Interpreting Latent Variables in Factor Models via Convex Optimization

Armeen Taeb [†] and Venkat Chandrasekaran [†,‡] [∗]

[†] Department of Electrical Engineering
[‡] Department of Computing and Mathematical Sciences
California Institute of Technology
Pasadena, Ca 91125

November 4, 2016

Abstract

Latent or unobserved phenomena pose a significant difficulty in data analysis as they induce complicated and confounding dependencies among a collection of observed variables. Factor analysis is a prominent multivariate statistical modeling approach that addresses this challenge by identifying the effects of (a small number of) latent variables on a set of observed variables. However, the latent variables in a factor model are purely mathematical objects that are derived from the observed phenomena, and they do not have any interpretation associated to them. A natural approach for attributing semantic information to the latent variables in a factor model is to obtain measurements of some additional plausibly useful covariates that may be related to the original set of observed variables, and to associate these auxiliary covariates to the latent variables. In this paper, we describe a systematic approach for identifying such associations. Our method is based on solving computationally tractable convex optimization problems, and it can be viewed as a generalization of the minimum-trace factor analysis procedure for fitting factor models via convex optimization. We analyze the theoretical consistency of our approach in a high-dimensional setting as well as its utility in practice via experimental demonstrations with real data.

1 Introduction

A central goal in data analysis is to identify concisely described models that characterize the statistical dependencies among a collection of variables. Such concisely parametrized models avoid problems associated with overfitting, and they are often useful in providing meaningful interpretations of the relationships inherent in the underlying variables. Latent or unobserved phenomena complicate the task of determining concisely parametrized models as they induce confounding dependencies among the observed variables that are not easily or succinctly described. Consequently, significant efforts over many decades have been directed towards the problem of accounting for the effects of latent phenomena in statistical modeling. A common shortcoming of approaches to latent-variable modeling is that the latent variables are typically mathematical constructs that

---

[∗]Email: ataeb@caltech.edu, venkatc@caltech.edu

1

are derived from the originally observed data, and these variables do not directly have semantic information linked to them. Discovering interpretable meaning underlying latent variables would clearly impact a range of contemporary problem domains throughout science and technology. For example, in data-driven approaches to scientific discovery, the association of semantics to latent variables would lead to the identification of new phenomena that are relevant to a scientific process, or would guide data-gathering exercises by providing choices of variables for which to obtain new measurements.

In this paper, we focus for the sake of concreteness on the challenge of interpreting the latent variables in a factor model [20]. *Factor analysis* is perhaps the most widely used latent-variable modeling technique in practice. The objective with this method is to fit observations of a collection of random variables $y \in \mathbb{R}^p$ to the following linear model:

$$y = \mathcal{B}\zeta + \epsilon, \tag{1.1}$$

where $\mathcal{B} \in \mathbb{R}^{p \times k}, k \ll p$. The random vectors $\zeta \in \mathbb{R}^k, \epsilon \in \mathbb{R}^p$ are independent of each other, and they are normally distributed as[1] $\zeta \sim \mathcal{N}(0, \Sigma_\zeta), \epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$, with $\Sigma_\zeta \succ 0, \Sigma_\epsilon \succ 0$ and Σ_ϵ being diagonal. Here the random vector ζ represents a small number of unobserved, latent variables that impact all the observed variables y, and the matrix $\mathcal{B}$ specifies the effect that the latent variables have on the observed variables. However, the latent variables ζ themselves do not have any interpretable meaning, and they are essentially a mathematical abstraction employed to fit a concisely parameterized model to the conditional distribution of $y|\zeta$ (which represents the remaining uncertainty in y after accounting for the effects of the latent variables ζ) – this conditional distribution is succinctly described as it is specified by a model consisting of independent variables (as the covariance of the Gaussian random vector ϵ is diagonal).

A natural approach to attributing semantic information to the latent variables ζ in a factor model is to obtain measurements of some additional plausibly useful covariates $x \in \mathbb{R}^q$ (the choice of these variables is domain-specific), and to link these to the variables ζ. However, defining and specifying such a link in a precise manner is challenging. Indeed, a fundamental difficulty that arises in establishing this association is that the variables ζ in the factor model (1.1) are not identifiable. In particular, for any non-singular matrix $\mathcal{W} \in \mathbb{R}^{k \times k}$, we have that $\mathcal{B}\zeta = (\mathcal{B}\mathcal{W}^{-1})(\mathcal{W}\zeta)$. In this paper, we describe a systematic and computationally tractable methodology based on convex optimization that integrates factor analysis and the task of interpreting the latent variables. Our convex relaxation approach generalizes the *minimum-trace factor analysis* technique, which has received much attention in the mathematical programming community over the years [10, 17, 18, 19, 16].

1.1 A Composite Factor Model

We begin by making the observation that the column space of $\mathcal{B}$ – which specifies the k-dimensional component of y that is influenced by the latent variables ζ – is invariant under transformations of the form $\mathcal{B} \to \mathcal{B}\mathcal{W}^{-1}$ for non-singular matrices $\mathcal{W} \in \mathbb{R}^{k \times k}$. Consequently, we approach the problem of associating the covariates x to the latent variables ζ by linking the effects of x on y to the column space of $\mathcal{B}$. Conceptually, we seek a decomposition of the column space of $\mathcal{B}$ into transverse subspaces $\mathfrak{H}_x, \mathfrak{H}_u \subset \mathbb{R}^p$, $\mathfrak{H}_x \cap \mathfrak{H}_u = \{0\}$ so that column-space$(\mathcal{B}) \approx \mathfrak{H}_x \oplus \mathfrak{H}_u$ – the subspace $\mathfrak{H}_x$ specifies those components of y that are influenced by the latent variables ζ and are also affected by the covariates x, and the subspace $\mathfrak{H}_u$ represents any unobserved residual effects on y due to ζ that

---

[1]The mean vector does not play a significant role in our development, and therefore we consider zero-mean random variables throughout this paper.

are not captured by x. To identify such a decomposition of the column space of $\mathcal{B}$, our objective is to split the term $\mathcal{B}\zeta$ in the factor model (1.1) as

$$\mathcal{B}\zeta \approx \mathcal{A}x + \mathcal{B}_u\zeta_u, \qquad (1.2)$$

where the column space of $\mathcal{A} \in \mathbb{R}^{p \times q}$ is the subspace $\mathfrak{H}_x$ and the column space of $\mathcal{B}_u \in \mathbb{R}^{p \times \dim(\mathfrak{H}_u)}$ is the subspace $\mathfrak{H}_u$, i.e., $\dim(\text{column-space}(\mathcal{A})) + \dim(\text{column-space}(\mathcal{B}_u)) = \dim(\text{column-space}(\mathcal{B}))$ and column-space$(\mathcal{A}) \cap$ column-space$(\mathcal{B}_u) = \{0\}$. Since the number of latent variables ζ in the factor model (1.1) is typically much smaller than p, the dimension of the column space of $\mathcal{A}$ is also much smaller than p; as a result, if the dimension q of the additional covariates x is large, the matrix $\mathcal{A}$ has small rank. Hence, the matrix $\mathcal{A}$ plays two important roles: its column space (in $\mathbb{R}^p$) identifies those components of the subspace $\mathcal{B}$ that are influenced by the covariates x, and its rowspace (in $\mathbb{R}^q$) specifies those components of (a potentially large number of) the covariates x that influence y. Thus, *the projection of the covariates x onto the rowspace of $\mathcal{A}$ represents the interpretable component of the latent variables ζ.* The term $\mathcal{B}_u\zeta_u$ in (1.2) represents, in some sense, the effects of those phenomena that continue to remain unobserved despite the incorporation of the covariates x.

Motivated by this discussion, we fit observations of $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ to the following *composite factor model* that incorporates the effects of the covariates x as well as of additional unobserved latent phenomena on y:

$$y = \mathcal{A}x + \mathcal{B}_u\zeta_u + \bar{\epsilon} \qquad (1.3)$$

where $\mathcal{A} \in \mathbb{R}^{p \times q}$ with $\text{rank}(\mathcal{A}) \ll \min\{p, q\}$, $\mathcal{B}_u \in \mathbb{R}^{p \times k_u}$ with $k_u \ll p$, and the variables $\zeta_u, \bar{\epsilon}$ are independent of each other (and of x) and normally distributed as $\zeta_u \sim \mathcal{N}(0, \Sigma_{\zeta_u}), \bar{\epsilon} \sim \mathcal{N}(0, \Sigma_{\bar{\epsilon}})$, with $\Sigma_{\zeta_u} \succ 0, \Sigma_{\bar{\epsilon}} \succ 0$ and $\Sigma_{\bar{\epsilon}}$ being a diagonal matrix. The matrix $\mathcal{A}$ may also be viewed as the map specifying the best linear estimate of y based on x. In other words, the goal is to identify a low-rank matrix $\mathcal{A}$ such that the conditional distribution of $y|x$ (and equivalently of $y|\mathcal{A}x$) is specified by a standard factor model of the form (1.1).

1.2 Composite Factor Modeling via Convex Optimization

Next we describe techniques to fit observations of $y \in \mathbb{R}^p$ to the model (1.3). This method is a key subroutine in our algorithmic approach for associating semantics to the latent variables in a factor model (see Section 1.3 for a high-level discussion of our approach and Section 3 for a more detailed experimental demonstration). Fitting observations of $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ to the composite factor model (1.3) is accomplished by identifying a Gaussian model over (y, x) with the covariance matrix of the model satisfying certain algebraic properties. For background on multivariate Gaussian statistical models, we refer the reader to [9].

The covariance matrix of y in the factor model is decomposable as the sum of a low-rank matrix $\mathcal{B}\Sigma_\zeta\mathcal{B}'$ (corresponding to the $k \ll p$ latent variables ζ) and a diagonal matrix Σ_ϵ. Based on this algebraic structure, a natural approach to factor modeling is to find the smallest rank (positive semidefinite) matrix such that the difference between this matrix and the empirical covariance of the observations of y is close to being a diagonal matrix (according to some measure of closeness, such as in the Frobenius norm). This problem is computationally intractable to solve in general due to the rank minimization objective [13]. As a result, a common heuristic is to replace the matrix rank by the trace functional, which results in the minimum trace factor analysis problem [10, 17, 18, 19]; this problem is convex and it can be solved efficiently. The use of the trace of a positive semidefinite matrix as a surrogate for the matrix rank goes back many decades, and this topic has received much renewed interest over the past several years [12, 7, 15, 3].

3

In attempting to generalize the minimum-trace factor analysis approach to the composite factor model, one encounters a difficulty that arises due to the parametrization of the underlying Gaussian model in terms of covariance matrices. Specifically, with the additional covariates $x \in \mathbb{R}^q$ in the composite model (1.3), our objective is to identify a Gaussian model over $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ with the joint covariance $\Sigma = \begin{pmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma'_{yx} & \Sigma_x \end{pmatrix} \in \mathbb{S}^{p+q}$ satisfying certain structural properties. One of these properties is that the conditional distribution of $y|x$ is specified by a factor model, which implies that the conditional covariance of $y|x$ must be decomposable as the sum of a low-rank matrix and a diagonal matrix. However, this conditional covariance is given by the Schur complement $\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma'_{yx}$, and specifying a constraint on the conditional covariance matrix in terms of the joint covariance matrix Σ presents an obstacle to obtaining computationally tractable optimization formulations.

A more convenient approach to parameterizing conditional distributions in Gaussian models is to consider models specified in terms of inverse covariance matrices, which are also called *precision matrices*. Specifically, the algebraic properties that we desire in the joint covariance matrix Σ of (y, x) in a composite factor model can also be stated in terms of the joint precision matrix $\Theta = \Sigma^{-1}$ via conditions on the submatrices of $\Theta = \begin{pmatrix} \Theta_y & \Theta_{yx} \\ \Theta'_{yx} & \Theta_x \end{pmatrix}$. First, the precision matrix of the conditional distribution of $y|x$ is specified by the submatrix Θ_y; as the covariance matrix of the conditional distribution of $y|x$ is the sum of a diagonal matrix and a low-rank matrix, the Woodbury matrix identity implies that the submatrix Θ_y is the difference of a diagonal matrix and a low-rank matrix. Second, the rank of the submatrix $\Theta_{yx} \in \mathbb{R}^{p \times q}$ is equal to the rank of $\mathcal{A} \in \mathbb{R}^{p \times q}$ in non-degenerate models (i.e., if $\Sigma \succ 0$) because the relation between $\mathcal{A}$ and Θ is given by $\mathcal{A} = -[\Theta_y]^{-1}\Theta_{yx}$. Based on this algebraic structure desired in Θ, we propose the following natural convex relaxation for fitting a collection of observations $\mathcal{D}_n^+ = \{(y^{(i)}, x^{(i)})\}_{i=1}^n \subset \mathbb{R}^{p+q}$ to the composite model (1.3):

$$(\hat{\Theta}, \hat{D}_y, \hat{L}_y) = \arg \min_{\substack{\Theta \in \mathbb{S}^{p+q},\ \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}} \quad -\ell(\Theta; \mathcal{D}_+^n) + \lambda_n[\gamma\|\Theta_{yx}\|_\star + \text{trace}(L_y)]$$

$$\text{s.t.} \quad \Theta_y = D_y - L_y,\ L_y \succeq 0, D_y \text{ is diagonal} \tag{1.4}$$

The term $\ell(\Theta; \mathcal{D}_+^n)$ is the Gaussian log-likelihood function that enforces fidelity to the data, and it is given as follows (up to some additive and multiplicative terms):

$$\ell(\Theta; \mathcal{D}_+^n) = \log\det(\Theta) - \text{trace}\left[\Theta \cdot \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} y^{(i)} \\ x^{(i)} \end{pmatrix} \begin{pmatrix} y^{(i)} \\ x^{(i)} \end{pmatrix}'\right]. \tag{1.5}$$

This function is concave as a function of the joint precision matrix[2] Θ. The matrices D_y, L_y represent the diagonal and low-rank components of Θ_y. As with the idea behind minimum-trace factor analysis, the role of the trace norm penalty on L_y is to induce low-rank structure in this matrix. Based on a more recent line of work originating with the thesis of Fazel [7, 15, 3], the nuclear norm penalty $\|\Theta_{yx}\|_\star$ on the submatrix Θ_{yx} (which is in general a non-square matrix) is useful for promoting low-rank structure in that submatrix of Θ. The parameter γ provides a tradeoff between the observed/interpretable and the unobserved parts of the composite factor model (1.3), and the parameter λ_n provides a tradeoff between the fidelity of the model to the data and the overall complexity of the model (the total number of observed and unobserved components in the

---

[2]An additional virtue of parameterizing our problem in terms of precision matrices rather than in terms of covariance matrices is that the log-likelihood function in Gaussian models is not concave over the cone of positive semidefinite matrices when viewed as a function of the covariance matrix.

composite model (1.3)). In summary, for $\lambda_n, \gamma \geq 0$ the regularized maximum-likelihood problem (1.4) is a convex program. From the optimal solution $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ of (1.4), we can obtain estimates for the parameters of the composite factor model (1.3) as follows:

$$\hat{\mathcal{A}} = -[\hat{\Theta}_y]^{-1} \hat{\Theta}_{yx}$$
$$\hat{\mathcal{B}}_u = \text{any squareroot of } (\hat{D}_y - \hat{L}_y)^{-1} - \hat{D}_y^{-1} \text{ such that } \hat{\mathcal{B}}_u \in \mathbb{R}^{p \times \text{rank}(\hat{L}_y)}, \tag{1.6}$$

with the covariance of ζ_u being the identity matrix of appropriate dimensions and the covariance of $\bar{\epsilon}$ being $\hat{D}_y^{-1}$. The convex program (1.4) is log-determinant semidefinite programs that can be solved efficiently using existing numerical solvers such as the LogDetPPA package [21].

1.3 Algorithmic Approach for Interpreting Latent Variables in a Factor Model

Our discussion has led us to a natural (meta-) procedure for interpreting latent variables in a factor model. Suppose that we are given a factor model underlying $y \in \mathbb{R}^p$. The analyst proceeds by obtaining simultaneous measurements of the variables y as well as some additional covariates $x \in \mathbb{R}^q$ of plausibly relevant phenomena. Based on these joint observations, we identify a suitable composite factor model (1.3) via the convex program (1.4). In particular, we sweep over the parameters λ_n, γ in (1.4) to identify composite models that achieve a suitable decomposition – in terms of effects attributable to the additional covariates x and of effects corresponding to remaining unobserved phenomena – of the effects of the latent variables in the factor model given as input.

To make this approach more formal, consider a composite factor model (1.3) $y = \mathcal{A}x + \mathcal{B}_u \zeta_u + \epsilon$ underlying a pair of random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$, with $\text{rank}(\mathcal{A}) = k_x$, $\mathcal{B}_u \in \mathbb{R}^{p \times k_u}$, and column-space$(\mathcal{A}) \cap$ column-space$(\mathcal{B}_u) = \{0\}$. As described in Section 1.2, the algebraic aspects of the underlying composite factor model translate to algebraic properties of submatrices of $\Theta \in \mathbb{S}^{p+q}$. In particular, the submatrix Θ_{yx} has rank equal to k_x and the submatrix Θ_y is decomposable as $D_y - L_y$ with D_y being diagonal and $L_y \succeq 0$ having rank equal to k_u. Finally, the transversality of column-space$(\mathcal{A})$ and column-space$(\mathcal{B}_u)$ translates to the fact that column-space$(\Theta_{yx}) \cap$ column-space$(L_y) = \{0\}$ have a transverse intersection. One can simply check that the factor model underlying the random vector $y \in \mathbb{R}^p$ that is induced upon marginalization of x is specified by the precision matrix of y given by $\tilde{\Theta}_y = D_y - [L_y + \Theta_{yx}(\Theta_x)^{-1}\Theta_{xy}]$. Here, the matrix $L_y + \Theta_{yx}(\Theta_x)^{-1}\Theta_{xy}$ is a rank $k_x + k_u$ matrix that captures the effect of latent variables in the factor model. This effect is decomposed into $\Theta_{yx}(\Theta_x)^{-1}\Theta_{xy}$ – a rank k_x matrix representing the component of this effect attributed to x, and L_y – a matrix of rank k_u representing the effect attributed to residual latent variables.

