdable up to $\alpha = 0.5$, under these $n < p$ settings sample thresholdability (Figure 2.11(g)) is violated starting at $\alpha = 0$ for $n = 250, 500$, and at $\alpha = 0.25$ for $n = 1000$. Comparing the thresholdability statistics to the $F_1(\hat{\Lambda})$ scores we can see that structure learning is still fairly robust to thresholdability violations in the sample correlations. For example, when $\alpha = 0.5$, the sample thresholdability was close to zero for all three sample sizes, while the F_1 score of the learned structure was close to 0.75 for $n = 250$ and close to 1 for $n = 500, 1000$. Note that for $\alpha \leq 0.5$, $F_1(\hat{\Lambda})$ is quite close to $F_1(\hat{E}_B)$, suggesting that the best threshold that defines $\hat{E}_B$, the sample version of E_B (2.48), is close to one of the 40 input threshold values. Finally, we see that the computation time of the algorithm increases with α and

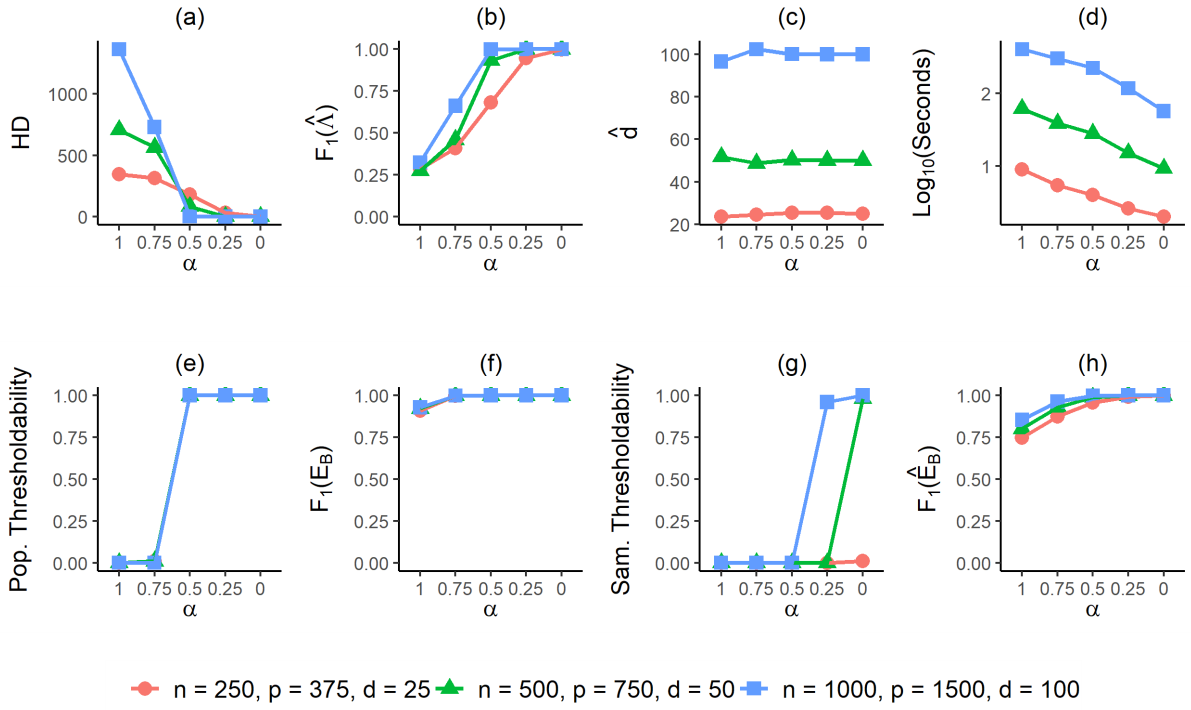

Figure 2.11: Average HD, $F_1(\hat{\Lambda})$, $\hat{d}$, running time, population and sample thresholdability, and $F_1(E_B)$ and $F_1(\hat{E}_B)$ statistics for the high-dimensional thresholdability study.

n. This may be attributable to the fact that the computational complexity of finding independent maximal cliques is $O(kp^2)$ (see Lemma 1 for details).

We display box plots of the evaluation metrics in Figure 2.12 to assess variability. For the HD statistics, we generally little to no variability when $\alpha \in \{0.5, 0.25, 0\}$. At $\alpha = 0.75$ however, we see a moderate to large amount of variability, with the variability decreasing once again at $\alpha = 1$. The higher for the $\alpha \in \{1, 0.75\}$ conditions, we see that variability increases with (n, p, d). For $F_1(\hat{\Lambda})$, we see the same general pattern as HD. Variability is moderate at $\alpha = 1$, then increases at $\alpha = 0.75$, and decreases to very small amounts at $\alpha \in \{0.5, 0.25, 0\}$. This affect appears to be moderated (n, p, d), with higher (n, p, d) conditions seeing more pronounced increasing and decreasing trends. For $\hat{d}$, we see a monotonically decreasing pattern of variability from $\alpha = 1$ to $\alpha = 0$, with overall variability being smaller for lower (n, p, d). The variability of $\log_{10}$(seconds) of computation time was fairly uniform across all α and (n, p, d). And finally, the F_1 scores of E_B and $\hat{E}_B$ showed very little variability across all conditions, with $F_1(\hat{E}_B)$ being slightly more variable.

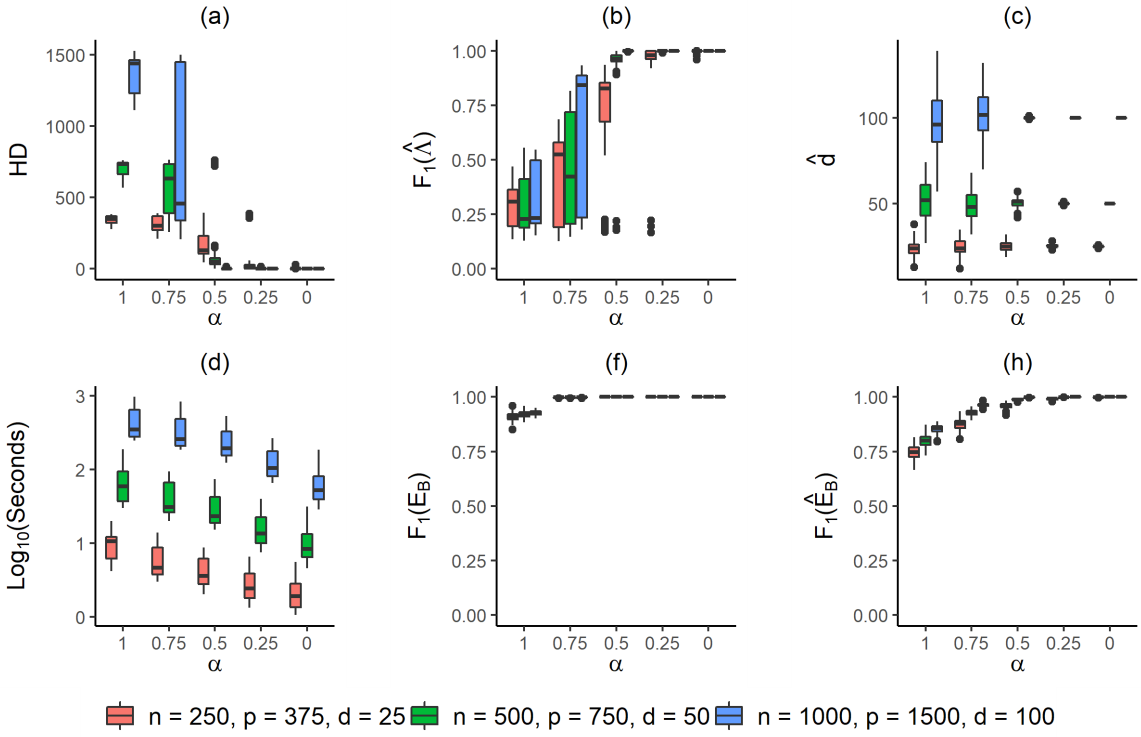

Figure 2.12: Box plots for HD, $F_1(\hat{\Lambda})$, $\hat{d}$, running time, $F_1(E_B)$, and $F_1(\hat{E}_B)$ of the high-dimensional thresholdability study are displayed.

We also use these results to demonstrate how the thresholdability gap γ (2.29) may be affected by α. In Figure 2.13 we plot the solution paths of three example data sets of size $n = 1000$ while varying $\alpha \in \{0.5, 0.25, 0\}$. In these paths we show how $F_1(\hat{\Lambda}_k)$ changes across various values of τ_k. We can see that the number of τ_k for which $F_1(\hat{\Lambda}_k) = 1$ is the greatest when $\alpha = 0$, which yielded six such cutoffs with a range of [0.128, 0.256], and thus $\gamma \approx 0.064$. In the $\alpha = 0.25$ dataset, the number of such τ_k shrinks to four cutoffs, with a smaller range of [0.205, 0.282] suggesting $\gamma \approx 0.039$. And finally, the $\alpha = 0.5$ dataset had only one such τ_k at 0.282.

2.4.4 High-Dimensional Unique Child Condition Study

Once again we study the high-dimensional setting, now turning our attention to violations of the unique child condition. As in the previous simulation, we allowed p and d to grow proportionally with n, with always having $n < p$. We studied how robust the CT algorithm to violations of the unique child condition under several settings of such (n, p, d).

We generated data under the same (n, p, d) as the previous study: $n \in \{250, 500, 1000\}$,

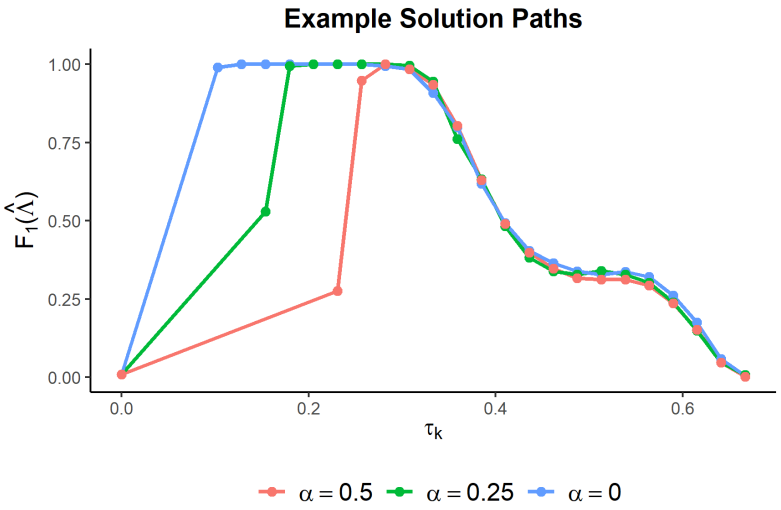

Figure 2.13: Example solution paths from the high-dimensional thresholdability study. All three examples are taken from the $n = 1000$, $p = 1500$, $d = 100$ condition.

$p = 1.5n$, and $d = 0.1n$. To begin with, the structure of Λ followed an independent cluster structure (one non-zero entry per row), for which the unique child condition trivially holds for every latent variable. We will call these latent variables the *main parent* of these observed variables. To isolate the effect of unique child conditions from that of thresholdability, we ensured thresholdability was always met in the population by setting $\Phi = I_d$ (Corollary 1). Then to dictate to what degree the unique child condition was violated, we defined a parameter $\beta \in \{1, 0.75, 0.5, 0.25, 0\}$, which was the randomly selected proportion of latent variables that would have no unique children. If a latent variable was deemed to have no unique children, we generated an extra path between the children of these latent variables to another random latent variable. Thus, if a latent variable was chosen to violate the unique child condition, all of its children were given an extra parent at random. We will call these extra parents the *alternative parent*.

For each X, we drew an $R^2 \sim \text{Uniform}(0.36, 0.64)$ as the proportion of variance in X explained by L. The range of (0.36, 0.64) is analogous to the range of path coefficients we were using in previous simulations which was (0.6, 0.8). If a given X only had a main parent and no alternative parent, then that X had a single path coefficient of $\sqrt{R^2}$ from its main parent. However, if a given X also had an alternative parent, then the R^2 was split using a 5:1 or 3:1 ratio between the main parent and the alternate parent, and the

path coefficients were calculated to reflect this accordingly.

As before, since the MLE and penalized EFA methods are prohibitively slow, we only examined the learned structures from the CT algorithm. That is, we examined the set of $(\hat{d}_k, \mathcal{A}(\hat{\Lambda}_k))$ for the set of thresholds $\{\tau_k,\ k \in [m]\}$. The selected model was the structure with the minimum HD among all $\{(\hat{d}_k, \mathcal{A}(\hat{\Lambda}_k)), k \in [m]\}$. Accordingly, we collected HD, $F_1(\hat{\Lambda})$, $\hat{d}$, elapsed time, sample thresholdability, and $F_1(\hat{E}_B)$ statistics as our study outcomes.

To assess the effect of violating the unique child condition, we considered additional metrics. Recall that E_B is the graph closest to E (in terms of HD) by checking all possible thresholds $\tau \in (0, 1)$. Let $\mathcal{A}(\Lambda_B)$ be a structure generated by performing the structure learning portion of the CT algorithm (Steps 2 through 7) on E_B. Subsequently, let $F_1(\Lambda_B)$ be the F_1 score of Λ_B. We use the $F_1(\Lambda_B)$ scores to show how adversely affected the best thresholded graphs might be to violations of the unique child condition.

