Structure Learning of Latent Factors via Clique Search on Correlation Thresholded Graphs

Dale S. Kim [1] **Qing Zhou** [1]

Abstract

Despite the widespread application of latent factor analysis, existing methods suffer from 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. We address these concerns by proposing a fast correlation thresholding (CT) algorithm that simultaneously learns the number of latent factors and a rotationally identifiable model structure. Our novel 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. Our clique analysis technique scales well up to thousands of variables, while competing methods are not applicable in a reasonable amount of running time. We establish a finite-sample error bound and high-dimensional consistency for the structure learning of our method. Through a series of simulation studies and a real data example, we show that the CT algorithm is an accurate method for learning the structure of factor analysis models and is robust to violations of its assumptions.

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.

To learn the parameters of factor analysis models, three problems must be addressed: (1) the number of latent factors must be determined, (2) the support of the coefficients must be found, and (3) a unique solution must be determined from rotationally equivalent parameters. Prior work on learning factor analysis models typically use a constraint-based or a score-based approach. Constraint-based methods involve 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.

Score-based methods are generally more amenable to single model outputs. Traditional Exploratory Factor Analysis (EFA) typically maximizes the likelihood, restricting the latent factors to be orthogonal. An oblique factor solution can be extracted by rotating the orthogonal solution, subject to the model constraints. There are numerous procedures for such rotations, which typically yield different solutions (for a review of such methods see Browne, 2001). After rotation, additional structure may be learned by setting small elements of Λ to zero if they are below an ad-hoc threshold (Ford et al., 1986; Howard, 2016). The major criticisms of EFA are the subjective use of these rotation criteria and thresholding steps, and requiring the number of latent factors to be known a priori.

As a potential solution to these problems in EFA, penalized methods also have been developed. Most relevant to oblique factor analysis models are adding LASSO (Tibshirani, 1996) and MCP (Zhang, 2010) penalties to the likelihood, which were developed by Hirose and Yamamoto (2014b). Instead of rotating factor coefficients after maximizing the likelihood, penalized EFA can achieve sparse solutions by directly maximizing a penalized likelihood. This requires the use of tuning parameters, followed by model selection with the Bayesian Information Criterion (BIC) or cross-

[1]Department of Statistics and Data Science, University of California, Los Angeles, United States. Correspondence to: Qing Zhou <zhou@stat.ucla.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

validation (CV; Scharf & Nestler, 2019). However, a search over a large set of tuning parameters is often computationally intense. Furthermore, the number of latent variables is still required as an input, and theoretical guarantees for rotational identifiability and structural estimation consistency have yet to be established.

For choosing the number of latent factors, many ad-hoc methods have been suggested, but suffer from poor performance, poor theoretical motivation, or both. Most classical methods are related to the eigenvalues of the sample correlation matrix among the observed variables. Famous examples include the Kaiser-Guttman criterion (Guttman, 1954; Kaiser, 1960), the Scree Test (Cattell, 1966; Raîche et al., 2013), and variants thereof (Horn, 1965; Glorfeld, 1995). On the other hand, modern methods use a model selection approach (Preacher et al., 2013). However, none of these methods are without controversy, and a great deal of literature has been devoted to criticisms on both empirical and theoretical grounds (Browne, 1968; Ford et al., 1986; Zwick & Velicer, 1986; Velicer & Jackson, 1990; Howard, 2016; Auerswald & Moshagen, 2019).

In summary, all methods of learning factor analysis must address three fundamental issues: (1) determine the number of factors, (2) learn the structure of the model, and (3) resolve the rotational nonidentifiability issue. As we have reviewed, an overabundance of literature has been dedicated to addressing these issues *separately*, all with varying degrees of success. In contrast, we seek to address all three aforementioned issues *simultaneously* under a unified framework. We do this by making use of thresholded correlation graphs of the observed correlation matrix, and exploiting two common assumptions in factor analysis designs. First, we assume that the correlation between variables that share latent factor parents is higher than the correlation between variables that do not. Second, we assume that each latent variable has at least one observed variable of which it is the sole parent. Under these conditions, there is a perfect correspondence between latent factors and a specific type of maximal clique from these graphs, which we call independent maximal clique (defined in Section 3.1). Therefore, the structure learning problem is converted to a search for all independent maximal cliques in the graph. We leverage this key relation to make the following contributions:

1. We propose a computationally efficient algorithm for learning the number of latent factors and the support of the coefficients simultaneously.
2. We establish high-dimensional consistency of our algorithm for learning the structure of the model.
3. We demonstrate the efficacy and practical uses of our algorithm on both real and simulated data, including high-dimensional settings.

There is another recent study that has taken a clique analysis approach to learning the structure of independent latent factors (Markham & Grosse-Wentrup, 2020). This work utilizes the maximal cliques of a conditional independence graph for structure learning. In contrast, we allow for correlated latent factors and our method is much more computationally efficient through the use of independent maximal cliques. Further, we establish theoretical guarantees for our method in high-dimensions.

Notation throughout this article will be as follows. Define $[n] := \{1, \ldots, n\}$. 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 $\mathcal{A}(M) := \{(i,j) : m_{ij} \neq 0\}$. 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.

For graph theoretic notation, 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. 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.

2. The Factor Analysis Model

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

$$X = \Lambda L + \epsilon, \tag{1}$$

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. 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. We are considering the more general case of oblique factor analysis models in this study, where the L variables may be correlated.

The model stated in Equation (1) implies a covariance structure Σ for X as follows:

$$\Sigma(\theta) := \text{Var}(X) = \text{Var}(\Lambda L + \epsilon) = \Lambda\Phi\Lambda^T + \Omega, \quad (2)$$

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 $D_\sigma = \text{diag}(\Sigma)^{1/2}$, 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:

$$\widetilde{X} := D_\sigma^{-1} X = D_\sigma^{-1}(\Lambda L + \epsilon) = \widetilde{\Lambda} L + \widetilde{\epsilon}, \quad (3)$$

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:

$$\widetilde{\Sigma}(\theta) := D_\sigma^{-1}\Sigma D_\sigma^{-1} = \widetilde{\Lambda}\Phi\widetilde{\Lambda}^T + \widetilde{\Omega}, \quad (4)$$

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 $X \sim \mathcal{N}_p(0, \Sigma(\theta))$. However, there is no closed-form solution for the MLE (Jöreskog, 1967). Therefore, iterative algorithms, such as Newton-Raphson (Jennrich & Robinson, 1969) or Expectation-Maximization (Rubin & 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 (2) are in general not identifiable, often referred to as rotational non-identifiability in the literature (Anderson & Rubin, 1956). This issue must be taken care of with additional criteria for parameter estimation or restrictions on the model structure.

