

# Interpreting Latent Variables in Factor Models via Convex Optimization

Armeen Taeb <sup>†</sup> and Venkat Chandrasekaran <sup>†,‡,\*</sup>

<sup>†</sup> Department of Electrical Engineering

<sup>‡</sup> Department of Computing and Mathematical Sciences  
California Institute of Technology  
Pasadena, Ca 91125

January 5, 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

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 [19]. *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<sup>1</sup>  $\zeta \sim \mathcal{N}(0, \Sigma_\zeta), \epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$ , with  $\Sigma_\zeta \succ 0, \Sigma_\epsilon \succ 0$  and  $\Sigma_\epsilon$  being diagonal. Here the random vector  $\zeta$  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  $\zeta$  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  $\zeta$ ) – 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  $\epsilon$  is diagonal).

A natural approach to attributing semantic information to the latent variables  $\zeta$  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  $\zeta$ . 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  $\zeta$  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, 16, 17, 18, 15].

## 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  $\zeta$  – is invariant under transformations of the form  $\mathcal{B} \rightarrow \mathcal{B}\mathcal{W}^{-1}$  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  $\zeta$  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  $\text{column-space}(\mathcal{B}) \approx \mathfrak{H}_x \oplus \mathfrak{H}_u$  – the subspace  $\mathfrak{H}_x$  specifies those components of  $y$  that are influenced by the latent variables  $\zeta$  and are also affected by the covariates  $x$ , and the subspace  $\mathfrak{H}_u$  represents any unobserved residual effects on  $y$  due to  $\zeta$  that

---

<sup>1</sup>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

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

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  $\text{column-space}(\mathcal{A}) \cap \text{column-space}(\mathcal{B}_u) = \{0\}$ . Since the number of latent variables  $\zeta$  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  $\zeta$* . 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  $y$ :

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

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 (1.3).

## 1.2 Composite Factor Modeling via Convex Optimization

Next we describe techniques to fit observations of  $y \in \mathbb{R}^p$  and of  $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$  to the models (1.1) and (1.3) respectively. These methods are key subroutines in our algorithmic approach for associating semantics to the latent variables in a factor model (see Section 3 for a high-level discussion of our approach and Section 3 for a more detailed experimental demonstration). Fitting observations of  $y \in \mathbb{R}^p$  (or of  $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ ) to the factor model (1.1) (or to (1.3)) is accomplished by identifying a Gaussian model over  $y$  (or 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  $\zeta$ ) and a diagonal matrix  $\Sigma_{\epsilon}$ . 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 [12]. 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, 16, 17, 18]; 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 [11, 7, 14, 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  $\Sigma$  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  $\Sigma$  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  $\Theta_y$ ; 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  $\Theta_y$  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  $\Theta$  is given by  $\mathcal{A} = -[\Theta_y]^{-1}\Theta_{yx}$ . Based on this algebraic structure desired in  $\Theta$ , we propose the following natural convex relaxation for fitting a collection of observations  $\{(y^{(i)}, x^{(i)})\} \subset \mathbb{R}^{p+q}$  to the composite model (1.3):

$$\begin{aligned} (\hat{\Theta}, \hat{D}_y, \hat{L}_y) = \arg \min_{\substack{\Theta \in \mathbb{S}^{p+q}, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}} & -\ell(\Theta; \{y^{(i)}, x^{(i)}\}_{i=1}^n) + \lambda_n[\gamma\|\Theta_{yx}\|_* + \text{trace}(L_y)] \\ \text{s.t.} & \quad \Theta_y = D_y - L_y, L_y \succeq 0, D_y \text{ is diagonal} \end{aligned} \quad (1.4)$$

The term  $\ell(\Theta; \{y^{(i)}, x^{(i)}\}_{i=1}^n)$  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):

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

This function is concave as a function of the joint precision matrix<sup>2</sup>  $\Theta$ . The matrices  $D_y, L_y$  represent the diagonal and low-rank components of  $\Theta_y$ . 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, 14, 3], the nuclear norm penalty  $\|\Theta_{yx}\|_*$  on the submatrix  $\Theta_{yx}$  (which is in general a non-square matrix) is useful for promoting low-rank structure in that submatrix of  $\Theta$ . The parameter  $\gamma$  provides a tradeoff between the observed/interpretable and the unobserved parts of the composite factor model (1.3), and the parameter  $\lambda_n$  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

<sup>2</sup>An additional virtue of parametrizing 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:

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

with the covariance of  $\zeta_u$  being the identity matrix of appropriate dimensions and the covariance of  $\bar{\epsilon}$  being  $\hat{D}_y^{-1}$ .

One can specialize the convex relaxation (1.4) for the composite factor model to obtain an approach for fitting observations  $\{y^{(i)}\}_{i=1}^n$  to the factor model (1.1) without additional covariates:

$$\begin{aligned} (\hat{D}_y, \hat{L}_y) &= \arg \min_{\substack{\tilde{D}_y, \tilde{L}_y \in \mathbb{S}^p \\ \tilde{D}_y - \tilde{L}_y \succ 0}} -\ell(\tilde{D}_y - \tilde{L}_y; \{y^{(i)}\}_{i=1}^n) + \tilde{\lambda}_n \text{trace}(\tilde{L}_y) \\ &\text{s.t.} \quad \tilde{L}_y \succeq 0, \tilde{D}_y \text{ is diagonal.} \end{aligned} \quad (1.7)$$

The parameter  $\tilde{\lambda}_n$  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, 16, 17, 18]– the regularized maximum-likelihood convex program (1.7) fits factor models by decomposing a precision matrix as the difference between a diagonal matrix and a low-rank matrix. As with the composite factor model, one can obtain estimates for the parameters of the factor model (1.1) based on the optimal solution  $(\hat{D}_y, \hat{L}_y)$  of (1.7) as follows:

$$\hat{B} = \text{any squareroot of } \left( \hat{D}_y - \hat{L}_y \right)^{-1} - \hat{D}_y^{-1} \text{ such that } \hat{B} \in \mathbb{R}^{p \times \text{rank}(\hat{L}_y)}, \quad (1.8)$$

with the covariance of  $\zeta$  being the identity matrix of appropriate dimensions and the covariance of  $\epsilon$  being  $\hat{D}_y^{-1}$ .

The convex programs (1.4) and (1.7) are log-determinant semidefinite programs that can be solved efficiently using existing numerical solvers such as the LogDetPPA package [20].

### 1.3 Our Results

Our discussion thus far has assumed that a factor model underlying a collection of variables  $y \in \mathbb{R}^p$  is given, and the objective is to obtain semantic interpretation of the latent variables. However, in many situations, a factor model underlying  $y \in \mathbb{R}^p$  may not be available in advance, and it must be learned from observations of  $y \in \mathbb{R}^p$ . We consider this latter more general setting for our theoretical development as well as our experimental demonstration.

Conceptually, a natural (meta-)procedure for learning a factor model and then interpreting the latent variables in this factor model is to proceed in two stages. In the first step, the analyst identifies a factor model (1.1) based solely on observations of the variables  $y \in \mathbb{R}^p$  using the convex relaxation (1.7), which results in an estimate for a factor model underlying the variables  $y$ . To attribute interpretable meaning to the latent variables in this factor model, the analyst then obtains 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, the second step consists of identifying a suitable composite factor model (1.3) via the convex program (1.4). In

particular, we sweep over the parameters  $\lambda_n, \gamma$  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 obtained in the first stage.

In Section 2 we carry out a theoretical analysis to investigate whether a two-stage framework as outlined above can possibly succeed. If one identifies a factor model for a collection of variables in the first step, are there conditions under which one can identify composite models satisfying the requirements of the second stage, which would then serve to provide interpretations of the latent variables? To address such questions, we discuss a model problem, 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  $x$ . Now consider the following two questions:

1. Given observations jointly of  $(y, x) \in \mathbb{R}^{p+q}$ , does the convex relaxation (1.4) (for suitable choices of regularization parameters  $\lambda_n, \gamma$ ) estimate the composite factor model underlying these two random vectors accurately?
2. Given observations of *only* the variables  $y$ , does the convex relaxation (1.7) (for an appropriate choice of regularization  $\tilde{\lambda}_n$ ) estimate the factor model underlying  $y$  accurately?

An affirmative answer to both of these questions demonstrates the success of the two-stage procedure outlined in the preceding paragraph. First, assume for simplicity that we are given the factor model underlying  $y$  (obtained by marginalizing the covariates  $x$ ). In other words, we are supposing for now that we do not need to perform stage 1 in the approach outlined above. A positive answer to Question (1) 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, stage 2 in the algorithmic framework above is successful. Next, a positive answer to Question (2) implies that we can in fact estimate the factor model underlying  $y$  from observations of the only the variables  $y$  (i.e., we do not need some sort of oracle knowledge of this factor model in advance), and therefore stage 1 of our method described above can be performed successfully.

In Section 2, we show that under suitable identifiability conditions on the population model of the combined random vector  $(y, x)$ , the convex programs (1.4) and (1.7) succeed in solving these two questions. Our analysis is carried out in a high-dimensional asymptotic scaling setup in which the dimensions  $p, q$ , the number of observations  $n$ , and other model parameters may all grow simultaneously [2, 22].

The implementation of such a two-stage approach in practice requires the specification of several further details, especially for the second step. In Section 3 we give a full algorithmic description of our methodology as well as a concrete demonstration with experiments on real-world financial data. Specifically, we consider as our variables  $y$  the monthly averaged stock prices of  $p = 66$  companies from the Standard and Poor index over the period June 1990 to July 2014, and we identify a factor model (1.1) over  $y$  with 13 latent variables influencing the stock return. We then obtain observations of  $q = 11$  covariates on quantities related to oil trade, employment levels, etc. (see Section 3 for the full list), as these plausibly influence the stock returns. By suitably employing the convex relaxation (1.4) for composite factor modeling, we identify a two-dimensional projection of these 11 covariates that represent an interpretable component of the 13 latent variables in the factor model, as well as a remaining set of 11 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 inflation rate are the most relevant of the 11 covariates considered in our experiment, while gold prices and oil exports are less useful. See Section 3 for complete details.

## 1.4 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 (1.7) 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 [22] 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.

## 1.5 Notation

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

$$\mathcal{F}(M, N, K, O) \triangleq \begin{pmatrix} M - N & K \\ K^T & O \end{pmatrix}, \quad \mathcal{F}^\dagger \begin{pmatrix} Q & K \\ K^T & O \end{pmatrix} \triangleq (Q, Q, K, O) \quad (1.9)$$

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 theorems that address Questions (1) and (2) from Section 1.3. These theorems require assumptions on the population precision matrix, which are discussed in Section 2.1, with the theorem statements given in Section 2.2. The proofs of these theorems are given in Section 4 with some details deferred to the supplementary material. We confirm the theoretical predictions of with numerical simulations on synthetic data in Section 2.3.

### 2.1 Technical Setup

As discussed in Section 1.3, our theorems are premised on the existence of a population 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  $\text{column-space}(\mathcal{A}^*) \cap \text{column-space}(\mathcal{B}_u^*) = \{0\}$ . As the convex relaxations (1.4) and (1.7) are solved in the precision matrix parametrization, the conditions for

our theorems are more naturally stated in terms of the joint precision matrix  $\Theta^* \in \mathbb{S}^{p+q}$ ,  $\Theta^* \succ 0$  of  $(y, x)$ . The algebraic aspects of the parameters underlying the factor model translate to algebraic properties of submatrices of  $\Theta^*$ . In particular, the submatrix  $\Theta_{yx}^*$  has rank equal to  $k_x$ , and the submatrix  $\Theta_y^*$  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  $\text{column-space}(\mathcal{A})$  and  $\text{column-space}(\mathcal{B}_u)$  translates to the fact that  $\text{column-space}(\Theta_{yx}^*) \cap \text{column-space}(L_y^*) = \{0\}$  have a transverse intersection.

To address the requirements raised in Question (1) in Section 1.3, we seek an estimate  $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$  from the convex relaxation (1.4) such that  $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^*)$ ,  $\text{rank}(\hat{L}_y) = \text{rank}(L_y^*)$ , and that  $\|\hat{\Theta} - \Theta^*\|_2$  is small. To satisfy the requirements of Question (2) in Section 1.3, we first need to 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  $\tilde{\Theta}_y^* = D_y^* - L_y^* - \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*$ , and Question (2) asks whether we can obtain an estimate  $(\hat{D}_y, \hat{L}_y)$  such that  $\text{rank}(\hat{L}_y) = \text{rank}(L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*)$ , and the errors  $\|(\hat{D}_y - D_y^*)\|_2, \|\hat{L}_y - [L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*]\|_2$  are small.