The results of this simulation are displayed in Figures 2.14 for the 5:1 ratio and 2.15 for the 3:1 ratio. The patterns between the two ratio conditions are generally the same, thus we will focus on the 5:1 ratio condition. From the plots for HD and $F_1(\hat{\Lambda})$, we can see that the accuracy of structure recovery has a strong correspondence with the proportion of latent variables that violate the unique child condition (β). This is expected since violations of the unique child condition no longer guarantee a bijective mapping between the latent variable structure and the independent maximal cliques (Lemma 3). However, the $F_1(\hat{\Lambda})$ remained fairly strong even at $\beta = 1$, where it ranges from 0.358 to 0.655. This shows very good robustness of CT algorithm in learning the structure despite violations to the unique child condition. This robustness may be explained by an interesting contrast between the patterns exhibited by $F_1(\hat{\Lambda})$ and that of $F_1(\hat{\Lambda}_B)$ and $F_1(\Lambda_B)$. The $F_1(\hat{\Lambda}_B)$ and $F_1(\Lambda_B)$ statistics show that the accuracy of the structure learning procedure have a nearly linear relation with β if $\hat{E}_B$ or E_B are used as the thresholded graph. However, the accuracy of $F_1(\hat{\Lambda})$ is higher than that $F_1(\hat{\Lambda}_B)$ and $F_1(\Lambda_B)$, showing that there exists some $E(\tau)$ on the solution path that yields a better approximation to the true structure. In other words, even if E does not follow the unique child condition, using approximations

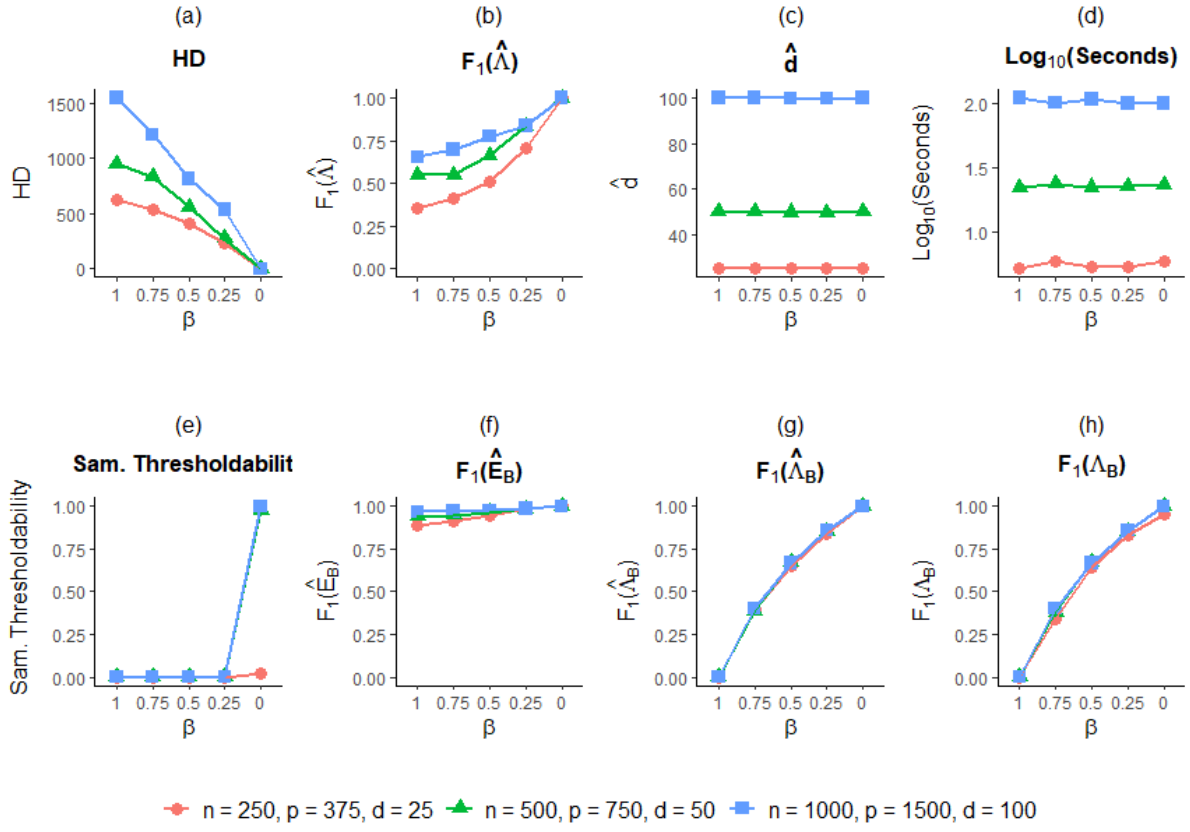

Figure 2.14: Average HD, $F_1(\hat{\Lambda})$, $\hat{d}$, running time, sample thresholdability, $F_1(\hat{E}_B)$, $F_1(\hat{\Lambda}_B)$, and $F_1(\Lambda_B)$ statistics for the 5:1 ratio high-dimensional unique child condition study.

$E(\tau)$ which still follow the unique child assumption can yield reasonably close structures.

2.5 Real Data Application

We examined a widely used factor analysis dataset comprised of intelligence test scores of $n = 301$ middle school students (Holzinger and Swineford, 1939). The data consist of 9 variables designed to measure 3 factors of intelligence. These were a spatial factor L_1 (visual perception tasks), a verbal factor L_2 (paragraph comprehension, sentence completion, and word meaning), and a speed factor L_3 (speed tests of addition, counting groups of dots, and discrimination of straight and curved capitals). The hypothesized structure is shown in Figure 2.16(a), and we applied the CT algorithm, EFA, EFA-LASSO, and EFA-MCP methods to the data. Again, for a fair comparison, we input the same set of d values produced in the CT algorithm to each of the EFA methods as we did in the simulation studies.

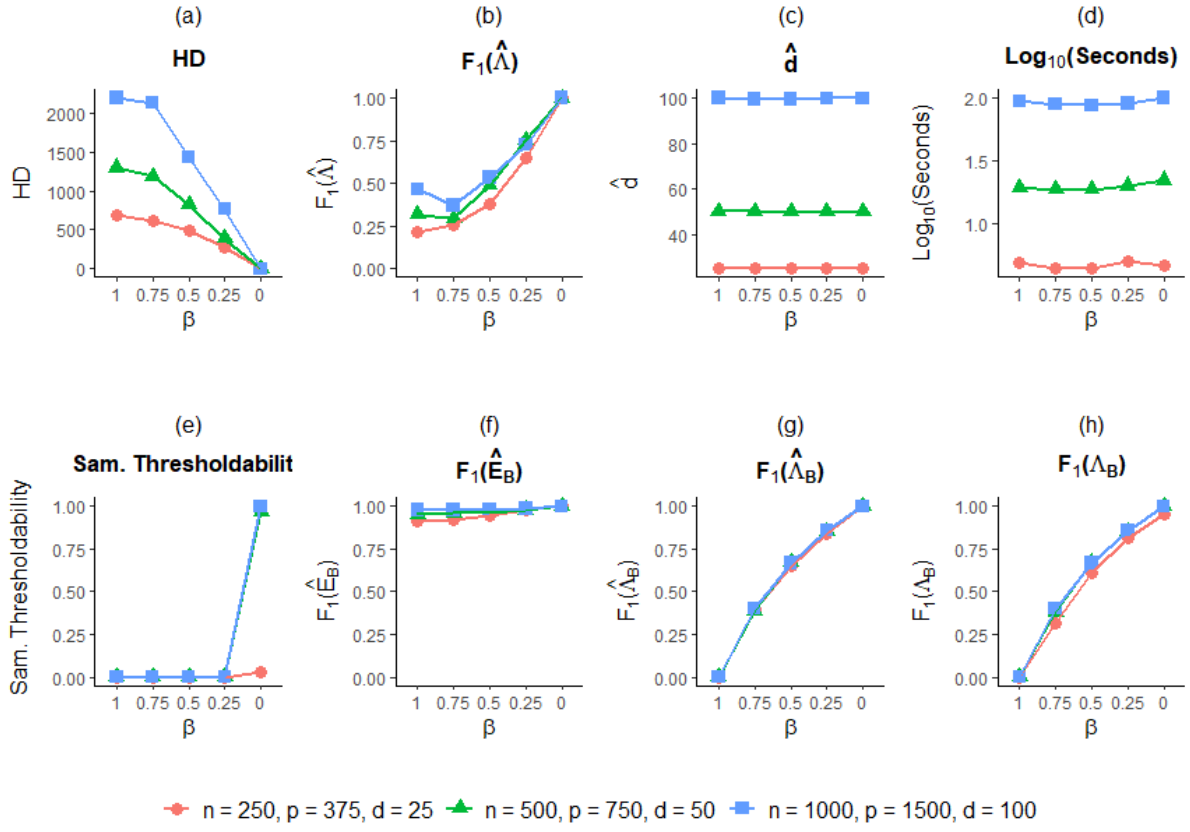

Figure 2.15: Average HD, $F_1(\hat{\Lambda})$, $\hat{d}$, running time, sample thresholdability, $F_1(\hat{E}_B)$, $F_1(\hat{\Lambda}_B)$, and $F_1(\Lambda_B)$ statistics for the 3:1 ratio high-dimensional unique child condition study.

We first checked the HD between the solution path of a method and the hypothesized model structure. The minimum HD over the solution path was zero for CT algorithm, 6 for EFA and EFA-LASSO, and 3 for EFA-MCP. This indicates that the hypothesized model was perfectly recovered within the solution path of the CT algorithm, but not for any of the other methods. Moreover, the CT algorithm identified the hypothesized structure with a much smaller set of candidate models (number of models in Table 2.1). The number of models checked by the CT algorithm was 13, compared to the 120 models checked by EFA-LASSO, and the 1080 models checked by EFA-MCP.

As in the simulation studies, we selected a model via BIC for each method and used 10-fold CV to evaluate the estimated models. The estimate of d and test data log-likelihood in 10-fold CV are reported Table 2.1. In terms of the test data log-likelihood, the results are similar across the CT algorithm, EFA-LASSO, and EFA-MCP methods, all three being much better than EFA. Despite the comparable performance between the

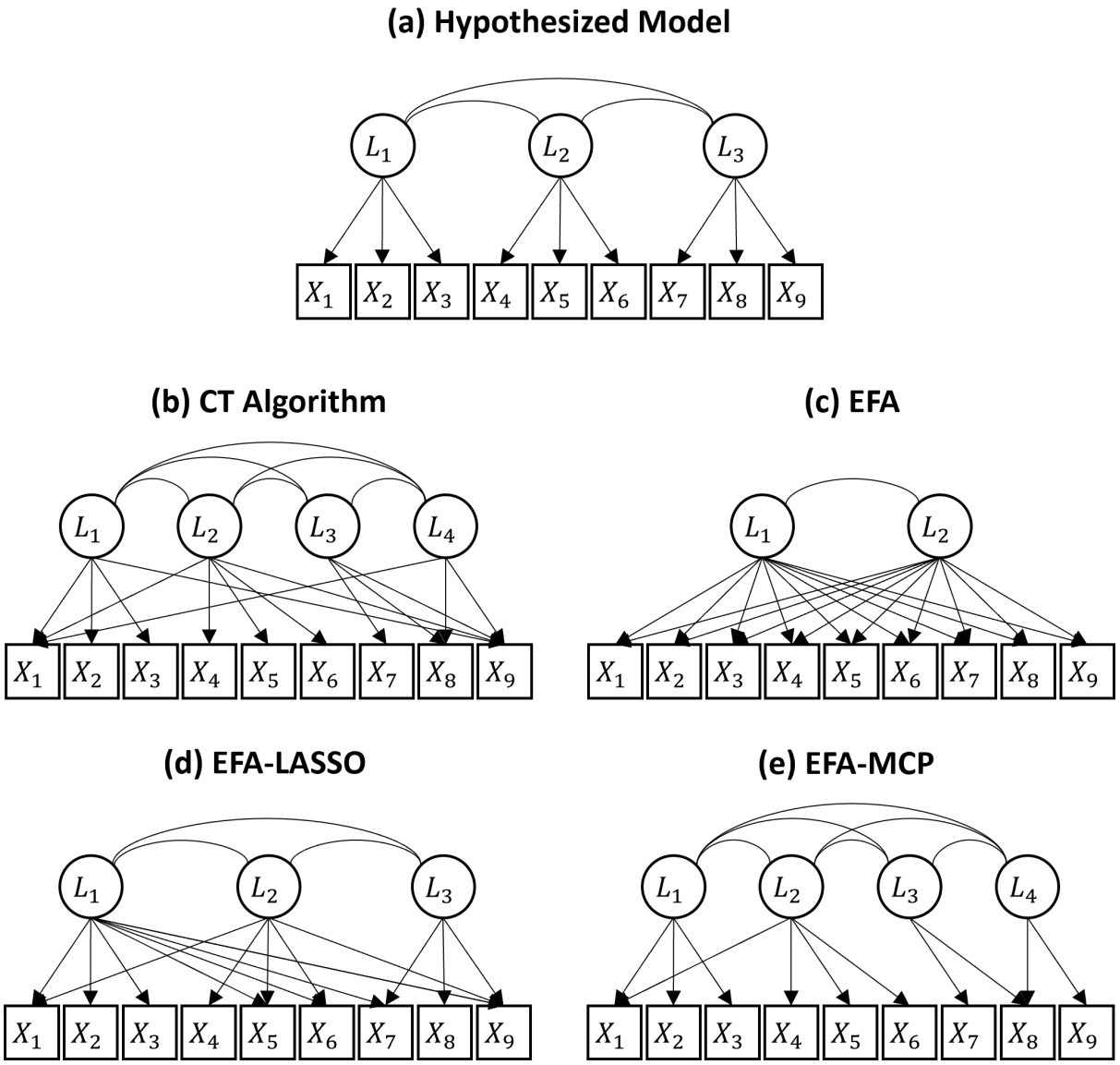

Figure 2.16: Estimated model structure by each method in the real data example. Variables X_1, X_2, and X_3 were visual perception tasks, variables X_4, X_5, and X_6 were verbal/reading tasks, and variables X_7, X_8, and X_9 were speed tests.