These observations motivate the following algorithmic procedure. Suppose we are given a factor model that specifies the precision matrix of y as the difference $\hat{\tilde{D}}_y - \hat{\tilde{L}}_y$, where $\hat{\tilde{D}}_y$ is diagonal and $\hat{\tilde{L}}_y$ is low rank. Then the composite factor model of (y, x) with estimates $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ offers an interpretation of the latent variables of the given factor model if (i) $\text{rank}(\hat{\tilde{L}}_y) = \text{rank}(\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy})$, (ii) column-space$(\hat{\Theta}_{yx}) \cap$ column-space$(\hat{L}_y) = \{0\}$, and
(iii) $\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2 / \|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - [\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2 / \|\hat{\tilde{L}}_y\|_2\}$ is small. The full algorithmic procedure for attributing meaning to latent variables of a factor model is outlined below:

5

Algorithm 1 Interpreting Latent Variables in a Factor Model

1: **Input**: A collection of observations $\mathcal{D}_n^+ = \{(y^{(i)}, x^{(i)})\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}^q$ of the variables y and of some auxiliary covariates x; Factor model with parameters $(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$.

2: **Composite Factor Modeling**: For each $d = 1, \dots, q$, sweep over parameters (λ_n, γ) in the convex program (1.4) (with $\mathcal{D}_n^+$ as input) to identify composite models with estimates $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ that satisfy the following three properties: (i) $\mathrm{rank}(\hat{\Theta}_{yx}) = d$, (ii) $\mathrm{rank}(\hat{\tilde{L}}_y) = \mathrm{rank}(\hat{L}_y) + \mathrm{rank}(\hat{\Theta}_{yx})$, and (iii) $\mathrm{rank}(\hat{\tilde{L}}_y) = \mathrm{rank}(\hat{L}_y + \mathrm{rank}(\hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}))$.

3: **Identifying Subspace**: For each $d = 1, \dots, q$ and among the candidate composite models (from the previous step), choose the composite factor model that minimizes the quantity $\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2 / \|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - [\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2 / \|\hat{\tilde{L}}_y\|_2\}$.

4: **Output**: For each $d = 1, \dots q$, the d-dimensional projection of x into the row-space of $\hat{\Theta}_{yx}$ represents the interpretable component of the latent variables in the factor model.

The effectiveness of Algorithm 1 is dependent on the size of the quantity $\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2 / \|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - \hat{L}_y - \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2 / \|\hat{\tilde{L}}_y\|_2\}$. The smaller this quantity, the better the composite factor model fits to the given factor model. Finally, recall from Section 1.1 that the projection of covariates x onto to the row-space of $\mathcal{A}$ (from the composite model (1.3)) represents the interpretable component of the latent variables of the factor model. Because of the relation $\mathcal{A} = -[\Theta_y]^{-1}\Theta_{yx}$, this interpretable component is obtained by projecting the covariates x onto the row-space of Θ_{yx}. This observation explains the final step of Algorithm 1.

The input to Algorithm 1 is a factor model underlying a collection of variables $y \in \mathbb{R}^p$, and the algorithm proceeds to obtain semantic interpretation of the latent variables of the factor model. However, in many situations, a factor model underlying $y \in \mathbb{R}^p$ may not be available in advance, and must be learned in a data-driven fashion based on observations of $y \in \mathbb{R}^p$. In our experiments (see Section 3), we learn a factor model using a specialization of the convex program (1.4). It is reasonable to ask whether one might directly fit to a composite model to the covariates and responses jointly without reference to the underlying factor model based on the responses. However, in our experience with applications, it is often the case that observations of the responses y are much more plentiful than of joint observations of responses y and covariates x. As an example, consider a setting in which the responses are a collection of financial asset prices (such as stock return values); observations of these variables are available at a very fine time-resolution on the order of seconds. On the other hand, some potentially useful covariates such as GDP, government expenditures, federal debt, and consumer rate are available at a much coarser scale (usually on the order of months or quarters). As another example, consider a setting in which the responses are reservoir volumes of California; observations of these variables are available at a daily scale. On the other hand, reasonable covariates that one may wish to associate to the latent variables underlying California reservoir volumes such as agricultural production, crop yield rate, average income, and population growth rate are available at a much coarser time scale (e.g. monthly). In such settings, the analyst can utilize the more abundant set of observations of the responses y to learn an accurate factor model first. Subsequently, one can employ our approach to associate semantics to the latent variables in this factor model based on the potentially limited number of observations of the responses y and the covariates x.

1.4 Our Results

In Section 2 we carry out a theoretical analysis to investigate whether the framework outlined in Algorithm 1 can succeed. We discuss a model problem setup, which serves as the basis for the main theoretical result in Section 2. Suppose we have Gaussian random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ that are related to each other via a composite factor model (1.3). Note that this composite factor model induces a factor model underlying the variables $y \in \mathbb{R}^p$ upon marginalization of the covariates x. In the subsequent discussion, we assume that the factor model that is supplied as input to Algorithm 1 is the factor model underlying the responses y.

Now we consider the following question: Given observations jointly of $(y, x) \in \mathbb{R}^{p+q}$, does the convex relaxation (1.4) (for suitable choices of regularization parameters λ_n, γ) estimate the composite factor model underlying these two random vectors accurately? An affirmative answer to this question demonstrates the success of Algorithm 1. In particular, a positive answer to this question implies that we can decompose the effects of the latent variables in the factor model underlying y using the convex relaxation (1.4), as the accurate estimation of the composite model underlying (y, x) implies a successful decomposition of the effects of the latent variables in the factor model underlying y. That is, steps 2-3 in the Algorithm are successful. In Section 2, we show that under suitable identifiability conditions on the population model of the joint random vector (y, x), the convex program (1.4) succeeds in solving this question. Our analysis is carried out in a high-dimensional asymptotic scaling regime in which the dimensions p, q, the number of observations n, and other model parameters may all grow simultaneously [2, 23].

We give concrete demonstration of Algorithm 1 with experiments on synthetic data and real-world financial data. For the financial asset problem, we consider as our variables y the monthly averaged stock prices of 45 companies from the Standard and Poor index over the period March 1982 to March 2016, and we identify a factor model (1.1) over y with 10 latent variables (the approach we use to fit a factor model is described in Section 3). We then obtain observations of $q = 13$ covariates on quantities related to oil trade, GDP, government expenditures, etc. (See Section 3 for the full list), as these plausibly influence stock returns. Following the steps outlined in Algorithm 1, we use the convex program (1.4) to identify a two-dimensional projection of these 13 covariates that represent an interpretable component of the 10 latent variables in the factor model, as well as a remaining set of 8 latent variables that constitute phenomena not observed via the covariates x. In further analyzing the characteristics of the two-dimensional projection, we find that EUR to USD exchange rate and government expenditures are the most relevant of the 13 covariates considered in our experiment, while mortgage rate and oil imports are less useful. See Section 3 for complete details.

1.5 Related Work

Elements of our approach bear some similarity with *canonical correlations analysis* [8], which is a classical technique for identifying relationships between two sets of variables. In particular, for a pair of jointly Gaussian random vectors $(y, x) \in \mathbb{R}^{p \times q}$, canonical correlations analysis may be used as a technique for identifying the most relevant component(s) of x that influence y. However, the composite factor model (1.3) allows for the effect of further unobserved phenomena not captured via observations of the covariates x. Consequently, our approach in some sense incorporates elements of both canonical correlations analysis and factor analysis. It is important to note that algorithms for factor analysis and for canonical correlations analysis usually operate on covariance and cross-covariance matrices. However, we parametrize our regularized maximum-likelihood problem (1.4)

in terms of precision matrices, which is a crucial ingredient in leading to a computationally tractable convex program.

The nuclear-norm heuristic has been employed widely over the past several years in a range of statistical modeling tasks involving rank minimization problems; see [23] and the references therein. The proof of our main result in Section 2 incorporates some elements from the theoretical analyses in these previous papers, along with the introduction of some new ingredients. We give specific pointers to the relevant literature in Section 4.

1.6 Notation

Given a matrix $U \in \mathbb{R}^{p_1 \times p_2}$, and the norm $\|U\|_2$ denotes the spectral norm (the largest singular value of U). We define the linear operators $\mathcal{F} : \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q \to \mathbb{S}^{(p+q)}$ and its adjoint $\mathcal{F}^\dagger : \mathbb{S}^{(p+q)} \to \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ as follows:

$$\mathcal{F}(M, N, K, O) \triangleq \left(\begin{array}{cc} M - N & K \\ K^T & O \end{array} \right), \qquad \mathcal{F}^\dagger \left(\begin{array}{cc} Q & K \\ K^T & O \end{array} \right) \triangleq (Q, Q, K, O) \qquad (1.7)$$

Similarly, we define the linear operators $\mathcal{G} : \mathbb{S}^p \times \mathbb{R}^{p \times q} \to \mathbb{S}^{(p+q)}$ and its adjoint $\mathcal{G}^\dagger : \mathbb{S}^{(p+q)} \to \mathbb{S}^p \times \mathbb{R}^{p \times q}$ as follows:

$$\mathcal{G}(M, K) \triangleq \left(\begin{array}{cc} M & K \\ K^T & 0 \end{array} \right), \qquad \mathcal{G}^\dagger \left(\begin{array}{cc} Q & K \\ K^T & O \end{array} \right) \triangleq (Q, K) \qquad (1.8)$$

Finally, for any subspace $\mathfrak{H}$, the projection onto the subspace is denoted by $\mathcal{P}_{\mathfrak{H}}$.

2 Theoretical Results

In this section, we state a theorem to prove the consistency of convex program (1.4). This theorem requires assumptions on the population precision matrix, which are discussed in Section 2.2. We provide examples of population composite factor models (1.4) that satisfy these conditions. The theorem statement is given in Section 2.4 and the proof of the theorem is given in Section 4 with some details deferred to the appendix.

2.1 Technical Setup

As discussed in Section 1.4, our theorems are premised on the existence of a population composite factor model (1.3) $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \epsilon$ underlying a pair of random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$, with $\mathrm{rank}(\mathcal{A}^\star) = k_x$, $\mathcal{B}_u^\star \in \mathbb{R}^{p \times k_u}$, and column-space$(\mathcal{A}^\star) \cap$ column-space$(\mathcal{B}_u{}^\star) = \{0\}$. As the convex relaxation (1.4) is solved in the precision matrix parametrization, the conditions for our theorems are more naturally stated in terms of the joint precision matrix $\Theta^\star \in \mathbb{S}^{p+q}$, $\Theta^\star \succ 0$ of (y, x). The algebraic aspects of the parameters underlying the factor model translate to algebraic properties of submatrices of $\Theta^\star$. In particular, the submatrix $\Theta_{yx}^\star$ has rank equal to k_x, and the submatrix $\Theta_y^\star$ is decomposable as $D_y^\star - L_y^\star$ with $D_y^\star$ being diagonal and $L_y^\star \succeq 0$ having rank equal to k_u. Finally, the transversality of column-space$(\mathcal{A}^\star)$ and column-space$(\mathcal{B}_u{}^\star)$ translates to the fact that column-space$(\Theta_{yx}^\star) \cap$ column-space$(L_y^\star) = \{0\}$ have a transverse intersection.

To address the requirements raised in Section 1.4, we seek an estimate $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ from the convex relaxation (1.4) such that $\mathrm{rank}(\hat{\Theta}_{yx}) = \mathrm{rank}(\Theta_{yx}^\star)$, $\mathrm{rank}(\hat{L}_y) = \mathrm{rank}(L_y^\star)$, and that $\|\hat{\Theta} - \Theta^\star\|_2$ is small. Building on both classical statistical estimation theory [1] as well as the recent literature on high-dimensional statistical inference [2, 23], a natural set of conditions for obtaining accurate

parameter estimates is to assume that the curvature of the likelihood function at $\Theta^\star$ is bounded in certain directions. This curvature is governed by the Fisher information at $\Theta^\star$:

$$\mathbb{I}^\star \triangleq \Theta^{\star-1} \otimes \Theta^{\star-1} = \Sigma^\star \otimes \Sigma^\star.$$

Here $\otimes$ denotes a tensor product between matrices and $\mathbb{I}^\star$ may be viewed as a map from $\mathbb{S}^{(p+q)}$ to $\mathbb{S}^{(p+q)}$. We impose conditions requiring that $\mathbb{I}^\star$ is well-behaved when applied to matrices of the form $\Theta - \Theta^\star = \begin{pmatrix} (D_y - D_y^\star) - (L_y - L_y^\star) & \Theta_{yx} - \Theta_{yx}^\star \\ \Theta_{yx}{}' - \Theta_{yx}^\star{}' & \Theta_x - \Theta_x^\star \end{pmatrix}$, where (L_y, Θ_{yx}) are in a neighborhood of $(L_y^\star, \Theta_{yx}^\star)$ restricted to sets of low-rank matrices. These local properties of $\mathbb{I}^\star$ around $\Theta^\star$ are conveniently stated in terms of *tangent spaces* to the algebraic varieties of low-rank matrices. In particular, the tangent space at a rank-r matrix $N \in \mathbb{R}^{p_1 \times p_2}$ with respect to the algebraic variety of $p_1 \times p_2$ matrices with rank less than or equal to r is given by[3]:

$$
\begin{aligned}
T(N) \quad \triangleq \quad & \{N_R + N_C | N_R, N_C \in \mathbb{R}^{p_1 \times p_2}, \\
& \text{row-space } (N_R) \subseteq \text{row-space } (N), \\
& \text{column-space } (N_C) \subseteq \text{column-space } (N)\}
\end{aligned}
$$

In the next section, we describe conditions on the population Fisher information $\mathbb{I}^\star$ in terms of the tangent spaces $T(L_y^\star)$, and $T(\Theta_{yx}^\star)$; under these conditions, we present a theorem in Section 2.4 showing that the convex program (1.4) obtains accurate estimates.

2.2   Fisher Information Conditions

Given a norm $\| \cdot \|_\Upsilon$ on $\mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$, we first consider a classical condition in statistical estimation literature, which is to control the minimum gain of the Fisher information $\mathbb{I}^\star$ restricted to a subspace $\mathbb{H} \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ as follows:

$$\chi(\mathbb{H}, \| \cdot \|_\Upsilon) \triangleq \min_{\substack{Z \in \mathbb{H} \\ \|Z\|_\Upsilon = 1}} \|\mathcal{P}_\mathbb{H} \mathcal{I}^\dagger \mathbb{I}^\star \mathcal{I} \mathcal{P}_\mathbb{H}(Z)\|_\Upsilon, \tag{2.1}$$

where $\mathcal{P}_\mathbb{H}$ denotes the projection operator onto the subspace $\mathbb{H}$ and the linear maps $\mathcal{I}$ and $\mathcal{I}^\dagger$ are defined in (1.7). The quantity $\chi(\mathbb{H}, \| \cdot \|_\Upsilon)$ being large ensures that the Fisher information $\mathbb{I}^\star$ is well-conditioned restricted to image $\mathcal{I}\mathbb{H} \subseteq \mathbb{S}^{p+q}$. The remaining conditions that we impose on $\mathbb{I}^\star$ are in the spirit of irrepresentibility-type conditions [11, 24, 22, 5, 2] that are frequently employed in high-dimensional estimation. In the subsequent discussion, we employ the following notation to denote restrictions of a subspace $\mathbb{H} = H_1 \times H_2 \times H_3 \times H_4 \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ (here H_1, H_2, H_3, H_4 are subspaces in $\mathbb{S}^p, \mathbb{S}^p, \mathbb{R}^{p \times q}, \mathbb{S}^q$, respectively) to its individual components. The restriction to the second components of $\mathbb{H}$ is given by $\mathbb{H}[2] = H_2$. The restriction to the second and third component of $\mathbb{H}$ is given by $\mathbb{H}[2, 3] = H_2 \times H_3 \subset \mathbb{S}^p \times \mathbb{R}^{p \times q}$. Given a norm $\|.\|_\Pi$ on $\mathbb{S}^p \times \mathbb{R}^{p \times q}$, we control the gain of $\mathbb{I}^\star$ restricted to $\mathbb{H}[2, 3]$

$$\Xi(\mathbb{H}, \| \cdot \|_\Pi) \triangleq \min_{\substack{Z \in \mathbb{H}[2,3] \\ \|Z\|_\Pi = 1}} \|\mathcal{P}_{\mathbb{H}[2,3]} \mathcal{G}^\dagger \mathbb{I}^\star \mathcal{G} \mathcal{P}_{\mathbb{H}[2,3]}(Z)\|_\Pi. \tag{2.2}$$

Here, the linear maps $\mathcal{G}$ and $\mathcal{G}^\dagger$ are defined in (1.8). In the spirit of irrepresentability conditions, we control the inner-product between elements in $\mathcal{G}\mathbb{H}[2, 3]$ and $\mathcal{G}\mathbb{H}[2, 3]^\perp$, as quantified by the metric

---

[3]We also consider the tangent space at a symmetric low-rank matrix with respect to the algebraic variety of symmetric low-rank matrices. We use the same notation 'T' to denote tangent spaces in both the symmetric and non-symmetric cases, and the appropriate tangent space is clear from the context.

induced by $\mathbb{I}^\star$ via the following quantity

$$\varphi(\mathbb{H}, \|\cdot\|_\Pi) \triangleq \max_{\substack{Z \in \mathbb{H}[2,3] \\ \|Z\|_\Pi = 1}} \|\mathcal{P}_{\mathbb{H}[2,3]^\perp} \mathcal{G}^\dagger \mathbb{I}^\star \mathcal{G} \mathcal{P}_{\mathbb{H}[2,3]} (\mathcal{P}_{\mathbb{H}[2,3]} \mathcal{G}^\dagger \mathbb{I}^\star \mathcal{G} \mathcal{P}_{\mathbb{H}[2,3]})^{-1}(Z)\|_\Pi. \tag{2.3}$$

The operator $(\mathcal{P}_{\mathbb{H}[2,3]} \mathcal{G}^\dagger \mathbb{I}^\star \mathcal{G} \mathcal{P}_{\mathbb{H}[2,3]})^{-1}$ in (2.3) is well-defined if $\Xi(\mathbb{H}) > 0$, since this latter condition implies that $\mathbb{I}^\star$ is injective restricted to $\mathcal{G}\mathbb{H}[2,3]$. The quantity $\varphi(\mathbb{H}, \|\cdot\|_\Upsilon)$ being small implies that any element of $\mathcal{G}\mathbb{H}[2,3]$ and any element of $\mathcal{G}\mathbb{H}[2,3]^\perp$ have a small inner-product (in the metric induced by $\mathbb{I}^\star$). The reason that we restrict this inner product to the second and third components of $\mathbb{H}$ in the quantity $\varphi(\mathbb{H}, \|.\|_\Upsilon)$ is that the regularization terms in the convex program (1.4) are only applied to the matrices L_y and Θ_{yx}.