Building on both classical statistical estimation theory [1] as well as the recent literature on high-dimensional statistical inference [2, 22], we describe our conditions in terms of assumptions on the *Fisher information* at the population model parametrized by the precision matrix  $\Theta^*$ :

$$\mathbb{I}^* = \Theta^{*-1} \otimes \Theta^{*-1}.$$

Here the symbol  $\otimes$  refers to the tensor product, and the Fisher information may be viewed as an operator from the parameter space  $\mathbb{S}^{p+q}$  to itself. From a mathematical programming perspective, the importance of the Fisher information can be seen from the fact that the Hessian of the negative-log-likelihood function (1.4) evaluated at  $\Theta^*$  is given by  $\mathbb{I}^*$ .

To ensure that the error term  $\|\hat{\Theta} - \Theta^*\|_2$  is small, a classical condition from the statistical estimation literature is to control the minimum gain of the Fisher information  $\mathbb{I}^*$  [1]:

$$\eta_1^* \triangleq \min_{M \in \mathbb{S}^{p+q}, \|M\|_2=1} \|\mathbb{I}^* M\|_2. \quad (2.1)$$

From an optimization viewpoint, the condition that  $\eta_1^*$  is large is useful in ensuring that the negative-log-likelihood function at  $\Theta^*$  is sufficiently curved.

To further satisfy the requirements that  $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^*)$ ,  $\text{rank}(\hat{L}_y) = \text{rank}(L_y^*)$  with the convex relaxation (1.4) for the composite approach and that  $\text{rank}(\hat{L}_y) = \text{rank}(L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*)$  with the relaxation (1.7) for the factor modeling approach, bounding the minimum gain quantity  $\eta_1^*$  (2.1) from below is insufficient. To this end, we need to control two additional quantities associated with the Fisher information  $\mathbb{I}^*$ . The first of these is the maximum inner-product between orthogonal elements in  $\mathbb{S}^{p+q}$  in the metric induced by the Fisher information  $\mathbb{I}^*$ :

$$\eta_2^* \triangleq \max_{\substack{\mathbb{W} \subset \mathbb{S}^{p+q} \\ \mathbb{W} \text{ is a subspace}}} \max_{M \in \mathbb{W}, \|M\|_2 \leq 1} \|\mathcal{P}_{\mathbb{W}^\perp} \mathbb{I}^* \mathcal{P}_{\mathbb{W}}(M)\|_2 \quad (2.2)$$

Assuming that  $\eta_2^*$  is small ensures that errors in the estimation of the submatrix  $\Theta_{yx}^*$  do not impact the estimation of the  $\Theta_y^*$  submatrix (and vice versa). Indeed, in the absence of an upper bound on  $\eta_2^*$ , the effect of the term  $\mathcal{A}^*x$  (represented by the column space of the submatrix  $\Theta_{yx}^*$ ) would not be distinguishable from the effects of the conditional factor model  $\mathcal{B}_u^* \zeta_u + \epsilon$  (represented by the submatrix  $\Theta_y^*$ ).

The final parameter associated to the Fisher information that we need to control for our main theorem is motivated by two concerns. First, we need to ensure that the diagonal and low-rank components  $D_y^*$  and  $L_y^*$  that compose the submatrix  $\Theta_y^*$  can be distinguished from each other



(for Question (2)). For example, if the matrix  $L_y^*$  consists of just a single nonzero in one of the diagonal entries and is zero elsewhere, identifying such a low-rank matrix from the difference  $D_y^* - L_y^*$  is impossible. Second, we require that the components  $D_y^*$  and  $L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*$  are identifiable given  $D_y^* - L_y^* - \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*$ . These identifiability issues arising in the decomposition of sums of diagonal (and, more generally, sparse) and low-rank matrices have been investigated thoroughly [4, 15]. Specifically, the geometric insights in these papers imply that a natural condition to ensure identifiability in such decomposition problems is to assume the transversality of the intersection between the subspace of diagonal matrices in  $\mathbb{S}^p$  and the *tangent space* with respect to the algebraic variety of low-rank matrices at  $L_y^*$  (for the composite factor model, i.e. Question (1)) or at  $L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*$  (for the factor model, i.e., Question (2)). In particular, the tangent space at a rank- $r$  matrix  $N$  with respect to the algebraic variety of  $p_1 \times p_2$  matrices with rank less than or equal to  $r$  is given by:

$$T(N) \triangleq \{N_R + N_C \mid N_R, N_C \in \mathbb{R}^{p_1 \times p_2}, \\ \text{row-space } N_R \subseteq \text{row-space } N, \text{ column-space } N_C \subseteq \text{column-space } N\}$$

The tangent space at  $L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*$  contains inside it the tangent space at  $L_y^*$  as the row/column spaces of  $L_y^*$  are contained inside the row/column spaces of  $L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*$  (due to the assumption above that  $\text{column-space}(\Theta_{yx}^*) \cap \text{column-space}(L_y^*) = \{0\}$ ); consequently, we assume that the set of diagonal matrices has a transverse intersection with the tangent space at  $L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*$  (which allows us to address the identifiability issues for both questions) by controlling the following parameter associated to the Fisher information  $\mathbb{I}^*$  for some  $\omega > 0$  and for all subspaces  $\tilde{T} \subset \mathbb{S}^p$  “close to  $T(L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*)$ ”:

$$\eta_3(\tilde{T}; \omega) = \max \left\{ \max_{\substack{M \in \tilde{T} \\ \|M\|_2 \leq 1}} \|\mathcal{P}_{\text{diag}} \mathbb{I}_Y^* M\|_2, \max_{\substack{M \text{ is diagonal} \\ \|M\|_2 \leq 1}} \|\mathcal{P}_{\tilde{T}} \mathbb{I}_Y^* M\|_2 \right\} \quad (2.3)$$

Here the operator  $\mathcal{P}_{\text{diag}}$  represents projection onto the space of diagonal matrices, and  $\mathbb{I}_y^* = \tilde{\Theta}_y^{-1} \otimes \tilde{\Theta}_y^{-1}$  represents the Fisher information with respect to the precision matrix  $\tilde{\Theta}_y$  of the random vector  $y$ . The reason for considering the Fisher information with respect to the marginal precision matrix corresponding to  $y$  is that the transversality conditions pertain only to the components of the precision matrix of  $y$ . We bound the quantity  $\eta_3(\tilde{T}; \omega)$  for *all* spaces  $\tilde{T}$  that are within a small distortion of  $T(L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*)$ :

$$\eta_3^*(\omega) = \max_{\rho(\tilde{T}, T(L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*)) \leq \omega} \eta_3(\tilde{T}; \omega),$$

where the distortion  $\rho$  is measured 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 reason for considering such distortions around the tangent space  $T(L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*)$  is that the variety of low-rank matrix are locally curved around their smooth points. Consequently, the tangent spaces at matrices in a neighborhood around  $L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*$  having the same rank are generally not the same as  $T(L_y^* + \Theta_{yx}^*(\Theta_x^*)^{-1}\Theta_{xy}^*)$ . In particular, the estimate  $\hat{\tilde{L}}_y$  produced by

the convex program (1.7) for the population low-rank matrix  $L_y^* + \Theta_{yx}^* x \Theta_x^{-1} \Theta_{xy}^*$  may be such that  $\hat{\tilde{L}}_y \approx L_y^* + \Theta_{yx}^* x \Theta_x^{-1} \Theta_{xy}^*$  and  $\text{rank}(\hat{\tilde{L}}_y) = \text{rank}(L_y^* + \Theta_{yx}^* x \Theta_x^{-1} \Theta_{xy}^*)$ , but it is generically the case that  $T(\hat{\tilde{L}}_y) \neq T(L_y^* + \Theta_{yx}^* x \Theta_x^{-1} \Theta_{xy}^*)$ . Consequently, it is critical to control the quantity  $\eta_3(\tilde{T}; \omega)$  in (2.3) for all  $\tilde{T}$  near the tangent space  $T(L_y^* + \Theta_{yx}^* (\Theta_{yx}^*)^{-1} \Theta_{xy}^*)$ .

## 2.2 Theorem Statements

We now describe the performance of the regularized maximum-likelihood programs (1.4) and (1.7) under suitable conditions on the quantities introduced in the previous section. Before formally stating our main result, we introduce some notation. Let  $\sigma_y$  denote the minimum nonzero singular value of  $L_y^*$ , let  $\sigma_{yx}$  denote the minimum nonzero singular value of  $\Theta_{yx}^*$ , and finally let  $\sigma$  denote the minimum nonzero singular value of  $L_y^* + \Theta_{yx}^* \Theta_x^{-1} \Theta_{xy}^*$ . In the following theorem statements, suppose that there exists  $\alpha > 0$ ,  $\beta \geq 8$ , and  $\omega \in (0, 1)$  such that the population Fisher information  $\mathbb{I}^*$  satisfies the following properties: (i)  $\eta_1^* \geq 3\alpha$ , (ii)  $\eta_2^* \leq \frac{8\alpha}{3\beta}$ , and (iii)  $\eta_3^*(\omega) \leq \frac{2\alpha}{\beta}$ . Theorem 2.1 pertains to the consistency of the estimator (1.4), and Theorem 2.2 relates to the consistency of the estimator (1.7).

**Theorem 2.1.** *Let  $m \triangleq \max\{1, \frac{1}{\gamma}\}$ ,  $\bar{m} \triangleq \max\{1, \gamma\}$ , and  $\psi \triangleq \|(\Theta^*)^{-1}\|_2$ . Further,  $C_1 = \frac{24}{\alpha} + \frac{1}{\psi^2}$ ,  $C_2 = \frac{4}{\alpha}(\frac{1}{3\beta} + 1)$ ,  $C_{\sigma_Y} = C_1^2 \psi^2 \max\{12\beta + 1, \frac{2}{C_2 \psi^2} + 1\}$ ,  $C_{\sigma_{YX}} = C_1^2 \psi^2 \max\{18\beta, \frac{2}{C_2 \psi^2} + 6\beta\}$ ,  $C_{samp} = \max\{\frac{1}{48\psi\beta}, 48\beta\psi^3 C_1^2, 8\psi C_2, \frac{64\psi^3 C_2}{\alpha}\}$ , and  $\lambda_{upper} = \frac{1}{m\bar{m}^2 C_{samp}}$ . Suppose that the following conditions hold:*

1.  $n \geq \frac{4608\psi^2 \beta^2 m^2 (p+q)}{\lambda_{upper}^2}$ ; that is  $n \gtrsim \left[\frac{\beta^4}{\alpha^2} m^4 \bar{m}^4\right] (p+q)$
2.  $\lambda_n \in \left[\sqrt{\frac{4608\psi^2 \beta^2 m^2 (p+q)}{n}}, \lambda_{upper}\right]$ ; e.g.  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$
3.  $\gamma \in \left[1, \frac{8\alpha}{3\beta\eta_2^*}\right]$
4.  $\sigma_Y \geq \frac{m}{\omega} C_{\sigma_Y} \lambda_n$ ; that is  $\sigma_Y \gtrsim \frac{\beta^2}{\alpha^2 \omega} m \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$
5.  $\sigma_{YX} \geq m^2 C_{\sigma_{YX}} \gamma^2 \lambda_n$ ; that is  $\sigma_{YX} \gtrsim \beta^2 \gamma^2 \alpha^2 m^2 \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$

Then with probability greater than  $1 - 2\exp\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\}$ , the optimal solution  $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$  of (1.4) with i.i.d. observations  $\{y^{(i)}, x^{(i)}\}_{i=1}^n$  of  $(y, x)$  satisfies the following properties:

1.  $\text{rank}(\hat{L}_y) = \text{rank}(L_y^*)$ ,  $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^*)$
2.  $\|\hat{D}_y - D_y^*\|_2 \leq C_1 \lambda_n$ ,  $\|\hat{L}_y - L_y^*\|_2 \leq C_1 \lambda_n$ ,  $\|\hat{\Theta}_{yx} - \Theta_{yx}^*\|_2 \leq C_1 \lambda_n \gamma$ , and  $\|\hat{\Theta}_x - \Theta_x^*\|_2 \leq C_1 \lambda_n$ ; that is  $\|\hat{D}_y - D_y^*\|_2 \lesssim \frac{\beta}{\alpha} m \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{L}_y - L_y^*\|_2 \lesssim \frac{\beta}{\alpha} m \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{\Theta}_{yx} - \Theta_{yx}^*\|_2 \lesssim \frac{\beta}{\alpha} \gamma m \sqrt{\frac{p+q}{n}}$ ,  $\|\hat{\Theta}_x - \Theta_x^*\|_2 \lesssim \frac{\beta}{\alpha} m \sqrt{\frac{p+q}{n}}$  if  $\lambda_n \sim \beta m \sqrt{\frac{p+q}{n}}$ .