CT algorithm and the sparse EFA methods, the CT algorithm obtained these results with much improved computational efficiency, as discussed above on the numbers of candidate models estimated by these methods. Further, if we assume the hypothesized factor model to be the true model, the $F_1(\hat{E}_B)$ metric yielded a score of 0.8, indicating that the sample correlation matrix cannot be properly thresholded. This demonstrates how much thresholdability may be violated in practice, and how robust the CT algorithm is despite this assumption violation.

All methods slightly differed in the structure learned, as shown in Figure 2.16. Note

Method	HD(min.)	$\hat{d}$	test data log-likelihood	Number of Models
CT Algorithm	0	4	-3749.60	13
EFA	6	2	-3823.14	4
EFA-LASSO	6	3	-3751.82	120
EFA-MCP	3	4	-3751.37	1080

Table 2.1: Results of real data example. HD(min.) denotes the minimum HD to the hypothesized structure across all solutions.

that these structures, selected via BIC, are in general different from those with the minimum HD to the hypothesized model, probably due to the relatively small sample size of this dataset. The CT algorithm and EFA-LASSO methods were able to recover all of the original hypothesized paths, however learned some extra paths. From the visual factor L_1, the CT algorithm introduced one extra path to a speed task (X_9), while EFA-LASSO introduced four extra paths to the visual and speed tasks (X_5, X_6, X_7, X_9). In addition, an extra path was learned by both methods from the verbal factor L_2 to each of the visual and speed tasks. No extra path was learned by either method from the speed factor L_3. Both the CT algorithm and EFA-MCP learned a fourth factor, which had paths to two of the speed tasks. In the model learned by the CT algorithm, the fourth factor seems to supplement the already existing speed factor, while with the EFA-MCP method, its speed factor seems to have been split into two.

2.6 Concluding Remarks

Overall, the CT algorithm is a promising method for learning factor analysis structures. In this article, we motivated the algorithm using thresholded correlation graphs, and established the conditions for structural identifiability, parameter uniqueness, and asymptotic consistency. In addition, the CT algorithm yields a method of learning d, which the EFA counterparts lack. In our simulation studies, the CT algorithm performed nearly perfectly when the assumption of thresholdability was met, and showed robust results when this assumption was violated. Further, the computational efficiency of the CT algorithm is unrivaled relative to the EFA-LASSO and EFA-MCP methods, as it checks substantially less models.

There are some limitations of the CT algorithm to keep in mind. The most obvious is the reliance on the thresholdability assumption. Even though we demonstrated that the CT algorithm is robust to thresholdability violations in practice, EFA-MCP outperforms the CT algorithm in these cases. Additionally, our statistical consistency results depend on this assumption being true in the population. Future work can focus on relaxing the thresholdability assumption.

Similarly, the CT algorithm is restricted to structures fulfilling the unique child condition. However, in practice this assumption is a relaxation of many common factor analysis designs as discussed in Section 2.3.2. In many factor analysis applications, the observed variables are designed a priori as measurements of the latent variables. In these cases, unique child variables can always be designed beforehand to use with the CT algorithm.

We also note some computational limitations for the high-dimensional ($n < p$) regime for parameter estimation. Both penalized and traditional MLE estimation procedures have fairly long computation routines. Since the CT algorithm relies on external existing estimation method to provide parameter estimates, it is subsequently limited by the existing technology in this area. Thus the estimation portion of our algorithm will also benefit from computational advances on this topic. We may also develop alternative methods to select a solution from the candidate models generated by the CT algorithm without maximizing the likelihood, such as ideas similar to stability selection (Meinshausen and Bühlmann, 2010). This is certainly an interesting and promising future direction that will further broaden the application of this work.

Chapter 3

Piecewise Linear Splines for Non-Linear Factor Analysis

In the previous chapters we considered the factor analysis model where X is comprised of multiple linear influences from a multidimensional L. In this chapter, we consider the case where X is comprised of non-linear influences from L. In general, a non-linear factor analysis model can be written as:

$$X = g(L; \beta) + \epsilon, \tag{3.1}$$

where X, L, and ϵ are observed, latent, and error variables, respectively, as before, β are the parameters, and $g(\cdot)$ may be a vector-valued non-linear function of L, β, or both. We will continue to assume that $L \sim \mathcal{N}_d(0, \Phi)$, however the specification of X and ϵ may vary depending on the application.

3.1 Prior Work

Prior work in non-linear factor analysis has nearly exclusively focused on the case where $g(\cdot)$ is specified a priori. One of the most basic specifications are polynomial forms of L, while remaining linear in β (McDonald, 1965, 1967; Etezadi-Amoli, 1983). More generally, frameworks have also been presented where the elements of $g(\cdot)$ can be linear functions of L or β, or non-linear in both (Yalcin and Amemiya, 2001). However, one of the most popular non-linear specification is the sigmoid curve, either logit or probit in nature.

Much of the field of psychometrics is dedicated to the study such methods, called *Item Response Theory* (IRT; Cai et al., 2016; Baker and Kim, 2004), and is typically used for modeling binary or polytomous data such as questionnaire responses and educational assessments (Embretson and Reise, 2000). Overall, these methods can be computationally challenging, as polynomial forms of L often require estimation of higher-order moments, and the non-linear models of β (such as IRT) require numerical methods of integrating the density of L (Cudeck et al., 2009).

On the other hand, very little work has been done with methods that learn the function $g(\cdot)$. The most similar methods are latent variable interpretations of manifold learning techniques. One closely related example are so-called principal curves and surfaces (Hastie and Stuetzle, 1989; Tibshirani, 1992). Analogous to principal components analysis, these methods are motivated by finding the curve through the "middle" of a dataset such that the sum of squared deviations from all variables to the curve are minimized (subject to flexibility constraints). Likewise, other manifold learning techniques can be interpreted from a latent variable perspective. For example, Carreira-Perpiñán and Lu (2007) use Laplacian eigenmaps and kernel density estimators to provide non-linear estimates of L. The drawback to these techniques are that the variances of the errors are either assumed to be homogeneous, or not readily available at all. And finally, other examples come from neural network based approaches, such as generative networks (Goodfellow et al., 2014; Han et al., 2017) and variational auto-encoders (Kingma and Welling, 2019). These methods generally utilize neural networks to learn $g(\cdot)$ and L with the purpose of reconstructing X.

In contrast, we will focus on methods of learning an interpretable $g(\cdot)$ which can easily assign substantive meanings to L, while retaining the important factor analysis property of heterogeneous error variances. To this end, we will explore a simple piecewise linear spline method, motivated by an underlying generative factor analysis model, estimated using the EM-algorithm.

3.1.1 Model

Consider a latent variable scalar $L \sim \mathcal{N}(0,1)$, with a set of m knot locations c_j for $j \in [m]$, and let $c_0 = -\infty$ and $c_{m+1} = \infty$. We then consider a piecewise linear model of the form:

$$X = \beta_0 L + \sum_{j=1}^{m} \beta_j (L - c_j)_+ + \epsilon, \tag{3.2}$$

where $X \in \mathbb{R}^p$ is a vector of observed variables, $\epsilon \sim \mathcal{N}_p(0, \Omega)$ is a random vector of errors (with Ω being a diagonal matrix), and $\beta \in \mathbb{R}^p$ are the regression coefficients. The term $(L - c_j)_+$ is the positive part of $(L - c_j)$ and can be thought of as a threshold function:

$$(L - c_j)_+ \coloneqq \max(L - c_j, 0) = \begin{cases} L - c_j & \text{if } L > c_j \\ 0 & \text{if } L \leq c_j. \end{cases} \tag{3.3}$$

It will also be useful to consider the model in Equation 3.2 conditional on the event that $L \in (c_j, c_{j+1}]$. To do this let us define the random variable $Z = z \Leftrightarrow c_z < L \leq c_{z+1}$, for $z \in \{0, \ldots, m\}$. This can be thought of as a bin label. Then notice that we have the following decomposition of Equation 3.2 about z:

$$\begin{aligned} X &= \beta_0 L + \sum_{j=1}^{m} \beta_j (L - c_j)_+ + \epsilon \\ &= \beta_0 L + \sum_{j=1}^{z} \beta_j (L - c_j)_+ + \sum_{j=z+1}^{m} \beta_j (L - c_j)_+ + \epsilon. \end{aligned} \tag{3.4}$$

Now, if we condition on the event $Z = z$, we obtain the following conditional model:

$$\begin{aligned} X &= \beta_0 L + \sum_{j=1}^{z} \beta_j (L - c_j) + \epsilon && \text{if } Z = z \\ &= \left(-\sum_{j=1}^{z} \beta_j c_j \right) + \left(\sum_{j=0}^{z} \beta_j \right) L + \epsilon && \text{if } Z = z. \end{aligned} \tag{3.5}$$

Thus, given $Z = z$, we have our usual factor analysis model. To see this, define $\alpha_z \coloneqq$

$-\sum_{j=1}^{z}\beta_j c_j$ and $\Lambda_z = \sum_{j=0}^{z}\beta_j$. Then we can write Equation 3.5 as

$$\begin{aligned} X &= \alpha_z + \Lambda_z L + \epsilon \quad \text{if } Z = z \\ \Rightarrow \Sigma_z &= \phi_z \Lambda_z \Lambda_z + \Omega \quad \text{if } Z = z \end{aligned} \tag{3.6}$$

which yields a conditional covariance factor analytic structure, with $\phi_z \coloneqq \text{Var}(L|Z)$.

Finally, in subsequent calculations it will be also convenient to consider a traditional regression form of Equation 3.2. We can re-write it as

$$X = \boldsymbol{\beta} d(L) + \epsilon, \tag{3.7}$$

where $\boldsymbol{\beta}$ is a $p \times (m+1)$ matrix containing all the regression coefficients and $d(L)$ is an $(m+1) \times 1$ *design function* that maps L to a *design vector* (or feature vector) as

$$d(L) = \begin{bmatrix} L & (L-c_1)_+ & \dots & (L-c_m)_+ \end{bmatrix}^T. \tag{3.8}$$

We note that our choice to begin with a latent variable scalar is motivated from a substantive perspective. In particular, it may be the case that a set of observed variables have been designed to function with a certain latent variable dimensionality in mind (typically $d = 1$) but unanticipated non-linearities may interfere with statistical fit and introduce parameter bias. Rather than add more linear latent factors to improve the model (which may not be interpretable), it has been argued to specify non-linear functions to maintain the integrity of interpretation (McDonald, 1965; Ferguson, 1941). As such, the starting point of $d = 1$ has very wide applicability, however we will present some preliminary work and outline some challenges associated with the multivariable case in Chapter 4.

3.2 Estimation

3.2.1 EM Algorithm

The EM algorithm is a two-step iterative procedure for obtaining parameter estimates for models with missing data (Dempster et al., 1977). Considering X as our observed data and L as missing data, the steps are as follows:

***E*-Step.** For any iteration t, define a Q-function given an initial parameter start value $\theta^{(0)}$:

$$\begin{aligned} Q_{\theta^{(t)}}(\theta) &= \mathbb{E}_{\theta^{(t)}}\left[\log \mathbb{P}_\theta(X, L)|X\right] \\ &= \int_L \log \mathbb{P}_\theta(X, L)\mathbb{P}_{\theta^{(t)}}(L|X)\, dL. \end{aligned} \tag{3.9}$$

***M*-Step.** Maximize the Q-function with respect to θ and set the result as $\theta^{(t+1)}$:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q_{\theta^{(t)}}(\theta), \tag{3.10}$$

Hence, this is an iterative procedure that maximizes the expectation of the complete data log-likelihood, given the observed data. It is known to converge to a local maximum of the likelihood function under very general conditions (Wu, 1983).

