The Annals of Statistics
2012, Vol. 40, No. 4, 1935–1967
DOI: 10.1214/11-AOS949

LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION[1]

BY VENKAT CHANDRASEKARAN, PABLO A. PARRILO AND ALAN S. WILLSKY

California Institute of Technology, Massachusetts Institute of Technology and Massachusetts Institute of Technology

Suppose we observe samples of a *subset* of a collection of random variables. No additional information is provided about the number of latent variables, nor of the relationship between the latent and observed variables. Is it possible to discover the number of latent components, and to learn a statistical model over the entire collection of variables? We address this question in the setting in which the latent and observed variables are jointly Gaussian, with the conditional statistics of the observed variables conditioned on the latent variables being specified by a graphical model. As a first step we give natural conditions under which such latent-variable Gaussian graphical models are identifiable given marginal statistics of only the observed variables. Essentially these conditions require that the conditional graphical model among the observed variables is sparse, while the effect of the latent variables is "spread out" over most of the observed variables. Next we propose a tractable convex program based on regularized maximum-likelihood for model selection in this latent-variable setting; the regularizer uses both the ℓ_1 norm and the nuclear norm. Our modeling framework can be viewed as a combination of dimensionality reduction (to identify latent variables) and graphical modeling (to capture remaining statistical structure not attributable to the latent variables), and it consistently estimates both the number of latent components and the conditional graphical model structure among the observed variables. These results are applicable in the high-dimensional setting in which the number of latent/observed variables grows with the number of samples of the observed variables. The geometric properties of the algebraic varieties of sparse matrices and of low-rank matrices play an important role in our analysis.

Received August 2010; revised November 2011.

Discussed in 10.1214/12-AOS979, 10.1214/12-AOS980, 10.1214/12-AOS981, 10.1214/12-AOS984, 10.1214/12-AOS985 and 10.1214/12-AOS1001; rejoinder at 10.1214/12-AOS1020.

[1]Supported in part by AFOSR Grant FA9550-08-1-0180, in part under a MURI through AFOSR Grant FA9550-06-1-0324, in part under a MURI through AFOSR Grant FA9550-06-1-0303 and in part by NSF FRG 0757207.

MSC2010 subject classifications. 62F12, 62H12.

Key words and phrases. Gaussian graphical models, covariance selection, latent variables, regularization, sparsity, low-rank, algebraic statistics, high-dimensional asymptotics.

1. Introduction and setup. Statistical model selection in the high-dimensional regime arises in a number of applications. In many data analysis problems in geophysics, radiology, genetics, climate studies, and image processing, the number of samples available is comparable to or even smaller than the number of variables. As empirical statistics in these settings may not be well-behaved (see [17, 22]), high-dimensional model selection is therefore both challenging and of great interest. A model selection problem that has received considerable attention recently is the estimation of covariance matrices in the high-dimensional setting. As the sample covariance matrix is poorly behaved in such a regime, some form of *regularization* of the sample covariance is adopted based on assumptions about the true underlying covariance matrix [1, 2, 12, 14, 20, 36].

Graphical models. A number of papers have studied covariance estimation in the context of *Gaussian graphical model selection*. A *Gaussian graphical model* [19, 30] (also commonly referred to as a Gauss–Markov random field) is a statistical model defined with respect to a graph, in which the nodes index a collection of jointly Gaussian random variables and the edges represent the conditional independence relations (Markov structure) among the variables. In such models the sparsity pattern of the inverse of the covariance matrix, or the *concentration* matrix, directly corresponds to the graphical model structure. Specifically, consider a Gaussian graphical model in which the covariance matrix is given by a positive-definite Σ^* and the concentration matrix is given by $K^* = (\Sigma^*)^{-1}$. Then an edge $\{i, j\}$ is present in the underlying graphical model if and only if $K^*_{i,j} \neq 0$. In particular the absence of an edge between two nodes implies that the corresponding variables are independent conditioned on all the other variables. The model selection method usually studied in such a Gaussian graphical model setting is ℓ_1-regularized maximum-likelihood, with the ℓ_1 penalty applied to the entries of the concentration matrix to induce sparsity. The consistency properties of such an estimator have been studied [18, 26, 29], and under suitable conditions [18, 26] this estimator is also "sparsistent," that is, the estimated concentration matrix has the same sparsity pattern as the true model from which the samples are generated. An alternative approach to ℓ_1-regularized maximum-likelihood is to estimate the sparsity pattern of the concentration matrix by performing regression separately on each variable [23]; while such a method consistently estimates the sparsity pattern, it does not directly provide estimates of the covariance or concentration matrix.

In many applications throughout science and engineering (e.g., psychology, computational biology, and economics), a challenge is that one may not have access to observations of all the relevant phenomena, that is, some of the relevant variables may be latent or unobserved. In general latent variables pose a significant difficulty for model selection because one may not know the number of relevant latent variables, nor the relationship between these variables and the observed variables. Typical algorithmic methods that try to get around this difficulty usually fix the number of latent variables as well as the structural relationship between

latent and observed variables (e.g., the graphical model structure between latent and observed variables), and use the EM algorithm to fit parameters [9]. This approach suffers from the problem that one optimizes nonconvex functions, and thus one may get stuck in suboptimal local minima. An alternative suggestion [13] is one based on a greedy, local, combinatorial heuristic that assigns latent variables to groups of observed variables, via some form of clustering of the observed variables; however, this approach has no consistency guarantees.

Our setup. In this paper we study the problem of latent-variable graphical model selection in the setting where all the variables, both observed and latent, are jointly Gaussian. More concretely, X is a Gaussian random vector in $\mathbb{R}^{p+h}$, O and H are disjoint subsets of indices in $\{1, \ldots, p+h\}$ of cardinalities $|O| = p$ and $|H| = h$, and the corresponding subvectors of X are denoted by X_O and X_H, respectively. Let the covariance matrix underlying X be denoted by $\Sigma^*_{(O\,H)}$. The marginal statistics corresponding to the observed variables X_O are given by the marginal covariance matrix Σ^*_O, which is simply a submatrix of the full covariance matrix $\Sigma^*_{(O\,H)}$. However, suppose that we parameterize our model by the concentration matrix $K^*_{(O\,H)} = (\Sigma^*_{(O\,H)})^{-1}$, which as discussed above reveals the connection to graphical models. Here the submatrices K^*_O, $K^*_{O,H}$, K^*_H specify (in the full model) the dependencies among the observed variables, between the observed and latent variables, and among the latent variables, respectively. In such a parameterization, the *marginal concentration matrix* $(\Sigma^*_O)^{-1}$ corresponding to the observed variables X_O is given by the Schur complement [16] with respect to the block K^*_H:

$$\tilde{K}^*_O = (\Sigma^*_O)^{-1} = K^*_O - K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}. \tag{1.1}$$

Thus if we only observe the variables X_O, we only have access to Σ^*_O (or $\tilde{K}^*_O$). The two terms that compose $\tilde{K}^*_O$ above have interesting properties. The matrix K^*_O specifies the concentration matrix of the *conditional statistics* of the observed variables given the latent variables. If these conditional statistics are given by a sparse graphical model, then K^*_O is *sparse*. On the other hand, the matrix $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ serves as a *summary* of the effect of marginalization over the latent variables X_H. This matrix has small rank if the number of latent, unobserved variables X_H is small relative to the number of observed variables X_O. Therefore the marginal concentration matrix $\tilde{K}^*_O$ is generally *not sparse* due to the additional low-rank term $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$. Hence standard graphical model selection techniques applied directly to the observed variables X_O are not useful.

A modeling paradigm that infers the effect of the latent variables X_H would be more suitable in order to provide a concise explanation of the underlying statistical structure. Hence we approximate the sample covariance by a model in which the concentration matrix *decomposes* into the sum of a sparse matrix and a low-rank matrix, which reveals the conditional graphical model structure in the observed

variables as well as the number of and effect due to the unobserved latent variables. Such a method can be viewed as a blend of principal component analysis and graphical modeling. In standard graphical modeling one would directly approximate a concentration matrix by a sparse matrix to learn a sparse graphical model, while in principal component analysis the goal is to explain the statistical structure underlying a set of observations using a small number of latent variables (i.e., approximate a covariance matrix as a low-rank matrix). In our framework we learn a sparse graphical model among the observed variables *conditioned* on a few (additional) latent variables. These latent variables are *not* principal components, as the conditional statistics (conditioned on these latent variables) are given by a graphical model. Therefore we refer to these latent variables informally as *latent components*.

Contributions. Our first contribution in Section 3 is to address the fundamental question of *identifiability* of such latent-variable graphical models given the marginal statistics of only the observed variables. The critical point is that we need to tease apart the correlations induced due to marginalization over the latent variables from the conditional graphical model structure among the observed variables. As the identifiability problem is one of *uniquely* decomposing the sum of a sparse matrix and a low-rank matrix into the individual components, we study the algebraic varieties of sparse matrices and low-rank matrices. An important theme in this paper is the connection between the tangent spaces to these algebraic varieties and the question of identifiability. Specifically let $\Omega(K_O^*)$ denote the tangent space at K_O^* to the algebraic variety of sparse matrices, and let $T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ denote the tangent space at $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ to the algebraic variety of low-rank matrices. Then the *statistical* question of identifiability of K_O^* and $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ given $\tilde{K}_O^*$ is determined by the *geometric* notion of *transversality* of the tangent spaces $\Omega(K_O^*)$ and $T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$. The study of the transversality of these tangent spaces leads to natural conditions for identifiability. In particular we show that latent-variable models in which (1) the sparse matrix K_O^* has a small number of nonzeros per row/column, and (2) the low-rank matrix $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ has row/column spaces that are not closely aligned with the coordinate axes, are identifiable. These conditions have natural statistical interpretations. The first condition ensures that there are no densely connected subgraphs in the conditional graphical model structure among the observed variables, that is, that these conditional statistics are indeed specified by a sparse graphical model. Such statistical relationships may otherwise be mistakenly attributed to the effect of marginalization over some latent variable. The second condition ensures that the effect of marginalization over the latent variables is "spread out" over many observed variables; thus, the effect of marginalization over a latent variable is not confused with the conditional graphical model structure among the observed variables. In fact the first condition is often assumed in standard graphical model selection without latent variables (e.g., [26]).

As our next contribution we propose a *regularized maximum-likelihood decomposition* framework to approximate a given sample covariance matrix by a model in which the concentration matrix decomposes into a sparse matrix and a low-rank matrix. Based on the effectiveness of the ℓ_1 norm as a tractable convex relaxation for recovering sparse models [5, 10, 11] and the nuclear norm for low-rank matrices [4, 15, 27], we propose the following penalized likelihood method given a sample covariance matrix Σ_O^n formed from n samples of the observed variables:

$$(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} -\ell(S-L;\Sigma_O^n) + \lambda_n\big(\gamma\|S\|_1 + \operatorname{tr}(L)\big) \quad \text{s.t. } S-L \succ 0, L \succeq 0. \tag{1.2}$$

The constraints $\succ 0$ and $\succeq 0$ impose positive-definiteness and positive-semidefiniteness. The function ℓ represents the Gaussian log-likelihood $\ell(K;\Sigma) = \log\det(K) - \operatorname{tr}(K\Sigma)$ for $K \succ 0$, where tr is the trace of a matrix and det is the determinant. Here $\hat{S}_n$ provides an estimate of K_O^*, which represents the conditional concentration matrix of the observed variables; $\hat{L}_n$ provides an estimate of $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$, which represents the effect of marginalization over the latent variables. The regularizer is a combination of the ℓ_1 norm applied to S and the nuclear norm applied to L (the nuclear norm reduces to the trace over the cone of symmetric, positive-semidefinite matrices), with γ providing a trade-off between the two terms. This variational formulation is a *convex optimization* problem, and it is a regularized max-det program that can be solved in polynomial time using general-purpose solvers [33].