**Theorem 2.2.** *Denote  $\tilde{\psi} \triangleq \|(\tilde{\Theta}^*)^{-1}\|_2$ . Let  $C_1 = \frac{24}{\alpha} + \frac{1}{\tilde{\psi}^2}$ ,  $C_2 = \frac{4}{\alpha}(\frac{1}{3\beta} + 1)$ ,  $C_\sigma = C_1^2 \tilde{\psi}^2 \max\{12\beta + 1, \frac{2}{C_2 \tilde{\psi}^2} + 1\}$ ,  $C_{samp} = \max\{\frac{1}{48\psi\beta}, 48\beta\tilde{\psi}^3 C_1^2, 8\tilde{\psi} C_2, \frac{64\tilde{\psi}^3 C_2}{\alpha}\}$ , and  $\lambda_{upper} = \frac{1}{C_{samp}}$ . Suppose that the following conditions hold:*

1.  $n \geq \frac{4608\tilde{\psi}^2\beta^2m^2p}{\lambda_{upper}^2}$ ; that is  $n \gtrsim \frac{\beta^4}{\alpha^2}p$
2.  $\tilde{\lambda}_n \in \left[ \sqrt{\frac{4608\tilde{\psi}^2\beta^2p}{n}}, \lambda_{upper} \right]$ ; e.g.  $\tilde{\lambda}_n \sim \beta\sqrt{\frac{p}{n}}$
3.  $\sigma \geq \frac{1}{\omega}C_\sigma\tilde{\lambda}_n$ ; that is  $\sigma \gtrsim \frac{\beta^2}{\alpha^2\omega}\sqrt{\frac{p}{n}}$  if  $\tilde{\lambda}_n \sim \beta\sqrt{\frac{p}{n}}$

Then with probability greater than  $1 - 2 \exp\{-\frac{n\tilde{\lambda}_n^2}{4608\beta^2\tilde{\psi}^2}\}$ , the optimal solution  $(\hat{D}_y, \hat{L}_y)$  of (1.7) with i.i.d. observations  $\{y^{(i)}\}_{i=1}^n$  of  $y$  satisfies the following properties:

1.  $\text{rank}(\hat{L}_y) = \text{rank}(L_y^* + \Theta_{yx}^* \Theta_x^{*-1} \Theta_{yx}^{*-1})$
2.  $\|\hat{D}_y - D_y^*\|_2 \leq C_1\tilde{\lambda}_n$ ;  $\|\hat{L}_y - L_y^* - \Theta_{yx}^* \Theta_x^{*-1} \Theta_{yx}^{*-1}\|_2 \leq C_1\tilde{\lambda}_n$ ; that is  $\|\hat{D}_y - D_y^*\|_2 \lesssim \frac{\beta}{\alpha}\sqrt{\frac{p}{n}}$  and  $\|\hat{L}_y - L_y^* - \Theta_{yx}^* \Theta_x^{*-1} \Theta_{yx}^{*-1}\|_2 \lesssim \frac{\beta}{\alpha}\sqrt{\frac{p}{n}}$ .

We outline the proof of Theorem 2.1 in Section 4. We omit the proof of Theorem 2.2 as it follows a very similar sequence of steps to that of Theorem 2.1. The quantities  $\alpha, \beta, \omega$  as well as the choices of parameters  $\gamma$  play a prominent role in our results. Indeed larger values of  $\alpha, \omega$  and smaller values of  $\beta$  (leading to a better conditioned Fisher information even for large distortions around the tangent space  $T(L_y^* + \Theta_{yx}^* \Theta_x^{-1} \Theta_{yx}^*)$ ) lead to less stringent requirements on the sample complexity, on the minimum nonzero singular value of  $\sigma_y$  of  $L_y^*$ , on the minimum nonzero singular value of  $\sigma$  of  $L_y^* + \Theta_{yx}^* \Theta_x^{-1} \Theta_{yx}^*$ , and on the minimum nonzero singular value  $\sigma_{yx}$  of  $\Theta_{yx}^*$ .

### 2.3 Synthetic Simulations

In this section, we give experimental evidence for the consistency of our convex program (1.4) and (1.7) on synthetic examples. Our choices of regularization parameters  $\lambda_n$  and  $\gamma$  (in the convex program (1.4)) and  $\tilde{\lambda}_n$  (in the convex program (1.7)) are guided by Theorem 2.1 and Theorem 2.2. Specifically, the parameters  $\lambda_n$  and  $\tilde{\lambda}_n$  are to be set in a manner that depends on the number of observations  $n$ , and we choose these to be proportional to  $\sqrt{\frac{p+q}{n}}$  and  $\sqrt{\frac{p}{n}}$ , respectively. Further, the theorems suggest that we should expect the ranks of the estimated solutions  $\hat{L}_y$  and  $\hat{\Theta}_{yx}$  to be the same for a range of values of  $\gamma$  when the correct underlying model structure is identified. Therefore, we solve the convex program (1.4) for several values of  $\gamma$ , and choose a value for which the structure of the estimated model remains the same for the largest contiguous range of values of  $\gamma$ . To solve the convex programs (1.4) and (1.7) numerically, we use the special-purpose solver LogDetPPA [20].

We generate the population model  $\Theta^* = \begin{pmatrix} D_y^* - L_y^* & \Theta_{yx}^* \\ \Theta_{yx}^{*'} & \Theta_x^* \end{pmatrix}$  as follows: we fix the number of responses to be  $p = 40$  and the number of covariates to be  $q = 10$ . We then generate a random Gaussian matrix  $W \in \mathbb{R}^{p \times k_u}$  and let  $L_y^* = WW'$ . We let  $D_y^* = \mu I_p$  where  $\mu$  is chosen to be twice the largest eigenvalue of  $L_y^*$ . We also generate random Gaussian matrices  $\tilde{W} \in \mathbb{R}^{p \times k_u}$  and  $Z \in \mathbb{R}^{q \times k_u}$  and let  $\Theta_{yx}^* = \tilde{W}Z'$ . Finally, we let  $\Theta_x^* = \Theta_{xy}^* \Theta_y^{*-1} \Theta_{yx}^* + \delta I_q$  where  $\delta$  is chosen large enough so that the overall matrix  $\Theta^*$  has a condition number that is relatively small (in our examples, we chose  $\delta$  so that the condition number is less than 20). This approach generates a composite factor model (1.3) with  $\text{rank}(\mathcal{A}) = k_x$  and  $\tilde{\mathcal{B}} \in \mathbb{R}^{p \times k_u}$ , and a factor model (1.1) with  $k = k_x + k_u$  latent factors. We obtain three models with  $(k_x, k_u) = (1, 1)$ ,  $(k_x, k_u) = (1, 2)$ , and  $(k_x, k_u) = 2$ . For each

model, we generate  $n$  samples of responses  $y$  and covariates  $x$ , and use these observations as input to convex programs (1.4) and (1.7). Figure 1(a) shows the probability of obtaining structurally correct estimates of the factor model parameters using (1.7) (i.e.  $\text{rank}(\hat{\hat{L}}_y) = \text{rank}(L_y^* + \Theta_{yx}^* \Theta_x^{-1} \Theta_{yx}^{* \prime})$ ) and Figure 1(b) shows the probability of obtaining structurally correct estimates of the composite factor model (i.e.  $\text{rank}(\hat{L}_y) = \text{rank}(L_y^*)$  and  $\text{rank}(\hat{\Theta}_{yx}^*) = \text{rank}(\hat{\Theta}_{yx})$ ). These probabilities are evaluated over 10 experiments for each value of  $n$ . These results agree with our theoretical results that given (sufficiently many) samples of responses/covariates, the convex programs (1.7) and (1.4) provide structurally correct estimates of a factor model (1.1) and composite factor model (1.3), respectively.

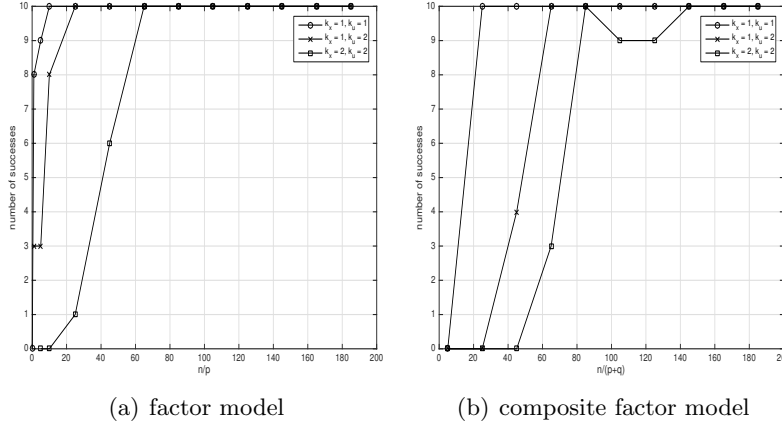


Figure 1: Synthetic data: plot shows probability of correct structure recovery in factor model and composite factor model. The three models studied are (i)  $(k_x, k_u) = (1, 1)$ , (ii)  $(k_x, k_u) = (2, 1)$ , and (iii)  $(k_x, k_u) = (2, 2)$ . For each plotted point, the probability of structurally correct estimation is obtained over 10 trials.

### 3 Identifying Latent Factors Influencing Stock Returns

Recall from Section 1.3 that our approach to associate semantics to latent variables in a factor model is a two-stage process. In the first stage, we identify a factor model based on observations of  $y$  using the convex relaxation (1.7), which results in an estimate  $(\hat{\hat{D}}_y, \hat{\hat{L}}_y)$ . In the second step, we use simultaneous observations of  $y$  and some additional covariates  $x$  to identify a composite factor model using the convex relaxation (1.4) with the resulting estimates being  $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ . As discussed in Section 2 the composite factor model of  $(y, x)$  offers an interpretation of the latent variables of the factor model underlying  $y$  if (i)  $\text{rank}(\hat{\hat{L}}_y) = \text{rank}(\hat{L}_y + \hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}_{xy})$ , (ii)  $\text{column-space}(\hat{\Theta}_{yx}) \cap \text{column-space}(\hat{L}_y) = \{0\}$ , and (iii)  $\max\{\|\hat{\hat{D}}_y - \hat{D}_y\|_2 / \|\hat{\hat{D}}_y\|_2, \|\hat{\hat{L}}_y - \hat{L}_y - \hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}_{xy}\|_2 / \|\hat{\hat{L}}_y\|_2\}$  is small. These observations naturally lead to the following algorithmic approach:

---

**Algorithm 1** Interpreting Latent Variables in a Factor Model

---

- 1: **Input:** A collection of observations  $\{(y^{(i)}, x^{(i)})\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}^q$  of the variables  $y$  and of some auxiliary covariates  $x$ .
  - 2: **Factor Modeling:** Supply observations  $\{y^{(i)}\}_{i=1}^n$  to the convex program (1.7) to learn a factor model with parameters  $(\hat{D}_y, \hat{L}_y)$ .
  - 3: **Composite Factor Modeling:** For each  $d = 1, \dots, q$ , sweep over parameters  $(\lambda_n, \gamma)$  in the convex program (1.4) (with  $\{y^{(i)}, x^{(i)}\}_{i=1}^n$  as input) to identify composite models with estimates  $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$  that satisfy the following three properties: (i)  $\text{rank}(\hat{\Theta}_{yx}) = d$ , (ii)  $\text{rank}(\hat{L}_y) = \text{rank}(\hat{L}_y) + \text{rank}(\hat{\Theta}_{yx})$ , and (iii)  $\text{rank}(\hat{L}_y) = \text{rank}(\hat{L}_y + \text{rank}(\hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}_{xy}))$ .
  - 4: **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  $\max\{\|\hat{D}_y - \hat{D}_y\|_2 / \|\hat{D}_y\|_2, \|\hat{L}_y - \hat{L}_y - \hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}_{xy}\|_2 / \|\hat{L}_y\|_2\}$ .
  - 5: **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.
- 

We note that in some cases, a factor model is obtained off-line based on scientific considerations. In these situations, one would proceed to step 3 of the algorithm. Typically, however, factor models are learned in a data-driven approach based on observations of responses  $\{y^{(i)}\}_{i=1}^n$ . We from step 2 of the algorithm that our approach to learn a factor model via the convex program (1.7) requires the specification of the regularization parameter  $\tilde{\lambda}_n$ . In our experimental results on the financial asset dataset, we choose this parameter via cross-validation techniques. Further, we note that the effectiveness of this algorithm in identifying semantics to latent variables in the input factor model is largely dependent on the size of the quantity  $\max\{\|\hat{D}_y - \hat{D}_y\|_2 / \|\hat{D}_y\|_2, \|\hat{L}_y - \hat{L}_y - \hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}_{xy}\|_2 / \|\hat{L}_y\|_2\}$ . Indeed, the smaller this quantity, the more confidence one should have that the composite factor model has attributed meaning to some component of the latent variables in the factor model.