An important property of the EM-algorithm is that each iteration will never decrease the observed data log-likelihood. To see why this is the case, consider a decomposition of the observed data log-likelihood as follows:

$$\begin{aligned} \mathbb{P}_\theta(X, L) &= \mathbb{P}_\theta(L|X)\mathbb{P}_\theta(X) \\ \log \mathbb{P}_\theta(X, L) &= \log \mathbb{P}_\theta(L|X) + \log \mathbb{P}_\theta(X) \\ \Rightarrow \log \mathbb{P}_\theta(X) &= \log \mathbb{P}_\theta(X, L) - \log \mathbb{P}_\theta(L|X) \\ &= Q_{\theta^{(t)}}(\theta) - H_{\theta^{(t)}}(\theta), \end{aligned} \tag{3.11}$$

where we took the expectation with respect to $\mathbb{P}_\theta(L|X)$, and as such

$$H_{\theta^{(t)}}(\theta) \coloneqq \mathbb{E}_{\theta^{(t)}}\left[\log \mathbb{P}_\theta(L|X)|X\right] = \int_L \log \mathbb{P}_\theta(L|X)\mathbb{P}_{\theta^{(t)}}(L|X)\, dL. \tag{3.12}$$

Then the difference between successive iterations of the observed data log-likelihood is

$$\begin{aligned}\log \mathbb{P}_{\theta^{(t+1)}} - \log \mathbb{P}_{\theta^{(t)}} &= \left(Q_{\theta^{(t)}}(\theta^{(t+1)}) - H_{\theta^{(t)}}(\theta^{(t+1)})\right) - \left(Q_{\theta^{(t)}}(\theta^{(t)}) - H_{\theta^{(t)}}(\theta^{(t)})\right) \\ &= \left(Q_{\theta^{(t)}}(\theta^{(t+1)}) - Q_{\theta^{(t)}}(\theta^{(t)})\right) + \left(H_{\theta^{(t)}}(\theta^{(t)}) - H_{\theta^{(t)}}(\theta^{(t+1)})\right).\end{aligned} \tag{3.13}$$

The first term must be positive, as by the definition of $\theta^{(t+1)}$ as the maximizer of $Q_{\theta^{(t)}}(\theta)$ we have

$$\begin{aligned} & Q_{\theta^{(t)}}(\theta^{(t+1)}) \geq Q_{\theta^{(t)}}(\theta^{(t)}) \\ \Rightarrow\ & Q_{\theta^{(t)}}(\theta^{(t+1)}) - Q_{\theta^{(t)}}(\theta^{(t)}) \geq 0.\end{aligned} \tag{3.14}$$

Then for the second term we have

$$\begin{aligned} H_{\theta^{(t)}}(\theta^{(t)}) - H_{\theta^{(t)}}(\theta^{(t+1)}) &= \mathbb{E}_{\theta^{(t)}}\left[\log \mathbb{P}_{\theta^{(t)}}(L|X)|X\right] - \mathbb{E}_{\theta^{(t)}}\left[\log \mathbb{P}_{\theta^{(t+1)}}(L|X)|X\right] \\ &= \mathbb{E}_{\theta^{(t)}}\left[-\log \frac{\mathbb{P}_{\theta^{(t+1)}}(L|X)}{\mathbb{P}_{\theta^{(t)}}(L|X)}\middle| X\right] \\ &\geq -\log\left(\mathbb{E}_{\theta^{(t)}}\left[\frac{\mathbb{P}_{\theta^{(t+1)}}(L|X)}{\mathbb{P}_{\theta^{(t)}}(L|X)}\middle| X\right]\right) \\ &= -\log\left(\int_L \frac{\mathbb{P}_{\theta^{(t+1)}}(L|X)}{\mathbb{P}_{\theta^{(t)}}(L|X)} \mathbb{P}_{\theta^{(t)}}(L|X)\, dL\right) \\ &= -\log(1) \\ &= 0,\end{aligned} \tag{3.15}$$

which follows from Jensen's inequality and the convexity of $-\log(\cdot)$. Hence, successive log-likelihoods of the observed data are non-decreasing in the iterations of the EM-algorithm.

3.2.2 Maximization of the Q-function

Given a sample from the model in Equation 3.2, the Q-function is as follows:

$$\begin{aligned} Q_{\theta^{(t)}}(\theta) &= \sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[\log \mathbb{P}_\theta(x_i, l_i)|x_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[\log \mathbb{P}_\theta(x_i|l_i) + \log \mathbb{P}_\theta(l_i)|x_i\right] \\ &= -\frac{n}{2}\log|\Omega| - \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[(x_i - \boldsymbol{\beta} d_i)^T \Omega^{-1}(x_i - \boldsymbol{\beta} d_i)\middle| x_i\right] + c.\end{aligned} \tag{3.16}$$

For maximizing with respect to $\boldsymbol{\beta}$, the relevant terms of the Q-function are:

$$Q_{\theta^{(t)}}(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[(x_i - \boldsymbol{\beta} d_i)^T \Omega^{-1}(x_i - \boldsymbol{\beta} d_i)\middle| x_i\right] + c, \tag{3.17}$$

where we abbreviate $d_i \coloneqq d(l_i)$. The first partial derivative is then

$$\frac{\partial Q_{\theta^{(t)}}(\beta)}{\partial \beta} = \sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[\Omega^{-1}(x_i - \boldsymbol{\beta} d_i) d_i^T \middle| x_i\right], \tag{3.18}$$

which we set to zero and obtain the first-order condition of

$$\boldsymbol{\beta}\left(\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[d_i d_i^T \middle| x_i\right]\right) = \sum_{i=1}^{n} x_i \mathbb{E}_{\theta^{(t)}}\left[d_i^T \middle| x_i\right] \tag{3.19}$$

Thus our maximizer of $Q_{\theta^{(t)}}(\boldsymbol{\beta})$ is

$$\boldsymbol{\beta}^{(t+1)} = \left(\sum_{i=1}^{n} x_i \mathbb{E}_{\theta^{(t)}}\left[d_i^T \middle| x_i\right]\right)\left(\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[d_i d_i^T \middle| x_i\right]\right)^{-1}. \tag{3.20}$$

For Ω the Q-function is

$$Q_{\theta^{(t)}}(\Omega) = -\frac{n}{2}\log|\Omega| - \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[(x_i - \boldsymbol{\beta} d_i)^T \Omega^{-1}(x_i - \boldsymbol{\beta} d_i)\middle| x_i\right] + c', \tag{3.21}$$

which for convenience we will write the Q-function in terms of ω_j^2 as

$$Q_{\theta^{(t)}}(\omega_j^2) = -\frac{n}{2}\sum_{j=1}^{p} \log \omega_j^2 - \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\mathbb{E}_{\theta^{(t)}}\left[(x_{ij} - \boldsymbol{\beta}_{j\bullet} d_i)^2 | x_{ij}\right]}{2\omega_j^2} + c', \tag{3.22}$$

where $\boldsymbol{\beta}_{j\bullet}$ is the jth row of $\boldsymbol{\beta}$. The first partial derivative is then

$$\frac{\partial Q_{\theta^{(t)}}(\omega_j^2)}{\partial \omega_j^2} = -\frac{n}{2\omega_j^2} + \frac{\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[(x_{ij} - \boldsymbol{\beta}_{j\bullet} d_i)^2 | x_{ij}\right]}{2\omega_j^4}. \tag{3.23}$$

Setting this equal to zero gives use a first-order condition of

$$\frac{n}{2\omega_j^2} = \frac{\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}\left[(x_{ij} - \boldsymbol{\beta}_{j\bullet} d_i)^2 | x_{ij}\right]}{2\omega_j^4}, \tag{3.24}$$

yielding a maximizer as

$$\omega_j^{2,(t+1)} = \frac{\sum_{i=1}^n \mathbb{E}_{\theta^{(t)}}\left[(x_{ij} - \boldsymbol{\beta}_{j\bullet}^{(t+1)} d_i)^2 \middle| x_{ij}\right]}{n}, \tag{3.25}$$

and re-combining the ω_j^2 into matrix form

$$\Omega^{t+1} = \text{diag}\left(\frac{\sum_{i=1}^n \mathbb{E}_{\theta^{(t)}}\left[(x_i - \boldsymbol{\beta}^{t+1} d_i)(x_i - \boldsymbol{\beta}^{t+1} d_i)^T \middle| x_i\right]}{n}\right). \tag{3.26}$$

For computational efficiency we may consider an alternate form for $\Omega^{(t+1)}$. Define $b \coloneqq \sum_{i=1}^n x_i \mathbb{E}_{\theta^{(t)}}\left[d_i^T \middle| x_i\right]$ and $A \coloneqq \sum_{i=1}^n \mathbb{E}_{\theta^{(t)}}\left[d_i d_i^T \middle| x_i\right]$ which yields $\boldsymbol{\beta}^{t+1} = bA^{-1}$. Then we can expand the outer product in Equation 3.26 and simplify as follows

$$\begin{aligned}
\Omega^{t+1} &= \text{diag}\left(\frac{\sum_{i=1}^n \mathbb{E}_{\theta^{(t)}}\left[(x_i x_i^T - \boldsymbol{\beta}^{t+1} d_i x_i^T - x_i d_i^T (\boldsymbol{\beta}^{t+1})^T + \boldsymbol{\beta}^{t+1} d_i d_i^T (\boldsymbol{\beta}^{t+1})^T \middle| x_i\right]}{n}\right) \\
&= \text{diag}\left(\frac{\sum_{i=1}^n x_i x_i^T - bA^{-1}b^T - bA^{-1}b^T + bA^{-1}AA^{-1}b^T}{n}\right) \\
&= \text{diag}\left(\frac{\sum_{i=1}^n x_i x_i^T - bA^{-1}b^T}{n}\right) \\
&= \text{diag}\left(\frac{\sum_{i=1}^n x_i x_i^T - \boldsymbol{\beta}^{t+1} \sum_{i=1}^n \mathbb{E}_{\theta^{(t)}}\left[d_i | x_i\right] x_i^T}{n}\right)
\end{aligned} \tag{3.27}$$

3.2.3 Conditional Expectations

We can see from the maximizers of the Q-function that the required conditional expectations are $\mathbb{E}[d_i|x_i]$ and $\mathbb{E}[d_i d_i^T|x_i]$. The main complication is that d_i is a vector that consists of threshold functions depending on the value of l_i. That is we need

$$\mathbb{E}[d_i|x_i] = \begin{bmatrix}\mathbb{E}[l_i|x_i] & \mathbb{E}[(l_i - c_j)_+|x_i] & \dots & \mathbb{E}[(l_i - c_j)_+|x_i]\end{bmatrix}^T. \tag{3.28}$$

However, we can address this using the law of iterated expectation as follows

$$\begin{aligned}
\mathbb{E}[d_i|x_i] &= \mathbb{E}[\mathbb{E}[d_i|x_i, z_i]|x_i] \\
&= \sum_{z=0}^{m} \mathbb{E}[d_i|x_i, z_i]\mathbb{P}(z_i|x_i).
\end{aligned} \tag{3.29}$$

which simplifies the problem since:

$$\mathbb{E}[(l_i - c_j)_+ | x_i, z_i] = \begin{cases} \mathbb{E}[l_i | x_i, z_i] - c_j & \text{if } z_i \geq j \\ 0 & \text{if } z_i < j \end{cases}. \tag{3.30}$$

The remaining quantity to compute is $\mathbb{E}[l_i | x_i, z_i]$. We will show in the following section that $\mathbb{P}(L|X, Z)$ is a truncated normal, thus $\mathbb{E}[l_i | x_i, z_i]$ can be calculated by using known expressions for the one-dimensional truncated normal moments. Turning our attention to $\mathbb{E}[d_i d_i^T | x_i]$, notice we have

$$d_i d_i^T = \begin{bmatrix} l_i^2 & l_i(l_i - c_1)_+ & \cdots & l_i(l_i - c_m)_+ \\ l_i(l_i - c_1)_+ & (l_i - c_1)_+^2 & \cdots & (l_i - c_1)_+(l_i - c_m)_+ \\ \vdots & \vdots & \ddots & \vdots \\ l_i(l_i - c_m)_+ & (l_i - c_1)_+(l_i - c_m)_+ & \cdots & (l_i - c_m)_+^2 \end{bmatrix}. \tag{3.31}$$

Once again, by law of iterated expectation, we can use $\mathbb{E}[d_i d_i^T | x_i, z_i]$ to make the expectations tractable. In general, we have three cases to consider. These are

$$\begin{aligned} \mathbb{E}[l_i(l_i - c_j)_+ | x_i, z_i] &= \begin{cases} \mathbb{E}[l_i^2 | x_i, z_i] - c_j \mathbb{E}[l_i | x_i, z_i] & \text{if } z_i \geq j \\ 0 & \text{if } z_i < j \end{cases} \\ \mathbb{E}[(l_i - c_j)_+^2 | x_i, z_i] &= \begin{cases} \mathbb{E}[l_i^2 | x_i, z_i] - 2c_j \mathbb{E}[l_i | x_i, z_i] + c_j^2 & \text{if } z_i \geq j \\ 0 & \text{if } z_i < j \end{cases} \\ \mathbb{E}[(l_i - c_j)_+(l_i - c_k)_+ | x_i, z_i] &= \begin{cases} \mathbb{E}[l_i^2 | x_i, z_i] - (c_j + c_k)\mathbb{E}[l_i | x_i, z_i] + c_j c_k & \text{if } z_i \geq j \text{ and } z_i \geq k \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.32}$$

Therefore $\mathbb{E}[d_i d_i^T | x_i]$ is just a function of $\mathbb{E}[l_i | x_i, z_i]$ and $\mathbb{E}[l_i^2 | x_i, z_i]$, which can also be easily obtained by known formulas for the one-dimensional truncated normal moments.