3. The Correlation Thresholding Algorithm

3.1. Preliminaries and Overview

The main idea behind our algorithm is that for several broad classes of factor analysis models, the correlation between observed variables that share parents is stronger than correlations between variables that do not (these classes of models are discussed in Section 4.4). Subsequently, the correlation graph amongst the variables that share parents yields much information about the structure of the model. We leverage these two ideas into an efficient algorithm to learn the structure.

Recall that ρ_{ij} is the correlation between X_i and X_j given by $\widetilde{\Sigma}(\theta)$ in Equation (4). Our first step is to define a *thresholded correlation graph* $\mathcal{G}(X, E(\tau))$ given some $\tau \in [0, 1]$, where the edge set

$$E(\tau) := \{(i, j) : |\rho_{ij}| > \tau\}. \quad (5)$$

In practice, given a sample of X, we can define an estimate of $E(\tau)$ as

$$\hat{E}(\tau) := \{(i, j) : |r_{ij}| > \tau\}, \quad (6)$$

where r_{ij} denotes the sample correlation.

Given a thresholded correlation graph, an implied structure can be extracted by examining the cliques of the graph. Specifically, there is a correspondence between the latent variable structures and a particular kind of maximal clique, which we term as *independent maximal clique*:

Definition 3.1 (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

$$C_i \nsubseteq \bigcup_{j \neq i} C_j. \quad (7)$$

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. In Section 4.1, we show that the each independent maximal clique corresponds to a latent variable, whose children are the members of those cliques. This transforms the result of the clique search into a factor analysis structure.

3.2. The Algorithm

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 thresholds $\tau_k \in [0, 1]$ and analyzing their respective thresholded correlation graphs $\mathcal{G}(X, \hat{E}(\tau_k))$ for independent maximal cliques. We exploit the correspondence between these cliques and the factor analysis structure to learn the number of latent variables and the support of Λ. When this is done over each candidate threshold, this yields a set of candidate models for which we can utilize model selection procedures (e.g., BIC) to select a final model. We formally describe these steps in Algorithm 1.

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

Algorithm 1 The Correlation Thresholding Algorithm

1: **Input:** Sample correlation matrix R and set of thresholds $\tau = \{\tau_k : k \in [m]\}$.
2: **Output:** Parameter estimates $\hat{\theta}$.
3: **for** $k \in [m]$ **do**
4: Calculate $\mathcal{G}(X, \hat{E}(\tau_k))$ and extract the set of independent maximal cliques: $\mathcal{C}_k = \{C_1, \dots C_{|\mathcal{C}_k|}\}$;
5: Set $\hat{d}_k = |\mathcal{C}_k|$;
6: Initialize $\hat{A}_k = \emptyset$;
7: **for** $(i, j) \in [p] \times [\hat{d}_k]$ **do**
8: If $i \in C_j$, add (i, j) to $\hat{A}_k$;
9: **end for**
10: Estimate $\hat{\theta}_k$ given $(\hat{d}_k, \hat{A}_k)$, i.e., subject to $\lambda_{ij} = 0$ for all $(i, j) \notin \hat{A}_k$;
11: **end for**
12: Select one of the m estimates from $\{\hat{\theta}_k : k \in [m]\}$ via a model selection procedure.

Lemma 3.2. *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 $X_i \in C$.*

In the worst case scenario, all independent maximal cliques can be found 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, allowing our algorithm to be used in high-dimensional settings. This is in sharp contrast to the exponential complexity in listing all (non-independent) maximal cliques in a graph (Eppstein et al., 2010).

After extracting these independent maximal cliques in Step 4, we learn the structure of the model in Steps 5 through 9. 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 $\hat{A}_k$.

In Step 10 we estimate each model given the learned structure, then in Step 12 we use a model selection procedure to select one of the models. We note that these steps are general in that any estimation and model selection method can be utilized here. In our implementation, we will prefer to use maximum likelihood estimation and BIC for model selection. Since this pair of methods are statistically consistent, this leads to the final output model having consistent parameters and model structure, as we will show in Section 4.2.

4. Theoretical Guarantees

In this section, we establish theoretical guarantees for the CT algorithm. We assume throughout that the factor analysis model in Equation (1) holds. Proofs of these results can be found in Appendix A.

4.1. Assumptions

We first present the main assumptions under which the structure for Λ can be recovered from the thresholded correlation graph. A discussion of these assumptions is provided in Section 4.4.

Let the *parent set* of X_i be $\Pi_i := \{j : \lambda_{ij} \neq 0, j \in [d]\}$. Then we formalize the set of pairs that share parents as

$$E_0 := \{(i, j) \in [p] \times [p] : \Pi_i \cap \Pi_j \neq \emptyset\}. \tag{8}$$

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

$$E_0^c = \{(i, j) \in [p] \times [p] : \Pi_i \cap \Pi_j = \emptyset\}. \tag{9}$$

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*:

Definition 4.1 (Thresholdable)**.** A set of parameters θ is called *thresholdable* if there exists a threshold τ_0 such that

$$\max_{(k,l) \in E_0^c} |\rho_{kl}| < \tau_0 < \min_{(i,j) \in E_0} |\rho_{ij}|. \tag{10}$$

Recall the use of independent maximal cliques (Definition 3.1) in the CT algorithm. Perfect model structure recovery can be achieved if there is a one-to-one correspondence between the latent variable structures and the independent maximal cliques. A simple sufficient condition for such a correspondence to hold is the *unique child condition*:

Condition 4.2 (Unique Child Condition)**.** Let the *child set* of a latent variable be denoted *ch*$(L_k) = \{i \in [p] : \lambda_{ik} \neq 0\}$. If

$$U_k := ch(L_k) - \bigcup_{j \neq k} ch(L_j) \neq \emptyset, \qquad \forall\, k \in [d], \tag{11}$$

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 4.3. *If the unique child condition holds in Λ (Condition 4.2), then the set $\{\text{ch}(L_k) : k \in [d]\}$ is identical to the set of independent maximal cliques in $\mathcal{G}(X, E_0)$.*

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 4.3 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 4.4. 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 (2).

4.2. Error Bounds and Consistency

In this section, we establish the consistency of the CT algorithm. We will call a structural estimate $(\hat{d}, \mathcal{A}(\hat{\Lambda}))$ consistent if

$$\lim_{n\to\infty} \mathbb{P}\Big[(\hat{d}, \mathcal{A}(\hat{\Lambda})) = (d, \mathcal{A}(\Lambda))\Big] = 1, \tag{12}$$

given an i.i.d. sample of size n from the model in Equation (1). By Lemma 4.3, 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. In what follows, it will be useful to define a gap of separation for a thresholdable θ as

$$\gamma := \frac{1}{2}\left[\min_{(i,j)\in E_0} |\rho_{ij}| - \max_{(i,j)\in E_0^c} |\rho_{ij}|\right]. \tag{13}$$

Theorem 4.5. *Assume the model described in Equation* (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*

$$\mathbb{P}(\hat{E}(\tau_0) \neq E_0) \leq Cp(p-1)(n-2)\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4} := \eta, \tag{14}$$

where $0 < C < \infty$ only depends on M. If additionally the unique child condition holds (Condition 4.2), then we have

$$\mathbb{P}((\hat{d}, \mathcal{A}(\hat{\Lambda})) = (d, \mathcal{A}(\Lambda))) \geq 1 - \eta, \tag{15}$$

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

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 14 to develop a consistency result with high-dimensional accommodations where the dimension $p = p_n \gg n$.