Our main result in Section 4 is a proof of the consistency of the estimator (1.2) in the high-dimensional regime in which both the number of observed variables and the number of latent components are allowed to grow with the number of samples (of the observed variables). We show that for a suitable choice of the regularization parameter λ_n, there exists a range of values of γ for which the estimates $(\hat{S}_n, \hat{L}_n)$ have the same sparsity (and sign) pattern and rank as $(K_O^*, K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ with high probability (see Theorem 4.1). The key technical requirement is an identifiability condition for the two components of the marginal concentration matrix $\tilde{K}_O^*$ with respect to the Fisher information (see Section 3.4). We make connections between our condition and the irrepresentability conditions required for support/graphical-model recovery using ℓ_1 regularization [26, 32, 37]. Our results provide numerous scaling regimes under which consistency holds in latent-variable graphical model selection. For example, we show that *under suitable identifiability conditions consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables* (*see Section* 4.2).

Related previous work. The problem of decomposing the sum of a sparse matrix and a low-rank matrix via convex optimization into the individual components

was initially studied in [7] by a superset of the authors of the present paper, with conditions derived under which the convex program exactly recovers the underlying components. In subsequent work Candès et al. [3] also studied this sparse-plus-low-rank decomposition problem, and provided guarantees for exact recovery using the convex program proposed in [7]. The problem setup considered in the present paper is quite different and is more challenging because we are only given access to an inexact sample covariance matrix, and we wish to produce an *inverse* covariance matrix that can be decomposed as the sum of sparse and low-rank components (preserving the sparsity pattern and rank of the components in the true underlying model). In addition to proving the consistency of the estimator (1.2), we also provide a statistical interpretation of our identifiability conditions and describe natural classes of latent-variable Gaussian graphical models that satisfy these conditions. As such our paper is closer in spirit to the many recent papers on covariance selection, but with the important difference that some of the variables are not observed.

Outline. Section 2 gives some background and a formal problem statement. Section 3 discusses the identifiability question, Section 4 states the main results of this paper, and Section 5 gives some proofs. We provide experimental demonstration of the effectiveness of our estimator on synthetic and real data in Section 6, and conclude with a brief discussion in Section 7. Some of our technical results are deferred to supplementary material [6].

2. Problem statement and background. We give a formal statement of the latent-variable model selection problem. We also briefly describe various properties of the algebraic varieties of sparse matrices and of low-rank matrices, and the properties of the Gaussian likelihood function.

The following matrix norms are employed throughout this paper. $\|M\|_2$ denotes the spectral norm, or the largest singular value of M. $\|M\|_\infty$ denotes the largest entry in magnitude of M. $\|M\|_F$ denotes the Frobenius norm, or the square root of the sum of the squares of the entries of M. $\|M\|_*$ denotes the nuclear norm, or the sum of the singular values of M (this reduces to the trace for positive-semidefinite matrices). $\|M\|_1$ denotes the sum of the absolute values of the entries of M. A number of *matrix operator* norms are also used. For example, let $\mathcal{Z}:\mathbb{R}^{p\times p}\to\mathbb{R}^{p\times p}$ be a linear operator acting on matrices. Then the induced operator norm is defined as $\|\mathcal{Z}\|_{q\to q}\triangleq\max_{N\in\mathbb{R}^{p\times p},\|N\|_q\leq 1}\|\mathcal{Z}(N)\|_q$. Therefore, $\|\mathcal{Z}\|_{F\to F}$ denotes the spectral norm of the operator $\mathcal{Z}$. The only vector norm used is the Euclidean norm, which is denoted by $\|\cdot\|$. Given any norm $\|\cdot\|_q$ (either a vector norm, a matrix norm or a matrix operator norm), the dual norm is given by $\|M\|_q^*\triangleq\sup\{\langle M,N\rangle|\|N\|_q\leq 1\}$.

2.1. *Problem statement.* In order to analyze latent-variable model selection methods, we need to define an appropriate notion of model selection con-

sistency for latent-variable graphical models. Given the two components K^*_O and $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ of the concentration matrix of the marginal distribution (1.1), there are *infinitely* many configurations of the latent variables [i.e., matrices $K^*_H \succ 0, K^*_{O,H} = (K^*_{H,O})^T$] that give rise to the *same* low-rank matrix $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$. Specifically for any nonsingular matrix $B \in \mathbb{R}^{|H|\times|H|}$, one can apply the transformations $K^*_H \to BK^*_HB^T$, $K^*_{O,H} \to K^*_{O,H}B^T$ and still preserve the low-rank matrix $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$. In *all* of these models the marginal statistics of the observed variables X_O remain the same upon marginalization over the latent variables X_H. The key *invariant* is the low-rank matrix $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, which summarizes the effect of marginalization over the latent variables. Consequently, from here on we use the notation $S^* = K^*_O$ and $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$. These observations give rise to the following notion of structure recovery.

DEFINITION 2.1. A pair of $|O| \times |O|$ symmetric matrices $(\hat{S}, \hat{L})$ is an *algebraically correct* estimate of a latent-variable Gaussian graphical model given by the concentration matrix $K^*_{(O\,H)}$ if the following conditions hold:

(1) The sign-pattern of $\hat{S}$ is the same as that of S^* [here $\text{sign}(0) = 0$]:

$$\text{sign}(\hat{S}_{i,j}) = \text{sign}(S^*_{i,j}) \qquad \forall i, j.$$

(2) The rank of $\hat{L}$ is the same as the rank of L^*:

$$\text{rank}(\hat{L}) = \text{rank}(L^*).$$

(3) The concentration matrix $\hat{S} - \hat{L}$ can be realized as the marginal concentration matrix of an appropriate latent-variable model:

$$\hat{S} - \hat{L} \succ 0, \qquad \hat{L} \succeq 0.$$

When a sequence of estimators is algebraically correct with probability approaching 1 in a suitable high-dimensional scaling regime, then we say that the estimators are *algebraically consistent*. The first condition ensures that $\hat{S}$ provides the correct structural estimate of the conditional graphical model of the observed variables conditioned on the latent components. This property is the same as the "sparsistency" property studied in standard graphical model selection [18, 26]. The second condition ensures that the number of latent components is properly estimated. Finally, the third condition ensures that the pair of matrices $(\hat{S}, \hat{L})$ leads to a realizable latent-variable model. In particular, this condition implies that there exists a valid latent-variable model in which (a) the conditional graphical model structure among the observed variables is given by $\hat{S}$, (b) the number of latent variables is equal to the rank of $\hat{L}$, and (c) the extra correlations induced due to marginalization over the latent variables are equal to $\hat{L}$. Any method for matrix

factorization (e.g., [35]) can be used to further factorize $\hat{L}$, depending on the property that one desires in the factors (e.g., sparsity).

We also study estimation error rates in the usual sense, that is, we show that one can produce estimates $(\hat{S}, \hat{L})$ that are close in various norms to the matrices (S^*, L^*). Notice that bounding the estimation error in some norm does not in general imply that the support/sign-pattern and rank of $(\hat{S}, \hat{L})$ are the same as those of (S^*, L^*). Therefore bounded estimation error is different from algebraic correctness, which requires that $(\hat{S}, \hat{L})$ have the same support/sign-pattern and rank as (S^*, L^*).

Goal. Let $K^*_{(O\,H)}$ denote the concentration matrix of a Gaussian model. Suppose that we have n samples $\{X^i_O\}_{i=1}^n$ of the observed variables X_O. We would like to produce estimates $(\hat{S}_n, \hat{L}_n)$ that, with high probability, are algebraically correct and have bounded estimation error (in some norm).

Our approach. We propose the regularized likelihood convex program (1.2) to produce estimates $(\hat{S}_n, \hat{L}_n)$. Specifically, the sample covariance matrix Σ^n_O in (1.2) is defined as

$$\Sigma^n_O \triangleq \frac{1}{n}\sum_{i=1}^{n} X^i_O {X^i_O}^T.$$

We give conditions on the underlying model $K^*_{(O\,H)}$ and suitable choices for the parameters λ_n, γ under which the estimates $(\hat{S}_n, \hat{L}_n)$ are consistent (see Theorem 4.1).

2.2. *Likelihood function and Fisher information.* Given n samples $\{X^i\}_{i=1}^n$ of a finite collection of jointly Gaussian zero-mean random variables with concentration matrix K^*, it is easily seen that the log-likelihood function is given by:

$$\ell(K; \Sigma^n) = \log\det(K) - \operatorname{tr}(K\Sigma^n), \tag{2.1}$$

where $\ell(K; \Sigma^n)$ is a function of K. Notice that this function is strictly concave for $K \succ 0$. In the latent-variable modeling problem with sample covariance Σ^n_O, the likelihood function with respect to the parametrization (S, L) is given by $\ell(S - L; \Sigma^n_O)$. This function is *jointly concave* with respect to the parameters (S, L) whenever $S - L \succ 0$, and it is employed in our variational formulation (1.2) to learn a latent-variable model.

In the analysis of a convex program involving the likelihood function, the Fisher information plays an important role as it is the negative of the Hessian of the likelihood function and thus controls the curvature. As the first term in the likelihood function is linear, we need only study higher-order derivatives of the log-determinant function in order to compute the Hessian. In the latent-variable setting with the marginal concentration matrix of the observed variables given by

$\tilde{K}^*_O = (\Sigma^*_O)^{-1}$ [see (1.1)], the corresponding Fisher information matrix is

$$\mathcal{I}(\tilde{K}^*_O) = (\tilde{K}^*_O)^{-1} \otimes (\tilde{K}^*_O)^{-1} = \Sigma^*_O \otimes \Sigma^*_O. \tag{2.2}$$

Here $\otimes$ denotes the tensor product between matrices. Notice that this is precisely the $|O|^2 \times |O|^2$ submatrix of the full Fisher information matrix $\mathcal{I}(K^*_{(O\,H)}) = \Sigma^*_{(O\,H)} \otimes \Sigma^*_{(O\,H)}$ with respect to all the parameters $K^*_{(O\,H)} = (\Sigma^*_{(O\,H)})^{-1}$ (corresponding to the situation in which *all* the variables $X_{O\cup H}$ are observed). In Section 3.4 we impose various conditions on the Fisher information matrix $\mathcal{I}(\tilde{K}^*_O)$ under which our regularized maximum-likelihood formulation provides consistent estimates.

2.3. *Algebraic varieties of sparse and low-rank matrices.* The set of sparse matrices and the set of low-rank matrices can be naturally viewed as algebraic varieties (solution sets of systems of polynomial equations). Here we describe these varieties, and discuss some of their geometric properties such as the tangent space and local curvature at a (smooth) point.

Let $\mathcal{S}(k)$ denote the set of matrices with at most k nonzeros:

$$\mathcal{S}(k) \triangleq \{M \in \mathbb{R}^{p\times p} \,|\, |\text{support}(M)| \le k\}. \tag{2.3}$$

Here support denotes the locations of nonzero entries. The set $\mathcal{S}(k)$ is an algebraic variety, and can in fact be viewed as a union of $\binom{p^2}{k}$ subspaces in $\mathbb{R}^{p\times p}$. This variety has dimension k, and it is smooth everywhere except at those matrices that have support size strictly smaller than k. For any matrix $M \in \mathbb{R}^{p\times p}$, consider the variety $\mathcal{S}(|\text{support}(M)|)$; M is a smooth point of this variety, and the tangent space at M is given by

$$\Omega(M) = \{N \in \mathbb{R}^{p\times p} \,|\, \text{support}(N) \subseteq \text{support}(M)\}. \tag{2.4}$$

Next let $\mathcal{L}(r)$ denote the algebraic variety of matrices with rank at most r:

$$\mathcal{L}(r) \triangleq \{M \in \mathbb{R}^{p\times p} \,|\, \text{rank}(M) \le r\}. \tag{2.5}$$

It is easily seen that $\mathcal{L}(r)$ is an algebraic variety because it can be defined through the vanishing of all $(r+1)\times(r+1)$ minors. This variety has dimension equal to $r(2p-r)$, and it is smooth everywhere except at those matrices that have rank strictly smaller than r. Consider a rank-r matrix M with singular value decomposition (SVD) given by $M = UDV^T$, where $U, V \in \mathbb{R}^{p\times r}$ and $D \in \mathbb{R}^{r\times r}$. The matrix M is a smooth point of the variety $\mathcal{L}(\text{rank}(M))$, and the tangent space at M with respect to this variety is given by

$$T(M) = \{UY_1^T + Y_2V^T \,|\, Y_1, Y_2 \in \mathbb{R}^{p\times r}\}. \tag{2.6}$$

We view both $\Omega(M)$ and $T(M)$ as subspaces in $\mathbb{R}^{p\times p}$. In Section 3 we explore the connection between geometric properties of these tangent spaces and the identifiability problem in latent-variable graphical models.