3.2.4 Truncated Normal Expressions

We show that $\mathbb{P}(L|X,Z)$ is a truncated normal distribution. This can be done using a simple proportionality argument:

$$\begin{aligned}\mathbb{P}(L|X=x,Z=z) &\propto \mathbb{P}(L,X=x,Z=z)\\ &\propto \mathbb{P}(L,X=x|Z=z)\\ &\propto h_z(L,X=x)I(c_z<L\leq c_{z+1})\\ &\propto h_z(L|X=x)I(c_z<L\leq c_{z+1}),\end{aligned} \tag{3.33}$$

where $h_z(L,X)$ is the hypothetically non-truncated joint distribution of L and X, whose parameters are given by the conditional model dictated by the event $Z=z$. That is, we can take the conditional model parameters (Equation 3.6) and consider a non-truncated version with the distribution:

$$h_z(X,L)=\mathcal{N}_{p+1}\left(\begin{bmatrix}\alpha_z\\0\end{bmatrix},\begin{bmatrix}\Lambda_z\Lambda_z^T+\Omega & \Lambda_z\\ \Lambda_z^T & 1\end{bmatrix}\right), \tag{3.34}$$

and by standard conditional Gaussian properties, $h_z(L|X)$ is

$$\begin{aligned}h_z(L|X)&\sim\mathcal{N}\left(\mu_{z,L|X},\Sigma_{z,L|X}\right)\\ \mu_{z,L|X}&=\Lambda_z^T\left(\Lambda_z\Lambda_z^T+\Omega\right)^{-1}(x-\alpha_z)\\ \Sigma_{z,L|X}&=1-\Lambda_z^T\left(\Lambda_z\Lambda_z^T+\Omega\right)^{-1}\Lambda_z.\end{aligned} \tag{3.35}$$

Finally, the expressions for $\mathbb{E}[l_i|x_i,z_i]$ and $\mathbb{E}[l_i^2|x_i,z_i]$ can be obtained through the known mean and variance expressions for a one-dimensional truncated normal. Suppose we have a generic one-dimensional Gaussian random variable $Y\sim\mathcal{N}(\mu,\sigma^2)$ and a support range $(k_1,k_2]$. Then the mean and variance of the truncated version of Y are

$$\begin{aligned}\mathbb{E}[Y|k_1<Y\leq k_2]&=\mu-\sigma\frac{f(k_2)-f(k_1)}{F(k_2)-F(k_1)}\\ \mathrm{Var}(Y|k_1<Y\leq k_2)&=\sigma^2\left[\frac{\kappa_2 f(k_2)-\kappa_1 f(k_1)}{F(k_2)-F(k_1)}-\left(\frac{f(k_2)-f(k_1)}{F(k_2)-F(k_1)}\right)^2\right]\end{aligned} \tag{3.36}$$

where $\kappa_j \coloneqq (k_j - \mu)/\sigma$, and $f(\cdot)$ and $F(\cdot)$ is the PDF and CDF of Y, respectively (Johnson et al., 1994).

3.3 Simulation Study

We conducted a simple simulation study to examine the performance of the EM-algorithm for the spline factor analysis model. Data were generated according to Equation 3.2, choosing $\boldsymbol{\beta}$ such that a variety of patterns were displayed. These patterns are illustrated in Figure 3.1. Numerically speaking, these came from the following $\boldsymbol{\beta}$ matrix:

$$\boldsymbol{\beta} = \begin{bmatrix} 3 & 6 & 9 & -3 & -6 \\ -6 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 5 \\ -1 & -2 & -3 & -4 & -5 \end{bmatrix}, \tag{3.37}$$

using the knot values of $c = (-\infty, -0.842, -0.253, 0.253, 0.842, \infty)$, which correspond to the points on a Gaussian distribution such that $\mathbb{P}(Z = z) = 0.2$ for all z. Various values for the diagonal of Ω were also chosen as $\omega = (1, 2, 5, 8)$. The sample size was set to $n = 1000$ and the number of total datasets was also 1000. On each dataset, we estimated the parameters using the EM-Algorithm described above, and for a baseline measure we used the MLE solution treating L as observed.

The simulation results are displayed in Table 3.1. We show the empirical bias, variance, and mean square errors (MSE) averaged over each observed variable (X_i), as well as over all $\boldsymbol{\beta}$ parameters. Both methods showed very little empirical bias, with the EM-algorithm having an overall average of -0.006 and the known L MLE having less than 0.001. The average empirical variances per observed variable for the EM-algorithm ranged from 0.374 to 1.451, compared to the known L MLE's range of 0.110 to 0.793. The overall average variance of the EM-algorithm was 2.922 times higher than the known L MLE. For the average empirical MSEs, the EM-algorithm had a range of 1.066 to 3.836, where the known L MLE had a range of 0.110 to 0.794. The overall average MSE of the EM-algorithm was 4.888 times higher than the known L MLE.

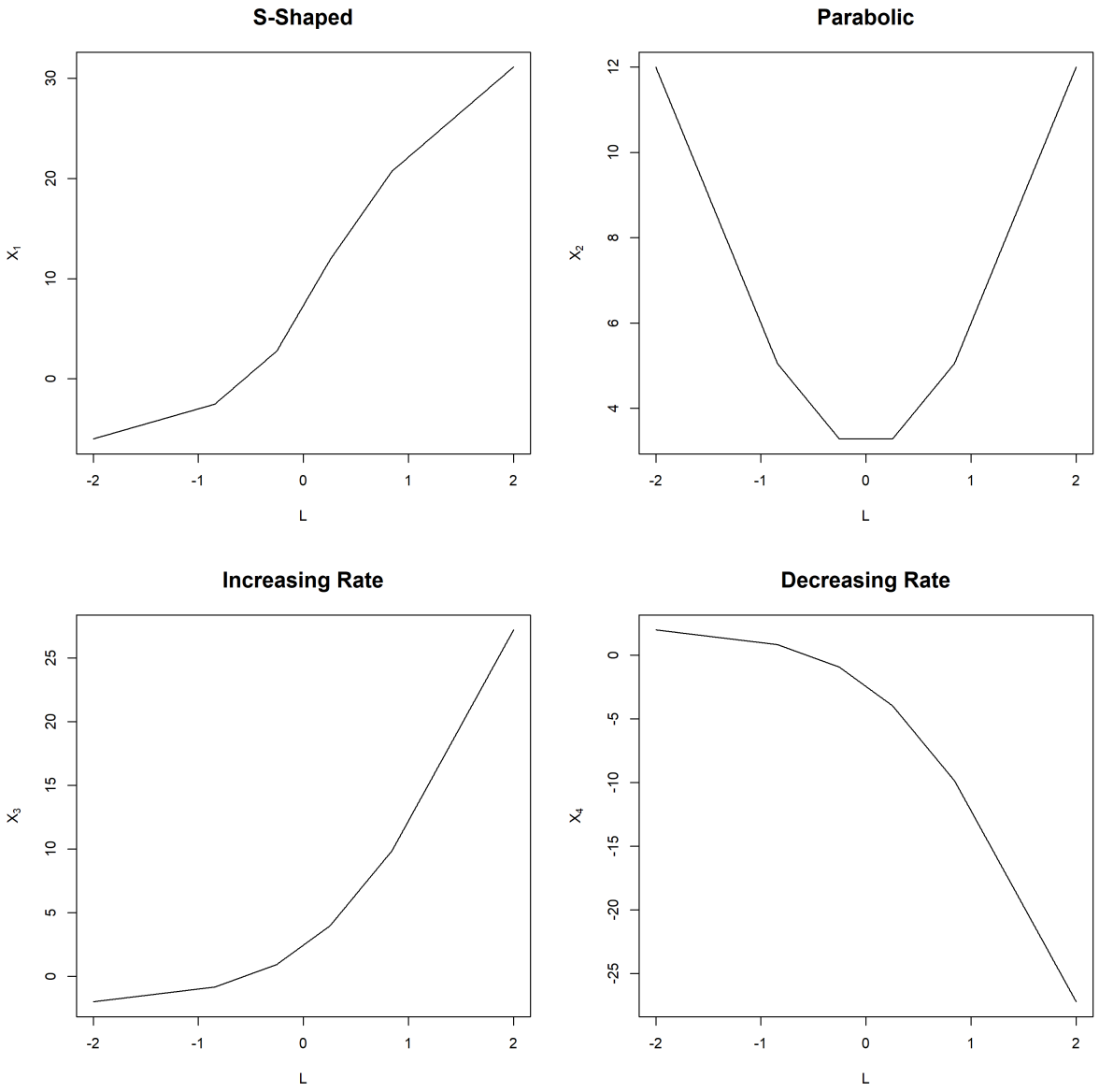

Figure 3.1: The data generating patterns for the piecewise linear factor analysis simulation. We use an S-shaped curve, a parabolic curve, a curve with an increasing rate of change and a curve with a decreasing rate of change.

3.3.1 Discussion and Future Work

In this chapter, we designed an EM-algorithm to estimate a non-linear factor analysis model using a piecewise linear construction. This work was done under the scalar L setting. Future work may extend this to multidimensional L (discussed in more detail in Chapter 4) and smoother curves, but both of these extensions comes with several challenges. Smoother curves may be achieved by generalizing the piecewise linear component with higher-order terms, such as cubic splines. However, one of the challenges of using higher

Method	Function	$\overline{\hat{\beta}_{ij} - \beta_{ij}}$	$\overline{\widehat{\text{Var}}(\hat{\beta}_{ij})}$	$\overline{(\hat{\beta}_{ij} - \beta_{ij})^2}$
EM-Algorithm	S-Shaped	0.006	1.363	3.836
	Parabolic	−0.030	0.374	0.533
	Increasing Rate	−0.075	0.988	1.066
	Decreasing Rate	0.074	1.451	1.555
	Overall	−0.006	1.044	1.748
Known L MLE	S-Shaped	0.000	0.110	0.110
	Parabolic	0.002	0.217	0.217
	Increasing Rate	−0.001	0.310	0.310
	Decreasing Rate	−0.002	0.793	0.794
	Overall	0.000	0.357	0.358

Table 3.1: Results of the piecewise linear factor analysis simulation. Displayed are the empirical biases, variances and mean square errors averaged across the $\boldsymbol{\beta}$ parameter per X_i, characterized by the "function" column.

order components is the computation of higher-order moments, which in the cubic case would result in the need to compute up to $E[L^6|X]$. This may result in greater estimation variance than their lower-order counterparts. A second challenge is that smoother curves may be more flexible, and thus require regularization to prevent overfitting. Additionally, if the curve is flexible enough, cases where the curve intersects itself may need to be carefully considered. These issues may be remedied by appropriately targeted regularization, or setting curvature constraints on the surface of L.

Chapter 4

Extensions and Miscellanea

In this chapter we describe some preliminary work and challenges on extending the piecewise linear spline method to the multiple factor case. The main challenge is that the region of each bin on L becomes hyper-rectangular on the multivariate Gaussian, which is notoriously difficult to integrate. For an iterative procedure to be used (such as the EM-algorithm), the integration must be very computationally fast as well as accurate, since each bin probability must be calculated at every parameter update. In this chapter, we briefly examine a variational EM-algorithm that obviates this issue. Sampling-based methods may also handle this issue, which are discussed in Section 4.5.

4.1 Model

We may generalize the model in Equation 3.2 to the multivariate L case as follows

$$X = \sum_{k=1}^{d} \left[\beta_{0k} L_k + \sum_{j=1}^{m_k} \beta_{jk} (L_k - c_{jk})_+ \right] + \epsilon, \tag{4.1}$$

where $X \in \mathbb{R}^p$ is a vector of observed variables, $L \sim \mathcal{N}_d(0, I_d)$ is a random vector of latent variables, $\epsilon \sim \mathcal{N}_p(0, \Omega)$ is a random vector of errors (with Ω being a diagonal matrix), $\beta_{jk} \in \mathbb{R}^p$ are the regression coefficients, and $c_{jk} \in \{1, \ldots, m_k\}$ are knot locations for the kth latent variable. The term $(L_k - c_{jk})_+$ is the positive part of $(L_k - c_{jk})$ and can be

thought of as a threshold function:

$$(L_k - c_{jk})_+ \coloneqq \max(L_k - c_{jk}, 0) = \begin{cases} L_k - c_{jk} & \text{if } L_k > c_{jk} \\ 0 & \text{if } L_k \leq c_{jk}. \end{cases} \tag{4.2}$$

4.1.1 Conditional Model Form

As before, it is useful to condition on the event $L_k \in (c_{jk} < L_k \leq c_{(j+1)k}]$. Hence we can define a random vector Z such that

$$Z_k = z_k \Leftrightarrow c_{zk} < L_k \leq c_{(z+1)k}, \text{ for } z_k \in \{0, \ldots, m_k\}, \tag{4.3}$$

where $c_{0k} = -\infty$ and $c_{m_k k} = \infty$. Then notice that we have the following decomposition of Equation (4.1) about z:

$$\begin{aligned} X &= \sum_{k=1}^{d} \left[\beta_{0k} L_k + \sum_{j=1}^{m_k} \beta_{jk} (L_k - c_{jk})_+ \right] + \epsilon \\ &= \sum_{k=1}^{d} \left[\beta_{0k} L_k + \sum_{j=1}^{z_k} \beta_j (L_k - c_{jk})_+ + \sum_{j=z_k+1}^{m_k} \beta_{jk} (L - c_{jk})_+ \right] + \epsilon. \end{aligned} \tag{4.4}$$

Now, if we condition on the event $Z = z$, we obtain the following conditional model:

$$\begin{aligned} X &= \sum_{k=1}^{d} \left[\beta_{0k} L_k + \sum_{j=1}^{z_k} \beta_{jk} (L_k - c_{jk}) \right] + \epsilon && \text{if } Z = z \\ &= \sum_{k=1}^{d} \beta_{0k} L_k + \sum_{k=1}^{d} \sum_{j=1}^{z_k} \beta_{jk} L_k - \sum_{k=1}^{d} \sum_{j=1}^{z_k} \beta_{jk} c_{jk} + \epsilon && \text{if } Z = z \\ &= \left(- \sum_{k=1}^{d} \sum_{j=1}^{z_k} \beta_{jk} c_{jk} \right) + \left(\sum_{k=1}^{d} \sum_{j=0}^{z_k} \beta_{jk} L_k \right) + \epsilon && \text{if } Z = z. \end{aligned} \tag{4.5}$$

From here, define

$$\begin{aligned} \alpha_z &= \left(- \sum_{k=1}^{d} \sum_{j=1}^{z_k} \beta_{jk} c_{jk} \right) \\ \Lambda_z &= \begin{bmatrix} \sum_{j=0}^{z_1} \beta_{j1} & \cdots & \sum_{j=0}^{z_d} \beta_{jd} \end{bmatrix}, \end{aligned} \tag{4.6}$$

and we can re-cast Equation 4.5 as a linear factor analysis model

$$\begin{aligned} X &= \alpha_z + \Lambda_z L + \epsilon \quad \text{if } Z = z \\ \Rightarrow \Sigma_z &= \Lambda_z \Phi_z \Lambda_z + \Omega \quad \text{if } Z = z \end{aligned} \tag{4.7}$$

where $\Phi_z \coloneqq \text{Var}(L|Z)$ which is a truncated normal variance.