Theorem 4.6. *Assume the model described in Equation* (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*

$$\lim_{n\to\infty} \mathbb{P}(\hat{E}(\tau_0) = E_0) = 1. \tag{16}$$

If additionally the unique child condition holds (Condition 4.2), then the structural estimate is consistent, as in Equation (12).

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 A.3 (a Gaussian sample correlation concentration bound) in the proofs of Theorems 4.5 and 4.6 with a bound for any non-Gaussian X of interest (see Appendices A.9 and A.10, respectively). So long as this bound is sufficiently well-behaved, the probability bounds in Theorem 4.5 will hold as will Theorem 4.6 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 $\{\tau_k\}$, the unique child condition 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 10) and a consistent model selection procedure (Step 12) 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.

4.3. Rotational Uniqueness

An important consideration with a factor analysis model is the identifiability of the parameters $\theta = \{\Lambda, \Phi, \Omega\}$. It is

well known that factor analysis models lack of rotational uniqueness, which implies that there may be many (Λ, Φ) pairs that exist such that $\Sigma(\theta) = \Lambda\Phi\Lambda^T + \Omega$. However, the solutions learned by the CT algorithm resolves this non-identifiability issue given that the zero constraints implied by $\mathcal{A}(\hat{\Lambda})$ are preserved. A formal definition of rotational uniqueness can be found in Appendix A.6.

Corollary 4.7. *If the unique child condition holds in Λ, then θ is locally rotationally unique (i.e., unique up to a polarity reversal on columns).*

Corollary 4.8. *Any $\hat{\theta}_k$ for $k \in [m]$, produced by Step 10 of the CT algorithm, is locally rotationally unique.*

First we note that all matrix factorizations will have a polarity reversal on columns or rows 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) makes no assumptions regarding the signs in Λ, local rotational uniqueness is the best type of rotational uniqueness that can be established. Second, note that Corollary 4.8 holds regardless if Condition 4.2 is true in the population structure. Thus the CT algorithm can be used as a model approximation tool for finding locally rotationally unique structures.

4.4. Discussion of Assumptions

We discuss the practicality of our thresholdability and unique child assumptions and how they relate to common factor analytic designs. Regarding the thresholdability of θ, several widely used factor analysis designs either meet the assumption outright, or under mild conditions. These stem from a technical necessary and sufficient condition for thresholdability presented in Lemma A.1 in the Appendix. Relevant to our discussion are the following corollaries to the lemma, which we discuss here.

Corollary 4.9. *If $\Phi = I_d$, then θ is thresholdable.*

That is, if we have the orthogonal factor analysis design, then thresholdability is met. Another common scenario is when Λ has exactly one non-zero entry per row. This is called "independent cluster structure" (Harris & Kaiser, 1964) or "perfect simple structure" (Jennrich, 2006). Such structures lead to a simplification of the thresholdability condition:

Corollary 4.10. *If Λ has exactly one non-zero entry per row, then θ is thresholdable if*

$$\max_{(k,l)\in E_0^c} |\widetilde{\lambda}_{ke}\widetilde{\lambda}_{lf}\phi_{ef}| < \min_{(i,j)\in E_0} |\widetilde{\lambda}_{ic}\widetilde{\lambda}_{jc}|, \tag{17}$$

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

Corollaries 4.9 and 4.10 involve desirable properties of factor analytic designs. 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.

As a common design in educational and psychological test construction (Hattie, 1985; Anderson & Gerbing, 1988), 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. In contrast, our unique child condition (Condition 4.2) is much more general, only requiring a single observed variable to serve as the sole measurement. Many other latent variable algorithms only focus on the independent cluster structure (Scheines et al., 1998; Jennrich, 2001; 2006; Silva et al., 2006), or require 3 to 4 observed variables to serve as unique children (Shimizu et al., 2009; Kummerfeld & Ramsey, 2016).

Additionally, we examine the unique child condition under a random graph model for $\mathcal{A}(\Lambda)$, in which edges are independently connected between any X_i and L_j with probability α. Let $m = d\alpha$ be the expected number of parents for any X_i. One can show that the unique child condition holds with probability $\geq 1 - d\exp[-\alpha p(1-\alpha)^d]$. Consequently, if $\log(\log d) \ll m \ll \min\{\log(mp/d), d\}$, then the unique child condition holds with high probability.

The assumption of the unique child condition does have a few limitations as it precludes certain structures from being perfectly discovered. Examples of these structures are illustrated in Appendix C (Figure 3). However, even under such settings, we will show through simulation that the CT algorithm will select a structure close to the true structure despite the unique child condition not holding.

5. Simulation Studies

5.1. Low-Dimensional Settings

Our first set of simulations were conducted in low-dimensional settings. Here, we compared the CT algorithm against three other methods: (1) EFA, (2) EFA-LASSO, and (3) EFA-MCP (all described in Section 1). Note that these EFA methods all require d as an input, thus we use the CT algorithm to give these EFA methods a set of d to work with. This was to make the comparison as fair as possible, rather than than using ad hoc choices. More specifically, we ran the CT algorithm to Step 5, where d is estimated from the number of independent maximal cliques. Thereafter, we replaced the support learning portion (Steps 6 through 9) of the algorithm with one of the EFA procedures. Then the support of the model was saved from the EFA methods and resumed the algorithm from Step 12, where the MLE was estimated from the support and used for model selection.

We generated data sets from a zero-mean Gaussian distribution, with a covariance matrix Σ parameterized by θ. 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 & Yamamoto, 2014a;b). The number of latent variables (d) was set to 3 with the number of children per latent variable set to 5 and $n = 1000$. The non-zero entries of Λ were drawn from a uniform distribution, $\lambda_{ij} \sim \text{Uniform}(0.6, 0.8)$. We varied the magnitude of the off-diagonals in Φ, as it is a key factor in whether or not θ is thresholdable for these structures, as shown by Corollary 4.10. Their entries began with the range of $[0.6, 0.8]$, with a low-magnitude setting scaling these by 0.25 and a high-magnitude setting scaling these by 0.75. The tuning parameters of the penalized EFA methods were left at the software package defaults, which were 30 tuning parameters for EFA-LASSO and a set of 270 tuning parameters for EFA-MCP. We conducted 100 replications per condition. Further details of the software and data generating settings can be found in Appendix B.