Curvature of rank variety. The sparse matrix variety $\mathcal{S}(k)$ has the property that it has *zero* curvature at any smooth point. The situation is more complicated for the low-rank matrix variety $\mathcal{L}(r)$, because the curvature at any smooth point is nonzero. We analyze how this variety curves locally, by studying how the tangent space changes from one point to a neighboring point. Indeed the amount of curvature at a point is directly related to the "angle" between the tangent space at that point and the tangent space at a neighboring point. For any linear subspace T of matrices, let $\mathcal{P}_T$ denote the projection onto T. Given two subspaces T_1, T_2 of the same dimension, we measure the "twisting" between these subspaces by considering the following quantity:

$$\rho(T_1, T_2) \triangleq \|\mathcal{P}_{T_1} - \mathcal{P}_{T_2}\|_{2\to 2} = \max_{\|N\|_2 \le 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_2. \tag{2.7}$$

In the supplement [6] we review relevant results from matrix perturbation theory, which suggest that the magnitude of the smallest nonzero singular value is closely tied to the local curvature of the variety. Therefore we control the twisting between tangent spaces at nearby points by bounding the smallest nonzero singular value away from zero.

3. Identifiability. In the absence of additional conditions, the latent-variable model selection problem is ill-posed. In this section we discuss a set of conditions on latent-variable models that ensure that these models are identifiable given marginal statistics for a subset of the variables. Some of the discussion in Sections 3.1 and 3.2 is presented in greater detail in [7].

3.1. *Structure between latent and observed variables.* Suppose that the low-rank matrix that summarizes the effect of the latent components is itself sparse. This leads to identifiability issues in the sparse-plus-low-rank decomposition problem. Statistically the additional correlations induced due to marginalization over the latent variables could be mistaken for the conditional graphical model structure of the observed variables. In order to avoid such identifiability problems the effect of the latent variables must be "diffuse" across the observed variables. To address this point the following quantity was introduced in [7] for any matrix M, defined with respect to the tangent space $T(M)$:

$$\xi(T(M)) \triangleq \max_{N \in T(M), \|N\|_2 \le 1} \|N\|_\infty. \tag{3.1}$$

Thus $\xi(T(M))$ being small implies that elements of the tangent space $T(M)$ cannot have their support concentrated in a few locations; as a result M cannot be too sparse. This idea is formalized in [7] by relating $\xi(T(M))$ to a notion of "incoherence" of the row/column spaces, where the row/column spaces are said to be incoherent with respect to the standard basis if these spaces are not aligned closely with any of the coordinate axes. Typically a matrix M with incoherent row/column spaces would have $\xi(T(M)) \ll 1$. This point is quantified precisely

in [7]. Specifically, we note that $\xi(T(M))$ can be as small as $\sim \sqrt{\frac{r}{p}}$ for a rank-r matrix $M \in \mathbb{R}^{p\times p}$ with row/column spaces that are almost maximally incoherent (e.g., if the row/column spaces span any r columns of a $p \times p$ orthonormal Hadamard matrix). On the other hand, $\xi(T(M)) = 1$ if the row/column spaces of M contain a standard basis vector.

Based on these concepts we roughly require that the low-rank matrix that summarizes the effect of the latent variables be *incoherent*, thereby ensuring that the extra correlations due to marginalization over the latent components cannot be confused with the conditional graphical model structure of the observed variables. Notice that the quantity ξ is not just a measure of the number of latent variables, but also of the overall effect of the correlations induced by marginalization over these variables.

Curvature and change in ξ: As noted previously, an important technical point is that the algebraic variety of low-rank matrices is locally curved at any smooth point. Consequently the quantity ξ changes as we move along the low-rank matrix variety smoothly. The quantity $\rho(T_1, T_2)$ introduced in (2.7) allows us to bound the variation in ξ as follows (proof in Section 5):

LEMMA 3.1. *Let T_1, T_2 be two linear subspaces of matrices of the same dimension with the property that $\rho(T_1, T_2) < 1$, where ρ is defined in* (2.7). *Then we have that*

$$\xi(T_2) \le \frac{1}{1-\rho(T_1,T_2)}[\xi(T_1) + \rho(T_1,T_2)].$$

3.2. *Structure among observed variables.* An identifiability problem also arises if the conditional graphical model among the observed variables contains a densely connected subgraph. These statistical relationships might be mistaken as correlations induced by marginalization over latent variables. Therefore we need to ensure that the conditional graphical model among the observed variables is sparse. We impose the condition that this conditional graphical model must have small "degree," that is, no observed variable is directly connected to too many other observed variables conditioned on the latent components. Notice that bounding the degree is a more refined condition than simply bounding the total number of nonzeros as the *sparsity pattern* also plays a role. In [7] the authors introduced the following quantity in order to provide an appropriate measure of the sparsity pattern of a matrix:

$$\mu(\Omega(M)) \triangleq \max_{N\in\Omega(M),\|N\|_\infty\le 1} \|N\|_2. \tag{3.2}$$

The quantity $\mu(\Omega(M))$ being small for a matrix implies that the spectrum of any element of the tangent space $\Omega(M)$ is not too "concentrated," that is, the singular values of the elements of the tangent space are not too large. In [7] it is shown that a sparse matrix M with "bounded degree" (a small number of nonzeros per

row/column) has small $\mu(M)$. Specifically, if $M \in \mathbb{R}^{p\times p}$ is any matrix with at most $\deg(M)$ nonzero entries per row/column, then we have that

$$\mu(\Omega(M)) \leq \deg(M).$$

3.3. *Transversality of tangent spaces.* Suppose that we have the sum of two vectors, each from two known subspaces. It is possible to uniquely recover the individual vectors from the sum if and only if the subspaces have a transverse intersection, that is, they only intersect at the origin. This simple observation leads to an appealing geometric notion of identifiability. Suppose now that we have the sum of a sparse matrix and a low-rank matrix, and that we are also given the tangent spaces at these matrices with respect to the algebraic varieties of sparse and low-rank matrices, respectively. Then a necessary and sufficient condition for identifiability with respect to the tangent spaces is that these spaces have a transverse intersection. This transverse intersection condition is also sufficient for local identifiability in a neighborhood around the sparse matrix and low-rank matrix with respect to the varieties of sparse and low-rank matrices (due to the inverse function theorem). It turns out that these tangent space transversality conditions are also sufficient for the convex program (1.2) to provide consistent estimates of a latent-variable graphical model (without any side information about the tangent spaces).

In order to quantify the level of transversality between the tangent spaces Ω and T we study the *minimum gain* with respect to some norm of the addition operator (which adds two matrices) $\mathcal{A}: \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ restricted to the cartesian product $\mathcal{Y} = \Omega \times T$. Then given any matrix norm $\|\cdot\|_q$ on $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$, the minimum gain of $\mathcal{A}$ restricted to $\mathcal{Y}$ is defined as

$$\varepsilon(\Omega, T, \|\cdot\|_q) \triangleq \min_{(S,L)\in\Omega\times T, \|(S,L)\|_q=1} \|\mathcal{P}_{\mathcal{Y}}\mathcal{A}^{\dagger}\mathcal{A}\mathcal{P}_{\mathcal{Y}}(S, L)\|_q,$$

where $\mathcal{P}_{\mathcal{Y}}$ denotes the projection onto $\mathcal{Y}$, and $\mathcal{A}^{\dagger}$ denotes the adjoint of the addition operator (with respect to the standard Euclidean inner-product). The "level" of transversality of Ω and T is measured by the magnitude of $\varepsilon(\Omega, T, \|\cdot\|_q)$, with transverse intersection being equivalent to $\varepsilon(\Omega, T, \|\cdot\|_q) > 0$. Note that $\varepsilon(\Omega, T, \|\cdot\|_F)$ is the square of the *minimum singular value* of the addition operator $\mathcal{A}$ restricted to $\Omega \times T$.

A natural norm with which to measure transversality is the dual norm of the regularization function in (1.2), as the subdifferential of the regularization function is specified in terms of its dual. The reasons for this will become clearer as we proceed through this paper. Recall that the regularization function used in the variational formulation (1.2) is given by

$$f_\gamma(S, L) = \gamma\|S\|_1 + \|L\|_*,$$

where the nuclear norm $\|\cdot\|_*$ reduces to the trace function over the cone of positive-semidefinite matrices. This function is a norm for all $\gamma > 0$. The dual norm of f_γ is given by

$$g_\gamma(S, L) = \max\left\{\frac{\|S\|_\infty}{\gamma}, \|L\|_2\right\}.$$

Next we define the quantity $\chi(\Omega, T, \gamma)$ as follows in order to study the transversality of the spaces Ω and T with respect to the g_γ norm:

$$\chi(\Omega, T, \gamma) \triangleq \max\left\{\frac{\xi(T)}{\gamma}, 2\mu(\Omega)\gamma\right\}. \tag{3.3}$$

Here μ and ξ are defined in (3.2) and (3.1). We then have the following result (proved in Section 5):

LEMMA 3.2. *Let $S \in \Omega, L \in T$ be matrices such that $\|S\|_\infty = \gamma$ and let $\|L\|_2 = 1$. Then we have that $g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_\mathcal{Y}(S, L)) \in [1 - \chi(\Omega, T, \gamma), 1 + \chi(\Omega, T, \gamma)]$, where $\mathcal{Y} = \Omega \times T$ and $\chi(\Omega, T, \gamma)$ is defined in (3.3). In particular we have that $1 - \chi(\Omega, T, \gamma) \leq \varepsilon(\Omega, T, g_\gamma)$.*

The quantity $\chi(\Omega, T, \gamma)$ being small implies that the addition operator is essentially isometric when restricted to $\mathcal{Y} = \Omega \times T$. Stated differently, the magnitude of $\chi(\Omega, T, \gamma)$ is a measure of the level of transversality of the spaces Ω and T. If $\mu(\Omega)\xi(T) < \frac{1}{2}$, then $\gamma \in (\xi(T), \frac{1}{2\mu(\Omega)})$ ensures that $\chi(\Omega, T, \gamma) < 1$, which in turn implies that the tangent spaces Ω and T have a transverse intersection.

Observation: Thus we have that the smaller the quantities $\mu(\Omega)$ and $\xi(T)$, the more transverse the intersection of the spaces Ω and T as measured by $\varepsilon(\Omega, T, g_\gamma)$.

3.4. *Conditions on Fisher information.* The main focus of Section 4 is to analyze the regularized maximum-likelihood convex program (1.2) by studying its optimality conditions. The log-likelihood function is well-approximated in a neighborhood by a quadratic form given by the Fisher information (which measures the curvature, as discussed in Section 2.2). Let $\mathcal{I}^* = \mathcal{I}(\tilde{K}_O^*)$ denote the Fisher information evaluated at the true marginal concentration matrix $\tilde{K}_O^*$ [see (1.1)]. The appropriate measure of transversality between the tangent spaces[2] $\Omega = \Omega(S^*)$ and $T = T(L^*)$ is then in a space in which the inner-product is given by $\mathcal{I}^*$. Specifically, we need to analyze the minimum gain of the operator $\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}$ restricted to the space $\mathcal{Y} = \Omega \times T$. Therefore we impose several conditions on the Fisher information $\mathcal{I}^*$. We define quantities that control the gains of $\mathcal{I}^*$ restricted to Ω and T separately; these ensure that elements of Ω and elements of T are

[2]We implicitly assume that these tangent spaces are subspaces of the space of *symmetric* matrices.

individually identifiable under the map $\mathcal{I}^*$. In addition we define quantities that, in conjunction with bounds on $\mu(\Omega)$ and $\xi(T)$, allow us to control the gain of $\mathcal{I}^*$ restricted to the direct-sum $\Omega \oplus T$.