4.1.2 Multivariate Regression Form

The model in Equation 4.1 can be described as a multivariate regression model using the following framework

$$X = \boldsymbol{\beta} d(L) + \epsilon, \tag{4.8}$$

where $\boldsymbol{\beta}$ is a $p \times M$ matrix of regression coefficients with $M = d + \sum_{k=1}^{d} m_k$ and $d(L)$ is a *design function* that maps L to a vector of *design variables* which serve as regressors to the model as

$$d(L) = \begin{bmatrix} L_1 & (L_1 - c_{11})_+ & \cdots & (L_1 - c_{m_1 1})_+ & \cdots & L_d & (L_d - c_{1d})_+ & \dots & (L_d - c_{m_d d})_+ \end{bmatrix}^T. \tag{4.9}$$

Some computational formulas are as follows. Define a computational version of c as

$$c_\alpha = \begin{bmatrix} 0 & c_{11} & \dots & c_{m_1 1} & \dots & 0 & c_{1d} & \dots & c_{m_d d} \end{bmatrix}, \tag{4.10}$$

which is the stacked vector of all c_{jk}, for all $j \in \{0, \dots, m_k\}$ and $k \in \{1, \dots d\}$, except that all $c_{0k} = 0$. Also, let us define a binary version of Z as follows.

$$\widetilde{Z} = \begin{bmatrix} 1 & I(c_{11} < L_1) & \dots & I(c_{m_1 1} < L_1) & \dots & 1 & I(c_{1d} < L_d) & \dots & I(c_{m_d d} < L_d) \end{bmatrix} \tag{4.11}$$

which is the stacked vector of all $I(c_{jk} < L_k)$, for all $j \in \{0, \dots, m_k\}$ and $k \in \{1, \dots d\}$, except that all the entry corresponding to c_{0k} is fixed to 1. Then we have the following

computational formulas

$$\begin{aligned}\alpha_z &= -\boldsymbol{\beta}\text{diag}(\widetilde{Z})c_\alpha \\ \Lambda_z &= \boldsymbol{\beta}\text{bdiag}(\widetilde{Z}),\end{aligned} \tag{4.12}$$

where $\text{bdiag}(\widetilde{Z})$ is a block-diagonal matrix with each block corresponding to the kth stack of $\widetilde{Z}$.

4.2 Conditional Expectations

The multivariate regression form described in Section 4.1.2 is convenient as its parameterization is identical to the one used for the Q-function in Section 3.2.2. Hence, the maximizers of the Q-function remain the same and we only need to generalize the conditional expectations to the multivariate case.

As before, for the conditional expectations we need $\mathbb{E}[d_i|x_i]$ and $E[d_i d_i^T|x_i]$. Since the threshold functions depend on Z, it will be easier to use the law of iterated expectation to obtain

$$\begin{aligned}\mathbb{E}[d_i|x_i] &= \mathbb{E}[\mathbb{E}[d_i|x_i, z_i]|x_i] \\ &= \sum_z \mathbb{E}[d_i|x_i, z_i]\mathbb{P}(z_i|x_i),\end{aligned} \tag{4.13}$$

and calculate a series of $\mathbb{E}[d_i|x_i, z_i]$ instead. Thus for a general threshold function $(l_{ik}-c_{jk})_+$ we have

$$\mathbb{E}[(l_{ik} - c_{jk})_+|x_i, z_i] = \begin{cases}\mathbb{E}[l_{ik}|x_i, z_i] - c_{jk} & \text{if } z_{ik} \geq j \\ 0 & \text{if } z_{ik} < j\end{cases}. \tag{4.14}$$

For $E[d_i d_i^T|x_i]$, note that we have to consider all possible cross-products among l_{ik} and

$(l_{ik} - c_{jk})$. For the expectations that contain threshold functions, we have

$$\mathbb{E}[l_{ik}(l_{ig} - c_{jg})_+|x_i, z_i] = \begin{cases} \mathbb{E}[l_{ik}l_{ig}|x_i, z_i] - c_{jg}\mathbb{E}[l_{ik}|x_i, z_i] & \text{if } z_{ig} \geq j \\ 0 & \text{if } z_{ig} < j \end{cases}$$

$$\mathbb{E}[(l_{ik} - c_{jk})_+^2|x_i, z_i] = \begin{cases} \mathbb{E}[l_{ik}^2|x_i, z_i] - 2c_{jk}\mathbb{E}[l_{ik}|x_i, z_i] + c_{jk}^2 & \text{if } z_{ik} \geq j \\ 0 & \text{if } z_{ik} < j \end{cases}.$$

$$\mathbb{E}[(l_{ik} - c_{jk})_+(l_{ig} - c_{ag})_+|x_i, z_i] = \begin{cases} \mathbb{E}[l_{ik}l_{ig}|x_i, z_i] - c_{ag}\mathbb{E}[l_{ik}|x_i, z_i] & \\ \qquad - c_{jk}\mathbb{E}[l_{ig}|x_i, z_i] + c_{jk}c_{ag} & \text{if } z_{ik} \geq j \text{ and } z_{ig} \geq a \\ 0 & \text{otherwise} \end{cases} \tag{4.15}$$

In what follows, we will use these expectations to set up the E-step of a *variational*-EM algorithm. As in the traditional EM-algorithm, this will involve taking the expectation with respect to $L|X$ over the complete data log-likelihood. Since L is written as a function of $d(L)$, the conditional expectations derived here allow us to construct $E[d(L)|X)]$ from $E[L|X, Z]$ and proceed with maximization in a straightforward manner.

4.3 Variational EM Algorithm

For multi-dimensional L, finding the probability of $c_{zk} < L_k \leq c_{(z+1)k}$ jointly over all $k \in [d]$ is a difficult integration problem over the multivariate Gaussian distribution. This makes the EM algorithm difficult to proceed with in the multiple L case. Therefore, we may perform a variational-EM algorithm (Beal, 2003), which maximizes a lower-bound of the log-likelihood as follows. Let $q(\cdot)$ be some distribution such that allows for the

tractability of the E-step, parameterized by γ:

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{n} \log \mathbb{P}_\theta(x_i) \\
&= \sum_{i=1}^{n} \log \int_{l_i} \mathbb{P}_\theta(x_i, l_i)\, dl_i \\
&= \sum_{i=1}^{n} \log \int_{l_i} q_{\gamma_i}(l_i) \frac{\mathbb{P}_\theta(x_i, l_i)}{q_{\gamma_i}(l_i)}\, dl_i \\
&= \sum_{i=1}^{n} \log \mathbb{E}_{\gamma_i} \left[\frac{\mathbb{P}_\theta(x_i, l_i)}{q_{\gamma_i}(l_i)} \right] \\
&\geq \sum_{i=1}^{n} \mathbb{E}_{\gamma_i} \left[\log \frac{\mathbb{P}_\theta(x_i, l_i)}{q_{\gamma_i}(l_i)} \right] \\
&:= \mathcal{V}(\theta, \gamma),
\end{aligned} \tag{4.16}$$

which follows from Jensen's inequality. This yields a variational EM algorithm as follows. Given an initial start value $\theta^{(0)}$:

$$\begin{aligned}
&\textbf{Step 1.}\ \ \gamma^{(t+1)} = \operatorname*{argmax}_{\gamma} \mathcal{V}(\theta^{(t)}, \gamma). \\
&\textbf{Step 2.}\ \ \theta^{(t+1)} = \operatorname*{argmax}_{\theta} \mathcal{V}(\theta, \gamma^{(t+1)}).
\end{aligned} \tag{4.17}$$

To make the problem of finding the joint probability of all $c_{zk} < L_k \leq c_{(z+1)k}$ tractable, we will use a $q_{\gamma_i}(L_i)$ that approximates $\mathbb{P}_\theta(L_i|X_i)$ with the constraint that $L_{ij}, \ldots, L_{id}$ are all mutually independent under $q_{\gamma_i}(L_i)$. Notice in Equation 4.16, if γ is considered a constant, then $\mathcal{V}$ is identical to the Q-function of Equation 3.9 up to an additive constant. Hence, Step 2 is identical to the EM steps of Chapter 3. Thus we will derive Step 1, the maximization of $\mathcal{V}$ with respect to γ.

4.3.1 Maximizing with Respect to γ

To maximize with respect to γ, we may equivalently write $\mathcal{V}(\theta, \gamma)$ as follows:

$$\begin{aligned}
\mathcal{V}(\theta,\gamma) &= \sum_{i=1}^{n} \mathbb{E}_{\gamma_i}\left[\log \frac{\mathbb{P}_\theta(x_i, l_i)}{q_{\gamma_i}(l_i)}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{\gamma_i}\left[\log \frac{\mathbb{P}_\theta(l_i|x_i)\mathbb{P}_\theta(x_i)}{q_{\gamma_i}(l_i)}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{\gamma_i}\left[\log \mathbb{P}_\theta(x_i) + \log \frac{\mathbb{P}_\theta(l_i|x_i)}{q_{\gamma_i}(l_i)}\right] \\
&= \sum_{i=1}^{n} \log \mathbb{P}_\theta(x_i) + \sum_{i=1}^{n} \mathbb{E}_{\gamma_i}\left[\log \frac{\mathbb{P}_\theta(l_i|x_i)}{q_{\gamma_i}(l_i)}\right] \\
&= \sum_{i=1}^{n} \log \mathbb{P}_\theta(x_i) - \sum_{i=1}^{n} \mathbb{E}_{\gamma_i}\left[\log \frac{q_{\gamma_i}(l_i)}{\mathbb{P}_\theta(l_i|x_i)}\right] \\
&= \sum_{i=1}^{n} \log \mathbb{P}_\theta(x_i) - \sum_{i=1}^{n} D_{KL}\left(q_{\gamma_i}(l_i) \parallel \mathbb{P}_\theta(l_i|x_i)\right) \\
&= -\sum_{i=1}^{n} D_{KL}\left(q_{\gamma_i}(l_i) \parallel \mathbb{P}_\theta(l_i|x_i)\right) + c.
\end{aligned} \tag{4.18}$$

Therefore, maximizing $\mathcal{V}(\theta, \gamma)$ with respect to γ is equivalent to minimizing the KL-divergences per observation.

In order to do this, we must choose an appropriate $q_{\gamma_i}(L_i)$ distribution ideally similar to $\mathbb{P}_\theta(L|X)$. First notice that $\mathbb{P}_\theta(L|X)$ can be characterized as a piecewise Gaussian in the following way. Let $S(Z)$ be the sample space of Z, which would be all permutations of Z_k.

$$\begin{aligned}
\mathbb{P}_\theta(L|X) &= \sum_{z \in S(Z)} \mathbb{P}(L|X, Z = z)\mathbb{P}(Z = z|X) \\
&= \sum_{z \in S(Z)} \frac{h_z(L|X)I(Z = z)}{\mathbb{P}(Z = z|X)}\mathbb{P}(Z = z|X) \\
&= \sum_{z \in S(Z)} h_z(L|X)I(Z = z) \\
&= \prod_{z \in S(Z)} h_z(L|X)^{I(Z=z)},
\end{aligned} \tag{4.19}$$

where we used the definition of a truncated distribution, and $h_z(L|X)$ is the hypothetically untruncated Gaussian distribution of $L|X$ as defined in Equation 3.35. Thus we can choose $q_{\gamma_i}(L_i)$ to also be a piecewise Gaussian, where γ_i contains the parameters of the underlying hypothetically untruncated Gaussian distributions of the pieces. That is, using

the logic from the previous equation, we could also write

$$q_{\gamma_i}(L_i) = \prod_{z_i \in S(Z_i)} q_{z_i}(L_i)^{I(Z_i = z_i)}, \tag{4.20}$$

where $q_{z_i}(L_i)$ is the hypothetically untruncated Gaussian distribution corresponding to the region z_i, with parameters $\mu_{\gamma_{z_i}}$ and $\Sigma_{\gamma_{z_i}}$. Additionally, we will introduce a *marginal bin* parameter:

$$\pi_{\gamma_{z_i k}} \coloneqq \mathbb{P}_{\gamma_i}(c_{z_i k} < L_{ik} \leq c_{(z_i+1)k}), \tag{4.21}$$

which is simply the probability of the kth latent variable falling in the marginal region $(c_{zk}, c_{(z+1)k}]$. Therefore, γ_i is the set that contains all the $\mu_{\gamma_{z_i}}$, $\Sigma_{\gamma_{z_i}}$, and $\pi_{\gamma_{z_i k}}$ parameters, for all $z \in S(Z)$ and $k \in [d]$.