We examined three outcomes to assess the performance of the methods: (1) The F_1 score of $\mathcal{A}(\hat{\Lambda})$, (2) the learned number of latent variables, and (3) the computational efficiency of each method. A precise definition of the F_1 score of $\mathcal{A}(\hat{\Lambda})$ can be found in Appendix B.2. To measure computational efficiency we simply counted the number of models each method estimated. This was to avoid idiosyncratic differences between the software implementations of each method. 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 results of this simulation are displayed in Figure 1. CT and EFA-MCP have the best F_1 scores (very close to 1.0), with EFA-LASSO at around 0.75 and EFA at 0.5 across both conditions. All methods learned the number of latent variables correctly in all data sets, and thus were omitted from the figure. For computational efficiency, the CT algorithm estimated a substantially less amount of models compared to the penalized EFA methods. EFA showed the best computational efficiency, but in contrast had the worst F_1 score. These results demonstrate that in low-dimensional settings, that the CT algorithm performs with near perfect accuracy along with EFA-MCP, however with substantial computational savings.

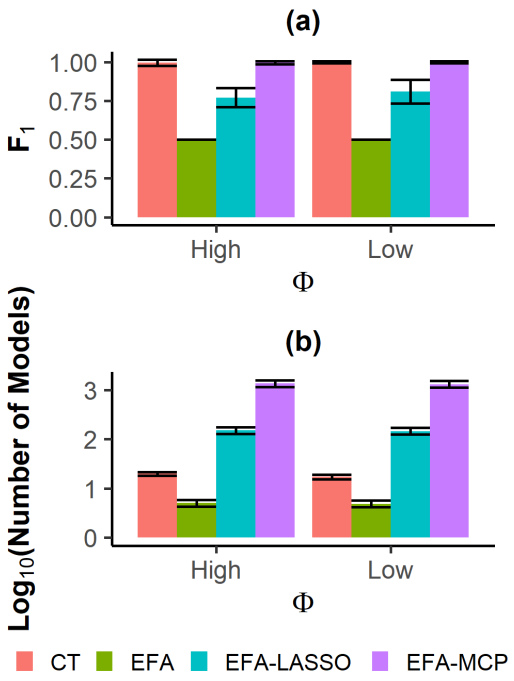

Figure 1. Averages of the performance metrics for the low-dimensional simulation. Error bars depict ± 1 standard deviation.

5.2. High-Dimensional Settings

For the high-dimensional settings, we examined the scenario where both p and d grow proportionally with n, and $n < p$. We examined three (n, p, d) settings where $n \in \{250, 500, 1000\}$, and set $p = 1.5n$ and $d = 0.1n$. In addition, we studied two conditions where the key assumptions of the CT algorithm would be violated: (1) thresholdability, which we violated using high-magnitude off-diagonals in Φ as in the previous simulation, and (2) the unique child condition. We violated the unique child condition by starting with the independent cluster structure, then randomly selecting 75% of the latent variables to have no unique children, whose children were all given an extra random parent. To isolate the effect of the unique child condition from that of thresholdability, we ensured thresholdability was always met in the latter set of simulations by setting $\Phi = I_d$ (Corollary 4.9). Further details regarding the simulation settings can be found in Appendix B and additional results using more varied assumption violations can be found in Appendix C (Figure 4).

Under these high p settings, both the EFA and penalized EFA methods are prohibitively slow, thus could not be used as comparisons for this study. Further, MLE routines also do not complete in a reasonable amount of running time, hence, we omitted the estimation step of the CT algorithm (Step 10). Rather, a final model structure was chosen as the one with the minimum Hamming distance (HD) among the candidate

thresholds $\{\tau_k, k \in [m]\}$ (a precise definition of HD can be found in Appendix B.2). As before, we examined the F_1 score, $\hat{d}$, and computational efficiency as the outcomes for this study.

The results are displayed in Figure 2. We first note that in the thresholdability violation condition, thresholdability was indeed violated in at least 99% of the data sets for each of the $n = 250$, 500, and 1000 configurations. However, we can see the F_1 score become more accurate with n despite these challenging conditions and the proportional growth in 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. Unsurprisingly, the computational time increased with p, but remained reasonable even at $p = 1500$.

6. Real Data Application

We examined a widely used factor analysis data set comprised of intelligence test scores of $n = 301$ middle school students (Holzinger & Swineford, 1939), and compared the performance of the CT algorithm with EFA and the penalized EFA methods. The data consist of 9 variables designed to measure 3 factors of intelligence. These were a spatial factor, a verbal factor, and a speed factor. The hypothesized structure of this design was an independent cluster structure between these three factors. 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.

We display the results in Table 1. 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 only for the CT algorithm, indicating that the hypothesized model was perfectly recovered within its solution path only, and not any other method. Moreover, the CT algorithm identified the hypothesized structure with a much smaller set of candidate models. We selected a structure via BIC for each method and used 10-fold CV to calculate a test-data log-likelihood and evaluate the structure. The results for the test data log-likelihood are similar across all methods except EFA, which was much worse. Despite the comparable performance between the CT algorithm and the sparse EFA methods, the CT algorithm obtained these results with much improved computational efficiency, measured by the number of candidate models evaluated.

As with most factor analytic designs, the hypothesized structure followed the unique child condition. An illustration of the hypothesized structure and all the selected structures by the four competing methods are displayed in Appendix C (Figure 5). Both the EFA-LASSO and EFA-MCP methods selected structures that followed the unique child condition, despite the fact that these methods are not developed under this assumption. These convergent results lend empirical support for the unique child condition holding for factor analysis structures within this data set, as designed.

Method	HD(min.)	$\hat{d}$	Test LL	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 1. Results of real data example. HD(min.) denotes the minimum HD to the hypothesized structure across all solutions, Test LL refers to the test-data log-likelihood, and Models denotes the number of models evaluated per method.

7. 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 conditions for the clique mapping procedure, 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 in low-dimensional settings, and showed robust results in high-dimensional settings. 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, mainly the assumptions of thresholdability and the unique child condition. While we demonstrated that the CT algorithm can be robust to violations of these assumptions in practice, our statistical consistency results depends on these assumptions being true in the population. Future work can focus on the relaxation of these assumptions.

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.

Acknowledgments

This work was supported by NSF grant DMS-1952929.