$\mathcal{I}^*$ restricted to Ω: The minimum gain of the operator $\mathcal{P}_\Omega \mathcal{I}^* \mathcal{P}_\Omega$ restricted to Ω is given by

$$\alpha_\Omega \triangleq \min_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_\Omega \mathcal{I}^* \mathcal{P}_\Omega(M)\|_\infty.$$

The maximum effect of elements in Ω in the orthogonal direction $\Omega^\perp$ is given by

$$\delta_\Omega \triangleq \max_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_{\Omega^\perp} \mathcal{I}^* \mathcal{P}_\Omega(M)\|_\infty.$$

The operator $\mathcal{I}^*$ is injective on Ω if $\alpha_\Omega > 0$. The ratio $\frac{\delta_\Omega}{\alpha_\Omega} \leq 1 - \nu$ implies the irrepresentability condition imposed in [26], which gives a sufficient condition for consistent recovery of graphical model structure using ℓ_1-regularized maximum-likelihood. Notice that this condition is a generalization of the usual Lasso irrepresentability conditions [32, 37], which are typically imposed on the covariance matrix. Finally we also consider the following quantity, which controls the behavior of $\mathcal{I}^*$ restricted to Ω in the spectral norm:

$$\beta_\Omega \triangleq \max_{M \in \Omega, \|M\|_2 = 1} \|\mathcal{I}^*(M)\|_2.$$

$\mathcal{I}^*$ restricted to T: Analogously to the case of Ω one could control the gains of the operators $\mathcal{P}_{T^\perp} \mathcal{I}^* \mathcal{P}_T$ and $\mathcal{P}_T \mathcal{I}^* \mathcal{P}_T$. However, as discussed previously, one complication is that the tangent spaces at nearby smooth points on the rank variety are in general different, and the amount of twisting between these spaces is governed by the local curvature. Therefore we control the gains of the operators $\mathcal{P}_{T'^\perp} \mathcal{I}^* \mathcal{P}_{T'}$ and $\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}$ for all tangent spaces T' that are "close to" the nominal T (at the true underlying low-rank matrix), measured by $\rho(T, T')$ (2.7) being small. The minimum gain of the operator $\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}$ restricted to T' (close to T) is given by

$$\alpha_T \triangleq \min_{\rho(T',T) \leq \xi(T)/2} \min_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}(M)\|_2.$$

Similarly, the maximum effect of elements in T' in the orthogonal direction $T'^\perp$ (for T' close to T) is given by

$$\delta_T \triangleq \max_{\rho(T',T) \leq \xi(T)/2} \max_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'^\perp} \mathcal{I}^* \mathcal{P}_{T'}(M)\|_2.$$

Implicit in the definition of α_T and δ_T is the fact that the outer minimum and maximum are only taken over spaces T' that are tangent spaces to the rank-variety. The operator $\mathcal{I}^*$ is injective on all tangent spaces T' such that $\rho(T', T) \leq \frac{\xi(T)}{2}$ if $\alpha_T > 0$. An irrepresentability condition (analogous to those developed for the sparse case) for tangent spaces near T to the rank variety would be that $\frac{\delta_T}{\alpha_T} \leq 1 - \nu$.

Finally we also control the behavior of $\mathcal{I}^*$ restricted to T' close to T in the ℓ_∞ norm:

$$\beta_T \triangleq \max_{\rho(T',T)\le\xi(T)/2} \max_{M\in T',\|M\|_\infty=1} \|\mathcal{I}^*(M)\|_\infty.$$

The two sets of quantities $(\alpha_\Omega, \delta_\Omega)$ and (α_T, δ_T) essentially control how $\mathcal{I}^*$ behaves when restricted to the spaces Ω and T *separately* (in the natural norms). The quantities β_Ω and β_T are useful in order to control the gains of the operator $\mathcal{I}^*$ restricted to the *direct sum* $\Omega \oplus T$. Notice that although the magnitudes of elements in Ω are measured most naturally in the ℓ_∞ norm, the quantity β_Ω is specified with respect to the spectral norm. Similarly, elements of the tangent spaces T' to the rank variety are most naturally measured in the spectral norm, but β_T provides control in the ℓ_∞ norm. These quantities, combined with $\mu(\Omega)$ and $\xi(T)$ [defined in (3.2) and (3.1)], provide the "coupling" necessary to control the behavior of $\mathcal{I}^*$ restricted to elements in the direct sum $\Omega \oplus T$. In order to keep track of fewer quantities, we summarize the six quantities as follows:

$$\alpha \triangleq \min(\alpha_\Omega, \alpha_T); \qquad \delta \triangleq \max(\delta_\Omega, \delta_T); \qquad \beta \triangleq \max(\beta_\Omega, \beta_T).$$

Main assumption: There exists a $\nu \in (0, \frac{1}{2}]$ such that

$$\frac{\delta}{\alpha} \le 1 - 2\nu.$$

This assumption is to be viewed as a generalization of the irrepresentability conditions imposed on the covariance matrix [32, 37] or the Fisher information matrix [26] in order to provide consistency guarantees for sparse model selection using the ℓ_1 norm. With this assumption we have the following proposition, proved in Section 5, about the gains of the operator $\mathcal{I}^*$ restricted to $\Omega \oplus T$. This proposition plays a fundamental role in the analysis of the performance of the regularized maximum-likelihood procedure (1.2). Specifically, it gives conditions under which a suitable primal-dual pair can be specified to certify optimality with respect to (1.2) (see Section 5.2 for more details).

PROPOSITION 3.3. *Let Ω and T be the tangent spaces defined in this section, and let $\mathcal{I}^*$ be the Fisher information evaluated at the true marginal concentration matrix. Further let α, β, ν be as defined above. Suppose that*

$$\mu(\Omega)\xi(T) \le \frac{1}{6}\left(\frac{\nu\alpha}{\beta(2-\nu)}\right)^2,$$

and that γ is in the following range:

$$\gamma \in \left[\frac{3\xi(T)\beta(2-\nu)}{\nu\alpha}, \frac{\nu\alpha}{2\mu(\Omega)\beta(2-\nu)}\right].$$

Then we have the following two conclusions for $\mathcal{Y} = \Omega \times T'$ with $\rho(T', T) \le \frac{\xi(T)}{2}$:

(1) *The minimum gain of $\mathcal{I}^*$ restricted to $\Omega \oplus T'$ is bounded below*:

$$\min_{(S,L)\in\mathcal{Y}, \|S\|_\infty=\gamma, \|L\|_2=1} g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \geq \frac{\alpha}{2}.$$

Specifically this implies that for all $(S, L) \in \mathcal{Y}$

$$g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \geq \frac{\alpha}{2} g_\gamma(S,L).$$

(2) *The effect of elements in $\mathcal{Y} = \Omega \times T'$ on the orthogonal complement $\mathcal{Y}^\perp = \Omega^\perp \times T'^\perp$ is bounded above*:

$$\|\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y})^{-1}\|_{g_\gamma\to g_\gamma} \leq 1-\nu.$$

Specifically this implies that for all $(S, L) \in \mathcal{Y}$

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \leq (1-\nu) g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)).$$

The last quantity we consider is the spectral norm of the marginal covariance matrix $\Sigma_O^* = (\tilde{K}_O^*)^{-1}$:

$$\psi \triangleq \|\Sigma_O^*\|_2 = \|(\tilde{K}_O^*)^{-1}\|_2. \tag{3.4}$$

A bound on ψ is useful in the probabilistic component of our analysis, in order to derive convergence rates of the sample covariance matrix to the true covariance matrix. We also observe that

$$\|\mathcal{I}^*\|_{2\to 2} = \|(\tilde{K}_O^*)^{-1} \otimes (\tilde{K}_O^*)^{-1}\|_{2\to 2} = \psi^2.$$

Remarks. The quantities α, β, δ bound the gains of the Fisher information $\mathcal{I}^*$ restricted to the spaces Ω and T (and tangent spaces near T). One can make stronger assumptions on $\mathcal{I}^*$ that are more easily interpretable. For example, $\alpha_\Omega, \beta_\Omega$ could bound the minimum/maximum gains of $\mathcal{I}^*$ for *all* matrices (rather than just those in Ω), and δ_Ω the $\mathcal{I}^*$-inner-product for *all* pairs of orthogonal matrices (rather than just those in Ω and $\Omega^\perp$). Similarly, α_T, β_T could bound the minimum/maximum gains of $\mathcal{I}^*$ for all matrices (rather than just those near T), and δ_T the $\mathcal{I}^*$-inner-product for all pairs of orthogonal matrices (rather than just those near T and $T^\perp$). Such bounds would apply in either the $\|\cdot\|_{2\to 2}$ norm (for $\alpha_T, \delta_T, \beta_\Omega$) or the $\|\cdot\|_{\infty\to\infty}$ norm (for $\alpha_\Omega, \delta_\Omega, \beta_T$). These modified assumptions are global in nature (not restricted just to Ω or near T) and are consequently stronger (they lower-bound the original α_Ω, α_T and they upper-bound the original $\beta_\Omega, \beta_T, \delta_\Omega, \delta_T$), and they essentially control the gains of the operator $\mathcal{I}^*$ in the $\|\cdot\|_{2\to 2}$ norm and the $\|\cdot\|_{\infty\to\infty}$ norm. In contrast, previous works on covariance selection [1, 2, 29] consider *well-conditioned* families of covariance matrices by

bounding the minimum/maximum eigenvalues (i.e., gain with respect to the spectral norm).

4. Consistency of regularized maximum-likelihood program.

4.1. *Main results.* Recall that $K^*_{(O\,H)}$ denotes the full concentration matrix of a collection of zero-mean jointly-Gaussian observed and latent variables. Let $p = |O|$ denote the number of observed variables, and let $h = |H|$ denote the number of latent variables. We are given n samples $\{X^i_O\}_{i=1}^n$ of the observed variables X_O. We consider the high-dimensional setting in which (p, h, n) are all allowed to grow simultaneously. We present our main result next demonstrating the consistency of the estimator (1.2), and then discuss classes of latent-variable graphical models and various scaling regimes in which our estimator is consistent. Recall from (1.2) that λ_n is a regularization parameter, and γ is a tradeoff parameter between the rank and sparsity terms. Notice from Proposition 3.3 that the choice of γ depends on the values of $\mu(\Omega(S^*))$ and $\xi(T(L^*))$. While these quantities may not be known a priori, we discuss a method to choose γ numerically in our experimental results (see Section 6). The following theorem shows that the estimates $(\hat{S}_n, \hat{L}_n)$ provided by the convex program (1.2) are consistent for a suitable choice of λ_n. In addition to the appropriate identifiability conditions (as specified by Proposition 3.3), we also impose lower bounds on the minimum magnitude nonzero entry θ of the sparse conditional graphical model matrix S^* and on the minimum nonzero singular value σ of the low-rank matrix L^* summarizing the effect of the latent variables. The theorem is stated in terms of the quantities α, β, ν, ψ, and we particularly emphasize the dependence on $\mu(\Omega(S^*))$ and $\xi(T(L^*))$ because these control the complexity of the underlying latent-variable graphical model given by $K^*_{(O\,H)}$. A number of quantities play a role in our theorem: let $D = \max\{1, \frac{\nu\alpha}{3\beta(2-\nu)}\}$, $C_1 = \psi(1 + \frac{\alpha}{6\beta})$, $C_2 = \frac{48}{\alpha} + \frac{1}{\psi^2}$, $C_{\text{samp}} = \frac{\alpha\nu}{32(3-\nu)D} \min\{\frac{1}{4C_1}, \frac{\alpha\nu}{256D(3-\nu)\psi C_1^2}\}$, $C_\lambda = \frac{48\sqrt{2}D\psi(2-\nu)}{\xi(T)\nu}$, $C_S = \max\{(\frac{6(2-\nu)}{\nu} + 1)C_2^2\psi^2 D, C_2 + \frac{3\alpha C_2^2(2-\nu)}{16(3-\nu)}\}$ and $C_L = \frac{C_2\nu\alpha}{\beta(2-\nu)}$.