To simplify the E-step, we will impose two constraints on γ_i. First, we will assume that all $\Sigma_{\gamma_{z_i}}$ are diagonal matrices. Second, we will constrain

$$\pi_{\gamma_{z_i k}} = \mathbb{P}_\theta(c_{z_i k} < L_{ik} \leq c_{(z_i+1)k} | x_i). \tag{4.22}$$

That is, the probability of the event $(c_{z_i k} < L_{ik} \leq c_{(z_i+1)k})$ is identical under $q_{\gamma_i}(L_i)$ and $\mathbb{P}_\theta(L|X = x_i)$. The simplification offered by this second constraint will be made apparent as we derive the E-step.

With the piecewise Gaussian forms of $q_{\gamma_i}(L_i)$ and $\mathbb{P}_\theta(L|X_i)$ in hand, our strategy will be to obtain the KL-divergence between these two distributions in terms of their

hypothetically untruncated Gaussian components. For any given observation i, we have

$$\begin{aligned}
D_{KL}\left(q_{\gamma_i}(l_i) \parallel \mathbb{P}_\theta(l_i|x_i)\right) &= \mathbb{E}_{\gamma_i}\left[\log \frac{q_{\gamma_i}(l_i)}{\mathbb{P}_\theta(l_i|x_i)}\right] \\
&= \mathbb{E}_{\gamma_i}\left[\log q_{\gamma_i}(l_i) - \log \mathbb{P}_\theta(l_i|x_i)\right] \\
&= \mathbb{E}_{\gamma_i}\left[\log \prod_{z_i \in S(Z_i)} q_{z_i}(l_i)^{I(Z_i=z_i)} - \log \prod_{z_i \in S(Z_i)} h_{z_i}(l_i|x_i)^{I(Z_i=z_i)}\right] \\
&= \mathbb{E}_{\gamma_i}\left[\sum_{z_i \in S(Z_i)} I(Z_i = z_i) \log q_{z_i}(l_i) - \sum_{z_i \in S(Z_i)} I(Z_i = z_i) \log h_{z_i}(l_i|x_i)\right] \\
&= \mathbb{E}_{\gamma_i}\left[\sum_{z_i \in S(Z_i)} I(Z_i = z_i) \log \frac{q_{z_i}(l_i)}{h_{z_i}(l_i|x_i)}\right] \\
&= \mathbb{E}_{\gamma_i, Z_i}\left[\mathbb{E}_{\gamma_i, L_i|Z_i}\left[\sum_{z_i \in S(Z_i)} I(Z_i = z_i) \log \frac{q_{z_i}(l_i)}{h_{z_i}(l_i|x_i)} \middle| Z_i\right]\right] \\
&= \mathbb{E}_{\gamma_i, Z_i}\left[\sum_{z_i \in S(Z_i)} I(Z_i = z_i) D_{KL}\left(q_{z_i}(l_i) \parallel h_{z_i}(l_i|x_i)\right)\right] \\
&= \sum_{z_i \in S(Z_i)} \mathbb{P}_{\gamma_i}(Z_i = z_i) D_{KL}\left(q_{z_i}(l_i) \parallel h_{z_i}(l_i|x_i)\right) \\
&= \sum_{z_i \in S(Z_i)} \prod_{k=1}^{d} \mathbb{P}_\theta(c_{z_i k} < L_k \leq c_{(z_i+1)k}|x_i) D_{KL}\left(q_{z_i}(l_i) \parallel h_{z_i}(l_i|x_i)\right),
\end{aligned} \tag{4.23}$$

where the final equality follows from the constraints imposed on γ_i. That is, due to all L_{ik} being independent within each piece, as well as the constraint characterized by Equation 4.22, we have

$$\begin{aligned}
\mathbb{P}_{\gamma_i}(Z_i = z_i) &= \prod_{k=1}^{d} \mathbb{P}_{\gamma_i}(c_{z_i k} < L_k \leq c_{(z_i+1)k}) \\
&= \prod_{k=1}^{d} \mathbb{P}_\theta(c_{z_i k} < L_k \leq c_{(z_i+1)k}|x_i).
\end{aligned} \tag{4.24}$$

From here, we can see that the KL-divergence between $q_{\gamma_i}(l_i)$ and $\mathbb{P}_\theta(l_i|x_i)$ is a weighted sum of the KL-divergences of their hypothetically untruncated Gaussian components, $q_{z_i}(l_i)$ and $h_{z_i}(l_i|x_i)$. Since each D_{KL} are comprised of independent parameters in γ_i, it is sufficient to minimize each D_{KL} individually per z_i. Let $\mu_{\theta_{z_i}}$ and $\Sigma_{\theta_{z_i}}$ denote the mean and covariance parameters, respectively, for $h_{z_i}(l_i|x_i)$. Then the KL-divergence for

each z_i is

$$D_{KL}\left(q_{z_i}(l_i) \parallel h_{z_i}(l_i|x_i)\right) = \frac{1}{2}\left[\text{tr}(\Sigma_{\theta_{z_i}}^{-1}\Sigma_{\gamma_{z_i}}) + (\mu_{\theta_{z_i}} - \mu_{\gamma_{z_i}})^T\Sigma_{\theta_{z_i}}^{-1}(\mu_{\theta_{z_i}} - \mu_{\gamma_{z_i}}) - d + \log\frac{|\Sigma_{\theta_{z_i}}|}{|\Sigma_{\gamma_{z_i}}|}\right]. \tag{4.25}$$

To minimize with respect to $\mu_{\gamma_{z_i}}$, we can collect the relevant terms as

$$D_{KL}(\mu_{\gamma_{z_i}}) \propto (\mu_{\theta_{z_i}} - \mu_{\gamma_{z_i}})^T\Sigma_{\theta_{z_i}}^{-1}(\mu_{\theta_{z_i}} - \mu_{\gamma_{z_i}}) + c, \tag{4.26}$$

then differentiating and setting equal to zero

$$\begin{aligned}\frac{\partial D_{KL}}{\partial \mu_{\gamma_{z_i}}} &\propto -2\Sigma_{\theta_{z_i}}^{-1}(\mu_{\theta_{z_i}} - \mu_{\gamma_{z_i}}) = 0 \\ &\Rightarrow \mu_{\theta_{z_i}} = \mu_{\gamma_{z_i}},\end{aligned} \tag{4.27}$$

trivially shows that the value that minimizes D_{KL} with respect to $\mu_{\gamma_{z_i}}$ is simply $\mu_{\theta_{z_i}}$. Now to minimize with respect to $\Sigma_{\gamma_{z_i}}$ we once again collect relevant terms as

$$\begin{aligned}D_{KL}(\Sigma_{\gamma_{z_i}}) &\propto \text{tr}(\Sigma_{\theta_{z_i}}^{-1}\Sigma_{\gamma_{z_i}}) + \log\frac{|\Sigma_{\theta_{z_i}}|}{|\Sigma_{\gamma_{z_i}}|} + c \\ &= \text{tr}(\Sigma_{\theta_{z_i}}^{-1}\Sigma_{\gamma_{z_i}}) - \log|\Sigma_{\gamma_{z_i}}| + c' \\ &= \text{tr}(\text{diag}(\Sigma_{\theta_{z_i}}^{-1})\Sigma_{\gamma_{z_i}}) - \log|\Sigma_{\gamma_{z_i}}| + c',\end{aligned} \tag{4.28}$$

which makes use of the fact that $\Sigma_{\gamma_{z_i}}$ is diagonal. Then differentiating and setting equal to zero

$$\begin{aligned}\frac{\partial D_{KL}}{\partial \Sigma_{\gamma_{z_i}}} &\propto \text{diag}(\Sigma_{\theta_{z_i}}^{-1}) - \Sigma_{\gamma_{z_i}}^{-1} = 0 \\ &\Rightarrow \text{diag}(\Sigma_{\theta_{z_i}}^{-1}) = \Sigma_{\gamma_{z_i}}^{-1}.\end{aligned} \tag{4.29}$$

Thus, the component-wise parameters in γ_i that minimize D_{KL} are simply the same as those in θ, except we truncate the off-diagonals of $\Sigma_{\theta_{z_i}}$ to zero. That is,

$$\begin{aligned}\mu_{\gamma_{z_i}}^{(t+1)} &= \underset{\mu_{\gamma_{z_i}}}{\text{argmax}}\, \mathcal{V}(\theta^{(t)}, \gamma) = \mu_{\theta_{z_i}}^{(t)} \\ \Sigma_{\gamma_{z_i}}^{(t+1)} &= \underset{\Sigma_{\gamma_{z_i}}}{\text{argmax}}\, \mathcal{V}(\theta^{(t)}, \gamma) = \text{diag}(\Sigma_{\theta_{z_i}}^{(t)}).\end{aligned} \tag{4.30}$$

Method	$\overline{\hat{\beta}_{ij} - \beta_{ij}}$	$\overline{\widehat{\text{Var}}(\hat{\beta}_{ij})}$	$\overline{(\hat{\beta}_{ij} - \beta_{ij})^2}$	Var Ratio	MSE Ratio
Regularized-VEM	-0.425	2.301	15.149	5.339	35.148
Known L MLE	-0.001	0.431	0.431	-	-

Table 4.1: Results of the piecewise linear factor analysis simulation with multiple L. Displayed are the empirical biases, variances and mean square errors averaged across all $\boldsymbol{\beta}$ parameters. Var ratio and MSE ratio are the ratios of empirical variances and MSE between the regularized variational-EM and the known L MLE, also averaged over all $\boldsymbol{\beta}$.

4.4 Simulation Study

To test the variational-EM approach, we conducted a small simulation study. The simulation was designed identically to the simulation in Section 3.3 with the following changes. First, rather than a latent variable scalar, we used a 3-dimensional latent variable $L \sim \mathcal{N}_3(0, I_3)$. Second, while we used the same non-linear functions as the 1-dimensional simulation, each latent variable added a different independent copy of each function. That is, if we denote the design function from the 1-dimensional simulation as $d_1(\cdot)$ and the current design function as $d_3(\cdot)$, we have a block vector

$$d_3(L) = \begin{bmatrix} d_1(L_1) & d_1(L_2) & d_1(L_3) \end{bmatrix}^T. \tag{4.31}$$

Third, we note that variational-EM algorithm was used (Equation 4.17), with the caveat that we added an ℓ_2 regularizer on the loss function in Equation 4.16. Regularization was needed to help the algorithm converge and a regularization constant was arbitrarily set to 5. We compared this method to the MLE estimates that used L as observed data.

The simulation results are displayed in Table 4.1. We show the empirical bias, variance, and mean square errors averaged over all the $\boldsymbol{\beta}$ parameters. In addition, we calculated the ratio of empirical variances and MSE between the regularized variational-EM method and known L MLE methods, also averaged over all $\boldsymbol{\beta}$. The results show that the regularized variational-EM estimates show a small amount of bias, which is to be expected with regularization. The empirical variance was 2.301, which was 5.339 times higher than the known L on average, which was 0.431. Further, the empirical MSE was 15.149, which was 35.148 times higher than the known L MLE, which was 0.431.

Comparing these variance and MSE ratios to the EM-algorithm method in 1-dimensional data, the regularized variational-EM ratios were moderately to much higher. The variational EM variance ratio was 5.339, compared to the EM-algorithm's 2.922, and the comparison of MSE ratio was 35.148 vs. 4.888. These ratios may be improved using an optimal choice of regularization constant, however this may be challenging since a complete run of the regularized variational-EM method took 9000 iterations (about 1 hour). Hence, a fast method of evaluating and choosing a regularization constant under the EM-framework is a topic of further study.

4.5 Sampling-Based Methods

In addition to the variational-EM method, we may also consider sampling-based methods. Recall that the key difficulty of the model in Equation 4.1 is calculating the probability of a given region ($Z = z$) over the multivariate Gaussian. Another way to obviate this problem is through sampling-based methods.

Within the EM-framework, one technique is *stochastic* EM (Celeux and Diebolt, 1985), which replaces the E-step with a sampling step. In the case of our model, this amounts to replacing the computation of $\mathbb{E}_{\theta^{(t)}}[d_i|x_i]$ with a draw from $\mathbb{P}_{\theta^{(t)}}(d_i|x_i)$, for all $i \in [n]$. Then the maximization step is then carried out acting as if the sampled data were observed. Essentially, latent variables are imputed with a random draws followed by a parameter update. In turn, the sequence of $\theta^{(t)}$ becomes a Markov Chain (which is ergodic under general conditions; Nielsen, 2000), and a final estimate can be obtained by averaging over a final set of iterations, after discarding an initial burn-in set.