### 3.1 Experimental Results on Financial Asset Data

We demonstrate the utility of our algorithmic approach in a financial asset data analysis problem. Specifically, we consider as our responses  $y$  the monthly stock return of  $p = 66$  companies from the Standard and Poor index over the period June 1990 to July 2014, which leads to a total of  $n = 277$  observations. We also obtain monthly observations over the same period of the following  $q = 11$  covariates, consumer price index, EUR to USD exchange rate, federal reserve rate, gold prices, industrial production, inflation rate, mortgage rate, oil exports, oil imports, treasury rate, and unemployment rate. These covariates were chosen because they plausibly influence the values of stock prices. For the purposes of our experiments, we set aside a random subset of  $n_{\text{train}} = 227$  of these observations as a training set and the remaining subset of  $n_{\text{test}} = 50$  as the test set. Similar to synthetic simulations in Section 2.3, we solve the convex program (1.4) and (1.7) numerically using the LogDetPPA package [20].

We begin by the second step of our algorithm which is to identify a factor model (1.1) that is well-suited for modeling stock returns. We find such a factor model by solving the convex program (1.7) where the regularization parameter  $\tilde{\lambda}_n$  is chosen via cross-validation. Concretely, for a particular choice of  $\tilde{\lambda}_n$ , we use the training set  $\{y_{\text{train}}^j\}_{j=1}^{222} \in \mathbb{R}^{67}$  as input to the convex

program (1.4), and solve (1.4) to obtain a factor model specified by  $(\hat{\hat{D}}_y, \hat{\hat{L}}_y)$ . We then compute the average log-likelihood over the testing set  $\{y_{\text{test}}^j\}_{j=1}^{50} \in \mathbb{R}^{67}$  using the distribution specified by the precision matrix  $\hat{\hat{D}}_y - \hat{\hat{L}}_y$ . We perform this procedure as we vary  $\tilde{\lambda}_n$  from 0.04 to 4 in increments of 0.004. Figure 2 shows a plot of  $\text{rank}(\hat{\hat{L}}_y)$  (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.  $\text{rank}(\hat{\hat{L}}_y)$  remains the same). As larger values of average log-likelihood are indicative of a better fit to test samples, these results suggest that 12 or 13 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 12 latent factors, and the factor model with the largest average log-likelihood performance that consists of 13 latent factors.

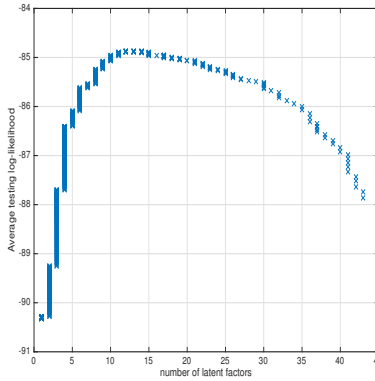


Figure 2: Number of latent factors vs. average log-likelihood over testing set. These results are obtained by sweeping over parameters  $\tilde{\lambda}_n \in [0.04, 4]$  in increments of 0.004 and solving the convex program (1.7)

We now proceed with the third step of our algorithm. Using joint observations of responses and covariates  $\{y_{\text{train}}^j, x_{\text{train}}^j\}_{j=1}^{227}$  as input to the convex program (1.4), we perform an exhaustive sweep over parameter space  $(\lambda_n, \gamma)$  to learn composite models with estimates  $(\hat{\hat{\Theta}}, \hat{\hat{D}}_y, \hat{\hat{L}}_y)$  such that  $\text{rank}(\hat{\hat{\Theta}}) = 1, 2, \dots, 11$ , and  $\text{rank}(\hat{\hat{L}}_y) = 1, 2, \dots, 12$ . As we are interested comparing these composite models to the factor model with 12 or 13 latent variables, we finely grid the parameter space  $(\lambda_n, \gamma)$  so that there are a large number of models for which  $\text{rank}(\hat{\hat{\Theta}}) + \text{rank}(\hat{\hat{L}}_y)$  is equal to 12 or 13. Among these models, we restrict to those that satisfy the conditions of step 3 of the algorithm. Table 1 shows the number of models that satisfy these conditions for  $\text{rank}(\hat{\hat{\Theta}}_{yx}) = 1, \dots, 5$ . For each  $d = 1, \dots, 11$ , we then identify the composite factor model which minimizes the quantity  $\max\{\|\hat{\hat{D}}_y - \hat{D}_y\|_2 / \|\hat{\hat{D}}_y\|_2, \|\hat{\hat{L}}_y - \hat{L}_y - \hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}'_{yx}\|_2 / \|\hat{\hat{L}}_y\|_2\}$ . Table 2 and Table 3 show the values of this quantity for  $\text{rank}(\hat{\hat{\Theta}}_{yx}) = 1, \dots, 5$  with respect to the factor model with 12 and 13 latent variables, respectively.

$(\text{rank}(\hat{\Theta}_{yx}), \text{rank}(\hat{L}_y))$	# models satisfying conditions of step 2.
(1,11)	261
(1,12)	174
(2,10)	84
(2,11)	126
(3,9)	112
(3,10)	84
(4,8)	144
(4,9)	72
(5,7)	4
(5,8)	64

Table 1: Number of composite factor models with  $\text{rank}(\hat{\Theta}_{yx}) = 1, \dots, 5$  that satisfy the requirements of step 2 in the algorithm description at the beginning of this section (for the factor model with 12 or 13 latent variables).

$(\text{rank}(\hat{\Theta}_{yx}), \text{rank}(\hat{L}_y))$	$\max\{\ \hat{D}_y - \hat{D}_y\ _2 / \ \hat{D}_y\ _2, \ \hat{L}_y - \hat{L}_y - \hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}'_{yx}\ _2 / \ \hat{L}_y\ _2\}$
(1,11)	0.08
(2,10)	0.17
(3,9)	0.26
(4,8)	0.31
(5,7)	0.43

Table 2: Deviation of the candidate composite factor model from the factor model consisting of 12 latent variables

$(\text{rank}(\hat{\Theta}_{yx}), \text{rank}(\hat{L}_y))$	$\max\{\ \hat{D}_y - \hat{D}_y\ _2 / \ \hat{D}_y\ _2, \ \hat{L}_y - \hat{L}_y - \hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}'_{yx}\ _2 / \ \hat{L}_y\ _2\}$
(1,12)	0.004
(2,11)	0.08
(3,10)	0.17
(4,9)	0.26
(5,8)	0.31

Table 3: Deviation of the candidate composite factor model from the factor model consisting of 13 latent variables

Focussing on the case corresponding to which we identified a 13-factor model underlying  $y$ , the results of Table 3 suggest that we identify a 2-dimensional interpretable component of the 13 latent variables as the deviation  $\max\{\|\hat{D}_y - \hat{D}_y\|_2 / \|\hat{D}_y\|_2, \|\hat{L}_y - \hat{L}_y - \hat{\Theta}_{yx} \hat{\Theta}_x^{-1} \hat{\Theta}'_{yx}\|_2 / \|\hat{L}_y\|_2\}$  on the right-hand-side of this table is small as long as  $\text{rank}(\hat{\Theta}_{yx}) = 1, 2$ . For  $\text{rank}(\hat{\Theta}_{yx}) = 3, 4, 5$ , the deviation appears to be quite large and may not lead to meaningful conclusions.

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 13-factor model. In particular, for the composite factor model with  $(\text{rank}(\hat{\Theta}_{yx}), \text{rank}(\hat{L}_y)) = (2, 11)$  (second row in Table 3), we let  $V \in \mathbb{R}^{11 \times 2}$  denote a matrix with orthogonal, unit-norm columns such that  $V$  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). Recall that  $Vx$  represents the component of the 13 latent variables that is interpretable via the covariates  $x$ . 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 4, 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, inflation rate, and oil imports seem to be the most relevant, and the covariates gold prices and oil exports seem to be the least relevant.

covariate	strength
CPI	0.07
exchange rate	0.18
federal reserve rate	0.06
gold prices	0.04
industrial production	0.09
inflation rate	0.15
mortgage rate	0.07
oil exports	0.02
oil imports	0.15
treasury rate	0.08
unemployment	0.07

Table 4: Strength of each covariate in the composite factor model with 2-dimensional projection of covariates and 11 latent variables

## 4 Proofs of Main Results

### 4.1 Proof Strategy

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  $\Theta_{yx}^*$  and  $L_y^*$ ) - 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 [21]. 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

$$\begin{aligned}
 (\hat{\Theta}, \hat{D}_y, \hat{L}_y) = \arg \min_{\substack{\Theta \in \mathbb{S}^{p+q}, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}} & -\ell(\Theta; \{y^{(i)}, x^{(i)}\}_{i=1}^n) + \lambda_n[\gamma \|\Theta_{yx}\|_* + \|L_y\|_*] \\
 \text{s.t.} & \Theta_y = D_y - L_y, D_y \text{ is diagonal}
 \end{aligned} \tag{4.1}$$

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 (6.5) reduces to the trace of  $L_y$ . We show in the appendix that with high probability, the matrix  $\tilde{L}_y$  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:

$$\begin{aligned} [\Sigma_n - \hat{\Theta}^{-1}]_y + \Lambda = 0 & \quad ; \quad [\Sigma_n - \hat{\Theta}^{-1}]_y \in \lambda_n \partial \|\hat{L}_y\|_* \\ [\Sigma_n - \hat{\Theta}^{-1}]_{yx} \in -\lambda_n \gamma \partial \|\hat{\Theta}_{yx}\|_* & \quad ; \quad [\Sigma_n - \hat{\Theta}^{-1}]_x = 0 \\ \hat{\Theta}_y = \hat{D}_y - \hat{L}_y; \quad \hat{D}_y \text{ is diagonal} & \quad ; \quad \Lambda_{i,i} = 0 \text{ for } i = 1, 2, \dots, p \end{aligned}$$

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:

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

where  $\mathcal{P}$  denote a projection operator. Let us denote the subspace  $W \in \mathbb{S}^p$  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  $\hat{\Theta}_{yx} = \check{U}\check{Q}\check{V}'$  respectively, and  $Z \triangleq (0, \lambda_n \bar{U}\bar{V}', -\lambda_n \gamma \check{U}\check{V}', 0)$ . Setting  $\Lambda = [\Sigma_n - \hat{\Theta}^{-1}]_{Y, \text{off diagonal}}$ , and letting  $\mathbb{H} = W \times T(\hat{L}_y) \times T(\hat{\Theta}_{yx}) \times \mathbb{S}^q$ , 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 (nonconvex) constraints that  $L_y$  and  $\Theta_{yx}$  belong to algebraic variety of low rank matrices specified by  $L_y^*$  and  $\Theta_{yx}^*$ . While this new program is nonconvex, it has a very interesting property that at the global optimal solution (and indeed at any locally optimal solution)  $\tilde{L}_y$  and  $\tilde{\Theta}_{yx}$  are smooth points of their respective algebraic varieties. This observation suggests that the Lagrange multipliers corresponding to the additional variety constraints belongs to  $T(\tilde{L}_y)^\perp$  and  $T(\tilde{\Theta}_{yx})^\perp$  respectively. We show under suitable conditions that  $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$  also satisfy the second optimality condition of (4.1) corresponding to the tangent spaces  $T(\tilde{L}_y)^\perp$  and  $T(\tilde{\Theta}_{yx})^\perp$ . Thus  $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$  is a unique solution of (1.4) and as constructed, is algebraically consistent (i.e.  $\text{rank}(\tilde{L}_y) = \text{rank}(L_y^*)$  and  $\text{rank}(\tilde{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^*)$ )