A natural approach to controlling the conditioning of the Fisher information around $\Theta^\star$ is to bound the quantities $\chi(\mathbb{H}^\star, \|\cdot\|_\Upsilon)$, $\Xi(\mathbb{H}^\star, \|\cdot\|_\Pi)$, and $\varphi(\mathbb{H}^\star, \|\cdot\|_\Upsilon)$ for $\mathbb{H}^\star = \mathcal{W} \times T(L_y^\star) \times T(\Theta_{yx}^\star) \times \mathbb{S}^q$ where $\mathcal{W} \in \mathbb{S}^p$ is the set of diagonal matrices. However, a complication that arises with this approach is that the varieties of low-rank matrices are locally curved around $L_y^\star$ and around $\Theta_{yx}^\star$. Consequently, the tangent spaces at points in neighborhoods around $L_y^\star$ and around $\Theta_{yx}^\star$ are not the same as $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$. In order to account for this curvature underlying the varieties of low-rank matrices, we bound the distance between nearby tangent spaces via the following induced norm:

$$\rho(T_1, T_2) \triangleq \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T_1} - \mathcal{P}_{T_2})(N)\|_2.$$

The quantity $\rho(T_1, T_2)$ measures the largest angle between T_1 and T_2. Using this approach for bounding nearby tangent spaces, we consider subspaces $\mathbb{H}' = \mathcal{W} \times T_y' \times T_{yx}' \times \mathbb{S}^q$ for all T_y' close to $T(L_y^\star)$ and for all T_{yx}' close to $T(\Theta_{yx}^\star)$, as measured by ρ [2]. For $\omega_y \in (0,1)$ and $\omega_{yx} \in (0,1)$, we bound $\chi(\mathbb{H}', \|\cdot\|_\Upsilon)$, $\Xi(\mathbb{H}', \|\cdot\|_\Pi)$, and $\varphi(\mathbb{H}', \|\cdot\|_\Pi)$ in the sequel for all subspaces $\mathbb{H}'$ in the following set:

$$U(\omega_y, \omega_{yx}) \triangleq \Big\{ \mathcal{W} \times T_y' \times T_{yx}' \times \mathbb{S}^q \mid \rho(T_y', T(L_y^\star)) \leq \omega_y$$

$$\rho(T_{yx}', T(\Theta_{yx}^\star)) \leq \omega_{yx} \Big\}. \tag{2.4}$$

We control the quantities $\Xi(\mathbb{H}', \|\cdot\|_\Pi)$ and $\varphi(\mathbb{H}', \|\cdot\|_\Pi)$ using the dual norm of the regularizer $\text{trace}(L_y) + \gamma\|\Theta_{yx}\|_\star$ in (1.4):

$$\Gamma_\gamma(L_y, \Theta_{yx}) \triangleq \max \left\{ \|L_y\|_2, \frac{\|\Theta_{yx}\|_2}{\gamma} \right\}. \tag{2.5}$$

Furthermore, we control the quantity $\chi(\mathbb{H}', \|\cdot\|_\Upsilon)$ using a slight variant of the dual norm:

$$\Phi_\gamma(D_y, L_y, \Theta_{yx}, \Theta_x) \triangleq \max \left\{ \|D_y\|_2, \|L_y\|_2, \frac{\|\Theta_{yx}\|_2}{\gamma}, \|\Theta_x\|_2 \right\}. \tag{2.6}$$

As the dual norm $\max \left\{ \|L_y\|_2, \frac{\|\Theta_{yx}\|_2}{\gamma} \right\}$ of the regularizer in (1.4) plays a central role in the optimality conditions of (1.4), controlling the quantities $\chi(\mathbb{H}', \Phi_\gamma)$, $\Xi(\mathbb{H}', \Gamma_\gamma)$, and $\varphi(\mathbb{H}', \Gamma_\gamma)$ leads to a natural set of conditions that guarantee the consistency of the estimates produced by (1.4). In summary, given a fixed set of parameters $(\gamma, \omega_y, \omega_{yx}) \in \mathbb{R}_+ \times (0,1) \times (0,1)$, we assume that $\mathbb{I}^\star$

satisfies the following conditions:

$$\text{Assumption 1} \quad : \quad \inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \chi(\mathbb{H}', \Phi_\gamma) \geq \alpha, \quad \text{for some } \alpha > 0 \tag{2.7}$$

$$\text{Assumption 2} \quad : \quad \inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \Xi(\mathbb{H}', \Gamma_\gamma) > 0 \tag{2.8}$$

$$\text{Assumption 3} \quad : \quad \sup_{\mathbb{H}' \in U(\omega_{yx}, \omega_{yx})} \varphi(\mathbb{H}', \Gamma_\gamma) \leq 1 - \frac{2}{\beta + 1} \quad \text{for some } \beta \geq 2. \tag{2.9}$$

For fixed $(\gamma, \omega_y, \omega_{yx})$, larger value of α and smaller value of β in these assumptions lead to a better conditioned $\mathbb{I}^\star$.

Assumptions 1, 2, and 3 are analogous to conditions that play an important role in the analysis of the Lasso for sparse linear regression, graphical model selection via the Graphical Lasso [5], and in several other approaches for high-dimensional estimation. As a point of comparison with respect to analyses of the Lasso, the role of the Fisher information $\mathbb{I}^\star$ is played by $A^T A$, where A is the underlying design matrix. In analyses of both the Lasso and the Graphical Lasso in the papers referenced above, the analog of the subspace $\mathbb{H}$ is the set of models with support contained inside the support of the underlying sparse population model. Assumptions 1, 2, and 3 are also similar in spirit to conditions employed in the analysis of convex relaxation methods for latent-variable graphical model selection [2].

2.3 When Do the Fisher Information Assumptions Hold?

In this section, we provide examples of composite models (1.3) that satisfy Assumptions 1, 2 and 3 in (2.7) (2.8), and (2.9) for some choices of $\alpha > 0$, $\beta \geq 2$, $\omega_y \in (0, 1)$, $\omega_{yx} \in (0, 1)$ and $\gamma > 0$. Specifically, consider a population composite factor model $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \bar{\epsilon}$, where $\mathcal{A}^\star \in \mathbb{R}^{p \times q}$ with $\text{rank}(\mathcal{A}^\star) = k_x$, $\mathcal{B}_u^\star \in \mathbb{R}^{p, k_u}$, column-space$(\mathcal{A}^\star) \cap$ column-space$(\mathcal{B}_u^\star) = \{0\}$, and the random variables $\zeta_u, \bar{\epsilon}, x$ are independent of each other and normally distributed as $\zeta_u \sim \mathcal{N}(0, \Sigma_{\zeta_u}), \bar{\epsilon} \sim \mathcal{N}(0, \Sigma_{\bar{\epsilon}})$. As described in Section 1.2, the properties of the composite factor model translate to algebraic properties on the underlying precision matrix $\Theta^\star \in \mathbb{S}^{p+q}$. Namely, the submatrix $\Theta_{yx}^\star$ has rank equal to k_x and the submatrix $\Theta_y^\star$ is decomposable as $D_y^\star - L_y^\star$ with $D_y^\star$ being diagonal and $L_y^\star \succeq 0$ having rank equal to k_u. Recall that the factor model underlying the random vector $y \in \mathbb{R}^p$ that is induced upon marginalization of x is specified by the precision matrix of y given by $\tilde{\Theta}_y^\star = D_y^\star - \left[L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star \right]$. Here, $L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ represents the effect of the latent variables in the underlying factor model. When learning a composite factor model, this effect is decomposed into: $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ – a rank k_x matrix representing the component of this affect attributed to x – and $L_y^\star$ – a matrix of rank k_u representing the effect of residual latent variables. There are two identifiability concerns that arise when learning a composite factor model. First, the low rank matrices $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ must be distinguishable from the diagonal matrix $D_y^\star$. Following previous literature in diagonal and low rank matrix decompositions [16, 2], this task can be achieved by ensuring that the column/row spaces of $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ are *incoherent* with respect to the standard basis. Specifically, given a subspace $U \subset \mathbb{R}^p$, the coherence of the subspace U is defined as:

$$\mu(U) = \max_{i=1,2...p} \|\mathcal{P}_U(e_i)\|_{\ell_2}^2$$

where $\mathcal{P}$ denotes a projection operation and $e_i \in \mathbb{R}^p$ denotes the i'th standard basis vector. It is not difficult to show that this incoherence parameter satisfies the following inequality:

$$\frac{\dim(U)}{p} \leq \mu(U) \leq 1.$$

A subspace U with small coherence is necessarily of small dimension and far from containing standard basis elements. As such, a symmetric matrix with incoherent row and column spaces is low-rank and quite different from being a diagonal matrix. Consequently, we require that the quantities $\mu(\text{column-space}(L_y^\star))$ and $\mu(\text{column-space}(\Theta_{yx}^\star \Theta_x^{\star -1} \Theta_{xy}^\star))$ are small [4]. The second identifiability issue that arises is distinguishing the low rank matrices $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ from one another. This task is made difficult when the row/column spaces of these matrices are nearly aligned. Thus, we must ensure that the row/column spaces of $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ are sufficiently transverse (i.e. have large angles).

These identifiability issues directly translate to conditions on the population composite factor model. Specifically, $\mu(\text{column-space}(L_y^\star))$ and $\mu(\text{column-space}(\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star))$ being small translates to $\mu(\text{column-space}(\mathcal{A}^\star))$ and $\mu(\text{column-space}(\mathcal{B}_u^\star))$ being small. Such a condition has another interpretation. It states that the effect of x and ζ_u must not concentrate on any one variable of y; otherwise, this effect can be absorbed by the random variable $\bar{\epsilon}$ in (1.3). The second identifiability assumption that the row/column spaces of $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ have a large angle translates to the angle between column spaces of $\mathcal{A}^\star$ and $\mathcal{B}_u^\star$ being large. This assumption ensures that the effect of x and ζ_u on y can be distinguished.

Having these identifiability concerns in mind, we give a stylized composite factor model (1.3) and check that the Fisher Information Assumptions 1,2, and 3 in (2.7), (2.8), and (2.9) are satisfied for appropriate choices of parameters. Specifically, we let $p = 60$, $q = 2$, $k_x = 1$, and $k_u = 1$. We let the random variables $x \in \mathbb{R}^q$, $\zeta_u \in \mathbb{R}^{k_u}$, $\bar{\epsilon} \in \mathbb{R}^p$ be distributed according to $x \sim \mathcal{N}(0, \mathcal{I}_{q \times q})$, $\zeta_u \sim \mathcal{N}(0, \mathcal{I}_{k_u \times k_u})$, and $\bar{\epsilon} \sim \mathcal{N}(0, \mathcal{I}_{p \times p})$. We generate matrices $J \in \mathbb{R}^{p \times k_x}$, $K \in \mathbb{R}^{q \times k_x}$ with i.i.d Gaussian entries, and let $\mathcal{A}^\star = JK^T$. Similarly, we generate $\mathcal{B}_u^\star \in \mathbb{R}^{p \times k_u}$ with i.i.d Gaussian entries. We scale matrices $\mathcal{A}^\star$ and $\mathcal{B}_u^\star$ to have spectral norm equal to 0.2. With this selection, the smallest angle between the column spaces of $\mathcal{A}^\star$ and $\mathcal{B}_u^\star$ is 87 degrees. Furthermore, the quantities $\mu(\text{column-space}(\mathcal{A}^\star))$ and $\mu(\text{column-space}(\mathcal{B}_u^\star))$ are 0.072 and 0.074 respectively, . Under this stylized setting, we numerically evaluate Assumptions 1, 2, and 3 in (2.7), (2.8), and (2.9) with a Fisher information $\mathbb{I}^\star$ that takes the form:

$$\mathbb{I}^\star = \begin{pmatrix} \mathcal{I} + \mathcal{A}^\star \mathcal{A}^{\star T} + \mathcal{B}_u^\star \mathcal{B}_u^{\star T} & \mathcal{A}^\star \\ \mathcal{A}^{\star T} & \mathcal{I} \end{pmatrix} \otimes \begin{pmatrix} \mathcal{I} + \mathcal{A}^\star \mathcal{A}^{\star T} + \mathcal{B}_u^\star \mathcal{B}_u^{\star T} & \mathcal{A}^\star \\ \mathcal{A}^{\star T} & \mathcal{I} \end{pmatrix}$$

We let $\omega_y = 0.03, \omega_{yx} = 0.03$ so that the largest angle between the pair of tangent spaces $T_y', T(L_y^\star)$ and tangent spaces $T_{yx}', T(\Theta_{yx}^\star)$ is less than 1.8 degrees. Letting $\gamma \in (1, 1.4)$, one can numerically check that
$\inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \chi(\mathbb{H}', \Phi_\gamma) > 0.2$, $\inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \Xi(\mathbb{H}', \Gamma_\gamma) > 0.4$ and
$\sup_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \varphi(\mathbb{H}', \Gamma_\gamma) < 0.8$. Thus, for $\omega_y = 0.03$, $\omega_{yx} = 0.03$, $\alpha = 0.2$, $\beta = 9$, and $\gamma \in (1, 1.4)$, the Fisher information condition Assumptions 1, 2, and 3 in (2.7), (2.8) and (2.9) are satisfied.

---

[4]We only need to control the coherence of the column spaces since these matrices are symmetric.

2.4 Theorem Statement

We now describe the performance of the regularized maximum-likelihood programs (1.4) under suitable conditions on the quantities introduced in the previous section. Before formally stating our main result, we introduce some notation. Let σ_y denote the minimum nonzero singular value of $L_y^\star$ and let σ_{yx} denote the minimum nonzero singular value of $\Theta_{yx}^\star$. We state the theorem based on essential aspects of the conditions required for the success of our convex relaxation (i.e. the Fisher information conditions) and omit complicated constants. We specify these constants in Section 4.

Theorem 2.1. *Suppose that there exists $\alpha > 0$, $\beta \geq 2$, $\omega_y \in (0,1)$, $\omega_{yx} \in (0,1)$, and the choice of parameter γ so that the population Fisher information $\mathbb{I}^\star$ satisfies Assumptions 1, 2, and 3 in (2.7), (2.8) and (2.9). Let $m \triangleq \max\{1, \frac{1}{\gamma}\}$, and $\bar{m} \triangleq \max\{1, \gamma\}$. Furthermore, suppose that the following conditions hold:*

1. $n \gtrsim \left[\frac{\beta^2}{\alpha^2} m^6\right](p+q)$

2. $\lambda_n \sim \left[\frac{\beta}{\alpha} m^2\right]\sqrt{\frac{p+q}{n}}$

3. $\sigma_y \gtrsim \left[\frac{\beta}{\alpha^5 \omega_y} m^4\right]\lambda_n$

4. $\sigma_{yx} \gtrsim \left[\frac{\beta}{\alpha^5 \omega_{yx}} m^5 \bar{m}^2\right]\lambda_n$

Then with probability greater than $1 - 2\exp\left\{-\tilde{C}_{prob}\frac{\alpha^2}{\beta^2 m^4} n\lambda_n^2\right\}$, the optimal solution $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ of (1.4) with i.i.d. observations $\mathcal{D}_n^+ = \{y^{(i)}, x^{(i)}\}_{i=1}^n$ of (y, x) satisfies the following properties:

1. $rank(\hat{L}_y) = rank(L_y^\star)$, $rank(\hat{\Theta}_{yx}) = rank(\Theta_{yx}^\star)$

2. $\|\hat{D}_y - D_y^\star\|_2 \lesssim \frac{m}{\alpha^2}\lambda_n$, $\|\hat{L}_y - L_y^\star\|_2 \lesssim \frac{m}{\alpha^2}\lambda_n$, $\|\hat{\Theta}_{yx} - \Theta_{yx}^\star\|_2 \lesssim \frac{m\bar{m}}{\alpha^2}\lambda_n$, $\|\hat{\Theta}_x - \Theta_x^\star\|_2 \lesssim \frac{m}{\alpha^2}\lambda_n$

We outline the proof of Theorem 6.4 in Section 4. The quantities $\alpha, \beta, \omega_y, \omega_{yx}$ as well as the choices of parameters γ play a prominent role in the results of Theorem 6.4. Indeed larger values of $\alpha, \omega_y, \omega_{yx}$ and smaller values of β (leading to a better conditioned Fisher information even for large distortions around the tangent space $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$ lead to less stringent requirements on the sample complexity, on the minimum nonzero singular value of σ_y of $L_y^\star$, and on the minimum nonzero singular value σ_{yx} of $\Theta_{yx}^\star$.