Alternatively, we may also consider a Gibbs sampling approach under a Bayesian framework (Geman and Geman, 1984). The multivariate L versions of the estimators described in Section 3.2.2 can be easily adapted into a Bayesian regression routine in the following way. Given start values $\boldsymbol{\beta}^{(0)}, \Omega^{(0)}$, with Gaussian and inverse-gamma priors, respectively:

1. Draw $L^{(t+1)}|X, \boldsymbol{\beta}^{(t)}, \Omega^{(t)}$ from a Gaussian distribution.

2. Compute $d^{(t+1)} = d(L^{(t+1)})$.

3. Draw $\boldsymbol{\beta}^{(t+1)}|X, d^{(t+1)}, \Omega^{(t)}$ from a Gaussian distribution.

4. Draw $\Omega^{(t+1)}|X, d^{(t+1)}, \boldsymbol{\beta}^{(t+1)}$ from an inverse-gamma distribution.

This procedure would provide empirical posterior distributions of $\boldsymbol{\beta}$ and Ω, which would also allow for the quantification of standard errors in a straightforward manner (Gelman et al., 2014).

4.6 Other Extensions

Akin to learning linear structures prior to estimation as in Chapter 2, we may augment the methods explored in Chapters 3 and 4 with a structure learning step as well. The clique search routines of Chapter 2 may readily be extended to non-linear correlation coefficients. That is, we may use coefficients that measure the degree of (non-linear) dependence between a pair of observed variables. This can be accomplished, for example, by examining the mutual information between pairs of variables (e.g., Reshef et al., 2011; Smith, 2015), or by analyzing and aggregating local linear dependencies (Delicado and Smrekar, 2009). Similar to the CT algorithm, this would allow for the number of latent variables and the presence of non-linear relations to be learned before carrying out estimation. The estimation routines described in Chapters 3 and 4 may then be carried out after specifying this learned structure. This would be a promising method for future research.

Bibliography

Adachi, K. and Trendafilov, N. T. (2018), "Sparsest factor analysis for clustering variables: A matrix decomposition approach," *Advances in Data Analysis and Classification*, 12, 559–585.

Anderson, J. C. and Gerbing, D. W. (1988), "Structural equation modeling in practice: A review and recommended two-step approach," *Psychological Bulletin*, 103, 411–423.

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

Baker, F. B. and Kim, S.-H. (2004), *Item response theory: Parameter estimation techniques*, New York, NY: CRC Press, 2nd ed.

Beal, M. J. (2003), *Variational algorithms for approximate Bayesian inference*, University of London, University College London (United Kingdom).

Bollen, K. A. (1980), "Issues in comparative measurement of political democracy," *American Sociological Review*, 45, 370–390.

Browne, M. W. (1968), "A comparison of factor analytic techniques," *Psychometrika*, 33, 1968.

— (2001), "An overview of analytic rotation in exploratory factor analysis," *Multivariate Behavioral Research*, 36, 111–150.

Cai, L., Choi, K., Hansen, M., and Harrell, L. (2016), "Item Response Theory," *Annual Review of Statistics and Its Application*, 3, 297–321.

Caner, M. and Han, X. (2014), "Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators," *Journal of Business & Economic Statistics*, 32, 359–374.

Carreira-Perpiñán, M. A. and Lu, Z. (2007), "The Laplacian Eigenmaps Latent Variable Model," in *Proceedings of Machine Learning Research*, vol. 2, p. 59–66.

Cattell, R. B. (1966), "The Scree Test for the number of factors," *Multivariate Behavioral Research*, 1, 245–276.

Celeux, G. and Diebolt, J. (1985), "The SEM algorithm: A probabalistic teacher algorithm derived from the EM algorithm for the mixture problem," *Computational Statistics Quarterly*, 2, 73–82.

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

Cliff, N. (1988), "The eigenvalues-greater-than-one rule and reliability of components," *Psychological Bulletin*, 103, 276–279.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012), "Learning high-dimensional directed acyclic graphs with latent and selection variables," *Annals of Statistics*, 40, 294–321.

Crawford, C. (1975), "A comparison of the direct oblimin and primary parsimony methods of oblique rotation," *British Journal of Mathematical and Statistical Psychology*, 28, 201–213.

Crawford, C. B. and Ferguson, G. A. (1970), "A general rotation criterion and its use in orthogonal rotation," *Psychometrika*, 35, 321–332.

Crawford, C. B. and Koopman, P. (1973), "A note on Horn's test for the number of factors in factor analysis," *Multivariate Behavioral Research*, 8, 117–125.

Cudeck, R., Harring, J. R., and du Toit, S. H. C. (2009), "Marginal maximum likelihood estimation of a latent variable model with interaction," *Journal of Educational and Behavioral Statistics*, 34, 131–144.

Delicado, P. and Smrekar, M. (2009), "Measuring non-linear dependence for two random variables distributed along a curve," *Statistics and Computing*, 19, 255–269.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society: Series B*, 39, 1–38.

Embretson, S. E. and Reise, S. P. (2000), *Item response theory for psychologists*, New York: Psychology Press, 1st ed.

Etezadi-Amoli, J. (1983), "A second generation nonlinear factor analysis," *Psychometrika*, 48, 315–342.

Ferguson, G. A. (1941), "The factorial interpretation of test difficulty," *Psychometrika*, 6, 323–329.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman & Hall.

Ford, J. K., MacCallum, R. C., and Tait, M. (1986), "The application of exploratory factor analysis in applied psychology: A critical review and analysis," *Personnel Psychology*, 39, 291–314.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014), *Bayesian Data Analysis*, Chapman & Hall/CRC Texts in Statistical Science, Boca Raton: CRC Press, third edition ed.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, 721–741.

Glorfeld, L. W. (1995), "An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain," *Educational and Psychological Measurement*, 55, 377–393.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014), "Generative adversarial nets," in *Advances in Neural Information Processing Systems*.

Guttman, L. (1954), "Some necessary conditions for common-factor analysis," *Psychometrika*, 19, 149–161.

Hakstian, A. R., Rogers, W. T., and Cattell, R. B. (1982), "The behavior of number-of-factors rules with simulated data," *Multivariate Behavioral Research*, 17, 193–219.

Han, T., Lu, Y., Song-Chun, Z., and Wu, Y. N. (2017), "Alternating back-propagation for generator network," in *Proceedings of the Thirty-First AAAI Conference on Artifical Intelligence*.

Harris, C. W. and Kaiser, H. F. (1964), "Oblique factor analytic solutions by orthogonal transformations," *Psychometrika*, 29, 347–362.

Hastie, T. and Stuetzle, W. (1989), "Principal curves," *Journal of the American Statistical Association*, 84, 502–516.

Hattie, J. (1985), "Methodology review: Assessing unidimensionality of tests and items," *Applied Psychological Measurement*, 9, 139–164.

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004), "Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis," *Organizational Research Methods*, 7, 191–205.

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

— (2014b), "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*.

Horn, J. L. (1965), "A rationale and test for the number of factors in factor analysis," *Psychometrika*, 30, 179–185.

Howard, M. C. (2016), "A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?" *International Journal of Human-Computer Interaction*, 32, 51–62.

Jennrich, R. I. (2001), "A simple general procedure for orthogonal rotation," *Psychometrika*, 66, 289–306.

— (2006), "Rotation to simple loadings using component loss functions: The oblique Case," *Psychometrika*, 71, 173–191.

Jennrich, R. I. and Robinson, S. M. (1969), "A Newton-Raphson algorithm for maximum likelihood factor analysis," *Psychometrika*, 34, 111–123.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions*, vol. 1 of *Wiley Series in Probability and Statistics*, New York: Wiley, 2nd ed.

Jöreskog, K. G. (1967), "Some contributions to maximum likelihood factor analysis," *Psychometrika*, 32, 443–482.

Kaiser, H. F. (1958), "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, 23, 187–200.

— (1960), "The application of electronic computers to factor analysis," *Educational and Psychological Measurement*, 20, 141–151.

Kalisch, M. and Bühlmann, P. (2007), "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *Journal of Machine Learning Research*, 8, 613–636.

Kingma, D. P. and Welling, M. (2019), "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, 12, 307–392.

Lance, C. E., Butts, M. M., and Michels, L. C. (2006), "The sources of four commonly reported cutoff criteria: What did they really say?" *Organizational Research Methods*, 9, 202–220.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford, UK: Oxford University Press.

Linn, R. L. (1968), "A Monte Carlo approach to the number of factors problem," *Psychometrika*, 33, 37–71.

Lovibond, P. and Lovibond, S. H. (1995), "The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories," *Behavioral Research and Therapy*, 33, 335–343.

Martínez, M. E., Marshall, J. R., and Sechrest, L. (1998), "Invited commentary: Factor analysis and the search for objectivity," *American Journal of Epidemiology*, 148, 17–19.

McCrae, R. R. and Costa, P. T. (1987), "Validation of the Five-Factor Model of Personality across instruments and observers," *Journal of Personality and Social Psychology*, 52, 81–90.

McDonald, R. P. (1965), "Difficulty factors and non-linear factor analysis," *The British Journal of Mathematical and Statistical Psychology*, 18, 11–23.

— (1967), "Numerical methods for polynomial models in nonlinear factor analysis," *Psychometrika*, 32, 77–112.

Meinshausen, N. and Bühlmann, P. (2010), "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473.

Nielsen, S. F. (2000), "The stochastic EM algorithm: Estimation and asymptotic results," *Bernoulli*, 6, 457.

Ning, L. and Georgiou, T. T. (2011), "Sparse factor analysis via likelihood and ℓ_1-regularization," in *Proceedings of the 50th IEEE Conference on Decision and Control*, pp. 5188–5192.

Nisenbaum, R., Reyes, M., Mawle, A. C., and Reeves, W. C. (1998), "Factor analysis of unexplained severe fatigue and interrelated symptoms: Overlap with criteria for chronic fatigue syndrome," *American Journal of Epidemiology*, 148, 72–77.

Patil, V. H., Singh, S. N., Mishra, S., and Donavan, D. T. (2008), "Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one' criterion," *Journal of Business Research*, 61, 162–170.

Peeters, C. F. W. (2012), "Rotational uniqueness conditions under oblique factor correlation metric," *Psychometrika*, 77, 288–292.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Raîche, G., Walls, T. A., Magis, D., Riopel, M., and Blais, J.-G. (2013), "Non-graphical solutions for Cattell's Scree Test," *Methodology*, 9, 23–29.

Reise, S. P., Waller, N. G., and Comrey, A. L. (2000), "Factor analysis and scale revision," *Psychological Assessment*, 12, 287–297.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011), "Detecting novel associations in large data sets," *Science*, 334, 1518–1524.

Revelle, W. (2019), *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois.

Richardson, T. S. and Spirtes, P. (2002), "Ancestral graph markov models," *Annals of Statistics*, 30, 962–1030.

Rosseel, Y. (2012), "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software*, 48, 1–36.

Rubin, D. B. and Thayer, D. T. (1982), "EM Algorithms for ML Factor Analysis," *Psychometrika*, 47, 69–76.

Scharf, F. and Nestler, S. (2019), "Should regularization replace simple structure rotation in exploratory factor analysis?" *Structural Equation Modeling*, 26, 576–590.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. (1998), "The TETRAD project: Constraint based aids to causal model specification," *Multivariate Behavioral Research*, 33, 65–117.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006), "Reporting structural equation modeling and confirmatory factor analysis results: A review," *Journal of Educational Research*, 99, 323–338.

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

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. (2006), "Learning the structure of linear latent variable models," *Journal of Machine Learning Research*, 7, 191–246.

Smith, R. (2015), "A mutual information approach to calculating nonlinearity: Measuring nonlinearity with mutual information," *Stat*, 4, 291–303.

Spirtes, P. (2001), "An anytime algorithm for causal inference," in *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, pp. 213–231.

Spirtes, P., Glymour, C., and Scheines, R. (2000), *Causation, Prediction, and Search*, Cambridge: MIT Press, 2nd ed.

Streiner, D. L. (1998), "Factors affecting reliability of interpretations of Scree Plots," *Psychological Reports*, 83, 687–694.

Tibshirani, R. (1992), "Principal curves revisited," *Statistics and Computing*, 2, 183–190.

— (1996), "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

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

Velicer, W. F. and Jackson, D. N. (1990), "Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure," *Multivariate Behavioral Research*, 25, 1–28.

Wayne, V. F., Eaton, C. A., and Fava, J. L. (2000), "Construct explication through factor or component analysis: A review and evaluation of alternative procedrues for determining the number of factors or components," in *Problems and Solutions in Human Assessment: Honoring Doublas N. Jackson at Seventy*, eds. Goffin, R. D. and E., H., Boston, MA: Kluwer, pp. 41–71.

Werts, C. E., Jöreskog, K. G., and Linn, R. L. (1973), "Identification and estimation in path analysis with unmeasured variables," *American Journal of Sociology*, 78, 1469–1484.

Whitely, S. E. (1983), "Construct validity: Construct representation versus nomothetic span," *Psychological Bulletin*, 93, 179–197.

Wu, C. F. J. (1983), "On the convergence properties of the EM algorithm," *The Annals of Statistics*, 11, 95–103.

Yalcin, I. and Amemiya, Y. (2001), "Nonlinear factor analysis as a statistical method," *Statistical Science*, 16, 275–294.

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

Zhang, J. (2008), "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias," *Artificial Intelligence*, 172, 1873–1896.

Zwick, W. R. and Velicer, W. F. (1986), "Comparison of five rules for determining the number of components to retain," *Psychological Bulletin*, 99, 432–442.