## 4.2 Results proved in the supplementary material

To ensure that the estimate  $\hat{\Theta}$  is close to the population quantity  $\Theta^*$ , the quantity  $E = \hat{\Theta} - \Theta^*$  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  $\Theta^{*-1}$ . Specifically, the Taylor series expansion of  $\hat{\Theta}^{-1}$  around  $\Theta^*$  is given by:

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

where,  $R_{\Sigma^*}(E) = \Sigma^* \left[ \sum_{k=2}^{\infty} (-E \Theta^*)^k \right]$ . Recalling that  $\mathbb{I}^* = \Theta^* \otimes \Theta^*$ , we note that  $\hat{\Theta}^{-1} - \Theta^{*-1} = \mathbb{I}^* E + R_{\Sigma^*}(E)$ . In Section 2, we imposed a set of conditions on  $\eta_1^*, \eta_2^*$  in (2.1) and (2.2) so that  $\mathbb{I}^*$  is globally well-conditioned, as well a condition on  $\eta_3^*(\omega)$  in (2.3) to address identifiability issues in the diagonal-minus-low-rank decomposition. These conditions allow us to control  $\mathbb{I}^*(E)$  when  $E$  is restricted to certain directions. We bound the remainder term  $R_{\Sigma^*}(E)$  in Proposition 4.1 where  $E$

is restricted to live in a certain space. Specifically, consider the following constrained optimization program:

$$\begin{aligned} (\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y) = & \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\operatorname{argmin}} & -\ell(\Theta; \{x^{(i)}, y^{(i)}\}_{i=1}^n) + \lambda_n [\|L_y\|_* + \gamma \|\Theta_{yx}\|_*] \\ \text{s.t.} & \Theta_y = D_y - L_y, (D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}' \end{aligned} \quad (4.2)$$

Here  $\mathbb{H}' = W \times T'_y \times T'_{yx} \times \mathbb{S}^q$ , where  $T'_y$  is a subspace in  $\mathbb{S}^p$ , and  $T'_{yx}$  is a subspace in  $\mathbb{R}^{p \times q}$ . Let  $\Delta = (\tilde{D}_y - D_y^*, \tilde{L}_y - L_y^*, \tilde{\Theta}_{yx} - \Theta_{yx}^*, \tilde{\Theta}_x - \Theta_x^*)$  denote the error in the estimated variables. In the following proposition, we bound the remainder term  $R_{\Sigma^*}(\mathcal{F}(\Delta))$  defined earlier. Before we proceed, we define the following norm on  $\mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$  that is useful in our analysis:

$$\Phi_\gamma(D_y, L_y, \Theta_{yx}, \Theta_x) = \max \left\{ \|D_y\|_2, \|L_y\|_2, \frac{1}{\gamma} \|\Theta_{yx}\|_2, \|\Theta_x\|_2 \right\}. \quad (4.3)$$

Notice this norm is a slight variant of the dual norm of the regularizer  $\|L_y\|_* + \gamma \|\Theta\|_*$  in (4.2).

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

Notice the bound on  $R_{\Sigma^*}(\mathcal{F}(\Delta))$  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. Specifically, suppose we let  $T'_y$  and  $T'_{yx}$  be tangent spaces to the low-rank matrix varieties and  $\rho(T'_y, T(L_y^*)) \leq \omega$ . Let  $E_n = \Sigma^* - \Sigma_n$  denote the difference between the true joint covariance and the sample covariance and let  $C_T = (0, \mathcal{P}_{T'_y \perp}(L_y^*), \mathcal{P}_{T'_{yx} \perp}(\Theta_{yx}^*), 0)$ . 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.** *Define:*

$$r = \max \left\{ \frac{4}{\alpha} \left( \Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} C_T] + \lambda_n \right), \Phi_\gamma[C_T] \right\} \quad (4.4)$$

*If  $r \leq \min\{\frac{1}{4C'}, \frac{\alpha}{16m\psi C'^2}\}$ , then  $\Phi_\gamma[\mathcal{F}^\dagger \Delta] \leq 2r$ .*

In the following proposition, we prove algebraic correctness of program (6.9). The statement theorem relies on the following constants:

$$C'_{\sigma_{yx}} = C_1^2 \psi^2 \max\left\{ 12\beta + \frac{6\beta}{\gamma}, \frac{2}{C_2 \psi^2} + \frac{6\beta}{\gamma} \right\}, \text{ and } C'_{samp} = \max\left\{ \frac{1}{48\psi\beta}, 4C_2 C', \frac{32m\psi C'^2 C_2}{\alpha}, 12\beta m\psi C'^2 C_2^2 \right\}.$$

**Proposition 4.3.** *Suppose  $\gamma$  is chosen in the range specified in Theorem 2.1 and  $\sigma_y \geq \frac{m}{\omega} C_{\sigma_y} \lambda_n$ ,  $\sigma_{yx} \geq m\gamma^2 C'_{\sigma_{yx}} \lambda_n$ . Further, suppose  $\lambda_n$  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 ( $k_u = \operatorname{rank}(L_y^*)$ ) and tangent space  $T'_{yx} \subset \mathbb{R}^{p \times q}$  in rank  $k_x$ -variety ( $k_x = \operatorname{rank}(\Theta_{yx}^*)$ ) where  $\rho(T'_y, T(L_y^*)) \leq \omega$  such that the corresponding solution  $(\hat{\Theta}, \hat{S}_y, \hat{L}_y)$  satisfies the following properties:*

1.  $\operatorname{rank}(\hat{L}_y) = \operatorname{rank}(L_y^*)$  and  $\operatorname{rank}(\hat{\Theta}_{yx}) = \operatorname{rank}(\Theta_{yx}^*)$
2. Letting  $C_T = (0, \mathcal{P}_{T'_y \perp}(L_y^*), \mathcal{P}_{T'_{yx} \perp}(\Theta_{yx}^*), 0)$ , we have that  $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F}(C_T)] \leq \frac{\lambda_n}{6\beta}$  and  $\Phi_\gamma[C_T] \leq \frac{16\alpha}{3\beta} \lambda_n$
3.  $\Phi_\gamma[\Delta] \leq 2C_1 \lambda_n$

Furthermore, suppose that  $\Phi_\gamma(A^\dagger E_n) \leq \frac{\lambda_n}{6\beta}$  and  $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*}(\mathcal{F}(\Delta))] \leq \frac{\lambda_n}{6\beta}$ . Then the tangent space constraints  $(D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}'$  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  $\Sigma_n$  converges to  $\Sigma^*$  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  $n \geq 4608\beta^2 m^2 \psi^2 C_{samp}^2 (p+q)$ , and the regularization parameter  $\lambda_n$  is chosen in the range specified by Theorem 1. Then, with probability greater than  $1 - 2\exp\left\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\right\}$ ,  $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{6\beta}$ .*

### 4.3 Proof of Theorem 1

Noting that  $C' \leq 4\psi\bar{m}$ , the constants  $C_{\sigma_{y'x'}}$  and  $C_{samp'}$  in Proposition 4.3 can be related to constants  $C_{\sigma_{yx}}$  and  $C_{samp}$  in Theorem 1 as follows:  $C_{\sigma_{y'x'}} \leq mC_{\sigma_{yx}}$  and  $C_{samp'} \leq m\bar{m}^2 C_{samp}$ . Using these relations, it is easy to check that the assumptions of Theorem 1 imply that the assumptions of Proposition 4.3 are satisfied. Thus we can conclude that the optimal solution  $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$  of (6.5) (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  $\Phi_\gamma(\mathcal{F}^\dagger E_n) \leq \frac{\lambda_n}{6\beta}$ . If we show that  $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*}(\Delta)] \leq \frac{\lambda_n}{6\beta}$ , then we conclude that the unique optimum  $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$  of the original convex program (1.4) has structurally correct structure (i.e.  $\text{rank}(\hat{L}_y) = \text{rank}(L_y^*)$  and  $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^*)$ ). To show that  $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*}(\Delta)] \leq \frac{\lambda_n}{6\beta}$ , we note that

$$\begin{aligned} \frac{4}{\alpha} \left( \Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} C_T] + \lambda_n \right) &\leq \frac{4}{\alpha} \left( \frac{\lambda_n}{6\beta} + \frac{\lambda_n}{6\beta} + \lambda_n \right) \leq \frac{16\alpha}{3\beta} \lambda_n \\ &\leq \min \left\{ \frac{1}{4C'}, \frac{\alpha}{16m\psi C'^2} \right\} \end{aligned}$$

Here, we used the bound on  $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} C_T]$  provided by Proposition 4.3 and the bound on  $\lambda_n$ . Furthermore, appealing to Proposition 4.3 once again, we have  $\Phi_\gamma[C_T] \leq \frac{16\alpha}{3\beta} \lambda_n \leq \min \left\{ \frac{1}{4C'}, \frac{\alpha}{16m\psi C'^2} \right\}$ . Thus Proposition 4.2 provides us with the bound  $\Phi_\gamma[\Delta] \leq \frac{32\alpha}{3\beta} \lambda_n \leq \frac{1}{2C'}$ . We subsequently apply the results of Proposition 4.1 to obtain:

$$\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*}(\mathcal{F}(\Delta))] \leq 2m\psi C'^2 \Phi_{\delta,\gamma}[\Delta]^2 \leq \left[ 2m\psi C'^2 \left( \frac{32\alpha}{3\beta} \right)^2 \lambda_n \right] \lambda_n \leq \frac{\lambda_n}{6\beta}$$

The last inequality follows from the bound on  $\lambda_n$ .

## 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] MESBAHI, M. & PAPAVALASSILOPOULOS, G. On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Transactions on Automatic Control*, 42:239-243, 1997.
- [12] NATARAJAN, B.K. Sparse approximate solutions to linear systems. *SIAM Journal of Computing*, 24:227-234, 1995.
- [13] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. & YU, B. High-dimensional covariance estimation by minimizing  $\ell_1$ -penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.
- [14] 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.
- [15] 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.
- [16] SHAPIRO, A. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, 47:187-199, 1982.
- [17] SHAPIRO, A. Weighted minimum trace factor analysis. *Psychometrika*, 47:243-264, 1982.

- [18] SHAPIRO, A. Identifiability of factor analysis: Some results and open problems. *Linear Algebra Applications*, 15:201-292, 1904.
- [19] SPEARMAN, C. ‘General intelligence’, objectively determined and measured. *American Journal of Psychology*, 15:201-292, 1904
- [20] 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/~mattokc/sdpt3.html>.
- [21] WAINWRIGHT, M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using  $\ell_1$ -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.
- [22] WAINWRIGHT, M. J. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and its Applications*. 1:233–253, 2014.

## 6 Supplementary Material

In the following Proposition, we appeal to conditions on  $\eta_1^*, \eta_2^*, \eta_3^*$  to prove a set of irrepresentability-type conditions on the population Fisher information  $\mathbb{I}^*$ .

**Proposition 6.1.** *Suppose that  $\eta_1^* \geq 3\alpha$ ,  $\eta_2^* \leq \frac{8\alpha}{3\beta}$  and  $\eta_3^*(\omega) \leq \frac{2\alpha}{\beta}$  (where these conditions were defined in Section 2.2). Further, suppose that the regularization parameter  $\gamma$  is chosen in the range specified in Theorem 2.1. Then we have that the following two conditions hold for  $\mathbb{H}' = W \times T'_y \times T'_{yx} \times \mathbb{S}^q \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$  where  $\rho(T(L_y^*), T'_y) \leq \omega$  and  $T'_{yx}$  is any subspace in  $\mathbb{R}^{p \times q}$ :*

1. *The minimum gain of  $\mathbb{I}^*$  restricted to  $\mathbb{H}'$  is bounded below:*

$$\min_{\substack{(D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}' \\ \Phi_\gamma(D_y, L_y, \Theta_{yx}, \Theta_x) = 1}} \Phi_\gamma[\mathcal{P}_{\mathbb{H}'} \mathcal{I}^\dagger \mathbb{I}^* \mathcal{I} \mathcal{P}_{\mathbb{H}'}(D_y, L_y, \Theta_{yx}, \Theta_x)] \geq \alpha \quad (6.1)$$

2. *The effect of elements in  $\mathbb{H}'$  on the orthogonal complement  $\mathbb{H}'^\perp$  is bounded above:*

$$\max_{Z \in \mathbb{H}'; \Phi_\gamma(Z) = 1} \Phi_\gamma[\mathcal{P}_{\mathbb{H}'^\perp} \mathcal{I}^\dagger \mathbb{I}^* \mathcal{I} \mathcal{P}_{\mathbb{H}'} (\mathcal{P}_{\mathbb{H}'} \mathcal{I}^\dagger \mathbb{I}^* \mathcal{I} \mathcal{P}_{\mathbb{H}'} )^{-1}(Z)] \leq 1 - \frac{3}{\beta + 1} \quad (6.2)$$

Although conditions (6.1) and (6.2) are satisfied for all subspaces  $T'_{yx} \subset \mathbb{R}^{p \times q}$ , we specialize these to tangent spaces of low-rank matrix variety. Conditions (6.1) and (6.2) 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  $\mathbb{I}^*$  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 and 2 are also similar in spirit to conditions employed in the analysis of convex relaxation methods for latent-variable graphical model selection [2].