THEOREM 4.1. *Let $K^*_{(O\,H)}$ denote the concentration matrix of a Gaussian model. We have n samples $\{X^i_O\}_{i=1}^n$ of the p observed variables denoted by O. Let $\Omega = \Omega(S^*)$ and $T = T(L^*)$ denote the tangent spaces at S^* and at L^* with respect to the sparse and low-rank matrix varieties, respectively.*

Assumptions: *Suppose that the quantities $\mu(\Omega)$ and $\xi(T)$ satisfy the assumption of Proposition 3.3 for identifiability, and γ is chosen in the range specified by Proposition 3.3. Further suppose that the following conditions hold*:

(1) *Let $n \geq \frac{p}{\xi(T)^4} \max\{\frac{128\psi^2}{C^2_{\text{samp}}}, 2\}$, that is, we require that $n \gtrsim \frac{p}{\xi(T)^4}$.*

(2) *Set $\lambda_n = \frac{48\sqrt{2}D\psi(2-\nu)}{\xi(T)\nu}\sqrt{\frac{p}{n}}$, that is, we require that $\lambda_n \asymp \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$.*

(3) *Let* $\sigma \geq \frac{C_L \lambda_n}{\xi(T)^2}$, *that is, we require that* $\sigma \gtrsim \frac{1}{\xi(T)^3}\sqrt{\frac{p}{n}}$.
(4) *Let* $\theta \geq \frac{C_S \lambda_n}{\mu(\Omega)}$, *that is, we require that* $\theta \gtrsim \frac{1}{\xi(T)\mu(\Omega)}\sqrt{\frac{p}{n}}$.

Conclusions: *Then with probability greater than* $1 - 2\exp\{-p\}$ *we have algebraic correctness and estimation error given by*:

(1) $\operatorname{sign}(\hat{S}_n) = \operatorname{sign}(S^*)$ *and* $\operatorname{rank}(\hat{L}_n) = \operatorname{rank}(L^*)$;
(2) $g_\gamma(\hat{S}_n - S^*, \hat{L}_n - L^*) \leq \frac{512\sqrt{2}(3-\nu)D\psi}{\nu\alpha\xi(T)}\sqrt{\frac{p}{n}} \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$.

The proof of this theorem is given in Section 5. The theorem essentially states that if the minimum nonzero singular value of the low-rank piece L^* and minimum nonzero entry of the sparse piece S^* are bounded away from zero, then the convex program (1.2) provides estimates that are both algebraically correct and have bounded estimation error (in the ℓ_∞ and spectral norms).

Notice that the condition on the minimum singular value of L^* is more stringent than the one on the minimum nonzero entry of S^*. One role played by these conditions is to ensure that the estimates $(\hat{S}_n, \hat{L}_n)$ do not have smaller support size/rank than (S^*, L^*). However, the minimum singular value bound plays the additional role of bounding the curvature of the low-rank matrix variety around the point L^*, which is the reason for this condition being more stringent. Notice also that the number of latent variables h does not explicitly appear in the bounds in Theorem 4.1, which only depend on p, $\mu(\Omega(S^*))$, $\xi(T(L^*))$. However, the dependence on h is implicit in the dependence on $\xi(T(L^*))$, and we discuss this point in greater detail in the following section.

Finally we note that consistency holds in Theorem 4.1 for a *range* of values of $\gamma \in [\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}]$. In particular the assumptions on the sample complexity, the minimum nonzero singular value of L^*, and the minimum magnitude nonzero entry of S^* are governed by the lower end of this range for γ. These assumptions can be weakened if we only require consistency for a smaller range of values of γ. The next result conveys this point with a specific example.

COROLLARY 4.2. *Consider the same setup and notation as in Theorem 4.1. Suppose that the quantities* $\mu(\Omega)$ *and* $\xi(T)$ *satisfy the assumption of Proposition 3.3 for identifiability, and that* $\gamma = \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}$ *(the upper end of the range specified in Proposition 3.3), that is,* $\gamma \asymp \frac{1}{\mu(\Omega)}$. *Further suppose that*: (1) $n \gtrsim \mu(\Omega)^4 p$; (2) $\lambda_n \asymp \mu(\Omega)\sqrt{\frac{p}{n}}$; (3) $\sigma \gtrsim \frac{\mu(\Omega)^2}{\xi(T)}\sqrt{\frac{p}{n}}$; (4) $\theta \gtrsim \sqrt{\frac{p}{n}}$. *Then with probability greater than* $1 - 2\exp\{-p\}$ *we have estimates* $(\hat{S}_n, \hat{L}_n)$ *that are algebraically correct, and with the error bounded as* $g_\gamma(\hat{S}_n - S^*, \hat{L}_n - L^*) \lesssim \mu(\Omega)\sqrt{\frac{p}{n}}$.

The proof of this corollary[3] is analogous to that of Theorem 4.1. We emphasize that in practice it is often beneficial to have consistent estimates for a range of values of γ (as in Theorem 4.1). Specifically, the stability of the sparsity pattern and rank of the estimates $(\hat{S}_n, \hat{L}_n)$ for a range of trade-off parameters is useful in order to choose a suitable value of γ, as prior information about the quantities $\mu(\Omega(S^*))$ and $\xi(T(L^*))$ is not typically available (see Section 6).

We remark here that the identifiability conditions of Proposition 3.3 are the main sufficient conditions required for Theorem 4.1 and Corollary 4.2 to hold. It would be interesting to obtain necessary conditions as well for these results, analogous to the necessity and sufficiency of the irrepresentability conditions for the Lasso [32, 37].

4.2. *Scaling regimes.* Next we consider classes of latent-variable models that satisfy the conditions of Theorem 4.1. Recall from Section 3.2 that $\mu(\Omega(S^*)) \leq \deg(S^*)$. Throughout this section, we consider latent-variable models in which the low-rank matrix L^* is almost maximally incoherent, that is, $\xi(T(L^*)) \sim \sqrt{\frac{h}{p}}$ so the effect of marginalization over the latent variables is diffuse across almost all the observed variables. We suppress the dependence on the quantities α, β, ν, ψ defined in Section 3.4 in our scaling results, and specifically focus on the trade-off between $\xi(T(L^*))$ and $\mu(\Omega(S^*))$ for consistent estimation (we also suppress the dependence of these quantities on n). Thus, based on Proposition 3.3 we study latent-variable models in which

$$\xi(T(L^*))\mu(\Omega(S^*)) = \mathcal{O}\left(\sqrt{\frac{h}{p}}\deg(S^*)\right) = \mathcal{O}(1).$$

As we describe next, there are nontrivial classes of latent-variable graphical models in which this condition holds.

Bounded degree: The first class of latent-variable models that we consider are those in which the conditional graphical model among the observed variables (given by K^*_O) has constant degree:

$$\deg(S^*) = \mathcal{O}(1), \qquad h \sim p.$$

Such models can be estimated consistently from $n \sim p$ samples. Thus consistent latent-variable model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables.

[3]By making stronger assumptions on the Fisher information matrix $\mathcal{I}^*$, one can further remove the factor of $\xi(T)$ in the lower bound for σ. Specifically, the lower bound $\sigma \gtrsim \mu(\Omega)^3\sqrt{\frac{p}{n}}$ suffices for consistent estimation if the bounds defined by the quantities $\alpha_T, \beta_T, \delta_T$ can be strengthened as described in the remarks at the end of Section 3.4.

Polylogarithmic degree: The next class of models that we consider are those in which the degree of the conditional graphical model of the observed variables grows polylogarithmically with p:

$$\deg(S^*) \sim \log(p)^q, \qquad h \sim \frac{p}{\log(p)^{2q}}.$$

Such latent-variable graphical models can be consistently estimated as long as $n \sim p\,\mathrm{polylog}(p)$.

For standard graphical model selection with no latent variables, ℓ_1-regularized maximum-likelihood is shown to be consistent with $n = \mathcal{O}(\log p)$ samples [26]. On the other hand, our results prove consistency in the setting with latent variables when $n = \mathcal{O}(p)$ samples. It would be interesting to study whether these rates are inherent to latent-variable model selection.

4.3. *Rates for covariance matrix estimation.* Theorem 4.1 gives conditions under which we can consistently estimate the sparse and low-rank parts that compose the marginal concentration matrix $\tilde{K}_O^*$. Here we state a corollary that gives rates for covariance matrix estimation, that is, the quality of the estimate $(\hat{S}_n - \hat{L}_n)^{-1}$ with respect to the "true" marginal covariance matrix Σ_O^*.

COROLLARY 4.3. *Under the same conditions as in Theorem* 4.1, *we have with probability greater than* $1 - 2\exp\{-p\}$ *that*

$$g_\gamma\big(\mathcal{A}^\dagger[(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma_O^*]\big) \leq \lambda_n\left[1 + \frac{\nu}{6(2-\nu)}\right].$$

This corollary implies that $\|(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma_O^*\|_2 \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$ based on the choice of λ_n in Theorem 4.1, and that $\|(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma_O^*\|_2 \lesssim \mu(\Omega)\sqrt{\frac{p}{n}}$ based on the choice of λ_n in Corollary 4.2.

5. Proofs.

5.1. *Proofs of Section* 3. Here we give proofs of the results stated in Section 3.

PROOF OF LEMMA 3.1. Since $\rho(T_1, T_2) < 1$, the largest principal angle between T_1 and T_2 is strictly less than $\frac{\pi}{2}$. Consequently, the mapping $\mathcal{P}_{T_2} : T_1 \to T_2$ restricted to T_1 is bijective (as it is injective, and the spaces T_1, T_2 have the same dimension). Consider the maximum and minimum gains of $\mathcal{P}_{T_2}$ restricted to T_1; for any $M \in T_1$, $\|M\|_2 = 1$:

$$\|\mathcal{P}_{T_2}(M)\|_2 = \|M + [\mathcal{P}_{T_2} - \mathcal{P}_{T_1}](M)\|_2 \in [1 - \rho(T_1, T_2), 1 + \rho(T_1, T_2)].$$

Therefore, we can rewrite $\xi(T_2)$ as follows:

$$\begin{aligned}
\xi(T_2) &= \max_{N\in T_2, \|N\|_2\le 1} \|N\|_\infty = \max_{N\in T_2, \|N\|_2\le 1} \|\mathcal{P}_{T_2}(N)\|_\infty \\
&\le \max_{N\in T_1, \|N\|_2\le 1/(1-\rho(T_1,T_2))} \|\mathcal{P}_{T_2}(N)\|_\infty \\
&\le \max_{N\in T_1, \|N\|_2\le 1/(1-\rho(T_1,T_2))} \big[\|N\|_\infty + \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_\infty\big] \\
&\le \frac{1}{1-\rho(T_1,T_2)}\Big[\xi(T_1) + \max_{N\in T_1, \|N\|_2\le 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_\infty\Big] \\
&\le \frac{1}{1-\rho(T_1,T_2)}\Big[\xi(T_1) + \max_{\|N\|_2\le 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_2\Big] \\
&\le \frac{1}{1-\rho(T_1,T_2)}[\xi(T_1) + \rho(T_1,T_2)].
\end{aligned}$$