3 Experimental Results

In this section, we demonstrate the utility of Algorithm 1 for interpreting latent variables in factor models both with synthetic and real financial asset data.

3.1 Synthetic Simulations

We give experimental evidence for the utility of Algorithm 1 on synthetic examples. Specifically, we generate a composite factor model (1.3) $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \bar{\epsilon}$ as follows: we fix $p = 40$ and $q = 10$. We let the random variables $x \in \mathbb{R}^q$, $\zeta_u \in \mathbb{R}^{k_u}$, $\bar{\epsilon} \in \mathbb{R}^p$ be distributed according to $x \sim \mathcal{N}(0, \mathcal{I}_{q \times q})$, $\zeta_u \sim \mathcal{N}(0, \mathcal{I}_{k_u \times k_u})$, and $\bar{\epsilon} \sim \mathcal{N}(0, \mathcal{I}_{p \times p})$. We generate matrices $J \in \mathbb{R}^{p \times k_x}, K \in \mathbb{R}^{q \times k_x}$ with i.i.d Gaussian entries, and let $\mathcal{A}^\star = JK^T$. Similarly, we generate $\mathcal{B}_u^\star \in \mathbb{R}^{p \times k_u}$ with i.i.d Gaussian

entries. This approach generates a factor model (1.1) with $k = k_x + k_u$. The composite factor model translates to a joint precision matrix $\Theta^\star$, with the submatrix $\Theta_y^\star = D_y^\star - L_y^\star$ where $D_y^\star$ is diagonal, $\text{rank}(L_y^\star) = k_u$, and $\text{rank}(\Theta_{yx}^\star) = k_x$. We scale matrices $\mathcal{A}^\star$ and $\mathcal{B}_u^\star$ to have spectral norm equal to τ. The value τ is chosen to be as large as possible without the condition number of $\Theta^\star$ exceeding 10 (this is imposed for the purposes of numerical conditioning). We obtain four models with $(k_x, k_u) = (1, 1), (k_x, k_u) = (2, 2)$, and $(k_x, k_u) = (4, 4)$, and $(k_x, k_u) = (6, 6)$.

For the purposes of this experiment, we assume that the input to Algorithm 1 is the oracle factor model specified by the parameters $(D_y^\star, L_y^\star + \Theta_{yx}^\star(\Theta_x)^{-1}\Theta_{xy}^\star)$, and demonstrate the success of steps 2-3 of Algorithm 1. In particular, for each model, we generate n samples of responses y and covariates x, and use these observations as input to the convex program (1.4). The regularization parameters λ_n, γ are chosen so that the estimates $(\hat{\Theta}, \hat{L}_y, \hat{D}_y)$ satisfy (i) $\text{rank}(L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star) = \text{rank}(\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy})$,
(ii) column-space$(\hat{\Theta}_{yx}) \cap$ column-space$(\hat{L}_y) = \{0\}$, and the deviation from the underlying factor model
$\max\{\|D_y^\star - \hat{D}_y\|_2/\|D_y^\star\|_2, \|L_y^\star - [\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2/\|L_y^\star\|_2\}$ is minimized. Figure 1(a) shows the magnitude of the deviation for different values of n. Furthermore, for each fixed n, we use the choice of regularization parameters (λ_n, γ) to compute the probability of obtaining structurally correct estimates of the composite model (i.e. $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\Theta_{yx}^\star) = \text{rank}(\hat{\Theta}_y)$). These probabilities are evaluated over 10 experiments and are shown in Figure 1(b). These results support Theorem 1 that given (sufficiently many) samples of responses/covariates, the convex program (1.4) provides accurate estimates of the composite factor model (1.3).

(a) composite factor model error   (b) composite factor model structural recovery
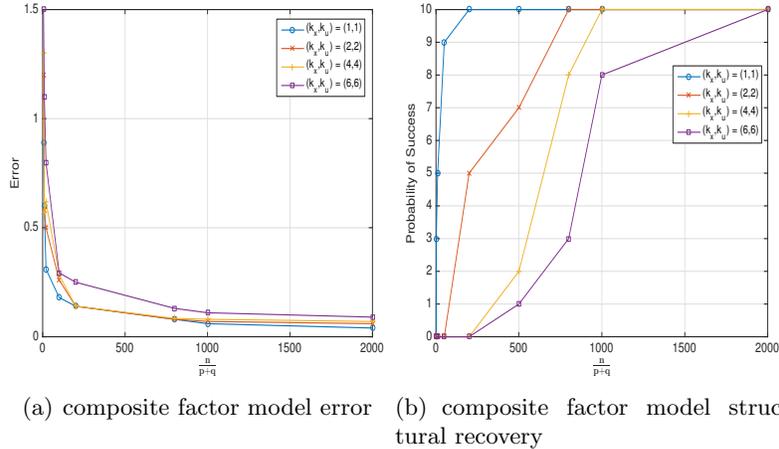
Figure 1: Synthetic data: plot shows the error (defined in the main text) and probability of correct structure recovery in composite factor models. The four models studied are (i) $(k_x, k_u) = (1, 1)$, (ii) $(k_x, k_u) = (2, 2)$, and (iii) $(k_x, k_u) = (4, 4)$, and (iv) $(k_x, k_u) = (6, 6)$. For each plotted point in (b), the probability of structurally correct estimation is obtained over 10 trials.

3.2   Experimental Results on Financial Asset Data

We consider as our responses y the monthly stock returns of $p = 45$ companies from the Standard and Poor index over the period March 1982 to March 2016, which leads to a total of $n = 408$ observations. We then obtain measurements of 13 covariates that can plausibly influence the values of stock prices: consumer price index, producer price index, EUR to USD exchange rate, federal debt

(normalized by GDP), federal reserve rate, GDP growth rate, government spending (normalized by GDP), home ownership rate, industrial production index, inflation rate, mortgage rate, oil import, and saving rate. Of these 13 covariates, the covariates federal debt, government spending, GDP growth rate, and home ownership rate are only available at a quarterly scale. Monthly observations are available for the remaining covariates. Evidently, many more observations of y are available than of (y, x) jointly. As described in Section 1.3, this scenario motivates us to first learn a factor model using the monthly observations of y. We then associate semantics to the latent variables of this factor model by fitting a composite factor model to the more limited joint observations of (y, x).

As a factor model is not available in advance, we begin with learning a factor model (1.1) using observations of y. In particular, we fit observations $\mathcal{D}_n = \{y^{(i)}\}_{i=1}^n$ to the factor model (1.1) using the following convex relaxation:

$$(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y) = \arg \min_{\substack{\tilde{D}_y, \tilde{L}_y \in \mathbb{S}^p \\ \tilde{D}_y - \tilde{L}_y \succ 0}} -\ell(\tilde{D}_y - \tilde{L}_y; \mathcal{D}_n) + \tilde{\lambda}_n \mathrm{trace}(\tilde{L}_y)$$

$$\text{s.t.} \qquad \tilde{L}_y \succeq 0, \tilde{D}_y \text{ is diagonal.} \tag{3.1}$$

We note that the convex program (3.1) is a specialization of the convex program (1.4) for learning a composite factor model. The parameter $\tilde{\lambda}_n$ in (3.1) provides a tradeoff between fidelity of the model to the observations and the complexity of the model (i.e., the number of latent variables). In contrast to minimum-trace factor analysis – in which the objective is to decompose a covariance matrix as the sum of a diagonal matrix and a low-rank matrix [10, 17, 18, 19]– the regularized maximum-likelihood convex program (3.1) fits factor models by decomposing a precision matrix as the difference between a diagonal matrix and a low-rank matrix. Although the focus of this paper is not about learning a factor model accurately, for the sake of completeness, we show in Section 5.6 of the appendix that under suitable conditions on the population model, the convex relaxation (3.1) provides an accurate estimate of the underlying factor model.

For the purpose of learning a factor model, we set aside a random subset of $n_{\text{train}} = 308$ of the total $n = 408$ observations as a training set and the remaining subset of $n_{\text{test}} = 100$ as the test set. We let $\mathcal{D}_{\text{train}} = \{y^{(i)}\}_{i=1}^{n_{\text{train}}}$ and $\mathcal{D}_{\text{test}} = \{y^{(i)}\}_{i=1}^{n_{\text{test}}}$ be the corresponding training and testing data sets respectively. We use the observations $\mathcal{D}_{\text{train}}$ as input to the convex program (3.1) where the regularization parameter $\tilde{\lambda}_n$ is chosen via cross-validation. Concretely, for a particular choice of $\tilde{\lambda}_n$, we supply $\mathcal{D}_{\text{train}}$ as input to the convex program (3.1), and solve (3.1) to obtain a factor model specified by $(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$. We then compute the average log-likelihood over the testing set $\mathcal{D}_{\text{test}}$ using the distribution specified by the precision matrix $\hat{\tilde{D}}_y - \hat{\tilde{L}}_y$. We perform this procedure as we vary $\tilde{\lambda}_n$ from 0.04 to 4 in increments of 0.004. Figure ?? shows a plot of $\mathrm{rank}(\hat{\tilde{L}}_y))$ (i.e. number of latent factors) vs. average log-likelihood performance on the testing set. Notice that fixing the number of latent factors does not lead to a unique factor model as varying the regularization parameter $\tilde{\lambda}_n$ may lead to a change in the estimated model, but no change in its structure (i.e. $\mathrm{rank}(\hat{\tilde{L}}_y)$ remains the same). As larger values of average log-likelihood are indicative of a better fit to test samples, these results suggest that 10 latent factors influence stock prices. We thus focus on associating semantics to the factor model with the largest average log-likelihood performance that consists of 10 latent factors.
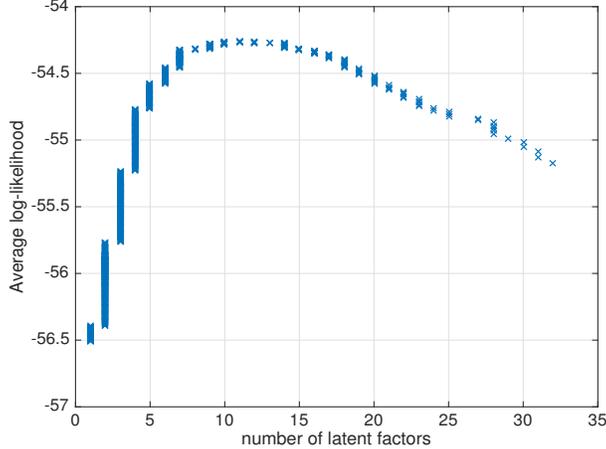
Figure 2: Number of latent factors vs. average log-likelihood over testing set. These results are obtained by sweeping over parameters $\tilde{\lambda}_n \in [0.04, 4]$ in increments of 0.004 and solving the convex program (3.1)

.

We now proceed with the steps 2-3 of Algorithm 1. To obtain a consistent set of joint observations (y, x) to employ as input to the convex program (1.4), we apply a 3-month averaging for each variable that is available at a monthly scale (i.e. the responses y and the covariates x with the exception of the four specified earlier) to obtain quarterly measurements. This leads to $n = 137$ quarterly measurements. We denote the quarterly responses and covariates by $\tilde{y}$ and $\tilde{x}$, respectively. We let $\mathcal{D}_n^+ = \{(\tilde{y}^{(i)}, \tilde{x}^{(i)})\}_{i=1}^n$ be the set of joint quarterly observations of response $\tilde{y}$ and covariates $\tilde{x}$. Using observations $\mathcal{D}_n^+$ as input to the convex program (1.4), we perform an exhaustive sweep over parameter space (λ_n, γ) to learn composite models with estimates $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ such that $\text{rank}(\hat{\Theta}) = 0, 1, 2, \ldots 10$, and $\text{rank}(\hat{L}_y) = 0, 1, 2, \ldots 10$. As we are interested comparing these composite models to the factor model with 10 latent variables, we finely grid the parameter space (λ_n, γ) so that there are a large number of models for which $\text{rank}(\hat{\Theta}) + \text{rank}(\hat{L}_y)$ is equal to 10. Among these models, we restrict to those that satisfy the conditions of step 3 of Algorithm 1. Table **??** shows the number of models that satisfy these conditions for $\text{rank}(\hat{\Theta}_{yx}) = 1, \ldots, 5$. For each $d = 1, \ldots, 5$, we then identify the composite factor model which minimizes the quantity $\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2 / \|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - \hat{L}_y - \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2 / \|\hat{\tilde{L}}_y\|_2\}$. Table 2 shows the values of this quantity for $\text{rank}(\hat{\Theta}_{yx}) = 1, \ldots, 5$ with respect to the factor model with 10 latent variables.

$(\text{rank}(\hat{\Theta}_{yx}), \text{rank}(\hat{L}_y))$	# models satisfying conditions of step 2.
(1,9)	167
(2,8)	196
(3,7)	218
(4,6)	110
(5,5)	98

Table 1: Number of composite factor models with $\text{rank}(\hat{\Theta}_{yx}) = 1, \ldots, 5$ that satisfy the requirements of step 2 in Algorithm 1(for the factor model with 10 latent variables).

$(\mathrm{rank}(\hat{\Theta}_{yx}), \mathrm{rank}(\hat{L}_y))$	$\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2/\|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - \hat{L}_y - \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2/\|\hat{\tilde{L}}_y\|_2\}$
(1,9)	0.39
(2,8)	0.40
(3,7)	0.47
(4,6)	0.51
(5,5)	0.55

Table 2: Deviation of the candidate composite factor model from the factor model consisting of 10 latent variables.

Examining Table 2, we note that there is large increase in deviation as $\mathrm{rank}(\hat{\Theta}_{yx})$ is increased above 2. Thus, we consider the composite factor model with $\mathrm{rank}(\hat{\Theta}_{yx}) = 2$ to be an acceptable approximation of the underlying factor model. As a final step of the algorithm, we investigate the properties of the two-dimensional row-space of $\hat{\Theta}_{yx}$ to shed some light on those covariates that appear to play a significant role in capturing some of the latent phenomena in the 10-factor model. In particular, for the composite factor model with $(\mathrm{rank}(\hat{\Theta}_{yx}), \mathrm{rank}(\hat{L}_y)) = (2, 8)$ (second row in Table 2), we let $V \in \mathbb{R}^{13 \times 2}$ denote a matrix with orthogonal, unit-norm columns such that the columns of V form a basis for the row space of $\hat{\Theta}_{yx}$ (such a matrix may be computed, for example, via the singular value decomposition). Thus, the projection of x onto the row-space of $\hat{\Theta}_{yx}$ – given by $V^T x$ – represents the interpretable component of the latent variables. We then consider the Euclidean-squared-norm of the i-th row of V, as this specifies the relative strength of the i-th covariate. As shown in Table 3, all covariates have some contribution (as we allow general linear combinations of the covariates x in the composite factor model (1.3)). However, the covariates exchange rate, government expenditures, and GDP growth rate seem to be the most relevant, and the covariates mortgage rate and oil import seem to be the least relevant.

covariate	strength
Exchange rate	0.18
Government expenditures	0.14
GDP growth rate	0.11
Home ownership rate	0.09
Industrial production rate	0.08
PPI	0.08
CPI	0.07
Federal debt	0.06
Saving rate	0.04
Inflation rate	0.04
Federal reserve rate	0.03
Oil import	0.03
Mortgage rate	0.01

Table 3: Strength of each covariate in the composite factor model with 2-dimensional projection of covariates and 8 latent variables

4 Proof Strategy of Theorem 1

We first begin by specifying the constants in Theorem 1. Let $\psi \triangleq \|\Theta^{\star-1}\|_2$, $\tilde{C} = 352\psi^3$, $\tilde{C}_0 = \max\{\frac{1}{192\psi}, 2\psi, \frac{1}{24\psi^2\max\{\frac{2}{\psi^2}+8, \frac{1}{\psi}\}}, \frac{\psi}{8}\}$, $\tilde{C}_{samp} = \tilde{C}\tilde{C}_0$, $\tilde{C}_1 = \frac{1}{6}(186\psi^2 + 56\psi^4)$, $\tilde{C}_\sigma = 6\psi^4(56\psi^4 + 186\psi^2)^2$, and $\tilde{C}_{prob} = \frac{1}{247808\psi^6}$. The precise conditions on the number of observations, the regularization parameter λ_n, minimum nonzero singular value of $L_y^\star$ and minimum nonzero singular value of $\Theta_{yx}^\star$ for Theorem 1 are given by:

1. $n \geq \tilde{C}_{samp}^2\left[\frac{\beta^4}{\alpha^2}m^6(p+q)\right]$

2. $\lambda_n \in \left[\tilde{C}\left\{\frac{\beta}{\alpha}m^2\sqrt{\frac{p+q}{n}}\right\}, \frac{1}{\beta m\tilde{C}_0}\right]$

3. $\sigma_y \geq \tilde{C}_\sigma\left[\frac{\beta}{\alpha^5\omega_y}m^4\lambda_n\right]$

4. $\sigma_{yx} \geq \tilde{C}_\sigma\left[\frac{\beta}{\alpha^5\omega_{yx}}m^5\bar{m}^2\lambda_n\right]$

Moreover, under these conditions, with probability greater than $1 - 2\exp\left(-\tilde{C}_{prob}\frac{\alpha^2}{m^4\beta^2}n\lambda_n^2\right)$, the optimal solution of the convex program (1.4) with estimates $(\hat{\Theta}, \hat{L}_y, \hat{D}_y)$ satisfies the following properties:

1. $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$, $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$

2. $\|\hat{D}_y - D_y^\star\|_2 \leq \tilde{C}_1\frac{m}{\alpha^2}\lambda_n$, $\|\hat{L}_y - L_y^\star\|_2 \leq \tilde{C}_1\frac{m}{\alpha^2}\lambda_n$, $\|\hat{\Theta}_{yx} - \Theta_{yx}^\star\|_2 \leq \tilde{C}_1\frac{m\bar{m}}{\alpha^2}\lambda_n$, $\|\hat{\Theta}_x - \Theta_x^\star\|_2 \leq \tilde{C}_1\frac{m}{\alpha^2}\lambda_n$

Now under assumptions of Theorem 1, we construct appropriate primal feasible variables $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ that satisfy the conclusions of the theorem - i.e., $\hat{\Theta}_{yx}$, $\hat{L}_y$ are low-rank (with the same ranks as the underlying population quantities $\Theta_{yx}^\star$ and $L_y^\star$) - and for which there exists a corresponding dual variable certifying optimality. This proof technique is sometimes also referred to as a primal-dual witness or certificate approach [22]. The high-level proof strategy is similar in spirit to the proofs of consistency results for sparse graphical model recovery [5] and latent variable graphical model recovery [2], although our convex program and the conditions required for its success are different from these previous results. Consider the following convex program

$$(\hat{\Theta}, \hat{D}_y, \hat{L}_y) = \arg\min_{\substack{\Theta \in \mathbb{S}^{p+q}, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}} -\ell(\Theta; \mathcal{D}_n^+) + \lambda_n[\gamma\|\Theta_{yx}\|_\star + \|L_y\|_\star]$$

$$\text{s.t.} \quad \Theta_y = D_y - L_y, D_y \text{ is diagonal} \tag{4.1}$$

Comparing (4.1) with the convex program (1.4), the difference is that we no longer constrain L_y to be a positive semidefinite matrix. In particular, if $L_y \succeq 0$, then the nuclear norm of the matrix L_y in the objective function of (4.1) reduces to the trace of L_y. We show in the appendix that with high probability, the matrix $\hat{L}_y$ is positive semidefinite. Standard convex analysis states that $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ is the solution of the convex program (4.1) if there exists a dual variable $\Lambda \in \mathbb{S}^p$ with the following optimality conditions being satisfied:

$$[\Sigma_n - \hat{\Theta}^{-1}]_y + \Lambda = 0 \quad ; \quad [\Sigma_n - \hat{\Theta}^{-1}]_y \in \lambda_n\partial\|\hat{L}_y\|_\star$$

$$[\Sigma_n - \hat{\Theta}^{-1}]_{yx} \in -\lambda_n\gamma\partial\|\hat{\Theta}_{yx}\|_\star \quad ; \quad [\Sigma_n - \hat{\Theta}^{-1}]_x = 0$$

$$\hat{\Theta}_y = \hat{D}_y - \hat{L}_y; \quad \hat{D}_y \text{ is diagonal} \quad ; \quad \Lambda_{i,i} = 0 \text{ for } i = 1, 2, \dots p$$

Recall that elements of the subdifferential with respect to nuclear norm at a matrix M have the key property that they decompose with respect to the tangent space $T(M)$. Specifically, the subdifferential with respect to the nuclear norm at a matrix M with (reduced) SVD given by $M = UQV^T$ is as follows:

$$N \in \partial \|M\|_\star \Leftrightarrow \mathcal{P}_{T(M)}(N) = UV^T \ , \ \|\mathcal{P}_{T(M)^\perp}(N)\|_2 \leq 1,$$

where $\mathcal{P}$ denote a projection operator. Let us denote the subspace $\mathcal{W} \in \mathbb{S}^p$ as the set of diagonal matrices with nonnegative entries. Let SVD of $\hat{L}_y$ and $\hat{\Theta}_{yx}$ be given by $\hat{L}_y = \bar{U}\bar{Q}\bar{V}'$ and $\hat{\Theta}_{yx} = \breve{U}\breve{Q}\breve{V}'$ respectively, and $Z \triangleq (0, \ \lambda_n \bar{U}\bar{V}', \ -\lambda_n \gamma \breve{U}\breve{V}', \ 0)$. Setting $\Lambda = [\Sigma_n - \hat{\Theta}^{-1}]_{Y,\text{off diagonal}}$, and letting $\mathbb{H} = \mathcal{W} \times T(\hat{L}_y) \times T(\hat{\Theta}_{yx}) \times \mathbb{S}^q$, the optimality conditions of (4.1) can be reduced to:

1. $\mathcal{P}_{\mathbb{H}} \mathcal{F}^\dagger(\Sigma_n - \hat{\Theta}^{-1}) = Z$

2. $\|\mathcal{P}_{T(\hat{L}_y)^\perp}(\Sigma_n - \hat{\Theta}^{-1})_y\|_2 < \lambda_n$; $\|\mathcal{P}_{T(\hat{\Theta}_{yx})^\perp}(\Sigma_n - \hat{\Theta}^{-1})_{yx}\|_2 < \lambda_n \gamma$

Our analysis proceeds by constructing variables $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ that satisfy the optimality conditions specified above. Consider the optimization program (4.1) with additional (non-convex) constraints that L_y and Θ_{yx} belong to algebraic variety of low rank matrices specified by $L_y^\star$ and $\Theta_{yx}^\star$. While this new program is non-convex, it has a very interesting property that at the global optimal solution (and indeed at any locally optimal solution) $\hat{L}_y$ and $\hat{\Theta}_{yx}$ are smooth points of their respective algebraic varieties. This observation suggests that the Lagrange multipliers corresponding to the additional variety constraints belongs to $T(\hat{L}_y)^\perp$ and $T(\hat{\Theta}_{yx})^\perp$ respectively. We show under suitable conditions that $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ also satisfy the second optimality condition of (4.1) corresponding to the tangent spaces $T(\hat{L}_y)^\perp$ and $T(\hat{\Theta}_{yx})^\perp$. Thus $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ is a unique solution of (1.4) and as constructed, is algebraically consistent (i.e. $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$)

4.1 Results Proved in the Appendix

To ensure that the estimate $\hat{\Theta}$ is close to the population quantity $\Theta^\star$, the quantity $E = \hat{\Theta} - \Theta^\star$ must be small. Since the optimality conditions of (4.1) are stated in terms of $\hat{\Theta}^{-1}$, we bound the deviation between $\hat{\Theta}^{-1}$ and $\Theta^{\star-1}$. Specifically, the Taylor series expansion of $\hat{\Theta}^{-1}$ around $\Theta^\star$ is given by:

$$\hat{\Theta}^{-1} = (\Theta^\star + E)^{-1} = \Theta^{\star-1} + \Theta^{\star-1}E\Theta^{\star-1} + R_{\Sigma^\star}(E)$$

where, $R_{\Sigma^\star}(E) = \Sigma^\star \left[\sum_{k=2}^\infty (-E\Theta^\star)^k \right]$. Recalling that $\mathbb{I}^\star = \Theta^{\star-1} \otimes \Theta^{\star-1}$, we note that $\hat{\Theta}^{-1} - \Theta^{\star-1} = \mathbb{I}^\star(E) + R_{\Sigma^\star}(E)$. In Section 2.2, we imposed assumptions 1,2, and 3 in (2.7), (2.8), and (2.9) on $\mathbb{I}^\star$. These assumptions allow us to control $\mathbb{I}^\star(E)$ when E is restricted to certain directions. We bound the remainder term $R_{\Sigma^\star}(E)$ in Proposition 4.1 where E is restricted to live in a certain space. Specifically, consider the following constrained optimization program:

$$(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y) = \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \ \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\text{argmin}} \quad -\ell(\Theta; \mathcal{D}_+^n) + \lambda_n[\|L_y\|_\star + \gamma\|\Theta_{yx}\|_\star]$$

$$\text{s.t.} \quad \Theta_y = D_y - L_y, \ (D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}' \tag{4.2}$$

Here $\mathbb{H}' = \mathcal{W} \times T'_y \times T'_{yx} \times \mathbb{S}^q$, where T'_y is a subspace in $\mathbb{S}^p$, and T'_{yx} is a subspace in $\mathbb{R}^{p \times q}$. Let $\Delta = (\tilde{D}_y - D_y^\star, \tilde{L}_y - L_y^\star, \tilde{\Theta}_{yx} - \Theta_{yx}^\star, \tilde{\Theta}_x - \Theta_x^\star)$ denote the error in the estimated variables. Furthermore, let $\Delta_1 = \tilde{D}_y - D_y^\star$, $\Delta_2 = \tilde{L}_y - L_y^\star$ and so forth. In the following proposition, we bound the remainder term $R_{\Sigma^\star}(\mathcal{F}(\Delta))$ defined earlier.

Proposition 4.1. *Let $\psi \triangleq \|\Theta^{\star-1}\|_2$ and $C' = (3+\gamma)\psi$. If $\Phi_\gamma[\Delta] \leq \frac{1}{2C'}$, then $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \leq 2m\psi C'^2 \Phi_\gamma[\Delta]^2$.*

Notice the bound on $R_{\Sigma^\star}(\mathcal{F}(\Delta))$ is dependent on the error term $\Phi_\gamma[\Delta]$. In the following proposition, we bound this error so that we can control the remainder term. Suppose that for $\alpha > 0$, $\beta \geq 2$, $\omega_y \in (0,1)$, and $\omega_{yx} \in (0,1)$, the Fisher information conditions (2.7), (2.8), and (2.9) are satisfied. Suppose we let T'_y and T'_{yx} be tangent spaces to the low-rank matrix varieties and $\rho(T'_y, T(L^\star_y)) \leq \omega_y$ and $\rho(T'_{yx}, T(\Theta^\star_{yx})) \leq \omega_{yx}$. Let $E_n = \Sigma^\star - \Sigma_n$ denote the difference between the true joint covariance and the sample covariance and let $C_T = (\mathcal{P}_{T'^\perp_y}(L^\star_y), \mathcal{P}_{T'^\perp_{yx}}(\Theta^\star_{yx}))$. The proof of the following result uses Brouwer's fixed-point theorem, and is inspired by the proof of a similar result in [5, 2].

Proposition 4.2. *Let $\kappa \triangleq \beta(3 + \frac{16}{\alpha}\psi^2 m)$. Consider the following two quantities:*

$$r_1 \triangleq \max\left\{ \frac{4}{\alpha}\left(\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} C_T] + \lambda_n\right), \ \Phi_\gamma[C_T]\right\} \tag{4.3}$$

$$r_2 \triangleq \frac{4}{\alpha}\left(\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} C_T]\right) \tag{4.4}$$

Define $r^u_1 \triangleq \max\left\{\frac{4}{\alpha}\left(\frac{2\lambda_n}{\kappa} + \lambda_n\right), \ \frac{\lambda_n}{\kappa}\right\}$ and $r^u_2 \triangleq \frac{8\lambda_n}{\alpha\kappa}$. Suppose that 1) $r_1 \leq r^u_1$, 2) $r_2 \leq r^u_2$, and 3) $r^u_1 \leq \min\{\frac{1}{4C'}, \frac{\alpha}{32\max\{1+\frac{\kappa}{2}, \frac{\alpha}{8}\}^2 m\psi C'^2}\}$, then $\max\{\|\Delta_2\|_2, \frac{1}{\gamma}\|\Delta_3\|_2\} \leq 2r^u_1$ and $\max\{\|\Delta_1\|_2, \|\Delta_4\|_2\} \leq r^u_2$. Consequently, $\Phi_\gamma(\Delta) \leq 2r^u_1$.

In the following proposition, we prove algebraic correctness of program (4.2). The statement of this proposition requires us to define some constants. Let $C'_1 = \frac{2m}{\alpha\kappa}\left(28(1 + \frac{4}{\alpha})\psi^2 + 2\kappa + 13\right)$, $C'_2 = \frac{4}{\alpha}(\frac{1}{2\kappa} + 1)$, $C'_{\sigma_y} = C'^2_1\psi^2 \max\{2\kappa+1, \frac{2}{C'_2\psi^2}+1\}$, $C'_{\sigma_{yx}} = C'^2_1\psi^2 \max\{2\kappa + \frac{\kappa}{\gamma}, \frac{2}{C'_2\psi^2} + \frac{\kappa}{\gamma}\}$, and $C'_{samp} = \max\{\frac{1}{8m\psi\kappa}, \frac{\alpha}{16C'(\frac{2}{\kappa}+1)}, \frac{\alpha^2}{128(\frac{2}{\kappa}+1)\max\{1+\frac{\kappa}{2}, \frac{\alpha}{8}\}^2 m\psi^2 C'^2}, \frac{1}{4C'_1C'}\}$

Proposition 4.3. *Suppose that $\sigma_y \geq \frac{m}{\omega}C'_{\sigma_y}\lambda_n$, $\sigma_{yx} \geq m\gamma^2 C'_{\sigma_{yx}}\lambda_n$. Further, suppose that λ_n is chosen so that $\lambda_n \leq \frac{1}{C'_{samp}}$. Then, there exists tangent space $T'_y \subset \mathbb{S}^p$ in the rank-k_u variety ($k_u = rank(L^\star_y)$) and tangent space $T'_{yx} \subset \mathbb{R}^{p\times q}$ in rank k_x-variety ($k_x = rank(\Theta^\star_{yx})$) where $\rho(T'_y, T(L^\star_y)) \leq \omega_y$, $\rho(T'_{yx}, T(\Theta^\star_{yx})) \leq \omega_{yx}$ such that the corresponding solution $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ of (4.2) satisfies the following properties:*

1. *$rank(\tilde{L}_y) = rank(L^\star_y)$ and $rank(\tilde{\Theta}_{yx}) = rank(\Theta^\star_{yx})$*

2. *Letting $C_T = (0, \ \mathcal{P}_{T'^\perp_y}(L^\star_y), \ \mathcal{P}_{T'^\perp_{yx}}(\Theta^\star_{yx}), \ 0)$, we have that $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F}(C_T)] \leq \frac{\lambda_n}{\kappa}$ and $\Phi_\gamma[C_T] \leq \frac{4}{\alpha}(1 + \frac{2}{\kappa})\lambda_n$*

3. *$\Phi_\gamma[\Delta] \leq 2C'_1\lambda_n$*

4. *$\tilde{L}_y \succeq 0$*

Furthermore, suppose that $\Phi_\gamma(\mathcal{I}^\dagger E_n) \leq \frac{\lambda_n}{\kappa}$ and $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \leq \frac{\lambda_n}{\kappa}$. Then the tangent space constraint $(D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}'$ in (4.2) is inactive, so that $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ is the unique solution of the original convex program (1.4).

Thus far, the analysis of the convex program so has been deterministic in nature. In the following proposition, we present the probabilistic component of our analysis by showing the rate at which the sample covariance matrix Σ_n converges to $\Sigma^\star$ in spectral norm. This result is well-known and is a specialization of a result proven by [3].

Proposition 4.4. *Suppose that the number of observed samples obeys*
$n \geq 64\kappa^2 m^2 \psi^2 C_{samp}'^2 (p+q)$, *and the regularization parameter* λ_n *is chosen so that:*
$\lambda_n \in [8\psi\kappa m \sqrt{\frac{p+q}{n}}, \frac{1}{C_{samp}'}]$. *Then, with probability greater than* $1 - 2\exp\left\{ -\frac{n\lambda_n^2}{128\kappa^2 m^2 \psi^2} \right\}$, $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{\kappa}$.