*Proof.* First, consider an arbitrary subspace  $\mathcal{S} \in \mathbb{S}^{p+q}$ . Let  $M \in \mathbb{W}$  with  $\|M\|_2 = 1$ . Then,

$$\|\mathcal{P}_{\mathcal{S}} \mathbb{I}^* M\|_2 \geq \|\mathbb{I}^*(M)\|_2 - \|\mathcal{P}_{\mathbb{W}^\perp} \mathbb{I}^* \mathcal{P}_{\mathbb{W}}(M)\|_2 \geq 3\alpha - \alpha \left(1 - \frac{3}{\beta + 1}\right) \geq 2\alpha$$

In the subsequent discussion in this section, 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 first component of  $\mathbb{H}$  is given by  $\mathbb{H}[1] = H_1 \times \{0\} \times \{0\} \times \{0\} \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ . The restrictions  $\mathbb{H}[2], \mathbb{H}[3], \mathbb{H}[4]$  to the other components of  $\mathbb{H}$  are defined in an analogous manner. Let  $\mathbb{H}' = W \times T'_y \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$  with  $\rho(T'_y, T(L_y^*)) \leq \omega$ . Recall, the subspace  $W$  is the set of diagonal matrices with nonnegative entries. Consider a set of variables  $(D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}'$  with  $\|D_y\|_2 \leq 1$ ,  $\|L_y\|_2 \leq 1$ ,  $\|\Theta_{yx}\|_2 \leq \gamma$ , and  $\|\Theta_x\|_2 \leq 1$ . Suppose equality holds in at least one of these set of inequalities so that  $\Phi_\gamma(D_y, L_y, \Theta_{yx}, \Theta_x) = 1$ . Then, at least one of the following cases is active (the following results use conditions on  $\eta_1^*, \eta_2^*$  and  $\eta_3^*(\omega)$  and Proposition 1 (main paper)) :

1. If  $\|D_y\|_2 = 1$ , then

$$\begin{aligned} \|\mathcal{P}_{\mathbb{H}'[1]}\mathbb{I}^*\mathcal{F}(D_y, L_y, \Theta_{yx}, \Theta_x)\|_2 &\geq \left[ \|\mathcal{P}_W\mathbb{I}_y^*(D_y)\|_2 - \|\mathcal{P}_W\mathbb{I}_y^*(L)\|_2 \right. \\ &\quad \left. - \|\mathcal{P}_{\mathbb{H}'[1]}\mathbb{I}^*\mathcal{F}(0, 0, \Theta_{yx}, \Theta_x)\|_2 \right] \geq 2\alpha - \eta_3^* - 2\eta_2^* \max\{\gamma, 1\} \\ &\geq 2\alpha - \frac{8\alpha}{\beta} \end{aligned}$$

2. If  $\|L_y\|_2 = 1$ , then

$$\begin{aligned} \|\mathcal{P}_{\mathbb{H}'[2]}\mathbb{I}^*\mathcal{F}(D_y, L_y, \Theta_{yx}, \Theta_x)\|_2 &\geq \left[ \|\mathcal{P}_{T_y'}\mathbb{I}_y^*(L_y)\|_2 - \|\mathcal{P}_{T_y'}\mathbb{I}_y^*(D_y)\|_2 \right. \\ &\quad \left. - \|\mathcal{P}_{\mathbb{H}'[2]}\mathbb{I}^*\mathcal{F}(0, 0, \Theta_{yx}, \Theta_x)\|_2 \right] \geq 2\alpha - \eta_3^* - 2\eta_2^* \max\{\gamma, 1\} \\ &\geq 2\alpha - \frac{8\alpha}{\beta} \end{aligned}$$

Similarly, one can show

3. If  $\|\Theta_{yx}\|_2 = \gamma$ , then  $\frac{1}{\gamma}\|\mathcal{P}_{\mathbb{H}'[3]}\mathbb{I}^*\mathcal{F}(D_y, L_y, \Theta_{yx}, \Theta_x)\|_2 \geq 2\alpha - \frac{8\alpha}{\beta}$

4. If  $\|\Theta_x\|_2 = 1$ , then  $\|\mathcal{P}_{\mathbb{H}'[4]}\mathbb{I}^*\mathcal{F}(D_y, L_y, \Theta_{yx}, \Theta_x)\|_2 \geq 2\alpha - \frac{8\alpha}{\beta}$

Combining these results, one can conclude that

$$\Phi_\gamma[\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^*\mathcal{A}\mathcal{P}_{\mathbb{H}'}(D_y, L_y, \Theta_{yx}, \Theta_x)] \geq 2\alpha - \frac{8\alpha}{\beta} \geq \alpha \quad (6.3)$$

Using a similar decoupling technique, it is easy to check that:

$$\Phi_\gamma\left[\mathcal{P}_{\mathbb{H}'^\perp}[\mathcal{F}^\dagger\mathbb{I}^*\mathcal{F}(D_y, L_y, \Theta_{yx}, \Theta_x)]\right] \leq \eta_2^* + \frac{8\alpha}{\beta} \leq \frac{8\alpha}{3\beta} + \frac{8\alpha}{\beta} \quad (6.4)$$

Letting  $Z = (D_y, L_y, \Theta_{yx}, \Theta_x)$ , we use (6.3) and (6.4) to conclude:

$$\Phi_\gamma[\mathcal{P}_{\mathbb{H}'^\perp}\mathcal{A}^\dagger\mathbb{I}^*\mathcal{A}\mathcal{P}_{\mathbb{H}'}(\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^*\mathcal{A}\mathcal{P}_{\mathbb{H}'}^{-1})(Z)] \leq \frac{\frac{8\alpha}{3\beta} + \frac{8\alpha}{\beta}}{2\alpha - \frac{8\alpha}{\beta}} \leq 1 - \frac{3}{1 + \beta}$$

## 6.1 Proof of main paper Proposition 1 – bounding curvature of the matrix inverse

Let  $(\tilde{D}_y, \tilde{L}_y, \tilde{\Theta}_{yx}, \tilde{\Theta}_x)$  be an estimate for the population quantities  $(D_y^*, L_y^*, \Theta_{yx}^*, \Theta_x^*)$ , and let  $\Delta = (\tilde{D}_y - D_y^*, \tilde{L}_y - L_y^*, \tilde{\Theta}_{yx} - \Theta_{yx}^*, \tilde{\Theta}_x - \Theta_x^*) \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ , recall that the Taylor expansion of the inverse of matrix perturbation is specified by:

$$\tilde{\Theta}^{-1} = (\Theta^* + \mathcal{F}(\Delta))^{-1} = \Theta^{*-1} + \Theta^{*-1}\mathcal{F}(\Delta)\Theta^{*-1} + R_{\Sigma^*}(\mathcal{F}(\Delta))$$

where,

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

The following proposition provides a bound on this second order term:

**Proposition 6.2.** *If  $\Phi_\gamma[\Delta] \leq \frac{1}{2C^\gamma}$ , then  $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*}(\mathcal{F}(\Delta))] \leq 2m\psi C'^2 \Phi_\gamma[\Delta]^2$ .*

*Proof.* We note that:

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

Using this observation and some algebra, we have that:

$$\begin{aligned} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*}(\mathcal{F}(\Delta))] &\leq m\psi \left[ \sum_{k=2}^{\infty} (\psi \|\Delta\|_2)^k \right] \leq m\psi^3 \frac{(3 + \gamma)^2 \Phi_\gamma[\Delta]^2}{1 - (3 + \gamma)\Phi_\gamma[\Delta]\psi} \\ &\leq 2m\psi C'^2 \Phi_\gamma[\Delta]^2 \end{aligned}$$

□

## 6.2 Proof of main paper Proposition 2

Next, we analyze the following convex program subject to certain additional tangent space constraints:

$$\begin{aligned} (\tilde{\Theta}, \tilde{S}_y, \tilde{L}_y) = \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\operatorname{argmin}} & -\ell(\Theta; \{X^{(i)}, Y^{(i)}\}_{i=1}^n) + \lambda_n [\|L_y\|_* + \gamma \|\Theta_{yx}\|_*] \\ \text{s.t.} & \quad \Theta_y = D_y - L_y, (D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}' \end{aligned} \quad (6.5)$$

where  $\mathbb{H}' = W \times T'_y \times T'_{yx} \times \mathbb{S}^q$  and  $T'_y \subset \mathbb{S}^p$  and  $T'_{yx} \subset \mathbb{R}^{p \times q}$  are subspaces. In the following proposition, we show that if  $T'_y$  and  $T'_{yx}$  are tangent spaces with respect to the variety of low-rank matrices with  $\rho(T'_y, T(L_y^*)) \leq \omega$ , then we can bound the error  $\Delta = (\tilde{D}_y - D_y^*, \tilde{L}_y - L_y^*, \tilde{\Theta}_{yx} - \Theta_{yx}^*, \tilde{\Theta}_x - \Theta_x^*)$ . We denote  $E_n = \Sigma^* - \Sigma_n$  as the difference between the population covariance matrix and the sample covariance matrix. Further,  $C_T = (0, \mathcal{P}_{T'_y \perp}(L_y^*), \mathcal{P}_{T'_{yx} \perp}(\Theta_{yx}^*), 0)$ . □

**Proposition 6.3.** *Finally, define:*

$$r = \max \left\{ \frac{4}{\alpha} \left( \Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} C_T] + \lambda_n \right), \Phi_\gamma[C_T] \right\} \quad (6.6)$$

*If  $r \leq \min\{\frac{1}{4C^\gamma}, \frac{\alpha}{16m\psi C'^2}\}$ , then  $\Phi_\gamma[\Delta] \leq 2r$ .*

*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 (6.5) suggest that there exist Lagrange multipliers  $Q_{D_y} \in W$ ,  $Q_{T_y} \in T'_y \perp$ , and  $Q_{T_{yx}} \in T'_{yx} \perp$  such that

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

Letting the SVD decomposition of  $\tilde{L}$  and  $\tilde{\Theta}_{yx}$  be given by  $\tilde{L}_y = \bar{U} \bar{D} \bar{V}'$  and  $\tilde{\Theta}_{yx} = \check{U} \check{D} \check{V}'$  respectively, and  $Z \triangleq (0, \lambda_n \bar{U} \bar{V}', -\lambda_n \gamma \check{U} \check{V}', 0)$ , we can restrict the optimality conditions 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  $\mathcal{P}_{\mathbb{H}\mathcal{M}} \mathcal{F}^\dagger(E_n - R_{\Sigma^*}(\Delta) + \mathbb{I}^* \mathcal{F}(\Delta)) = Z$ . Based on the Fisher information condition (6.1), the optimum of (6.5) 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. We now appeal to Brouwer's fixed-point theorem to bound  $\Phi_\gamma[\Delta]$ . Consider the following function  $G(\underline{\delta})$  restricted to  $\underline{\delta} \in W \times T'_y \times T'_{yx} \times \mathbb{S}^q$  with  $\rho(T(L_y^*), T'_y) \leq \omega$ :

$$G(\underline{\delta}) = \underline{\delta} - (\mathcal{P}_{\mathbb{H}} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}})^{-1} \left( \mathcal{P}_{\mathbb{H}} \mathcal{F}^\dagger [E_n - R_{\Sigma^*} \mathcal{F}(\underline{\delta} + C_T) + \mathbb{I}^* \mathcal{F}(\underline{\delta} + C_T)] - Z \right)$$