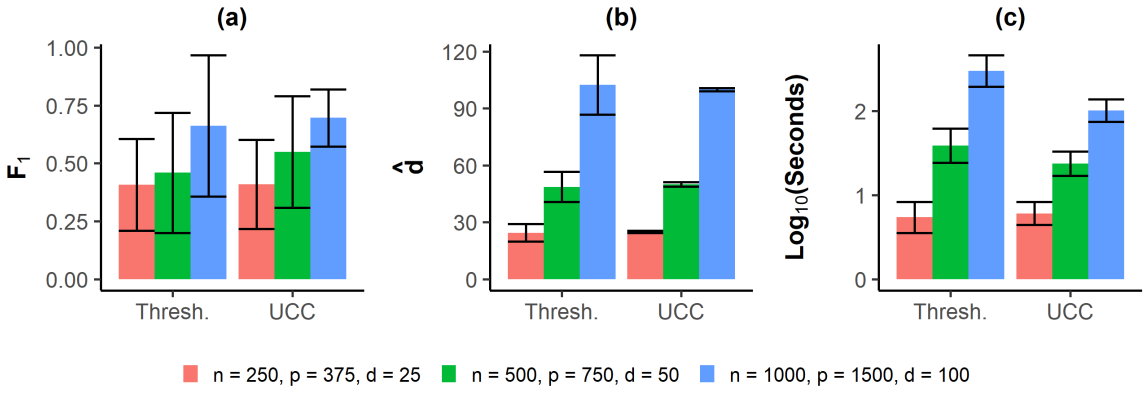

Figure 2. Averages of the performance metrics for the high-dimensional simulation. Error bars depict ± 1 standard deviation. "Thresh." refers to the high-magnitude Φ condition where thresholdability is violated, and "UCC" refers to the condition where the unique child condition is violated.

References

Anderson, J. C. and Gerbing, D. W. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3):411–423, 1988.

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

Auerswald, M. and Moshagen, M. How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, 24(4):468–491, 2019.

Browne, M. W. A comparison of factor analytic techniques. *Psychometrika*, 33(3):1968, 1968.

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

Caner, M. and Han, X. Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators. *Journal of Business & Economic Statistics*, 32(3):359–374, 2014.

Cattell, R. B. The Scree Test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

Crawford, C. A comparison of the direct oblimin and primary parsimony methods of oblique rotation. *British Journal of Mathematical and Statistical Psychology*, 28(2):201—-213, Nov 1975. ISSN 00071102. doi: 10.1111/j.2044-8317.1975.tb00563.x.

Eppstein, D., Löffler, M., and Strash, D. Listing all maximal cliques in sparse graphs in near-optimal time. In *International Symposium on Algorithms and Computation*, pp. 403–414, 2010.

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

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

Glorfeld, L. W. An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55(3):377–393, 1995.

Guttman, L. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2):149–161, 1954.

Harris, C. W. and Kaiser, H. F. Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 29(4):347–362, 1964.

Hattie, J. Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2):139–164, 1985.

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

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

Holzinger, K. J. and Swineford, F. A study in factor analysis: The stability of a bi-factor solution. *Supplementary Educational Monographs*, 1939.

Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

Howard, M. C. 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(1):51–62, 2016.

Jennrich, R. I. A simple general procedure for orthogonal rotation. *Psychometrika*, 66(2):289–306, 2001.

Jennrich, R. I. Rotation to simple loadings using component loss functions: The oblique Case. *Psychometrika*, 71(1): 173–191, 2006.

Jennrich, R. I. and Robinson, S. M. A Newton-Raphson algorithm for maximum likelihood factor analysis. *Psychometrika*, 34(1):111–123, 1969.

Jöreskog, K. G. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443–482, 1967.

Kaiser, H. F. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151, 1960.

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

Kummerfeld, E. and Ramsey, J. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1655–1664, San Francisco California USA, Aug 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939838. URL https://dl.acm.org/doi/10.1145/2939672.2939838.

Markham, A. and Grosse-Wentrup, M. Measurement dependence inducing latent causal models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124, 2020.

Martínez, M. E., Marshall, J. R., and Sechrest, L. Invited commentary: Factor analysis and the search for objectivity. *American Journal of Epidemiology*, 148(1):17–19, 1998.

Peeters, C. F. W. Rotational uniqueness conditions under oblique factor correlation metric. *Psychometrika*, 77(2): 288–292, 2012.

Preacher, K. J., Zhang, G., Kim, C., and Mels, G. Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, 48(1):28–56, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Raîche, G., Walls, T. A., Magis, D., Riopel, M., and Blais, J.-G. Non-graphical solutions for Cattell's Scree Test. *Methodology*, 9(1):23–29, 2013.

Reise, S. P., Waller, N. G., and Comrey, A. L. Factor analysis and scale revision. *Psychological Assessment*, 12(3):287–297, 2000.

Revelle, W. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2019. URL https://CRAN.R-project.org/package=psych.

Rosseel, Y. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012. URL http://www.jstatsoft.org/v48/i02/.

Rubin, D. B. and Thayer, D. T. EM Algorithms for ML Factor Analysis. *Psychometrika*, 47(1):69–76, 1982.

Scharf, F. and Nestler, S. Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling*, 26(4):576–590, 2019.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research*, 99(6):323–338, 2006.

Shimizu, S., Hoyer, P. O., and Hyvärinen, A. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7–9):2024—2027, Mar 2009. ISSN 09252312. doi: 10.1016/j.neucom.2008.11.018.

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

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

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

Werts, C. E., Jöreskog, K. G., and Linn, R. L. Identification and estimation in path analysis with unmeasured variables. *American Journal of Sociology*, 78(6):1469–1484, 1973.

Whitely, S. E. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1): 179–197, 1983.

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

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

A. Proofs and Additional Results

A.1. Proof of Lemma 3.2

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*$(X_i) = C$. □

A.2. Necessary and Sufficient Condition for Thresholdability

Lemma A.1. *Recall the definitions of E_0 and E_0^c in Equations* (8) *and* (9)*, respectively. A set of parameters θ is thresholdable if and only if:*

$$\max_{(k,l)\in E_0^c} |\widetilde{\Lambda}_{kE}\Phi_{EF}\widetilde{\Lambda}_{lF}^T| < \min_{(i,j)\in E_0} |\widetilde{\Lambda}_{iA}\Phi_{AB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iC}\Phi_{CB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iA}\Phi_{AC}\widetilde{\Lambda}_{jC}^T + \widetilde{\Lambda}_{iC}\Phi_{CC}\widetilde{\Lambda}_{jC}^T|, \tag{18}$$

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 $F = \Pi_l$.