4.2 Proof of Theorem 1

We first relate the constants $\tilde{C}_{samp}$, $\tilde{C}$, $\tilde{C}_0$, $\tilde{C}_1$, and $\tilde{C}_\sigma$ of Theorem 1 to the constants C_{samp}', C_1' C_{σ_y}', and $C_{\sigma_{yx}}'$. In particular, using the properties that $\beta \geq 2$ and $\frac{\psi^2}{\alpha} \geq \frac{1}{2}$ and $\bar{m}, m \geq 1$, one can check that: $\tilde{C}_{samp}^2 \geq \frac{64\psi^2\kappa^2\alpha^2}{\beta^4 m^6} C_{samp}'^2$, $\tilde{C}_0 \geq \frac{1}{\beta m} C_{samp}'$, $\tilde{C}_\sigma \geq \frac{\alpha^5}{\beta m^3} C_{\sigma_y}'$, $\tilde{C}_\sigma \geq \frac{\alpha^5}{\beta m^4} C_{\sigma_{yx}}'$, and $\tilde{C}_1 \geq \frac{\alpha^2}{m} C_1'$. Furthermore, we have that $\tilde{C} \geq \frac{\alpha}{\beta m} 8\psi\kappa$. Using these relations, one can also check that the assumptions of Theorem 1 imply that the assumptions of Proposition 4.3 and Proposition 4 are satisfied. Thus we can conclude that the optimal solution $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ of (4.2) (with a particular choice of tangent spaces T_y' and T_{yx}') satisfy results of Proposition 4.3. Further, by appealing to Proposition 4.4, we have that $\Phi_\gamma(\mathcal{F}^\dagger E_n) \leq \frac{\lambda_n}{\kappa}$. If we show that $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\Delta)] \leq \frac{\lambda_n}{\kappa}$, then we conclude that the unique optimum $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ of the original convex program (1.4) coincides with the optimum $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ of the convex program (4.2). Thus, we conclude that the estimates of (1.4) have structurally correct structure (i.e. $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$) and have their error bounded by $\Phi_\gamma(\Delta) \leq 2C_1'\lambda_n$. To show that $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\Delta)] \leq \frac{\lambda_n}{\kappa}$, we note that

$$\frac{4}{\alpha}\left(\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} C_T] + \lambda_n\right) \leq \frac{4}{\alpha}\left(\frac{\lambda_n}{\kappa} + \frac{\lambda_n}{\kappa} + \lambda_n\right) \leq \frac{4\lambda_n}{\alpha}\left(\frac{2}{\kappa} + 1\right)$$
$$\leq \min\left\{\frac{1}{4C'}, \frac{\alpha}{32\max\{1 + \frac{\kappa}{2}, \frac{\alpha}{8}\}^2 m\psi C'^2}\right\}$$

Here, we used the bound on $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} C_T]$ provided by Proposition 4.3 and the bound on λ_n. Furthermore, appealing to Proposition 4.3 once again, we have $\Phi_\gamma[C_T] \leq \frac{4}{\alpha}(1+\frac{2}{\kappa})\lambda_n \leq \min\{\frac{1}{4C'}, \frac{\alpha}{16m\psi C'^2}\}$. Thus Proposition 4.2 provides us with the bound $\Phi_\gamma[\Delta] \leq 2C_1'\lambda_n \leq \frac{1}{2C'}$. We subsequently apply the results of Proposition 4.1 to obtain:

$$\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \leq 2m\psi C'^2 \Phi_{\delta,\gamma}[\Delta]^2 \leq \left[2m\psi C'^2 C_1'^2 \lambda_n\right]\lambda_n \leq \frac{\lambda_n}{\kappa}$$

The last inequality follows from the bound on λ_n.

5 Discussion

In this paper we describe a new approach for interpreting the latent variables in a factor model. Our method proceeds by obtaining observations of auxiliary covariates that may plausibly be related to the observed phenomena, and then suitably associating these auxiliary covariates to the latent variables. The procedure involves the solutions of computationally tractable convex optimization problems, which are log-determinant semidefinite programs that can be solved efficiently. We give both theoretical as well as experimental evidence in support of our methodology. Our technique generalizes transparently to other families beyond factor models such as latent-variable graphical models [2], although we do not pursue these extensions in the present article.

References

[1] BICKEL, P. & DOKSUM, K. Mathematical Statistics, Basic Ideas and Selected Topics. *Prentice-Hall*, 2007.

[2] BÜHLMANN, P & VAN DE GEER, S.. *Statistics for high-dimensional data.* New York: Springer, 2011.

[3] CANDÈS, E. J. & RECHT, B. Exact matrix completion via convex optimization. *Foundation of Computational Mathematics*, 9:717–772, 2009.

[4] CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. & WILLSKY, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal of Optimization*, 21:572–596, 2011.

[5] CHANDRASEKARAN, V., PARRILO, P. A. & WILLSKY, A. S. Latent Variable Graphical Model Selection via Convex Optimization. *Annals of Statistics*, 40:1935–1967, 2012.

[6] DAVIDSON, K.R. & SZAREK, S.J. Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces*, 1:317–366, 2001.

[7] FAZEL, M. Matrix rank minimization with applications. *PhD thesis, Department of Electrical Engineering, Stanford University*, 2002.

[8] HOTELLING, H. Relations between two sets of variants. *Biometrika*, 28:321-377, 2002.

[9] KAY, S.M. Modern spectral estimation, theory and application. *Prentice-Hall*, 1988.

[10] LEDERMANN, W. On a problem concerning matrices with variable diagonal elements. *Proceeding of Royal Society Edinburgh*, 60:1-17, 1940.

[11] MEINHAUSEN, N. & B´ULMANN, AP. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436-1462, 2010.

[12] MESBAHI, M. & PAPAVASSILOPOULOS, G. On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Transactions on Automatic Control*, 42:239-243, 1997.

[13] NATARAJAN, B.K. Sparse approximate solutions to linear systems. *SIAM Journal of Computing*, 24:227-234, 1995.

[14] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. & YU, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.

[15] RECHT, B., FAZEL, M. & PARRILO, P. A. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review*, 52:471–501, 2010.

[16] SAUNDERSON, J., CHANDRASEKARAN, V., PARRILLO, P & WILLSKY, S. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis*, 33:1395-1416, 2012.

[17] SHAPIRO, A. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, 47:187-199, 1982.

[18] SHAPIRO, A. Weighted minimum trace factor analysis. *Psychometrika*, 47:243-264, 1982.

[19] SHAPIRO, A. Identifiability of factor analysis: Some results and open problems. *Linear Algebra Applications*, 15:201-292, 1904.

[20] SPEARMAN, C. 'General inteligence', objectively determined and measured. *American Journal of Psychology*, 15:201-292, 1904

[21] TOH, K. C, TODD, M. J. & TUTUNCU, R. H.. *SDPT3 - a MATLAB software package for semidefinite-quadratic-linear programming.* Available from http://www.math.nus.edu.sg/∼mattohkc/sdpt3.html.

[22] WAINWRIGHT, M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[23] WAINWRIGHT, M. J. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and its Applications.* 1:233–253, 2014.

[24] ZHAO, P & YU, B On model selection consistency of the lasso. *Journal of Machine Learnign Research.* 7:2541–2567, 2006.

6 Appendix

6.1 *Proof of Proposition 1 (main paper)*

Proof. We note that:

$$\|\Delta\|_2 \quad \leq \quad \|\Delta D_y\|_2 + \|\Delta L_y\|_2 + \|\Delta\Theta_{yx}\|_2 + \|\Delta\Theta_x\|_2 \leq (3+\gamma)\Phi_\gamma(\Delta)$$

Furthermore, recall that

$$R_{\Sigma^\star}(\mathcal{F}(\Delta)) = \Sigma^\star\Big[\sum_{k=2}^\infty (-\mathcal{F}(\Delta)\Sigma^{\star-1})^k\Big].$$

Using this observation and some algebra, we have that:

$$\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \leq m\psi\Big[\sum_{k=2}^\infty (\psi\|\Delta\|_2)^k\Big] \quad \leq \quad m\psi^3\frac{(3+\gamma)^2\Phi_\gamma[\Delta]^2}{1-(3+\gamma)\Phi_\gamma[\Delta]\psi}$$

$$\leq \quad 2m\psi C'^2\Phi_\gamma[\Delta]^2$$

$\square$

6.2 *Proof of Proposition 2 (main paper)*

Proof. The proof of this result uses Brouwer's fixed-point theorem, and is inspired by the proof of a similar result in [5, 2]. The optimality conditions of (4.2) suggest that there exist Lagrange multipliers $Q_{D_y} \in \mathcal{W}$, $Q_{T_y} \in T_y'^\perp$, and $Q_{T_{yx}} \in T_{yx}'^\perp$ such that

$$[\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{D_y} = 0; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{T_y} \quad \in \quad \lambda_n\partial\|\tilde{L}_y\|_\star$$

$$[\Sigma_n - \tilde{\Theta}^{-1}]_{yx} + Q_{T_{yx}} \in -\lambda_n\gamma\partial\|\tilde{\Theta}_{yx}\|_\star; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_x \quad = \quad 0$$

Letting the SVD of $\tilde{L}$ and $\tilde{\Theta}_{yx}$ be given by $\tilde{L}_y = \bar{U}\bar{D}\bar{V}'$ and $\tilde{\Theta}_{yx} = \breve{U}\breve{D}\breve{V}'$ respectively, and $Z \triangleq (0, \ \lambda_n\bar{U}\bar{V}', \ -\lambda_n\gamma\breve{U}\breve{V}', \ 0)$, we can restrict the optimality conditions of (4.2)(main paper) to the space $\mathbb{H}'$ to obtain, $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger(\Sigma_n - \tilde{\Theta}^{-1}) = Z$. Further, by appealing to the matrix inversion lemma, this condition can be restated as $\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger(E_n - R_{\Sigma^\star}(\mathcal{F}\Delta) + \mathbb{I}^\star\mathcal{F}(\Delta)) = Z$. Based on the Fisher information assumption 1 in (2.7) (main paper), the optimum of (4.2)(main paper) is unique (this is because the Hessian of the negative log-likelihood term is positive definite restricted to the tangent space constraints). Moreover, using standard Lagrangian duality, one can show that the set of variables $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ that satisfy the restricted optimality conditions are unique. Consider the following function $S(\underline{\delta})$ restricted to $\underline{\delta} \in \mathcal{W} \times T_y' \times T_{yx}' \times \mathbb{S}^q$ with $\rho(T(L_y^\star), T_y') \leq \omega_y$ and $\rho(T(\Theta_{yx}^\star), T_{yx}') \leq \omega_{yx}$:

$$S(\underline{\delta}) = \underline{\delta} - (\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'})^{-1}\Big(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n \quad - \quad R_{\Sigma^\star}\mathcal{F}(\underline{\delta} + C_T)$$

$$+\mathbb{I}^\star\mathcal{F}(\underline{\delta} + C_T)] - Z\Big)$$

The function $S(\underline{\delta})$ is well-defined since the operator $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'}$ is bijective due to Fisher information assumption 1 in (2.7) (main paper). As a result, $\underline{\delta}$ is a fixed point of $S(\underline{\delta})$ if and only if $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta} + C_T)) + \mathbb{I}^\star\mathcal{F}(\underline{\delta} + C_T)] = Z$. Since the pair $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ are the unique solution to (4.2)(main paper), the only fixed point of S is $\mathcal{P}_{\mathbb{H}'}[\Delta]$. Next we show that this unique optimum

lives inside the ball $\mathbb{B}_{r_1^u, r_2^u} = \{\underline{\delta} \mid \max\{\|\delta_2\|_2, \frac{1}{\gamma}\|\delta_3\|_2\} \le r_1^u, \max\{\|\delta_1\|_2, \|\delta_4\|_2\} \le r_2^u \, \underline{\delta} \in \mathbb{H}'\}$. In particular, we show that under the map S, the image of $\mathbb{B}_{r_1^u, r_2^u}$ lies in $\mathbb{B}_{r_1^u, r_2^u}$ and appeal to Brouwer's fixed point theorem to conclude that $\mathcal{P}_{\mathbb{H}'}[\Delta] \in \mathbb{B}_{r_1^u, r_2^u}$. For $\underline{\delta} \in \mathbb{B}_{r_1^u, r_2^u}$, the first component of $S(\underline{\delta})$, denoted by $S(\underline{\delta})_1$, can be bounded as follows:

$$
\begin{aligned}
\|S(\underline{\delta})_1\|_2 &= \left\| \left[(\mathcal{P}_{\mathbb{H}'} \mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} \mathcal{P}_{\mathbb{H}'})^{-1} \left(\mathcal{P}_{\mathbb{H}'} \mathcal{F}^\dagger [E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta} + C_T)) \right. \right. \right. \\
&\quad + \left. \left. \left. \mathbb{I}^\star \mathcal{F} C_T] + Z \right) \right]_1 \right\|_2 \le \frac{2}{\alpha} \left[\Phi_\gamma[\mathcal{F}^\dagger(E_n + \mathbb{I}^\star \mathcal{F}(C_T))] \right] \\
&\quad + \frac{2}{\alpha} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \le \frac{r_2^u}{2} + \frac{2}{\alpha} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)]
\end{aligned}
$$

The first inequality holds because of Fisher Information Assumption 1 in (2.7) (main paper), and the properties that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(.)] \le 2\Phi_\gamma(.)$ (since projecting into the tangent space of a low-rank matrix variety increases the spectral norm by a factor of at most two) and $\Phi_\gamma(Z) = \lambda_n$. Moreover, since $r_1^u \le \frac{1}{4C'}$, we have $\Phi_\gamma(\underline{\delta} + C_T) \le \Phi_\gamma(\underline{\delta}) + \Phi_\gamma(C_T) \le 2r_1^u \le \frac{1}{2C'}$. Moreover, $r_1^u \le r_2^u \max\{1 + \frac{\kappa}{2}, \frac{\alpha}{8}\}$. We can now appeal to Proposition 1 (main paper) to obtain:

$$
\begin{aligned}
\frac{2}{\alpha} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] &\le \frac{4}{\alpha} m\psi C'^2 [\Phi_\gamma(\underline{\delta} + C_T)]^2 \\
&\le \frac{16}{\alpha} m\psi C'^2 (r_2^u)^2 \max\{1 + \frac{\kappa}{2}, \frac{\alpha}{8}\}^2 \\
&\le \frac{r_2^u}{2}
\end{aligned}
$$

Thus, we conclude that $\|S(\delta)_1\|_2 \le r_2^u$. Similarly, we check that:

$$
\begin{aligned}
\|[S(\underline{\delta})_2]\|_2 &= \left\| \left[(\mathcal{P}_{\mathbb{H}'} \mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} \mathcal{P}_{\mathbb{H}'})^{-1} \left(\mathcal{P}_{\mathbb{H}'} \mathcal{F}^\dagger [E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta} + C_T)) \right. \right. \right. \\
&\quad + \left. \left. \left. \mathbb{I}^\star \mathcal{F} C_T] + Z \right) \right]_2 \right\|_2 \le \frac{2}{\alpha} \left[\Phi_\gamma[\mathcal{F}^\dagger(E_n + \mathbb{I}^\star \mathcal{F}(C_T)] + \lambda_n] \right] \\
&\quad + \frac{2}{\alpha} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \le \frac{r_1^u}{2} + \frac{2}{\alpha} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \le r_1^u
\end{aligned}
$$

Using a similar approach, we can conclude that $\frac{1}{\gamma}\|S(\delta)_3\|_2 \le r_1^u$ and $\|S(\delta)_3\|_2 \le r_2^u$. Therefore, Brouwer's fixed point theorem suggests that $\mathcal{P}_{\mathbb{H}'}(\Delta) \in \mathcal{B}_{r_1^u, r_2^u}$. Hence, $\|\Delta_1\|_2 \le r_2^u$, $\|\Delta_4\|_2 \le r_2^u$, $\|\Delta_2\|_2 \le \|\mathcal{P}_{\mathbb{H}'[2]}(\Delta_2)\|_2 + \|\mathcal{P}_{\mathbb{H}'[2]^\perp}(\Delta_2)\|_2 \le 2r_1^u$, and $\frac{1}{\gamma}\|\Delta_3\|_2 \le \frac{1}{\gamma}\|\mathcal{P}_{\mathbb{H}'[3]}(\Delta_3)\|_2 + \frac{1}{\gamma}\|\mathcal{P}_{\mathbb{H}'[3]^\perp}(\Delta_2)\|_2 \le 2r_1^u$. $\qquad\square$

6.3 Proof of Proposition 3 (main paper)

Below, we outline our proof strategy:

1. We proceed by analyzing (4.1) with additional constraints that the variables L_y, and Θ_{yx} belong to the algebraic varieties low-rank matrices (specified by rank of $L_y^\star$, and $\Theta_{yx}^\star$), and that the tangent spaces $T(L_y)$, $T(\Theta_{yx})$ are close to the nominal tangent spaces $T(L_y^\star)$, and $T(\Theta_{yx}^\star)$ respectively. We prove that under suitable conditions on the minimum nonzero singular value of $L_y^\star$, and minimum nonzero singular value of $\Theta_{yx}^\star$, any optimum pair of variables (Θ, D_y, L_y) of this non-convex program are smooth points of the underlying varieties; that is $\text{rank}(L_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\Theta_{yx}) = \text{rank}(\Theta_{yx}^\star)$. Further, we show that L_y has the same inertia as $L_y^\star$ so that $L_y \succeq 0$.

2. Conclusions of the previous step imply the the variety constraints can be "linearized" at the optimum of the non-convex program to obtain tangent-space constraints. Under the specified conditions on the regularization parameter λ_n, we prove that with high probability, the unique optimum of this "linearized" program coincides with the global optimum of the non-convex program.

3. Finally, we show that the tangent-space constraints of the linearized program are inactive at the optimum. Therefore the optimal solution of (4.1) has the property that with high probability: $\text{rank}(\bar{L}_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\bar{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$. Since $\bar{L}_y \succeq 0$, we conclude that the variables $(\bar{\Theta}, \bar{D}_y, \bar{L}_y)$ are the unique optimum of (1.4).