This concludes the proof of the lemma. □

PROOF OF LEMMA 3.2. We have that $\mathcal{A}^\dagger\mathcal{A}(S,L) = (S+L, S+L)$; therefore, $\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L) = (S+\mathcal{P}_\Omega(L), \mathcal{P}_T(S)+L)$. We need to bound $\|S+\mathcal{P}_\Omega(L)\|_\infty$ and $\|\mathcal{P}_T(S)+L\|_2$. First, we have

$$\begin{aligned}
\|S+\mathcal{P}_\Omega(L)\|_\infty &\in [\|S\|_\infty - \|\mathcal{P}_\Omega(L)\|_\infty, \|S\|_\infty + \|\mathcal{P}_\Omega(L)\|_\infty] \\
&\subseteq [\|S\|_\infty - \|L\|_\infty, \|S\|_\infty + \|L\|_\infty] \\
&\subseteq [\gamma - \xi(T), \gamma + \xi(T)].
\end{aligned}$$

Similarly, one can check that

$$\begin{aligned}
\|\mathcal{P}_T(S)+L\|_2 &\in [-\|\mathcal{P}_T(S)\|_2 + \|L\|_2, \|\mathcal{P}_T(S)\|_2 + \|L\|_2] \\
&\subseteq [1-2\|S\|_2, 1+2\|S\|_2] \\
&\subseteq [1-2\gamma\mu(\Omega), 1+2\gamma\mu(\Omega)].
\end{aligned}$$

These two bounds give us the desired result. □

PROOF OF PROPOSITION 3.3. Before proving the two parts of this proposition we make a simple observation about $\xi(T')$ using the condition that $\rho(T,T') \le \frac{\xi(T)}{2}$ by applying Lemma 3.1:

$$\xi(T') \le \frac{\xi(T)+\rho(T,T')}{1-\rho(T,T')} \le \frac{3\xi(T)/2}{1-\xi(T)/2} \le 3\xi(T).$$

Here we used the property that $\xi(T) \le 1$ in obtaining the final inequality. Consequently, noting that $\gamma \in [\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}]$ implies that

$$\chi(\Omega, T', \gamma) = \max\left\{\frac{\xi(T')}{\gamma}, 2\mu(\Omega)\gamma\right\} \le \frac{\nu\alpha}{\beta(2-\nu)}. \tag{5.1}$$

Part 1: The proof of this step proceeds in a similar manner to that of Lemma 3.2. First we have for $S \in \Omega$, $L \in T'$ with $\|S\|_\infty = \gamma$, $\|L\|_2 = 1$:

$$\|\mathcal{P}_\Omega \mathcal{I}^*(S+L)\|_\infty \geq \|\mathcal{P}_\Omega \mathcal{I}^* S\|_\infty - \|\mathcal{P}_\Omega \mathcal{I}^* L\|_\infty \geq \alpha\gamma - \|\mathcal{I}^* L\|_\infty \geq \alpha\gamma - \beta\xi(T').$$

Next, under the same conditions on S, L,

$$\|\mathcal{P}_{T'} \mathcal{I}^*(S+L)\|_2 \geq \|\mathcal{P}_{T'} \mathcal{I}^* L\|_2 - \|\mathcal{P}_{T'} \mathcal{I}^* S\|_2 \geq \alpha - 2\|\mathcal{I}^* S\|_2 \geq \alpha - 2\beta\mu(\Omega)\gamma.$$

Combining these last two bounds with (5.1), we conclude that

$$\min_{(S,L)\in\mathcal{Y}, \|S\|_\infty=\gamma, \|L\|_2=1} g_\gamma(\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_\mathcal{Y}(S,L))$$
$$\geq \alpha - \beta \max\left\{\frac{\xi(T')}{\gamma}, 2\mu(\Omega)\gamma\right\} \geq \alpha - \frac{\nu\alpha}{2-\nu} = \frac{2\alpha(1-\nu)}{2-\nu} \geq \frac{\alpha}{2},$$

where the final inequality follows from the assumption that $\nu \in (0, \frac{1}{2}]$.

Part 2: Note that for $S \in \Omega$, $L \in T'$ with $\|S\|_\infty \leq \gamma$, $\|L\|_2 \leq 1$,

$$\|\mathcal{P}_{\Omega^\perp} \mathcal{I}^*(S+L)\|_\infty \leq \|\mathcal{P}_{\Omega^\perp} \mathcal{I}^* S\|_\infty + \|\mathcal{P}_{\Omega^\perp} \mathcal{I}^* L\|_\infty \leq \delta\gamma + \beta\xi(T').$$

Similarly,

$$\|\mathcal{P}_{T'^\perp} \mathcal{I}^*(S+L)\|_2 \leq \|\mathcal{P}_{T'^\perp} \mathcal{I}^* S\|_2 + \|\mathcal{P}_{T'^\perp} \mathcal{I}^* L\|_2 \leq \beta\gamma\mu(\Omega) + \delta.$$

Combining these last two bounds with the bounds from the first part, we have that

$$\|\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_\mathcal{Y} (\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_\mathcal{Y})^{-1}\|_{g_\gamma \to g_\gamma}$$
$$\leq \frac{\delta + \beta \max\{\xi(T')/\gamma, 2\mu(\Omega)\gamma\}}{\alpha - \beta \max\{\xi(T')/\gamma, 2\mu(\Omega)\gamma\}} \leq \frac{\delta + \nu\alpha/(2-\nu)}{\alpha - \nu\alpha/(2-\nu)}$$
$$\leq \frac{(1-2\nu)\alpha + \nu\alpha/(2-\nu)}{\alpha - \nu\alpha/(2-\nu)} = 1 - \nu.$$

This concludes the proof of the proposition. □

5.2. *Proof strategy for Theorem* 4.1. Standard results from convex analysis [28] state that $(\hat{S}_n, \hat{L}_n)$ is a minimum of the convex program (1.2) if the zero matrix belongs to the subdifferential of the objective function evaluated at $(\hat{S}_n, \hat{L}_n)$ [in addition to $(\hat{S}_n, \hat{L}_n)$ satisfying the constraints]. Elements of the subdifferentials with respect to the ℓ_1 norm and the nuclear norm at a matrix M have the key property that they decompose with respect to the tangent spaces $\Omega(M)$ and $T(M)$ [34]. This decomposition property plays a critical role in our analysis. In particular it states that the optimality conditions consist of two parts, one part corresponding to the tangent spaces Ω and T and another corresponding to the normal spaces $\Omega^\perp$ and $T^\perp$.

Our analysis proceeds by constructing a primal-dual pair of variables that certify optimality with respect to (1.2). Consider the optimization problem (1.2) with

the additional (nonconvex) constraints that the variable S belongs to the algebraic variety of sparse matrices and that the variable L belongs to the algebraic variety of low-rank matrices. While this new optimization problem is nonconvex, it has a very interesting property. At a globally optimal solution (and indeed at any locally optimal solution) $(\tilde{S}, \tilde{L})$ such that $\tilde{S}$ and $\tilde{L}$ are smooth points of the algebraic varieties of sparse and low-rank matrices, the first-order optimality conditions state that the Lagrange multipliers corresponding to the additional variety constraints must lie in the *normal spaces* $\Omega(\tilde{S})^{\perp}$ and $T(\tilde{L})^{\perp}$. This basic observation, combined with the decomposition property of the subdifferentials of the ℓ_1 and nuclear norms, suggests the following high-level proof strategy: considering the solution $(\tilde{S}, \tilde{L})$ of the variety-constrained problem, we show under suitable conditions that the second part of the subgradient optimality conditions of (1.2) (without any variety constraints) corresponding to components in the normal spaces $\Omega(\tilde{S})^{\perp}$ and $T(\tilde{L})^{\perp}$ is also satisfied by $(\tilde{S}, \tilde{L})$. Thus, we show that $(\tilde{S}, \tilde{L})$ satisfies the optimality conditions of the *original convex program* (1.2). Consequently $(\tilde{S}, \tilde{L})$ is also the optimum of the convex program (1.2). As this estimate is obtained as the solution to the problem with the variety constraints, the algebraic correctness of $(\tilde{S}, \tilde{L})$ can be directly concluded. We emphasize here that the variety-constrained optimization problem is used solely as an analysis tool in order to prove consistency of the estimates provided by the convex program (1.2). The key technical complication is that the tangent spaces at $\tilde{L}$ and L^* are in general different. We bound the twisting between these tangent spaces by using the fact that the minimum nonzero singular value of L^* is bounded away from zero (as assumed in Theorem 4.1; see also the supplement [6]).

5.3. *Results proved in supplement.* In this section we give the statements of some results that are proved in a separate supplement [6]. These results are critical to the proof of our main theorem, but they deal mainly with nonstatistical aspects such as the curvature of the algebraic variety of low-rank matrices. Recall that $\Omega = \Omega(S^*)$ and $T = T(L^*)$. We also refer frequently to the constants defined in Theorem 4.1.

As the gradient of the log-determinant function is given by a matrix inverse, a key step in analyzing the properties of the convex program (1.2) is to show that the change in the inverse of a matrix due to small perturbations is well-approximated by the first-order term in the Taylor series expansion. Consider the Taylor series of the inverse of a matrix:

$$(M + \Delta)^{-1} = M^{-1} - M^{-1}\Delta M^{-1} + R_{M^{-1}}(\Delta),$$

where

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

This infinite sum converges for Δ sufficiently small. The following proposition provides a bound on the second-order term specialized to our setting:

PROPOSITION 5.1. *Suppose that γ is in the range given by Proposition 3.3. Further suppose $\Delta_S \in \Omega$, and let $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$. Then we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))) \leq \frac{2D\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2}{\xi(T)}.$$

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

$$(\hat{S}_\Omega, \hat{L}_{\tilde{T}}) = \arg\min_{S,L} \operatorname{tr}[(S-L)\Sigma_O^n] - \log\det(S-L) + \lambda_n[\gamma\|S\|_1 + \|L\|_*] \quad \text{s.t. } S - L \succ 0, S \in \Omega, L \in \tilde{T}, \tag{5.2}$$

for some subspace $\tilde{T}$. Comparing (5.2) with the convex program (1.2), we also do not constrain the variable L to be positive semidefinite in (5.2) for ease of proof of the next result (see the supplement [6] for more details; recall that the nuclear norm of a positive-semidefinite matrix is equal to its trace). We show that if $\tilde{T}$ is any tangent space to the low-rank matrix variety such that $\rho(T, \tilde{T}) \leq \frac{\xi(T)}{2}$, then we can bound the error $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{\tilde{T}})$. Let $\mathcal{C}_{\tilde{T}} = \mathcal{P}_{\tilde{T}^\perp}(L^*)$ denote the normal component of the true low-rank matrix at $\tilde{T}$, and let $E_n = \Sigma_O^n - \Sigma_O^*$ denote the difference between the true marginal covariance and the sample covariance. The proof of the following result uses Brouwer's fixed-point theorem [25], and is inspired by the proof of a similar result in [26] for standard sparse graphical model recovery without latent variables.

PROPOSITION 5.2. *Let the error (Δ_S, Δ_L) in the solution of the convex program (5.2) [with $\tilde{T}$ such that $\rho(\tilde{T}, T) \leq \frac{\xi(T)}{2}$] be as defined above, and define*

$$r = \max\left\{\frac{8}{\alpha}[g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{\tilde{T}}) + \lambda_n], \|\mathcal{C}_{T'}\|_2\right\}.$$

If $r \leq \min\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\}$ for γ as in Proposition 3.3, then $g_\gamma(\Delta_S, \Delta_L) \leq 2r$.

Finally we give a proposition that summarizes the algebraic component of our proof.