Proof. 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:

$$\begin{aligned} A &= \Pi_i - \Pi_j \\ B &= \Pi_j - \Pi_i \\ C &= \Pi_i \cap \Pi_j. \end{aligned} \tag{19}$$

Then we may re-cast Equation (1) for any pair $(\widetilde{X}_i, \widetilde{X}_j)$ as follows:

$$\begin{bmatrix} \widetilde{X}_i \\ \widetilde{X}_j \end{bmatrix} = \begin{bmatrix} \widetilde{\Lambda}_{iA} & \mathbf{0} & \widetilde{\Lambda}_{iC} \\ \mathbf{0} & \widetilde{\Lambda}_{jB} & \widetilde{\Lambda}_{jC} \end{bmatrix} \begin{bmatrix} L_A \\ L_B \\ L_C \end{bmatrix} + \begin{bmatrix} \widetilde{\epsilon}_i \\ \widetilde{\epsilon}_j \end{bmatrix}. \tag{20}$$

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

$$\text{Var}\left(\begin{bmatrix} \widetilde{X}_i \\ \widetilde{X}_j \end{bmatrix}\right) = \begin{bmatrix} \widetilde{\Lambda}_{iA} & \mathbf{0} & \widetilde{\Lambda}_{iC} \\ \mathbf{0} & \widetilde{\Lambda}_{jB} & \widetilde{\Lambda}_{jC} \end{bmatrix} \begin{bmatrix} \Phi_{AA} & \Phi_{AB} & \Phi_{AC} \\ \Phi_{BA} & \Phi_{BB} & \Phi_{BC} \\ \Phi_{CA} & \Phi_{CB} & \Phi_{CC} \end{bmatrix} \begin{bmatrix} \widetilde{\Lambda}_{iA}^T & \mathbf{0} \\ \mathbf{0} & \widetilde{\Lambda}_{jB}^T \\ \widetilde{\Lambda}_{iC}^T & \widetilde{\Lambda}_{jC}^T \end{bmatrix} + \begin{bmatrix} \widetilde{\omega}_i & 0 \\ 0 & \widetilde{\omega}_j \end{bmatrix}, \tag{21}$$

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

$$\rho_{ij} = \widetilde{\Lambda}_{iA}\Phi_{AB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iC}\Phi_{CB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iA}\Phi_{AC}\widetilde{\Lambda}_{jC}^T + \widetilde{\Lambda}_{iC}\Phi_{CC}\widetilde{\Lambda}_{jC}^T. \tag{22}$$

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 (22) to:

$$\rho_{ij} = \widetilde{\Lambda}_{iA}\Phi_{AB}\widetilde{\Lambda}_{jB}^T. \tag{23}$$

The result of Lemma A.1 follows by characterizing the definition of thresholdability (10) 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:

$$\max_{(k,l)\in E_0^c} |\widetilde{\Lambda}_{kE}\Phi_{EF}\widetilde{\Lambda}_{lF}^T| < \min_{(i,j)\in E_0} |\widetilde{\Lambda}_{iA}\Phi_{AB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iC}\Phi_{CB}\widetilde{\Lambda}_{jB}^T + \widetilde{\Lambda}_{iA}\Phi_{AC}\widetilde{\Lambda}_{jC}^T + \widetilde{\Lambda}_{iC}\Phi_{CC}\widetilde{\Lambda}_{jC}^T|. \tag{24}$$

□

A.3. Proof of Corollary 4.9

From Equation (18), 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 (18) reduces to

$$0 < \min_{(i,j)\in E_0} |\widetilde{\Lambda}_{iC}\widetilde{\Lambda}_{jC}^T|, \tag{25}$$

which trivially holds. □

A.4. Proof of Corollary 4.10

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 4.10 follows by simplifying Equation (18) with these reductions. □

A.5. Proof of Lemma 4.3

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 $\mathcal{G}$. □

A.6. Formal Definition of Rotational Uniqueness

Definition A.2 (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

$$\mathcal{M}_{CP} = \mathcal{M}_{CP}(\theta) \coloneqq \{M : \Sigma(\theta_M) = \Sigma(\theta), \mathcal{A}(\Lambda M) \subseteq \mathcal{A}(\Lambda), \text{diag}(M^{-1}\Phi M^{-T}) = I_d\}. \tag{26}$$

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.

Note that the condition $\mathcal{A}(\Lambda M) \subseteq \mathcal{A}(\Lambda)$ ensures that the zero constraints implied by $\mathcal{A}(\Lambda)$ are persevered. Two local rotational uniqueness properties relevant to the CT algorithm are described in Corollaries 4.7 and 4.8.

A.7. Proof of Corollary 4.7

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:

$$\Lambda = \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & 0 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{52} & \lambda_{53} \\ 0 & \lambda_{62} & 0 \\ 0 & 0 & \lambda_{73} \\ 0 & 0 & \lambda_{83} \\ \lambda_{91} & 0 & \lambda_{93} \end{bmatrix}, \quad \Lambda^{[1]} = \begin{bmatrix} \lambda_{42} & 0 \\ \lambda_{52} & \lambda_{53} \\ \lambda_{62} & 0 \\ 0 & \lambda_{73} \\ 0 & \lambda_{83} \end{bmatrix}, \quad \Lambda^{[2]} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{73} \\ 0 & \lambda_{83} \\ \lambda_{91} & \lambda_{93} \end{bmatrix}, \quad \Lambda^{[3]} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{62} \end{bmatrix}. \tag{27}$$

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 (11). For all $j, k \in [d]$, and $i \in [p]$ we can re-cast U_j as:

$$U_j = \{i : \lambda_{ij} \neq 0, \lambda_{ik} = 0, k \neq j\}, \tag{28}$$

and let the index of non-unique variables be:

$$\overline{U} = \{i : i \notin \cup_{j=1}^{d} U_j\}. \tag{29}$$

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:

$$P\Lambda = \begin{bmatrix} \Lambda_{U_1 1} & & \\ & \ddots & \\ & & \Lambda_{U_d d} \\ \Lambda_{\overline{U} 1} & \cdots & \Lambda_{\overline{U} d} \end{bmatrix}. \tag{30}$$

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 $(d-1)$. □

A.8. Proof of Corollary 4.8

Proof. As described in Section 3.2, Steps 6 through 9 of the CT algorithm construct the support $\hat{A}_k$ deterministically based on a set of independent maximal cliques $\mathcal{C}_k$ (from Step 5). Since by Definition 3.1 independent maximal cliques always have a unique node, the sparsity pattern in $\hat{A}_k$ is guaranteed to follow the unique child condition (Condition 4.2). By Corollary 4.7, $\hat{\theta}_k$ will be locally rotationally unique due to this pattern. □

A.9. Proof of Theorem 4.5

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:

$$\Big(\bigcap_{(i,j) \in E_0} |r_{ij}| > \tau_0 \Big) \cap \Big(\bigcap_{(i,j) \in E_0^c} |r_{ij}| < \tau_0 \Big). \tag{31}$$

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

$$\Big(\bigcup_{(i,j) \in E_0} |r_{ij}| \leq \tau_0 \Big) \cup \Big(\bigcup_{(i,j) \in E_0^c} |r_{ij}| \geq \tau_0 \Big), \tag{32}$$

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 & Bühlmann (2007). We re-state this as follows:

Lemma A.3. *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 $0 < \epsilon \leq 2$,*

$$\mathbb{P}(|r_{ij} - \rho_{ij}| \geq \epsilon) \leq C_0(n-2) \left(\frac{4 - \epsilon^2}{4 + \epsilon^2} \right)^{n-4}, \tag{33}$$