The function  $G(\underline{\delta})$  is well-defined since the operator  $\mathcal{P}_{\mathbb{H}'} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}'}$  is bijective due to Fisher information condition 1 in Proposition 2 (main paper). As a result,  $\underline{\delta}$  is a fixed point of  $G(\underline{\delta})$  if and only if  $\mathcal{P}_{\mathbb{H}'} \mathcal{F}^\dagger [E_n - R_{\Sigma^*} (\mathcal{F}(\underline{\delta} + C_T)) + \mathbb{I}^* \mathcal{F}(\underline{\delta} + C_T)] = Z$ . Since the pair  $(\tilde{\Theta}, \tilde{S}_y, \tilde{L}_y)$  are the unique solution to (6.5), the only fixed point of  $G$  is  $\mathcal{P}_{\mathbb{H}'}[\Delta]$ . Next we show that this unique optimum lives inside the ball  $\mathbb{B}_r = \{\underline{\delta} \mid \Phi_\gamma(\underline{\delta}) \leq r, \underline{\delta} \in \mathbb{H}'\}$ . In particular, we show that under the map  $G$ , the image of  $\mathbb{B}_r$  lies in  $\mathbb{B}_r$  and appeal to Brouwer's fixed point theorem to conclude that  $\mathcal{P}_{\mathbb{H}'}[\Delta] \in \mathbb{B}_r$ . For  $\underline{\delta} \in \mathbb{B}_r$ ,  $\Phi_\gamma[G(\underline{\delta})]$  can be bounded as follows:

$$\begin{aligned} \Phi_\gamma[G(\underline{\delta})] &= \Phi_\gamma \left[ (\mathcal{P}_{\mathbb{H}'} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}'})^{-1} \left( \mathcal{P}_{\mathbb{H}, \mathcal{M}} \mathcal{F}^\dagger [E_n - R_{\Sigma^*} (\mathcal{F}(\underline{\delta} + C_T)) + \mathbb{I}^* \mathcal{F} C_T] - Z \right) \right] \\ &\leq \frac{2}{\alpha} \left[ \Phi_\gamma[\mathcal{F}^\dagger (E_n + \mathbb{I}^* \mathcal{F} C_T)] + \lambda_n \right] \\ &+ \frac{2}{\alpha} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*} (\underline{\delta} + C_T)] \leq \frac{r}{2} + \frac{2}{\alpha} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*} (\underline{\delta} + C_T)] \end{aligned}$$

The first inequality holds because of Fisher information condition (6.1), and the properties that  $\Phi_\gamma[\mathcal{P}_{\mathbb{H}, \mathcal{M}}(\cdot)] \leq 2\Phi_\gamma(\cdot)$  (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  $r \leq \frac{1}{4C_T}$ , we have  $\Phi_\gamma(\underline{\delta} + C_T) \leq \Phi_\gamma(\underline{\delta}) + \Phi_\gamma(C_T) \leq 2r \leq \frac{1}{2C_T}$ . We can now appeal to Proposition 1 to obtain:

$$\frac{2}{\alpha} \Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*} (\underline{\delta} + C_T)] \leq \frac{4}{\alpha} m \psi C'^2 [\Phi_\gamma(\underline{\delta} + C_T)]^2 \leq \frac{r}{2}$$

Thus, we conclude that  $\Phi_\gamma[G(\underline{\delta})] \leq r$  and by Brouwer's fixed-point theorem,  $\Phi_\gamma[\mathcal{P}_{\mathbb{H}, \mathcal{M}}(\Delta)] \leq r$ . Furthermore,  $\Phi_\gamma[\Delta] \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}'}(\Delta)] + \Phi_\gamma(C_T) \leq 2r$   $\square$

### 6.3 Proof of main paper Proposition 4 - bounding deviation of population covariance matrix and sample covariance matrix

**Proposition 6.4.** *Suppose that the number of observed samples obeys  $n \geq 4608\beta^2 m^2 \psi^2 C_{\text{samp}}'^2 (p+q)$ , and the regularization parameter  $\lambda_n$  is chosen in the range specified by Theorem 1 (main paper). Then, with probability greater than  $1 - 2\exp\left\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\right\}$ ,  $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{6\beta}$ .*

*Proof.* First, note that  $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq m \|\Sigma_n - \Sigma^*\|_2$ . Using the results in [3] and the fact that  $\frac{\lambda_n}{6\beta} \leq 8\psi$  and  $n \geq \frac{2304(p+q)m^2\psi^2}{\lambda_n^2}$ , the following bound holds:  $\Pr[m \|\Sigma_n - \Sigma^*\|_2 \geq \frac{\lambda_n}{6\beta}] \leq 2\exp\left\{-\frac{n\lambda_n^2}{4608m^2\psi^2}\right\}$ . Thus,  $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{6\beta}$  with probability greater than  $1 - 2\exp\left\{-\frac{n\lambda_n^2}{4608\beta^2 m^2 \psi^2}\right\}$ .  $\square$

## 6.4 Proof of the main paper Proposition 3

Consider the following convex optimization program:

$$\begin{aligned}
(\bar{\Theta}, \bar{D}_y, \bar{L}_y) = & \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\operatorname{argmin}} & -\ell(\Theta; \{X^{(i)}, Y^{(i)}\}_{i=1}^n) + \lambda_n [\|L_y\|_* + \gamma \|\Theta_{yx}\|_*] \\
\text{s.t.} & & \Theta_y = D_y - L_y; D_y \text{ is diagonal}
\end{aligned} \tag{6.7}$$

Comparing (6.7) with the convex program (1.4) (main paper), 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 (6.7) reduces to the trace of  $L_y$ . We show that the unique optimum  $(\bar{\Theta}, \bar{S}_y, \bar{L}_y)$  of (6.7) has the property that with high probability,  $\bar{L}_y$  is positive semidefinite. As a result, with high probability, the variables  $(\bar{\Theta}, \bar{S}_y, \bar{L}_y)$  are also the optimum of (1.4). Below, we outline our proof strategy:

1. We proceed by analyzing (6.7) with additional constraints that the variables  $L_y$ , and  $\Theta_{yx}$  belong to the algebraic varieties low-rank matrices (specified by rank of  $L_y^*$ , and  $\Theta_{yx}^*$ ), and that the tangent spaces  $T(L_y)$ ,  $T(\Theta_{yx})$  are close to the nominal tangent spaces  $T(L_y^*)$ , and  $T(\Theta_{yx}^*)$  respectively. We prove that under suitable conditions on the minimum nonzero singular value of  $L_y^*$ , and minimum nonzero singular value of  $\Theta_{yx}^*$ , any optimum pair of variables  $(\Theta, D_y, L_y)$  of this non-convex program are smooth points of the underlying varieties; that is  $\operatorname{rank}(L_y) = \operatorname{rank}(L_y^*)$  and  $\operatorname{rank}(\Theta_{yx}) = \operatorname{rank}(\Theta_{yx}^*)$ . Further, we show that  $L_y$  has the same inertia as  $L_y^*$  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 suitable conditions on the regularization parameter  $\lambda_n$ , 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 (6.7) has the property that with high probability:  $\operatorname{rank}(\bar{L}_y) = \operatorname{rank}(L_y^*)$  and  $\operatorname{rank}(\bar{\Theta}_{yx}) = \operatorname{rank}(\Theta_{yx}^*)$ . 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).

### 6.4.1 Variety Constrained Program

We begin by considering a variety-constrained optimization program:

$$\begin{aligned}
(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}}) = & \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\operatorname{argmin}} & -\ell(\Theta; \{X^{(i)}, Y^{(i)}\}_{i=1}^n) + \lambda_n [\|L_y\|_* + \gamma \|\Theta_{yx}\|_*] \\
\text{s.t.} & & \Theta_y = D_y - L_y, (\Theta, D_y, L_y) \in \mathcal{M}.
\end{aligned} \tag{6.8}$$

Here, the set  $\mathcal{M}$  is given by:

$$\begin{aligned}
\mathcal{M} \triangleq & \left\{ (\Theta, D_y, L_y) \in \mathbb{S}^{(p+q)} \times \mathbb{S}^p \times \mathbb{S}^p \mid D_y \text{ is diagonal, } \operatorname{rank}(L_y) \leq \operatorname{rank}(L_y^*) \right. \\
& \operatorname{rank}(\Theta_{yx}) \leq \operatorname{rank}(\Theta_{yx}^*); \|\mathcal{P}_{T(L_y^*)^\perp}(L_y - L_y^*)\|_2 \leq \frac{\lambda_n}{2m\psi^2} \\
& \left. \|\mathcal{P}_{T(\Theta_{yx}^*)^\perp}(\Theta_{yx} - \Theta_{yx}^*)\|_2 \leq \frac{\lambda_n}{2m\psi^2}; \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \Delta] \leq 5\lambda_n \right\}
\end{aligned}$$

The optimization program (6.8) is non-convex due to the rank constraints  $\text{rank}(L_y) \leq \text{rank}(L_y^*)$  and  $\text{rank}(\Theta_{yx}) \leq \text{rank}(\Theta_{yx}^*)$  in the set  $\mathcal{M}$ . These constraints ensure that the matrices  $L_y$ , and  $\Theta_{yx}$  belong to appropriate varieties. The constraints in  $\mathcal{M}$  along  $T(L_y^*)^\perp$  and  $T(\Theta_{yx}^*)^\perp$  ensure that the tangent spaces  $T(L_y)$  and  $T(\Theta_{yx})$  are “close” to  $T(L_y^*)$  and  $T(\Theta_{yx}^*)$  respectively. Finally, the last condition roughly controls the error. We begin by proving the following useful proposition:

**Proposition 6.5.** *Let  $(\Theta, D_y, L_y)$  be a set of feasible variables of (6.8). Let  $\Delta = (D_y - D_y^*, L_y - L_y^*, \Theta_{yx} - \Theta_{yx}^*, \Theta_x - \Theta_x^*)$  and  $C_1 = \frac{12}{\alpha} + \frac{1}{\psi^2}$ . Then,  $\Phi_\gamma[\Delta] \leq C_1 \lambda_n$*

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

$$\begin{aligned} \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}^*}(\Delta)] &\leq \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F}(\Delta)] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}^* \perp}(\Delta)] \\ &\leq 5\lambda_n + m\psi^2 \left( \frac{\omega\lambda_n}{2m\psi^2} + \frac{\omega\lambda_n}{2m\psi^2} \right) \leq 6\lambda_n \end{aligned}$$

Since  $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^*}(\cdot)] \leq 2\Phi_\gamma(\cdot)$ , we have that  $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^*} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}^*}(\Delta)] \leq 12\lambda_n$ . Consequently, we apply the Fisher information condition (6.1) to conclude that  $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^*}(\Delta)] \leq \frac{12\lambda_n}{\alpha}$ . Moreover:

$$\Phi_\gamma[\Delta] \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}^*}(\Delta)] + \Phi_\gamma[\mathcal{P}_{\mathbb{H}^* \perp}(\Delta)] \leq \frac{12\lambda_n}{\alpha} + \frac{\lambda_n}{\psi^2} = C_1 \lambda_n$$

□

Proposition 6.5 leads to powerful implications. In particular, under additional conditions on the minimum nonzero singular values of  $L_y^*$  and  $\Theta_{yx}^*$ , any feasible set of variables  $(\Theta, D_y, L_y)$  of (6.8) has two key properties: (a) The variables  $(\Theta_{yx}, L_y)$  are smooth points of the underlying varieties, (b) The constraints in  $\mathcal{M}$  along  $T(L_y^*)^\perp$  and  $T(\Theta_{yx}^*)^\perp$  are locally inactive at  $\Theta_{yx}$  and  $L_y$ . These properties, among others, are proved in the following corollary.