6.3.1 Variety Constrained Program

We begin by considering a variety-constrained optimization program. Letting $(M, N, P, Q) \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$, we denote $\mathcal{P}_{[2,3]}(M, N, P, Q) = (N, P) \subset \mathbb{S}^p \times \mathbb{R}^{p \times q}$. The variety-constrained optimization program is given by:

$$(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}}) = \underset{\substack{\Theta \in \mathbb{S}^{q+p},\ \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\text{argmin}} \quad -\ell(\Theta; \mathcal{D}_n^+) + \lambda_n[\|L_y\|_\star + \gamma\|\Theta_{yx}\|_\star]$$

$$\text{s.t.} \quad \Theta_y = D_y - L_y, (\Theta, D_y, L_y) \in \mathcal{M}. \tag{6.1}$$

Here, the set $\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2$, where the sets $\mathcal{M}_1$ and $\mathcal{M}_2$ are given by:

$$\mathcal{M}_1 \triangleq \left\{ (\Theta, D_y, L_y) \in \mathbb{S}^{(p+q)} \times \mathbb{S}^p \times \mathbb{S}^p \Big| D_y \text{ is diagonal, } \text{rank}(L_y) \leq \text{rank}(L_y^\star) \right.$$

$$\text{rank} \quad (\Theta_{yx}) \leq \text{rank}(\Theta_{yx}^\star); \|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 \leq \frac{\lambda_n}{2\psi^2}$$

$$\left. \|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \leq \frac{\lambda_n}{2\psi^2} \right\}$$

$$\mathcal{M}_2 \triangleq \left\{ (\Theta, D_y, L_y) \in \mathbb{S}^{(p+q)} \times \mathbb{S}^p \times \mathbb{S}^p \right|$$

$$\Gamma_\gamma[\mathcal{P}_{[2,3]}(\mathbb{I}^\star \mathcal{F}(\Delta))] \leq \frac{\lambda_n}{\kappa}\left(20m(1 + \frac{4}{\alpha})\psi^2 + \kappa + 13\right)$$

$$\left. \|\mathbb{I}^\star \mathcal{F}(\Delta_1, 0, 0, \Delta_4)\|_2 \leq \frac{8\lambda_n}{\kappa}\left(1 + \frac{4}{\alpha}\right)\psi^2 \right\},$$

The optimization program (6.1) is non-convex due to the rank constraints $\text{rank}(L_y) \leq \text{rank}(L_y^\star)$ and $\text{rank}(\Theta_{yx}) \leq \text{rank}(\Theta_{yx}^\star)$ in the set $\mathcal{M}$. These constraints ensure that the matrices L_y, and Θ_{yx} belong to appropriate varieties. The constraints in $\mathcal{M}$ along $T(L_y^\star)^\perp$ and $T(\Theta_{yx}^\star)^\perp$ ensure that the tangent spaces $T(L_y)$ and $T(\Theta_{yx})$ are "close" to $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$ respectively. Finally, the last conditions roughly controls the error. We begin by proving the following useful proposition:

Proposition 6.1. *Let (Θ, D_y, L_y) be a set of feasible variables of (6.1). Let $\Delta = (D_y - D_y^\star, L_y - L_y^\star, \Theta_{yx} - \Theta_{yx}^\star, \Theta_x - \Theta_x^\star)$ and recall that $C_1' = \frac{2m}{\alpha\kappa}\left(28(1 + \frac{4}{\alpha})\psi^2 + 2\kappa + 13\right) + \frac{1}{\psi^2}$. Then, $\Phi_\gamma[\Delta] \leq C_1'\lambda_n$*

Proof. Let $\mathbb{H}^\star = \mathcal{W} \times T(L_y^\star) \times T(\Theta_{yx}^\star) \times \mathbb{S}^q$. Then,

$$
\begin{aligned}
\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} \mathcal{P}_{\mathbb{H}^\star}(\Delta)] &\leq \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F}(\Delta)] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} \mathcal{P}_{\mathbb{H}^{\star\perp}}(\Delta)] \\
&\leq \frac{8m\lambda_n}{\kappa}\Big(1 + \frac{4}{\alpha}\Big)\psi^2 + \frac{\lambda_n}{\kappa}\Big(20m(1+\frac{4}{\alpha})\psi^2 + \kappa + 13\Big) \\
&\quad + m\psi^2\Big(\frac{\omega_y \lambda_n}{2\psi^2} + \frac{\omega_{yx}\lambda_n}{2\psi^2}\Big) \\
&\leq \frac{m\lambda_n}{\kappa}\Big(28(1+\frac{4}{\alpha})\psi^2 + 2\kappa + 13\Big)
\end{aligned}
$$

Since $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\cdot)] \leq 2\Phi_\gamma(\cdot)$, we have that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \leq \frac{2m\lambda_n}{\kappa}\Big(28(1+\frac{4}{\alpha})\psi^2 + 2\kappa + 13\Big)$. Consequently, we apply Fisher Information Assumption 1 in (2.7) (main paper) to conclude that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \leq \frac{2m\lambda_n}{\alpha\kappa}\Big(28(1+\frac{4}{\alpha})\psi^2 + 2\kappa + 13\Big)$. Moreover:

$$
\begin{aligned}
\Phi_\gamma[\Delta] \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\Delta)] + \Phi_\gamma[\mathcal{P}_{\mathbb{H}^{\star\perp}}(\Delta)] &\leq \frac{2m\lambda_n}{\alpha\kappa}\Big(28(1+\frac{4}{\alpha})\psi^2 + 2\kappa + 13\Big) + \frac{\lambda_n}{\psi^2} \\
&= C_1'\lambda_n
\end{aligned}
$$

$\square$

Proposition 6.1 leads to powerful implications. In particular, under additional conditions on the minimum nonzero singular values of $L_y^\star$ and $\Theta_{yx}^\star$, any feasible set of variables (Θ, D_y, L_y) of (6.1) has two key properties: (*a*) The variables (Θ_{yx}, L_y) are smooth points of the underlying varieties, (*b*) The constraints in $\mathcal{M}$ along $T(L_y^\star)^\perp$ and $T(\Theta_{yx}^\star)^\perp$ are locally inactive at Θ_{yx} and L_y. These properties, among others, are proved in the following corollary.

Corollary 6.2. *Consider any feasible variables (Θ, D_y, L_y) of (6.1). Let σ_y be the smallest nonzero singular value of $L_y^\star$ and σ_{yx} be the smallest nonzero singular value of $\Theta_{yx}^\star$. Let $\mathbb{H}' = \mathcal{W} \times T(L_y) \times T(\Theta_{yx}) \times \mathbb{S}^q$ and $C_{T'} = \mathcal{P}_{\mathbb{H}'^\perp}(0, L_y^\star, \Theta_{yx}^\star, 0)$. Furthermore, recall that $C_1' = \frac{2m}{\alpha\kappa}\Big(28(1+\frac{4}{\alpha})\psi^2 + 2\kappa + 13\Big) + \frac{1}{\psi^2}$, $C_2' = \frac{4}{\alpha}(1+\frac{2}{\kappa})$, $C_{\sigma_y}' = C_1'^2\psi^2\max\{2\kappa+1, \frac{2}{C_2'\psi^2}+1\}$ and $C_{\sigma_{yx}}' = C_1'^2\psi^2\max\{2\kappa+\frac{\kappa}{\gamma}, \frac{2}{C_2'\psi^2}+\frac{\kappa}{\gamma}\}$. Suppose that the following inequalities are met: $\sigma_y \geq \frac{m}{\omega_y}C_{\sigma_y}'\lambda_n$,*
$\sigma_{yx} \geq \frac{m\gamma^2}{\omega_{yx}}C_{\sigma_{yx}}'\lambda_n$. *Then,*

1. L_y *and* Θ_{yx} *are smooth points of their underlying varieties, i.e.* $\mathrm{rank}(L_y) = \mathrm{rank}(L_y^\star)$, $\mathrm{rank}(\Theta_{yx}) = \mathrm{rank}(\Theta_{yx}^\star)$; *Moreover* L_y *has the same inertia as* $L_y^\star$.

2. $\|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 \leq \frac{\lambda_n\omega_y}{48m\psi^2}$ *and* $\|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \leq \frac{\lambda_n\omega_{yx}}{48m\psi^2}$

3. $\rho(T(L_y), T(L_y^\star)) \leq \omega_y$; $\rho(T(\Theta_{yx}), T(\Theta_{yx}^\star)) \leq \omega_{yx}$; *that is, the tangent spaces at* L_y *and* Θ_{yx} *is "close" to the tangent space* $L_y^\star$ *and* $\Theta_{yx}^\star$.

4. $\Phi_\gamma[C_{T'}] \leq \min\{\frac{\lambda_n}{\kappa\psi^2}, C_2'\lambda_n\}$

Proof. We note the following relations before proving each step: $C_1' \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2}$, $\omega_y, \omega_{yx} \in (0,1)$, and $\kappa \geq 6$. We also appeal to the results of regarding perturbation analysis of the low-rank matrix variety [1].

1. Based on the assumptions regarding the minimum nonzero singular values of $L_y^\star$ and $\Theta_{yx}^\star$, one can check that:

$$\sigma_y \;\geq\; \frac{C_1'^2\lambda_n}{\omega_y}m\psi^2(\kappa+1) \geq \frac{C_1'\lambda_n}{\omega_y}(2\kappa+1) \geq 8\|L - L_y^\star\|_2$$

$$\sigma_{yx} \;\geq\; \frac{C_1'^2\lambda_n}{\omega_{yx}}\gamma^2 m\psi^2\Big(\frac{6\beta}{\gamma}+2\kappa\Big) \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2$$

Combining these results and Proposition 6.1, we conclude that L_y and Θ_{yx} are smooth points of their respective varieties, i.e. $\text{rank}(L_y) = \text{rank}(L_y^\star)$, and $\text{rank}(\Theta_{yx}) = \text{rank}(\Theta_{yx}^\star)$. Furthermore, L_y has the same inertia as $L_y^\star$.

2. Since $\sigma_y \geq 8\|L_y - L_y^\star\|_2$, and $\sigma_{yx} \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2$, we can appeal to Proposition 2.2 of [2] to conclude that the constraints in $\mathcal{M}$ along $\mathcal{P}_{T(L_y^\star)^\perp}$ and $\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}$ are strictly feasible:

$$\|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 \;\leq\; \frac{\|L_y - L_y^\star\|_2^2}{\sigma_y} \leq \frac{\lambda_n}{48m\psi^2}$$

$$\|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \;\leq\; \frac{\|\Theta_{yx} - \Theta_{yx}^\star\|_2^2}{\sigma_{yx}} \leq \frac{\lambda_n}{48m\psi^2}$$

3. Appealing to Proposition 2.1 of [2], we prove that the tangent spaces $T(L_y)$ and $T(\Theta_{yx})$ are close to $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$ respectively:

$$\rho(T(L_y), T(L_y^\star)) \;\leq\; \frac{2\|L_y - L_y^\star\|_2}{\sigma_y} \leq \frac{2C_1'\lambda_n\omega}{C_1'^2\lambda_n m\psi^2(2\kappa+1)} \leq \omega_y$$

$$\rho(T(\Theta_{yx}), T(\Theta_{yx}^\star)) \;\leq\; \frac{2\|\Theta_{yx} - \Theta_{yx}^\star\|_2}{\sigma_{yx}} \leq \frac{2C_1'\lambda_n\gamma\omega_{yx}}{\frac{C_1'^2\lambda_n}{\omega_{yx}}\gamma^2 m\psi^2\Big(\frac{\kappa}{\gamma}+2\kappa\Big)} \leq \omega_{yx}$$

4. Letting σ_y' and σ_{yx}' be the minimum nonzero singular value of L_y and Θ_{yx} respectively, one can check that:

$$\sigma_y' \geq \sigma_y - \|L_y - L_y^\star\|_2 \geq 8C_1'\lambda_n \geq 8\|L_y - L_y^\star\|_2$$

$$\sigma_{yx}' \geq \sigma_{yx} - \|\Theta_{yx} - \Theta_{yx}^\star\|_2 \geq 8C_1'\lambda_n\gamma \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2$$

Once again appealing to Proposition 2.2 of [2] and simple algebra, we have:

$$\Phi_\gamma(C_{T'}) \;\leq\; m\|\mathcal{P}_{T(L_y)^\perp}(L_y - L_y^\star)\|_2 + m\|\mathcal{P}_{T(\Theta_{yx})^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2$$

$$\leq\; m\frac{\|L_y - L_y^\star\|_2^2}{\sigma_y'} + m\frac{\|\Theta_{yx} - \Theta_{yx}^\star\|_2^2}{\sigma_{yx}'} \leq \min\Big\{\frac{\lambda_n}{\kappa\psi^2}, C_2'\lambda_n\Big\}$$

$\square$

6.3.2 Variety Constrained Program to Tangent Space Constrained Program

Consider any optimal solution $(\Theta^\mathcal{M}, D_y^\mathcal{M}, L_y^\mathcal{M})$ of (6.1). In Corollary 6.2, we concluded that the variables $(\Theta_{yx}^\mathcal{M}, L_y^\mathcal{M})$ are smooth points of their respective varieties. As a result, the rank constraints $\text{rank}(L_y) \leq \text{rank}(L_y^\star)$ and $\text{rank}(\Theta_{yx}) \leq \text{rank}(\Theta_{yx}^\star)$ can be "linearized" to $L_y \in T(L_y^\mathcal{M})$ and $\Theta_{yx} \in T(\Theta_{yx}^\mathcal{M})$ respectively. Since all the remaining constraints are convex, the optimum of this linearized

program is also the optimum of (6.1). Moreover, we once more appeal to Corollary 6.2 to conclude that the constraints in $\mathcal{M}$ along $\mathcal{P}_{T(L_y^\star)^\perp}$ and $\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}$ are strictly feasible at $(\Theta^\mathcal{M}, D_y^\mathcal{M}, L_y^\mathcal{M})$. As a result, these constraints are locally inactive and can be removed without changing the optimum. Therefore the constraint $(\Theta^\mathcal{M}, D_y^\mathcal{M}, L_y^\mathcal{M}) \in \mathcal{M}_1$ is inactive and can be removed. We now argue that the constraint $(\Theta^\mathcal{M}, D_y^\mathcal{M}, L_y^\mathcal{M}) \in \mathcal{M}_2$ in (6.1) can also removed in this "linearized" convex program. In particular, letting $\mathbb{H}_\mathcal{M} \triangleq \mathcal{W} \times T(L_y^\mathcal{M}) \times T(\Theta_{yx}^\mathcal{M}) \times \mathbb{S}^q$, consider the following convex optimization program:

$$(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y) = \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \; \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\operatorname{argmin}} \quad -\ell(\Theta; \mathcal{D}_n^+) + \lambda_n[\|L_y\|_\star + \gamma\|\Theta_{yx}\|_\star]$$

$$\text{s.t.} \quad \Theta_y = D_y - L_y, \; (D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}_\mathcal{M} \tag{6.2}$$

We prove that under conditions imposed on the regularization parameter λ_n, the pair of variables $(\Theta^\mathcal{M}, D_y^\mathcal{M}, L_y^\mathcal{M})$ is the unique optimum of (6.2). That is, we show that

1. $\Gamma_\gamma[\mathcal{P}_{[2,3]}(\mathbb{I}^\star\mathcal{F}(\Delta))] < \frac{2\lambda_n}{\kappa}\left(20m(1+\frac{4}{\alpha})\psi^2 + \kappa + 13\right)$

2. $\|\mathbb{I}^\star\mathcal{F}(D_y^\mathcal{M} - D_y^\star, 0, 0, \Theta_x^\mathcal{M} - \Theta_x^\star)\|_2 < \frac{8\lambda_n}{\kappa}\left(1 + \frac{4}{\alpha}\right)\psi^2$

Appealing to Corollary 6.2 and Proposition 4.4, we have that $\Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \leq \frac{\lambda_n}{\kappa}$, $\Phi_\gamma[C_{T_\mathcal{M}}] \leq C_2'\lambda_n$ and (with high probability) $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{\kappa}$. Consequently, based on the bound on λ_n in assumption of Theorem 4.3, it is straightforward to show that $r_1^u \leq \min\left\{\frac{1}{4C'}, \frac{\alpha}{32\max\{1+\frac{\kappa}{2}, \frac{\alpha}{8}\}^2 m\psi C''^2}\right\}$. We further have that $r_2^u \leq \frac{2\lambda_n}{\kappa}\left(1 + \frac{4}{\alpha}\right)$. Hence by Proposition 2 (main paper), $\|D_y^\mathcal{M} - D_y^\star\|_2 \leq r_2^u$ and $\|\Theta_x^\mathcal{M} - \Theta_x^\star\|_2 \leq r_2^u$. Thus, $\|\mathbb{I}^\star\mathcal{F}(D_y^\mathcal{M} - D_y^\star, 0, 0, \Theta_x^\mathcal{M} - \Theta_x^\star)\|_2 \leq 2\psi^2 r_2^u < \frac{8\lambda_n}{\kappa}\left(1 + \frac{4}{\alpha}\right)\psi^2$. From Proposition 2 (main paper), we also have that $\Phi_\gamma[\Delta] \leq \frac{1}{2C'}$. Finally, we can appeal to Proposition 1 (main paper) and the bound on λ_n to conclude $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \leq 2m\psi C''^2\Phi_\gamma[\Delta]^2 \leq 2m\psi C''^2 C_1''^2\lambda_n^2 \leq \frac{\lambda_n}{\kappa}$. Based on the optimality condition of (6.2), the property that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}(.)] \leq 2\Phi_\gamma(.)$, we have:

$$\begin{aligned}
\Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}_\mathcal{M}}(\Delta)] &\leq 2\lambda_n + \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger R_{\Sigma^\star}(\Delta)] + \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \\
&+ \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger E_n] \\
&\leq 2\lambda_n + 2\lambda_n\left(\frac{1}{\kappa} + \frac{1}{\kappa} + \frac{1}{\kappa}\right) \leq \frac{2\lambda_n\kappa + 6\lambda_n}{\kappa}
\end{aligned}$$