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 $\gamma \leq \big||\rho_{ij}| - \tau_0\big|$ 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 (33) and setting $\epsilon = \gamma$, we have:

$$\begin{aligned} \mathbb{P}(|r_{ij} - \rho_{ij}| \geq \gamma) &\geq \mathbb{P}(|r_{ij}| - |\rho_{ij}| \geq \gamma) \\ &\geq \mathbb{P}(|r_{ij}| - |\rho_{ij}| \geq \tau_0 - |\rho_{ij}|) \\ &= \mathbb{P}(|r_{ij}| \geq \tau_0). \end{aligned} \tag{34}$$

Hence, $\mathbb{P}(|r_{ij}| \geq \tau_0)$ is bounded from above by the right-hand side of Equation (33) if $(i,j) \in E_0^c$. We can use the same strategy to conclude that, for $(i,j) \in E_0$,

$$\mathbb{P}(|r_{ij} - \rho_{ij}| \geq \gamma) \geq \mathbb{P}(|r_{ij}| \leq \tau_0). \tag{35}$$

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

$$B_{ij} = B(r_{ij}, \tau_0) := \begin{cases} |r_{ij}| \leq \tau_0 & \text{if } (i,j) \in E_0 \\ |r_{ij}| \geq \tau_0 & \text{if } (i,j) \in E_0^c \end{cases}. \tag{36}$$

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:

$$\begin{aligned} \mathbb{P}(\hat{E}(\tau_0) \neq E(\tau_0)) = \mathbb{P}\Big(\bigcup_{(i,j)} B_{ij} \Big) &\leq \sum_{(i,j)} \mathbb{P}\left(B_{ij}\right) \\ &\leq \frac{p(p-1)}{2} \max_{(i,j)} \{\mathbb{P}(B_{ij})\} \\ &\leq Cp(p-1)(n-2) \left(\frac{4 - \gamma^2}{4 + \gamma^2} \right)^{n-4}, \end{aligned} \tag{37}$$

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 A.3. Finally, this implies

$$\mathbb{P}(\hat{E}(\tau_0) = E_0) \geq 1 - Cp(p-1)(n-2)\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4} \tag{38}$$

and thus, (15) follows immediately under the unique child condition by Lemma 4.3. □

A.10. Proof of Theorem 4.6

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 (14), an upper bound on the decaying term with n can be derived as follows:

$$\begin{aligned}
\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4} &\leq \left(1 - \frac{\gamma^2}{4}\right)^{n-4} \\
&\leq \left(1 - \frac{c_1}{4(n-4)^b}\right)^{n-4} \\
&= \left(1 - \frac{c_1}{4(n-4)^b}\right)^{(n-4)^b(n-4)^{1-b}} \\
&= \left(\exp\left(-\frac{c_1}{4}\right) + o(1)\right)^{(n-4)^{1-b}} \\
&\leq \exp\left(-\frac{c_2(n-4)^{1-b}}{4}\right),
\end{aligned} \tag{39}$$

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 (14) as

$$\begin{aligned}
\mathbb{P}(\hat{E}(\tau_0) = E_0) &\geq 1 - Cp(p-1)(n-2)\left(\frac{4-\gamma^2}{4+\gamma^2}\right)^{n-4} \\
&\geq 1 - Cp_n(p_n-1)(n-2)\exp\left(-\frac{c_2(n-4)^{1-b}}{4}\right) \\
&= 1 - p(n)f(n),
\end{aligned} \tag{40}$$

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

$$\begin{aligned}
p_n^2 &= o\left(\exp\left[\frac{c_2(n-4)^{1-b}}{4} - \log n\right]\right) \\
\text{or if } p_n &= o\left(\exp\left[c(n-4)^{1-b}\right]\right),
\end{aligned} \tag{41}$$

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

B. Supplementary Details for Simulation Studies and Real Data Application

B.1. Simulation settings

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 10), 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 & Yamamoto, 2014a;b). 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. For estimating the number of latent factors in these 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 & Han, 2014).

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 Uniform$(0, 1)$ and rescaled it such that $A^T A$ had off-diagonals in the range of $[0.6, 0.8]$, the range 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. Then depending on the condition for the magnitude of Φ, the off-diagonals were scaled by 0.25 for the low Φ condition and by 0.75 for the high Φ condition.

For the data sets that violate the unique child condition, we began with 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 the unique child condition from that of thresholdability, we ensured thresholdability was always met in the population by setting $\Phi = I_d$ (Corollary 4.9). Then we chose 75% of the latent variables at random to have no unique children. If a latent variable was deemed to have no unique children, we generated an extra path between all the children of this latent variable to another random latent variable. We will call these parents the *extra parent*.

For each X_j, we drew an $R^2 \sim$ Uniform$(0.36, 0.64)$ as the proportion of variance in X_j 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_j only had a main parent and no extra parent, then that X_j had a single path coefficient of $\sqrt{R^2}$ from its main parent. However, if a given X_j also had an extra parent, then the R^2 was split using a 5:1 ratio between the main parent and the extra parent, and the path coefficients were calculated to reflect this accordingly.

B.2. Evaluation Metrics

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

$$\text{HD} := \min_P \left[|\mathcal{A}(\hat{\Lambda}P) \triangle \mathcal{A}(\Lambda)| \right], \tag{42}$$

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

$$F_1(\hat{\Lambda}) := \max_P \left[\frac{2|\mathcal{A}(\hat{\Lambda}P) \cap \mathcal{A}(\Lambda)|}{2|\mathcal{A}(\hat{\Lambda}P) \cap \mathcal{A}(\Lambda)| + |\mathcal{A}(\hat{\Lambda}P) \triangle \mathcal{A}(\Lambda)|} \right] \in [0, 1], \tag{43}$$

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

C. Additional Figures

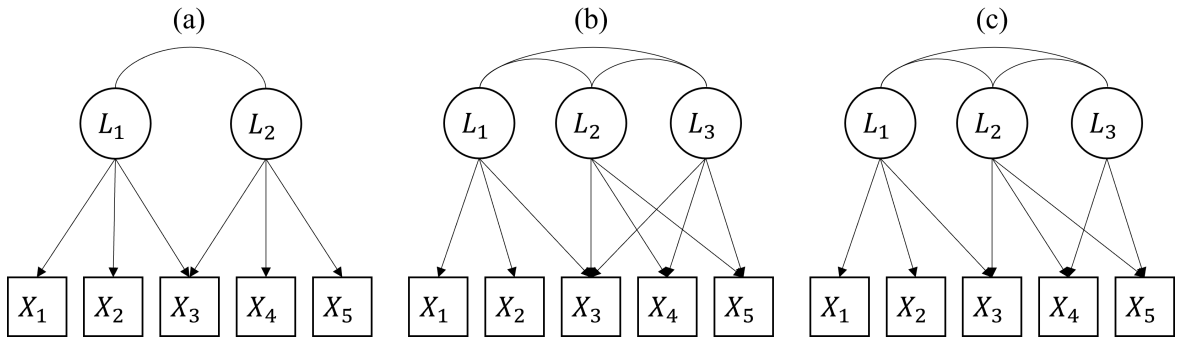

Figure 3. Example structures that may be precluded from the unique child condition. Figure (a) is a structure that meets the unique child condition, Figure (b) shows a structure where the children sets of L_2 and L_3 are identical, and Figure (c) shows a structure where the children set of L_3 is a proper subset of L_2. All three structures lead to the same thresholded correlation graph, and will contain two independent maximal cliques.