**Corollary 6.6.** *Consider any feasible variables  $(\Theta, D_y, L_y)$  of (6.8). Let  $\sigma_y$  be the smallest nonzero singular value of  $L_y^*$  and  $\sigma_{yx}$  be the smallest nonzero singular value of  $\Theta_{yx}^*$ . Let  $\mathbb{H}' = W \times T(L_y) \times T(\Theta_{yx}) \times \mathbb{S}^q$  and  $C_{T'} = \mathcal{P}_{\mathbb{H}' \perp}(0, L_y^*, \Theta_{yx}^*, 0)$ . Furthermore, let  $C_1 = \frac{12}{\alpha} + \frac{1}{\psi^2}$ ,  $C_2 = \frac{4}{\alpha}(1 + \frac{1}{3\beta})$ ,  $C_{\sigma_y} = C_1^2 \psi^2 \max\{12\beta + 1, \frac{2}{C_2 \psi^2} + 1\}$  and  $C'_{\sigma_{yx}} = C_1^2 \psi^2 \max\{12\beta + \frac{6\beta}{\gamma}, \frac{2}{C_2 \psi^2} + \frac{6\beta}{\gamma}\}$ . Suppose that the following inequalities are met:  $\sigma_y \geq \frac{m}{\omega} C_{\sigma_y} \lambda_n$ ,  $\sigma_{yx} \geq m\gamma^2 C'_{\sigma_{yx}} \lambda_n$ . Then,*

1.  $L_y$  and  $\Theta_{yx}$  are smooth points of their underlying varieties, i.e.  $\text{rank}(L_y) = \text{rank}(L_y^*)$ ,  $\text{rank}(\Theta_{yx}) = \text{rank}(\Theta_{yx}^*)$ ; Moreover  $L_y$  has the same inertia as  $L_y^*$ .
2.  $\|\mathcal{P}_{T(L_y^*)^\perp}(L_y - L_y^*)\|_2 \leq \frac{\lambda_n \omega}{48m\psi^2}$  and  $\|\mathcal{P}_{T(\Theta_{yx}^*)^\perp}(\Theta_{yx} - \Theta_{yx}^*)\|_2 \leq \frac{\lambda_n \omega}{48m\psi^2}$
3.  $\rho(T(L_y), T(L_y^*)) \leq \omega$ ; that is, the tangent spaces at  $L_y$  is “close” to the tangent spaces  $L_y^*$
4.  $\Phi_\gamma[C_{T'}] \leq \min\{\frac{\lambda_n}{6\beta\psi^2}, C_2 \lambda_n\}$

*Proof.* We note the following relations before proving each step:  $C_1 \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2}$ ,  $\omega \in (0, 1)$ , and  $\beta \geq 8$ . 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  $L_y^*$  and  $\Theta_{yx}^*$ , one can check that:

$$\begin{aligned}\sigma_y &\geq \frac{C_1^2 \lambda_n}{\omega} m \psi^2 (12\beta + 1) \geq \frac{C_1 \lambda_n}{\omega} (12\beta + 1) \geq 8 \|L - L_y^*\|_2 \\ \sigma_{yx} &\geq C_1^2 \lambda_n \gamma^2 m \psi^2 \left( \frac{6\beta}{\gamma} + 12\beta \right) \geq 8 \|\Theta_{yx} - \Theta_{yx}^*\|_2\end{aligned}$$

Combining these results and Proposition 6.5, we conclude that  $L_y$  and  $\Theta_{yx}$  are smooth points of their respective varieties, i.e.  $\text{rank}(L_y) = \text{rank}(L_y^*)$ , and  $\text{rank}(\Theta_{yx}) = \text{rank}(\Theta_{yx}^*)$ . Furthermore,  $L_y$  has the same inertia as  $L_y^*$ .

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

$$\begin{aligned}\|\mathcal{P}_{T(L_y^*)^\perp}(L_y - L_y^*)\|_2 &\leq \frac{\|L_y - L_y^*\|_2^2}{\sigma_y} \leq \frac{\lambda_n}{48m\psi^2} \\ \|\mathcal{P}_{T(\Theta_{yx}^*)^\perp}(\Theta_{yx} - \Theta_{yx}^*)\|_2 &\leq \frac{\|\Theta_{yx} - \Theta_{yx}^*\|_2^2}{\sigma_{yx}} \leq \frac{\lambda_n}{48m\psi^2}\end{aligned}$$

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

$$\rho(T(L_y), T(L_y^*)) \leq \frac{2\|L_y - L_y^*\|_2}{\sigma_y} \leq \frac{2C_1 \lambda_n \omega}{C_1^2 \lambda_n m \psi^2 (12\beta + 1)} \leq \omega$$

4. Letting  $\sigma'_y$  and  $\sigma'_{yx}$  be the minimum nonzero singular value of  $L$  and  $\Theta_{yx}$  respectively, one can check that:

$$\begin{aligned}\sigma'_y &\geq \sigma_y - \|L_y - L_y^*\|_2 \geq 8C_1 \lambda_n \geq 8\|L_y - L_y^*\|_2 \\ \sigma'_{yx} &\geq \sigma_{yx} - \|\Theta_{yx} - \Theta_{yx}^*\|_2 \geq 8C_1 \lambda_n \gamma \geq 8\|\Theta_{yx} - \Theta_{yx}^*\|_2\end{aligned}$$

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

$$\begin{aligned}\Phi_\gamma(C_{T'}) &\leq m \|\mathcal{P}_{T(L_y^*)^\perp}(L_y - L_y^*)\|_2 + m \|\mathcal{P}_{T(\Theta_{yx}^*)^\perp}(\Theta_{yx} - \Theta_{yx}^*)\|_2 \\ &\leq m \frac{\|L_y - L_y^*\|_2^2}{\sigma'_y} + m \frac{\|\Theta_{yx} - \Theta_{yx}^*\|_2^2}{\sigma'_{yx}} \leq \min\left\{ \frac{\lambda_n}{6\beta\psi^2}, C_2 \lambda_n \right\}\end{aligned}$$

□

#### 6.4.2 Variety Constrained Program to Tangent Space Constrained Program

Consider any optimal solution  $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}})$  of (6.8). In Corollary 6.6, 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  $\text{rank}(L_y) \leq \text{rank}(L_y^*)$  and  $\text{rank}(\Theta_{yx}) \leq \text{rank}(\Theta_{yx}^*)$  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 (6.8). Moreover, we once more appeal to Corollary 6.6 to conclude that the constraints in  $\mathcal{M}$  along  $\mathcal{P}_{T(L_y^*)^\perp}$  and  $\mathcal{P}_{T(\Theta_{yx}^*)^\perp}$  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. Finally, we claim that the constraint  $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \Delta] \leq 5\lambda_n$  in (6.8) can also be removed in this ‘‘linearized’’ convex program. In particular, letting  $\mathbb{H}_\mathcal{M} \triangleq W \times T(L_y^\mathcal{M}) \times T(\Theta_{yx}^\mathcal{M}) \times \mathbb{S}^q$ , consider the following convex optimization program with the constraint  $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \Delta] \leq 5\lambda_n$  removed :

$$\begin{aligned} (\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y) = & \underset{\substack{\Theta \in \mathbb{S}^{q+p}, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\operatorname{argmin}} & -\ell(\Theta; \{X^{(i)}, Y^{(i)}\}_{i=1}^n) + \lambda_n [\|L_y\|_* + \gamma \|\Theta_{yx}\|_*] \\ \text{s.t.} & \Theta_y = D_y - L_y, (D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}_\mathcal{M} \end{aligned} \quad (6.9)$$

We prove that under conditions imposed on the regularization parameter  $\lambda_n$ , the pair of variables  $(\Theta^\mathcal{M}, D_y^\mathcal{M}, L_y^\mathcal{M})$  is the unique optimum of (6.9). That is, we show that  $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} (\tilde{S}_y - D_y^*, \tilde{L}_y - L_y^*, \tilde{\Theta}_{yx} - \Theta_{yx}, \tilde{\Theta}_x - \Theta_x^*)] < 5\lambda_n$ . Appealing to Corollary 6.6 and Proposition 6.4, we have that  $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} C_{T_\mathcal{M}}] \leq \frac{\lambda_n}{6\beta}$ ,  $\Phi_\gamma[C_{T_\mathcal{M}}] \leq C_2\lambda_n$  and (with high probability)  $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{6\beta}$ . Consequently, based on the bound on  $\lambda_n$  in assumption of Theorem 4.3, it is straightforward to show that  $r \leq \min\{\frac{1}{4C'}\}$ ,  $\frac{\alpha}{16m\psi C'^2}$ . Hence by Proposition 6.3,  $\Phi_\gamma[\Delta] \leq \frac{1}{2C'}$ . Finally, we can appeal to Proposition 6.2 and the bound on  $\lambda_n$  to conclude  $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^*}(\mathcal{F}(\Delta))] \leq 2m\psi C'^2 \Phi_\gamma[\Delta]^2 \leq 2m\psi C'^2 C_1^2 \lambda_n^2 \leq \frac{\lambda_n}{6\beta}$ . Based on the optimality condition of (6.9), the property that  $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}(\cdot)] \leq 2\Phi_\gamma(\cdot)$ , and the fact that  $\beta \geq 8$ , we have:

$$\begin{aligned} \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}_\mathcal{M}}(\Delta)] & \leq 2\lambda_n + \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}} \mathcal{F}^\dagger R_{\Sigma^*}(\Delta)] + \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} C_{T_\mathcal{M}}] \\ & \quad + \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}} \mathcal{F}^\dagger E_n] \\ & \leq 2\lambda_n + 2\lambda_n \left( \frac{1}{6\beta} + \frac{1}{6\beta} + \frac{1}{6\beta} \right) \leq 2\lambda_n + \frac{\lambda_n}{\beta} \leq \frac{17\lambda_n}{8} \end{aligned}$$

Furthermore, by appealing to Fisher information condition (6.2), we have:

$$\begin{aligned} \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F}(\Delta)] & \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}_\mathcal{M}}(\Delta)] + \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}_\mathcal{M}}(\Delta)] \\ & \quad + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}(\Delta)] \\ & \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}_\mathcal{M}}(\Delta)] + \left(1 - \frac{3}{\beta + 1}\right) \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}_\mathcal{M}}(\Delta)] \\ & \quad + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} C_{T_\mathcal{M}}] \leq \frac{17\lambda_n}{8} + \frac{17\lambda_n}{8} \left(1 - \frac{3}{\beta + 1}\right) + \frac{\lambda_n}{6\beta} < 5\lambda_n \end{aligned}$$

### 6.4.3 From Tangent Space Constraints to the Original Problem

The optimality conditions of (6.9) suggest that there exist Lagrange multipliers  $Q_{D_y} \in W$ ,  $Q_{T_y} \in T(L_y^\mathcal{M})^\perp$ , and  $Q_{T_{yx}} \in T(\Theta_{yx}^\mathcal{M})^\perp$  such that

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

Letting the SVD decomposition of  $\tilde{L}_y$  and  $\tilde{\Theta}_{yx}$  be given by  $\tilde{L}_y = \bar{U} \bar{O} \bar{V}'$  and  $\tilde{\Theta}_{yx} = \check{U} \check{O} \check{V}'$  respectively, and  $Z \triangleq (0, \lambda_n \bar{U} \bar{V}', -\lambda_n \gamma \check{U} \check{V}', 0)$ , 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.  $\mathcal{P}_{\mathbb{H}_\mathcal{M}} \mathcal{F}^\dagger (\Sigma_n - \tilde{\Theta}^{-1}) = Z$

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

Here,  $UDV'$  is the SVD decomposition of  $\tilde{L}_y$  and  $\check{U}\check{D}\check{V}'$  is the SVD of  $\tilde{\Theta}_{yx}$ . It is clear that the first condition is satisfied since the pair  $(\tilde{\Theta}, \tilde{S}_y, \tilde{L}_y)$  is optimum for (6.9). To prove that the second condition, we prove a stronger statement that  $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^\perp} \mathcal{F}^\dagger(\Sigma_n - \tilde{\Theta}^{-1})] < \lambda_n$ . In particular, denoting  $\Delta = (\tilde{D}_y - D_y^*, \tilde{L}_y - L_y^*, \tilde{\Theta}_{yx} - \Theta_{yx}^*, \tilde{\Theta}_x - \Theta_x^*)$ , we show that:

$$\begin{aligned} \Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^\perp} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} \mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\Delta)] &< \lambda_n - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^\perp} \mathcal{F}^\dagger E_n] \\ &- \Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^\perp} \mathcal{F}^\dagger R_{\Sigma^*}(\mathcal{F}(\Delta))] - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^\perp} \mathcal{F}^\dagger \mathbb{I}^* \mathcal{F} C_{T_{\mathcal{M}}}] \end{aligned}$$

Using the first optimality condition and 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.  $\Phi_\gamma[\mathcal{P}_{\mathbb{H}'}(\cdot)] \leq 2\Phi_\gamma[\cdot]$ ), we have that:

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

Applying (6.2), we obtain:

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

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

## 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  $\ell_1$ -penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.