Recalling the notation $\mathbb{H}_\mathcal{M}[2,3] = L_y^\mathcal{M} \times \Theta_{yx}^\mathcal{M}$, we use the result above to conclude that:

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{H}_\mathcal{M}[2]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)\|_2 &\leq \frac{2\lambda_n\kappa + 6\lambda_n}{\kappa} + 4\psi^2 r_2^u \\
&\leq \frac{8\lambda_n}{\kappa}\left(1 + \frac{4}{\alpha}\right)\psi^2 + \frac{2\lambda_n\kappa + 6\lambda_n}{\kappa} \\
\frac{1}{\gamma}\|\mathcal{P}_{\mathbb{H}_\mathcal{M}[3]}F^\dagger\mathbb{I}^\star\mathcal{I}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)\|_2 &\leq \frac{2\lambda_n\kappa + 3\lambda_n}{\kappa} + \frac{4\psi^2 r_2^u}{\gamma} \\
&\leq \frac{1}{\gamma}\frac{8\lambda_n}{\kappa}\left(1 + \frac{4}{\alpha}\right)\psi^2 + \frac{2\lambda_n\kappa + 6\lambda_n}{\kappa}
\end{aligned}$$

Thus,

$$
\begin{aligned}
\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] &\leq \frac{4\lambda_n}{\kappa}\left(1+\frac{4}{\alpha}\right)\psi^2 + \frac{2\lambda_n\kappa+6\lambda_n}{\kappa} \\
&= \frac{\lambda_n}{\kappa}\left(8m(1+\frac{4}{\alpha})\psi^2+2\kappa+6\right)
\end{aligned}
$$

Furthermore, observing that the Fisher Information Assumption 3 in (2.9) (main paper) implies that
$\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] \leq \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)]$, we have:

$$
\begin{aligned}
\Gamma_\gamma(\mathcal{P}_{[2,3]}[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta)]) &\leq 2mr_2^u\psi^2 + \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] \\
&\quad + \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] + \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \\
&\leq \frac{\lambda_n}{\kappa}\left(20m(1+\frac{4}{\alpha})\psi^2+\kappa+13\right) \\
&< \frac{2\lambda_n}{\kappa}\left(20m(1+\frac{4}{\alpha})\psi^2+\kappa+13\right)
\end{aligned}
$$

6.3.3 From Tangent Space Constraints to the Original Problem

The optimality conditions of (6.2) suggest that there exist Lagrange multipliers $Q_{D_y} \in \mathcal{W}$, $Q_{T_y} \in T(L_y^\mathcal{M})^\perp$, and $Q_{T_{yx}} \in T(\Theta_{yx}^\mathcal{M})^\perp$ such that

$$
\begin{aligned}
[\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{D_y} = 0; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{T_y} &\in \lambda_n\partial\|\tilde{L}_y\|_\star \\
[\Sigma_n - \tilde{\Theta}^{-1}]_{yx} + Q_{T_{yx}} \in -\lambda_n\gamma\partial\|\tilde{\Theta}_{yx}\|_\star; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_x &= 0
\end{aligned}
$$

Letting the SVD of $\tilde{L}_y$ and $\tilde{\Theta}_{yx}$ be given by $\tilde{L}_y = \bar{U}\bar{O}\bar{V}'$ and $\tilde{\Theta}_{yx} = \breve{U}\breve{O}\breve{V}'$ respectively, and $Z \triangleq (0, \lambda_n\bar{U}\bar{V}', -\lambda_n\gamma\breve{U}\breve{V}', 0)$, we can restrict the optimality conditions to the space $\mathbb{H}_\mathcal{M}$ to obtain, $\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger(\Sigma_n - \tilde{\Theta}^{-1}) = Z$. We proceed by proving that the variables $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ satisfy the optimality conditions of the original convex program (1.4). That is:

1. $\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger(\Sigma_n - \tilde{\Theta}^{-1}) = Z$

2. $\max\left\{\|\mathcal{P}_{T_y'^\perp}(\Sigma_n - \tilde{\Theta}^{-1})_y\|_2, \frac{1}{\gamma}\|\mathcal{P}_{T_{yx}'^\perp}(\Sigma_n - \tilde{\Theta}^{-1})_{yx}\|_2\right\} < \lambda_n$

It is clear that the first condition is satisfied since the pair $(\tilde{\Theta}, \tilde{S}_y, \tilde{L}_y)$ is optimum for (6.2). To prove that the second condition, we must prove that $\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp[2,3]}\mathcal{G}^\dagger(\Sigma_n - \tilde{\Theta}^{-1})] < \lambda_n$. In particular, denoting $\Delta = (\tilde{D}_y - D_y^\star, \tilde{L}_y - L_y^\star, \tilde{\Theta}_{yx} - \Theta_{yx}^\star, \tilde{\Theta}_x - \Theta_x^\star)$ we show that:

$$
\begin{aligned}
\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] &< \lambda_n - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger E_n] \\
&\quad - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \\
&\quad - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \\
&\quad - \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1,0,0,\Delta_4)]
\end{aligned} \tag{6.3}
$$

30

Using the fact that $\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger(N)] \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger(N)]$ for any matrix $N \in \mathbb{S}^{p+q}$, this would in turn imply that:

$$
\begin{aligned}
\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] \;<\;& \lambda_n - \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger E_n] \\
& - \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \\
& - \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \\
& - \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1,0,0,\Delta_4)]
\end{aligned}
\tag{6.4}
$$

Indeed (26) implies that the second optimality condition is satisfied. So we focus on showing that (27) is satisfied. Using the first optimality condition, the fact that projecting into tangent spaces with respect to rank variety increase the spectral norm by at most a factor of two (i.e. $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}(\cdot)] \leq 2\Phi_\gamma(\cdot)$),the fact that $\Gamma_\gamma[\mathcal{G}^\dagger(\cdot)] \leq \Phi_\gamma[\mathcal{F}^\dagger(\cdot)]$, and that $\kappa = \beta(6 + \frac{16\psi^2 m}{\alpha})$, we have that:

$$
\begin{aligned}
\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] \;\leq\;& \lambda_n + 2\Gamma_\gamma[\mathcal{G}^\dagger R_{\Sigma^\star}(\Delta)] + 2\Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_M}] \\
& + 2\Gamma_\gamma[\mathcal{G}^\dagger E_n] + \Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1,0,0,\Delta_4)] \\
\leq\;& \lambda_n + 2\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\Delta)] + 2\Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_M}] \\
& + 2\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1,0,0,\Delta_4)] \\
\leq\;& \lambda_n + \frac{\lambda_n}{\beta}
\end{aligned}
$$

Applying Fisher Information Assumption 2 in (2.9) (main paper), we obtain:

$$
\begin{aligned}
\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] \;\leq\;& \frac{(\beta+1)\lambda_n}{\beta}\left(1 - \frac{2}{\beta+1}\right) = \lambda_n - \frac{\lambda_n}{\beta} \\
<\;& \lambda_n - \frac{\lambda_n}{2\beta} \\
\leq\;& \lambda_n - \Phi_\gamma[\mathcal{F}^\dagger R_\Sigma(\mathcal{F}(\Delta))] - \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \\
& - \Phi_\gamma[\mathcal{F}^\dagger E_n] - \Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1,0,0,\Delta_4)] \\
\leq\;& \lambda_n - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \\
& - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \\
& - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger E_n] \\
& - \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1,0,0,\Delta_4)]
\end{aligned}
$$

Here, we used the fact that $\|\mathcal{P}_{T^\perp}(.)\|_2 \leq \|.\|_2$ for a tangent space T of the low-rank matrix variety.

6.4 Proof of Proposition 4 (main paper)

We must study the rate of convergence of the sample covariance matrix to the population covariance matrix. The following result from [3] plays a key role in obtaining this result.

Proposition 6.3. *Given natural numbers n,p with $p \leq n$, Let Γ be a $p \times n$ matrix with i.i.d Gaussian entries that have zero-mean and variance $\frac{1}{n}$. Then the largest and smallest singular values $\sigma_1(\Gamma)$ and $\sigma_p(\Gamma)$ of Γ are such that:*

$$
\max\left\{Prob[\sigma_1(\Gamma) \leq 1 + \sqrt{\frac{p}{n}} + t], Prob[\sigma_p(\Gamma) \leq 1 - \sqrt{\frac{p}{n}} - t]\right\}
$$

31

We now proceed with proving Proposition 4 (main paper). First, note that $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq m\|\Sigma_n - \Sigma^\star\|_2$. Using Theorem 6.3 and the fact that $\frac{\lambda_n}{m\kappa} \leq 8\psi$ and $n \geq \frac{64\kappa^2(p+q)m^2\psi^2}{\lambda_n^2}$, the following bound holds: $\Pr[m\|\Sigma_n - \Sigma^\star\|_2 \geq \frac{\lambda_n}{\kappa}] \leq 2\exp\left\{-\frac{n\lambda_n^2}{128\kappa^2 m^2\psi^2}\right\}$. Thus, $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{\kappa}$ with probability greater than $1 - 2\exp\left\{-\frac{n\lambda_n^2}{128\kappa^2 m^2\psi^2}\right\}$

6.5   Consistency of the Convex Program (3.1) (main paper)

In this section, we prove the consistency of convex program (3.1) for estimating a factor model. We first introduce some notation. We define the linear operator: $\tilde{\mathcal{F}} : \mathbb{S}^p \times \mathbb{S}^p \to \mathbb{S}^p$ and its adjoint $\tilde{\mathcal{F}}^\dagger : \mathbb{S}^p \to \mathbb{S}^p \times \mathbb{S}^p$ as follows:

$$\tilde{\mathcal{F}}(M, K) \triangleq M - K, \qquad \tilde{\mathcal{F}}^\dagger(Q) \triangleq (Q, Q) \tag{6.5}$$

We consider a population composite factor model (1.3) $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \epsilon$ underlying a pair of random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$, with $\text{rank}(\mathcal{A}^\star) = k_x$, $\mathcal{B}_u^\star \in \mathbb{R}^{p \times k_u}$, and column-space($\mathcal{A}^\star$) $\cap$ column-space($\mathcal{B}_u^\star$) = $\{0\}$. As the convex relaxation (3.1) is solved in the precision matrix parametrization, the conditions for our theorems are more naturally stated in terms of the joint precision matrix $\Theta^\star \in \mathbb{S}^{p+q}$, $\Theta^\star \succ 0$ of (y, x). The algebraic aspects of the parameters underlying the factor model translate to algebraic properties of submatrices of $\Theta^\star$. In particular, the submatrix $\Theta_{yx}^\star$ has rank equal to k_x, and the submatrix $\Theta_y^\star$ is decomposable as $D_y^\star - L_y^\star$ with $D_y^\star$ being diagonal and $L_y^\star \succeq 0$ having rank equal to k_u. Finally, the transversality of column-space($\mathcal{A}^\star$) and column-space($\mathcal{B}_u^\star$) translates to the fact that column-space($\Theta_{yx}^\star$) $\cap$ column-space($L_y^\star$) = $\{0\}$ have a transverse intersection. We consider the factor model underlying the random vector $y \in \mathbb{R}^p$ that is induced upon marginalization of x. In particular, the precision matrix of y is given by $\tilde{\Theta}_y^\star = D_y^\star - [L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star]$. To learn an accurate factor model, we seek an estimate $(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$ from the convex program (3.1) (main paper) such that $\text{rank}(\hat{\tilde{L}}_y = \text{rank}(L_y^\star + \Theta_{yx}^\star\Theta_x^{\star-1}\Theta_{xy}^\star)$, and the errors $\|\hat{\tilde{D}}_y - D_y^\star\|_2, \|\hat{\tilde{L}}_y - [L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star]\|_2$ are small.

Following the same reasoning as the Fisher information conditions for consistency of the convex program (1.4)(main paper), A natural set of conditions on the population Fisher information at $\tilde{\Theta}_y^\star$ defined as $\mathbb{I}_y^\star = (\tilde{\Theta}_y^\star)^{-1} \otimes (\tilde{\Theta}_y^\star)^{-1}$ are given by:

$$\text{Assumption 4} \quad : \quad \inf_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\chi}(\mathbb{H}', \tilde{\Phi}) \geq \tilde{\alpha}, \quad \text{for some } \tilde{\alpha} > 0 \tag{6.6}$$

$$\text{Assumption 5} \quad : \quad \inf_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\Xi}(\mathbb{H}') > 0 \tag{6.7}$$

$$\text{Assumption 6} \quad : \quad \sup_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\varphi}(\mathbb{H}') \leq 1 - \frac{2}{\tilde{\beta} + 1} \quad \text{for some } \tilde{\beta} \geq 2, \tag{6.8}$$

where,

$$\tilde{\chi}(\mathbb{H}, \|.\|_\Upsilon) \triangleq \min_{\substack{Z \in \mathbb{H} \\ \|Z\|_\Upsilon = 1}} \|\mathcal{P}_\mathbb{H} \tilde{\mathcal{I}}^\dagger \mathbb{I}_y^\star \tilde{\mathcal{I}} \mathcal{P}_\mathbb{H}(Z)\|_\Upsilon$$

$$\tilde{\Xi}(\mathbb{H}) \triangleq \min_{\substack{Z \in \mathbb{H}[2] \\ \|Z\|_2 = 1}} \|\mathcal{P}_{\mathbb{H}[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]}(Z)\|_2$$

$$\tilde{\varphi}(\mathbb{H}) \triangleq \max_{\substack{Z \in \mathbb{H}[2] \\ \|Z\|_2 = 1}} \|\mathcal{P}_{\mathbb{H}^\perp[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]} (\mathcal{P}_{\mathbb{H}[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]})^{-1}(Z)\|_2$$

$$\tilde{U}(\tilde{\omega}_y) \triangleq \left\{ \mathcal{W} \times T' \mid \rho(T', T(L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star)) \le \tilde{\omega}_y \right\}$$

$$\tilde{\Phi}(D, L) \triangleq \max\{\|D\|_2, \|L\|_2\}.$$

Assumption 4 controls the gain of the Fisher information $\mathbb{I}_y^\star$ restricted to appropriate subspaces and Assumption 5 and 6 are in the spirit of irrepresentability conditions. As the variety of low-rank matrices is locally curved around $T(L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star)$, we control the Fisher information $\mathbb{I}_y^\star$ at nearby tangent spaces T' where $\rho(T', T(L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star) \le \tilde{\omega}_y$. We also note that measuring the gains of Fisher information $\mathbb{I}_y^\star$ with the norm $\tilde{\Phi}$ and $\|\cdot\|_2$ is natural as these are closely tied with dual norm of the regularizer trace($\tilde{L}_y$) in (3.1).

We present a theorem of consistency of the convex relaxation (3.1) under Assumptions 4, 5 and 6. We let σ denote the minimum nonzero singular value of $L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star$. The proof strategy is similar in spirit to the strategy for proving the consistency of the convex relaxation (1.4) and is left out for brevity.

Theorem 6.4. *Suppose that there exists $\tilde{\alpha} > 0$, $\tilde{\beta} \ge 2$, $\tilde{\omega}_y \in (0,1)$ so that the population Fisher information $\mathbb{I}_y^\star$ satisfies Assumptions 4, 5 and 6 in (6.6),(6.7) and (6.8). Suppose that the following conditions hold:*

1. $n \gtrsim \left[\frac{\tilde{\beta}^2}{\tilde{\alpha}^2}\right](p)$

2. $\tilde{\lambda}_n \sim \frac{\tilde{\beta}}{\tilde{\alpha}}\sqrt{\frac{p}{n}}$

3. $\sigma \gtrsim \frac{\tilde{\beta}}{\tilde{\alpha}^5 \tilde{\omega}_y} \tilde{\lambda}_n$

Then with probability greater than $1 - 2\exp\left\{-C\frac{\tilde{\alpha}}{\tilde{\beta}}n\tilde{\lambda}_n^2\right\}$, the optimal solution $(\hat{\Theta}, \hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$ of (3.1) with i.i.d. observations $\mathcal{D}_n = \{y^{(i)}\}_{i=1}^n$ satisfies the following properties:

1. $rank(\hat{\tilde{L}}_y) = rank(L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star)$

2. $\|\hat{\tilde{D}}_y - D_y^\star\|_2 \lesssim \frac{\tilde{\lambda}_n}{\tilde{\alpha}^2}$, $\|\hat{\tilde{L}}_y - L_y^\star - \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star\|_2 \lesssim \frac{\tilde{\lambda}_n}{\tilde{\alpha}^2}$

References

[1] BACH, F. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.

[2] CHANDRASEKARAN, V., PARRILO, P. A. & WILLSKY, A. S.. Latent Variable Graphical Model Selection via Convex Optimization. *Annals of Statistics*, 40:1935–1967, 2012.

[3] DAVIDSON, K.R. & SZAREK, S.J. Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces*, 1:317–366, 2001.

[4] KATO, T. Perturbation theory for linear operators. *Springer*, 1995.

[5] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. & YU, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.