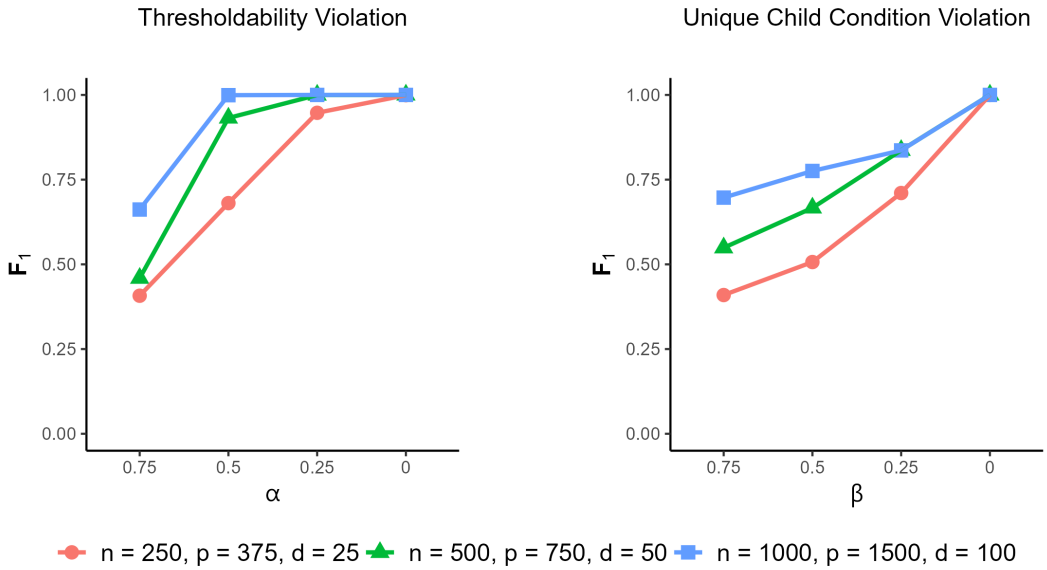

Figure 4. Trends for F_1 score when each of the assumptions are violated to varying degrees in the high-dimensional setting. Thresholdability was varied via the scaling factor $\alpha \in [0, 0.75]$ on the off-diagonals in Φ. The proportion of latent parents without unique children is represented by $\beta \in [0, 0.75]$.

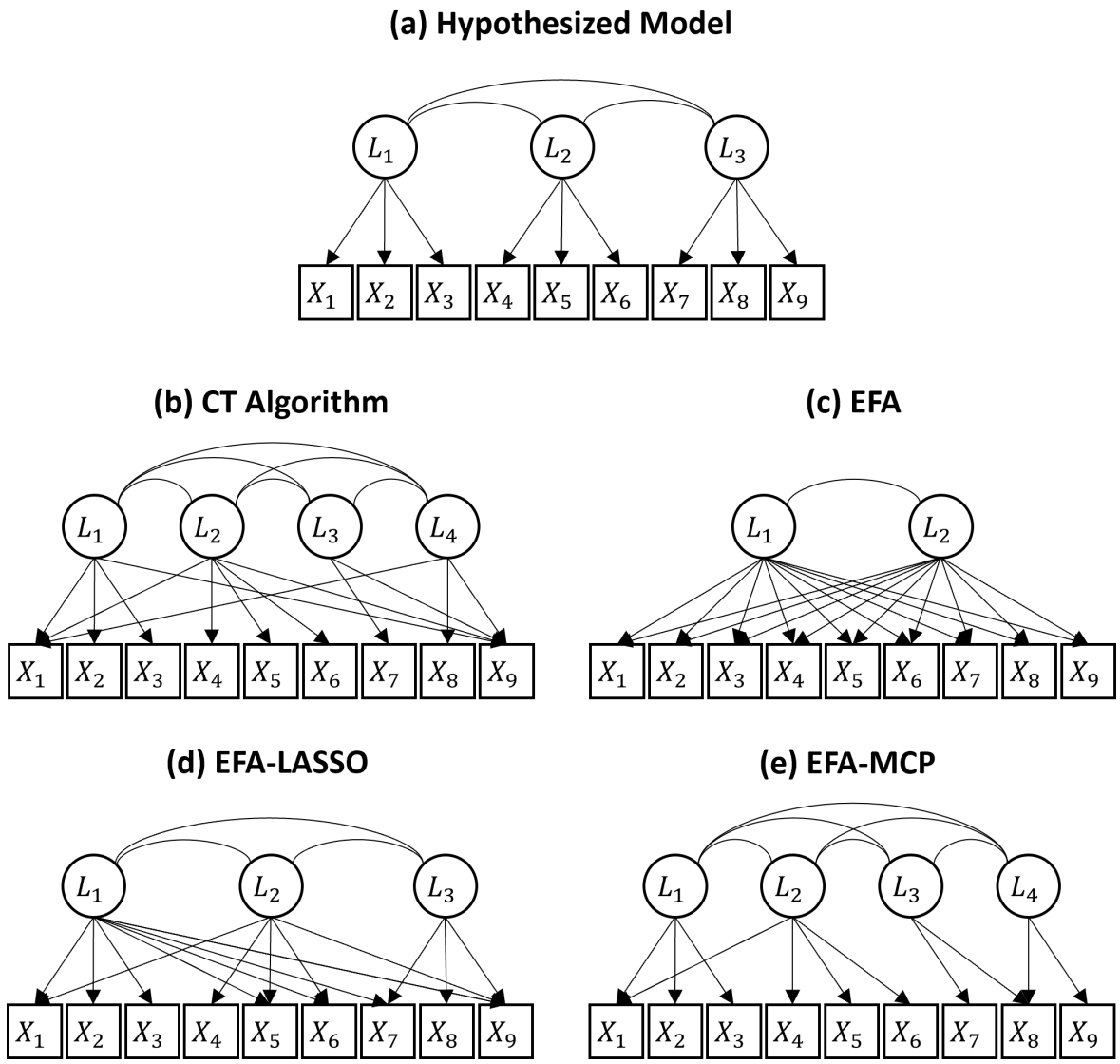

Figure 5. The hypothesized and estimated model structures by each method in the real data example. Variables X_1, X_2, and X_3 were visual perception tasks, variables X_4, X_5, and X_6 were verbal/reading tasks, and variables X_7, X_8, and X_9 were speed tests.