PROPOSITION 5.3. *Assume that γ is in the range specified by Proposition 3.3, $\sigma \geq \frac{C_L\lambda_n}{\xi(T)^2}$, $\theta \geq \frac{C_S\lambda_n}{\mu(\Omega)}$, $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$, and that $\lambda_n \leq \frac{3\alpha(2-\nu)}{16(3-\nu)}\min\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\}$. Then there exists a T' and a corresponding unique solution $(\hat{S}_\Omega, \hat{L}_{T'})$ of (5.2) with $\tilde{T} = T'$ with the following properties*:

(1) $\operatorname{sign}(\hat{S}_\Omega) = \operatorname{sign}(S^*)$ *and* $\operatorname{rank}(\hat{L}_{T'}) = \operatorname{rank}(L^*)$, *with* $\hat{L}_{T'} \succeq 0$. *Further* $T(\hat{L}_{T'}) = T'$ *and* $\rho(T, T') \leq \frac{\xi(T)}{4}$.

(2) *Letting* $\mathcal{C}_{T'} = \mathcal{P}_{T'^{\perp}}(L^*)$ *we have that* $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T'}) \le \frac{\lambda_n \nu}{6(2-\nu)}$, *and that* $\|\mathcal{C}_{T'}\|_2 \le \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$.

Further, if $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'}))) \le \frac{\lambda_n \nu}{6(2-\nu)}$, *then the tangent space constraints* $S \in \Omega, L \in T'$ *are inactive in* (5.2). *Consequently the unique solution of* (1.2) *is* $(\hat{S}_n, \hat{L}_n) = (\hat{S}_\Omega, \hat{L}_{T'})$.

5.4. *Probabilistic analysis.* The results given thus far in this section have been completely deterministic in nature. Here we present the probabilistic component of our proof by studying the rate at which the sample covariance matrix Σ_O^n converges to the true covariance matrix Σ_O^* in spectral norm. This result is well known and follows directly from Theorem II.13 in [8]; we mainly discuss it here for completeness and also to show explicitly the dependence on $\psi = \|\Sigma_O^*\|_2$ defined in (3.4). See the supplement [6] for a proof.

LEMMA 5.4. *Let* $\psi = \|\Sigma_O^*\|_2$. *Given any* $\delta > 0$ *with* $\delta \le 8\psi$, *let the number of samples* n *be such that* $n \ge \frac{64 p \psi^2}{\delta^2}$. *Then we have that*

$$\Pr[\|\Sigma_O^n - \Sigma_O^*\|_2 \ge \delta] \le 2\exp\left\{-\frac{n\delta^2}{128\psi^2}\right\}.$$

The following corollary relates the number of samples required for an error bound to hold with probability $1 - 2\exp\{-p\}$.

COROLLARY 5.5. *Let* Σ_O^n *be the sample covariance formed from* n *samples of the observed variables. Set* $\delta_n = \sqrt{\frac{128 p \psi^2}{n}}$. *If* $n \ge 2p$, *then*

$$\Pr[\|\Sigma_O^n - \Sigma_O^*\|_2 \le \delta_n] \ge 1 - 2\exp\{-p\}.$$

PROOF. Note that $n \ge 2p$ implies that $\delta_n \le 8\psi$, and apply Lemma 5.4. □

5.5. *Proof of Theorem 4.1 and Corollary 4.3.* We first combine the results obtained thus far to prove Theorem 4.1. Set $E_n = \Sigma_O^n - \Sigma_O^*$, set $\delta_n = \sqrt{\frac{128 p \psi^2}{n}}$, and then set $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$. This setting of λ_n is equivalent to the specification in the statement of Theorem 4.1.

PROOF OF THEOREM 4.1. We mainly need to show that the various sufficient conditions of Proposition 5.3 are satisfied. We condition on the event that $\|E_n\|_2 \le \delta_n$, which holds with probability greater than $1 - 2\exp\{-p\}$ from Corollary 5.5 as $n \ge 2p$ by assumption. Based on the bound on n, we also have that

$$\delta_n \le \xi(T)^2 \left[\frac{\alpha\nu}{32(3-\nu)D} \min\left\{\frac{1}{4C_1}, \frac{\alpha\nu}{256D(3-\nu)\psi C_1^2}\right\}\right].$$

In particular, these bounds imply that

$$\begin{aligned}\delta_n &\le \frac{\alpha\xi(T)\nu}{32(3-\nu)D}\min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\};\\ \delta_n &\le \frac{\alpha^2\xi(T)^2\nu^2}{8192\psi C_1^2(3-\nu)^2D^2}.\end{aligned} \tag{5.3}$$

Both these weaker bounds are used later.

Based on the assumptions of Theorem 4.1, the requirements of Proposition 5.3 on σ and θ are satisfied. Next we verify the bounds on λ_n and $g_\gamma(\mathcal{A}^\dagger E_n)$. Based on the setting of λ_n above and the bound on δ_n from (5.3), we have that

$$\lambda_n = \frac{6D(2-\nu)\delta_n}{\xi(T)\nu} \le \frac{3\alpha(2-\nu)}{16(3-\nu)}\min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.$$

Next we combine the facts that $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$ and that $\|E_n\|_2 \le \delta_n$ to conclude that

$$g_\gamma(\mathcal{A}^\dagger E_n) \le \frac{D\delta_n}{\xi(T)} = \frac{\lambda_n\nu}{6(2-\nu)}. \tag{5.4}$$

Thus, we have from Proposition 5.3 that there exists a T' and corresponding solution $(\hat{S}_\Omega, \hat{L}_{T'})$ of (5.2) with the prescribed properties. Next we apply Proposition 5.2 with $\tilde{T} = T'$ to bound the error $(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'})$. Noting that $\rho(T, T') \le \frac{\xi(T)}{4}$, we have that

$$\begin{aligned}\frac{8}{\alpha}[g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T'}) + \lambda_n] &\le \frac{8}{\alpha}\left[\frac{\nu}{3(2-\nu)} + 1\right]\lambda_n\\ &= \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)} &(5.5)\\ &= \frac{32(3-\nu)D}{\alpha\xi(T)\nu}\delta_n &(5.6)\\ &\le \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}. &(5.7)\end{aligned}$$

In the first inequality we used the fact that $g_\gamma(\mathcal{A}^\dagger E_n) \le \frac{\lambda_n\nu}{6(2-\nu)}$ (from above) and that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T'})$ is similarly bounded (from Proposition 5.3). In the second equality we used the relation $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$. In the final inequality we used the bound on δ_n from (5.3). This satisfies one of the requirements of Proposition 5.2. The second requirement of Proposition 5.2 on $\|\mathcal{C}_{T'}\|_2$ is also similarly satisfied as we have that $\|\mathcal{C}_{T'}\|_2 \le \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$ from Proposition 5.3, and we use the same sequence of inequalities as above. Thus we conclude from Proposition 5.2 and from

(5.5) that

$$g_\gamma(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'}) \le \frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)} \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}. \tag{5.8}$$

Here the last inequality follows from the bound on λ_n.

If we show that $(\hat{S}_n, \hat{L}_n) = (\hat{S}_\Omega, \hat{L}_{T'})$, we can conclude the proof of Theorem 4.1 since algebraic correctness of $(\hat{S}_\Omega, \hat{L}_{T'})$ holds from Proposition 5.3 and the estimation error bound follows from (5.8). In order to complete this final step, we again revert to Proposition 5.3 and prove the requisite bound on $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'})))$.

Since the bound (5.8) combined with the inequality (5.7) satisfies the condition of Proposition 5.1 [i.e., we have that $g_\gamma(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'}) \le \frac{1}{2C_1}$]:

$$\begin{aligned}
g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'}))) &\le \frac{2D\psi C_1^2}{\xi(T)} g_\gamma(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'})^2 \\
&\le \frac{2D\psi C_1^2}{\xi(T)}\left(\frac{64(3-\nu)D}{\alpha\xi(T)\nu}\right)^2 \delta_n^2 \\
&= \left[\frac{8192\psi C_1^2(3-\nu)^2 D^2}{\alpha^2\xi(T)^2\nu^2}\delta_n\right]\frac{D\delta_n}{\xi(T)} \\
&\le \frac{D\delta_n}{\xi(T)} \\
&= \frac{\lambda_n \nu}{6(2-\nu)}.
\end{aligned}$$

In the second inequality we used (5.6) and (5.8), in the final inequality we used the bound (5.3) on δ_n, and in the final equality we used the relation $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$. □

PROOF OF COROLLARY 4.3. Based on the optimality conditions of the modified convex program (5.2), we have that

$$g_\gamma\big(\mathcal{A}^\dagger[(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma_O^n]\big) \le \lambda_n.$$

Combining this with the bound (5.4) yields the desired result. □

6. Simulation results. In this section we give experimental demonstration of the consistency of our estimator (1.2) on synthetic examples, and its effectiveness in modeling real-world stock return data. Our choices of λ_n and γ are guided by Theorem 4.1. Specifically, we choose λ_n to be proportional to $\sqrt{\frac{p}{n}}$. For γ we observe that the support/sign-pattern and the rank of the solution $(\hat{S}_n, \hat{L}_n)$ are the same for a *range* of values of γ. Therefore one could solve the convex program

(1.2) for several values of γ, and choose a solution in a suitable range in which the sign-pattern and rank of the solution are stable (see [7] for details). In practical problems with real-world data these parameters may be chosen via cross-validation (it would be of interest to consider methods such as those developed in [24]). For small problem instances we solve the convex program (1.2) using a combination of YALMIP [21] and SDPT3 [31]. For larger problem instances we use the special-purpose solver LogdetPPA [33] developed for log-determinant semidefinite programs.

6.1. *Synthetic data*. In the first set of experiments we consider a setting in which we have access to samples of the observed variables of a latent-variable graphical model. We consider several latent-variable Gaussian graphical models. The first model consists of $p = 36$ observed variables and $h = 2$ latent variables. The conditional graphical model structure of the observed variables is a cycle with the edge partial correlation coefficients equal to 0.25; thus, this conditional model is specified by a sparse graphical model with degree 2. The second model is the same as the first one, but with $h = 3$ latent variables. The third model consists of $h = 1$ latent variable, and the conditional graphical model structure of the observed variables is given by a 6×6 nearest-neighbor grid (i.e., $p = 36$ and degree 4) with the partial correlation coefficients of the edges equal to 0.15. In all three of these models each latent variable is connected to a random subset of 80% of the observed variables (and the partial correlation coefficients corresponding to these edges are also random). Therefore the effect of the latent variables is "spread out" over most of the observed variables, that is, the low-rank matrix summarizing the effect of the latent variables is incoherent.

For each model we generate n samples of the observed variables, and use the resulting sample covariance Σ_O^n as input to our convex program (1.2). Figure 1 shows the probability of obtaining algebraically correct estimates as a function of n. This probability is evaluated over 50 experiments for each value of n. In all of these cases standard graphical model selection applied directly to the observed variables is not useful as the marginal concentration matrix of the observed variables is not well-approximated by a sparse matrix. These experiments agree with our theoretical results that the convex program (1.2) is an algebraically consistent estimator of a latent-variable model given (sufficiently many) samples of only the observed variables.

6.2. *Stock return data*. In the next experiment we model the statistical structure of monthly stock returns of 84 companies in the S&P 100 index from 1990 to 2007; we disregard 16 companies that were listed after 1990. The number of samples n is equal to 216. We compute the sample covariance based on these returns and use this as input to (1.2).

The model learned using (1.2) for suitable values of λ_n, γ consists of $h = 5$ latent variables, and the conditional graphical model structure of the stock returns conditioned on these latent components consists of 135 edges. Therefore

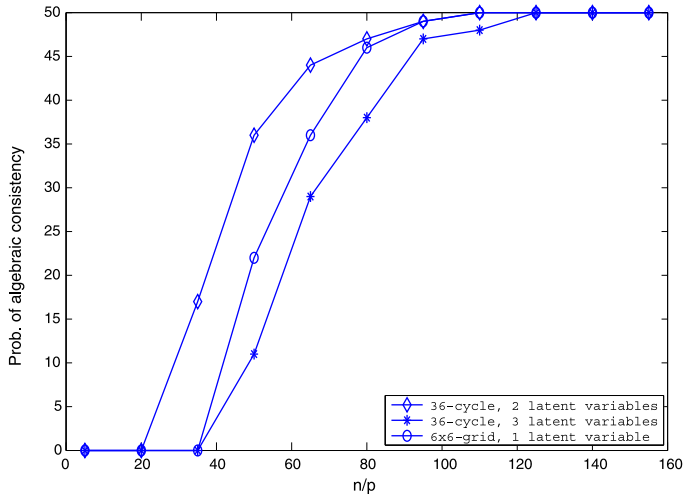


FIG. 1. *Synthetic data*: *plot showing probability of algebraically correct estimation*. *The three models studied are* (a) 36-*node conditional graphical model given by a cycle with* $h = 2$ *latent variables*, (b) 36-*node conditional graphical model given by a cycle with* $h = 3$ *latent variables and* (c) 36-*node conditional graphical model given by a* 6×6 *grid with* $h = 1$ *latent variable*. *For each plotted point*, *the probability of algebraically correct estimation is obtained over* 50 *random trials*.

the number of parameters in the model is $84 + 135 + (5 \times 84) = 639$. The resulting KL divergence between the distribution specified by this model and a Gaussian distribution specified by the sample covariance is 17.7. Figure 2 (left) shows the *conditional* graphical model structure. The strongest edges in this conditional graphical model, as measured by partial correlation, are between Baker Hughes–

FIG. 2. *Stock returns*: *the figure on the left shows the sparsity pattern* (*black denotes an edge*, *and white denotes no edge*) *of the concentration matrix of the conditional graphical model* (135 *edges*) *of the stock returns*, *conditioned on five latent variables*, *in a latent-variable graphical model* (*total number of parameters equals* 639). *This model is learned using* (1.2), *and the KL divergence with respect to a Gaussian distribution specified by the sample covariance is* 17.7. *The figure on the right shows the concentration matrix of the graphical model* (646 *edges*) *of the stock returns*, *learned using standard sparse graphical model selection based on solving an* ℓ_1*-regularized maximum-likelihood program* (*total number of parameters equals* 730). *The KL divergence between this distribution and a Gaussian distribution specified by the sample covariance is* 44.4.

Schlumberger, A.T.&T.–Verizon, Merrill Lynch–Morgan Stanley, Halliburton–Baker Hughes, Intel–Texas Instruments, Apple–Dell, and Microsoft–Dell. It is of interest to note that in the Standard Industrial Classification[4] system for grouping these companies, several of these pairs are in different classes. As mentioned in Section 2.1, our method estimates a low-rank matrix that summarizes the effect of the latent variables; in order to factorize this low-rank matrix, for example, into sparse factors, one could use methods such as those described in [35].

We compare these results to those obtained using a sparse graphical model learned using ℓ_1-regularized maximum-likelihood (see, e.g., [26]), without introducing any latent variables. Figure 2 (right) shows this graphical model structure. The number of edges in this model is 646 (the total number of parameters is equal to $646 + 84 = 730$), and the resulting KL divergence between this distribution and a Gaussian distribution specified by the sample covariance is 44.4.

These results suggest that a latent-variable graphical model is better suited than a standard sparse graphical model for modeling stock returns. This is likely due to the presence of global, long-range correlations in stock return data that are better modeled via latent variables.

7. Discussion. We have studied the problem of modeling the statistical structure of a collection of random variables as a sparse graphical model conditioned on a few additional latent components. As a first contribution we described conditions under which such latent-variable graphical models are identifiable given samples of only the observed variables. We also proposed a convex program based on ℓ_1 and nuclear norm regularized maximum-likelihood for latent-variable graphical model selection. Given samples of the observed variables of a latent-variable Gaussian model, we proved that this convex program provides consistent estimates of the number of latent components as well as the conditional graphical model structure among the observed variables conditioned on the latent components. Our analysis holds in the high-dimensional regime in which the number of observed/latent variables are allowed to grow with the number of samples of the observed variables. These theoretical predictions are verified via a set of experiments on synthetic data. We also demonstrate the effectiveness of our approach in modeling real-world stock return data.

Several questions arise that are worthy of further investigation. While (1.2) can be solved in polynomial time using off-the-shelf solvers, it is preferable to develop more efficient special-purpose solvers to scale to massive datasets by taking advantage of the structure of (1.2). It is also of interest to develop statistically consistent convex optimization methods for latent-variable modeling with non-Gaussian variables, for example, for categorical data.

[4]See the U.S. SEC website at http://www.sec.gov/info/edgar/siccodes.htm.

Acknowledgments. We would like to thank James Saunderson and Myung Jin Choi for helpful discussions, and Kim-Chuan Toh for kindly providing us specialized code to solve larger instances of our convex program.

SUPPLEMENTARY MATERIAL

Supplement to "Latent variable graphical model selection via convex optimization" (DOI: 10.1214/11-AOS949SUPP; .pdf). Due to space constraints, we have moved some technical proofs to a supplementary document [6].

REFERENCES

[1] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[2] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[3] CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000

[4] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

[5] CANDÈS, E. J., ROMBERG, J. and TAO, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* **52** 489–509. MR2236170

[6] CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2011). Supplement to "Latent variable graphical model selection via convex optimization." DOI:10.1214/11-AOS949SUPP.

[7] CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479

[8] DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces, Vol. I* 317–366. North-Holland, Amsterdam. MR1863696

[9] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

[10] DONOHO, D. L. (2006). For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* **59** 797–829. MR2217606

[11] DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

[12] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

[13] ELIDAN, G., NACHMAN, I. and FRIEDMAN, N. (2007). "Ideal parent" structure learning for continuous variable Bayesian networks. *J. Mach. Learn. Res.* **8** 1799–1833. MR2353820

[14] FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991

[15] FAZEL, M. (2002). Matrix rank minimization with applications. Ph.D. thesis, Dept. Elec. Eng., Stanford Univ.

[16] HORN, R. A. and JOHNSON, C. R. (1990). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1084815

[17] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[18] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

[19] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

[20] LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339

[21] LÖFBERG, J. (2004). YALMIP: A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference, Taiwan*. Available at http://control.ee.ethz.ch/~joloef/yalmip.php.

[22] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb.* **72** 507–536. MR0208649

[23] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[24] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

[25] ORTEGA, J. M. and RHEINBOLDT, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York. MR0273810

[26] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **4** 935–980.

[27] RECHT, B., FAZEL, M. and PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501. MR2680543

[28] ROCKAFELLAR, R. T. (1996). *Convex Analysis*. Princeton Univ. Press, Princeton, NJ. MR1451876

[29] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[30] SPEED, T. P. and KIIVERI, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14** 138–150. MR0829559

[31] TOH, K. C., TODD, M. J. and TUTUNCU, R. H. (1999). SDPT3—a MATLAB software package for semidefinite-quadratic-linear programming. Available at http://www.math.nus.edu.sg/~mattohkc/sdpt3.html.

[32] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[33] WANG, C., SUN, D. and TOH, K.-C. (2010). Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *SIAM J. Optim.* **20** 2994–3013. MR2735941

[34] WATSON, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.* **170** 33–45. MR1160950

[35] WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

[36] WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. MR2024760

[37] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

V. CHANDRASEKARAN
DEPARTMENT OF COMPUTING
AND MATHEMATICAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91106
USA
E-MAIL: venkatc@caltech.edu

P. A. PARRILO
A. S. WILLSKY
LABORATORY FOR INFORMATION
AND DECISION SYSTEMS
DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: parrilo@mit.edu
willsky@mit.edu

The Annals of Statistics
2012, Vol. 40, No. 4, 1968–1972
DOI: 10.1214/12-AOS979
Main article DOI: 10.1214/11-AOS949

DISCUSSION: LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION[1]

BY MING YUAN

Georgia Institute of Technology

I want to start by congratulating Professors Chandrasekaran, Parrilo and Willsky for this fine piece of work. Their paper, hereafter referred to as CPW, addresses one of the biggest practical challenges of Gaussian graphical models—how to make inferences for a graphical model in the presence of missing variables. The difficulty comes from the fact that the validity of conditional independence relationships implied by a graphical model relies critically on the assumption that all conditional variables are observed, which of course can be unrealistic. As CPW shows, this is not as hopeless as it might appear to be. They characterize conditions under which a conditional graphical model can be identified, and offer a penalized likelihood method to reconstruct it. CPW notes that with missing variables, the concentration matrix of the observables can be expressed as the difference between a sparse matrix and a low-rank matrix; and suggests to exploit the sparsity using an ℓ_1 penalty and the low-rank structure by a trace norm penalty. In particular, the trace norm penalty or, more generally, nuclear norm penalties, can be viewed as a convex relaxation to the more direct rank constraint. Its use oftentimes comes as a necessity because rank constrained optimization could be computationally prohibitive. Interestingly, as I note here, the current problem actually lends itself to efficient algorithms in dealing with the rank constraint, and therefore allows for an attractive alternative to the approach of CPW.

1. Rank constrained latent variable graphical Lasso. Recall that the penalized likelihood estimate of CPW is defined as

$$(\hat{S}_n, \hat{L}_n) = \underset{L\succeq 0, S-L\succ 0}{\arg\min} \{-\ell(S-L, \Sigma_O^n) + \lambda_n(\gamma\|S\|_1 + \operatorname{trace}(L))\},$$

where the vector ℓ_1 norm and trace/nuclear norm penalties are designated to induce sparsity among elements of S and low-rank structure of L respectively. Of course, we can attempt a more direct rank penalty as opposed to the nuclear norm penalty on L, leading to

$$(\hat{S}_n, \hat{L}_n) = \underset{L\succeq 0, S-L\succ 0}{\arg\min} \{-\ell(S-L, \Sigma_O^n) + \lambda_n(\gamma\|S\|_1 + \operatorname{rank}(L))\};$$

Received February 2012.
[1]Supported in part by NSF Career Award DMS-08-46234.

or for computational purposes, it is more convenient to consider the constrained version:

$$(\hat{\hat{S}}_n, \hat{\hat{L}}_n) = \underset{\substack{L\succeq 0, S-L\succ 0\\ \operatorname{rank}(L)\le r}}{\arg\min} \{-\ell(S-L, \Sigma_O^n) + \lambda_n \|S^\dagger\|_1\},$$

for some integer $0 \le r \le p$, where $S^\dagger = S - \operatorname{diag}(S)$, that is, $S^\dagger$ equals S except that its diagonals are replaced by 0. This slight modification reflects our intention to encourage sparsity on the off-diagonal entries of S only. The remaining discussion, however, can be easily adapted to deal with the original vector ℓ_1 penalty on S. It is clear that when $r = 0$, that is, $L = 0$, this new estimator reduces to the so-called graphical Lasso estimate (`glasso`, for short) of Yuan and Lin (2007). See also Banerjee, El Ghaoui and d'Aspremont (2008), Friedman, Hastie and Tibshirani (2008), and Rothman et al. (2008). Drawn to this similarity, I shall hereafter refer to this method as the latent variable graphical Lasso, or `LVglasso`, for short.

Common wisdom on $(\hat{\hat{S}}_n, \hat{\hat{L}}_n)$ is that it is infeasible to compute because of the nonconvexity of the rank constraint. Interestingly, though, this more direct approach actually allows for fast computation, thanks to a combination of EM algorithm and some recent advances in computing graphical Lasso estimates for high-dimensional problems.

2. An EM algorithm. The constraint $\operatorname{rank}(L) \le r$ amounts to postulating r latent variables. The latent variable model naturally has a missing data formulation. It is clear that when observing the complete data $X = (X_O^\mathsf{T}, X_H^\mathsf{T})^\mathsf{T}$, the `LVglasso` estimator becomes

$$\hat{K}_\lambda = \underset{K\in\mathbb{R}^{(p+r)\times(p+r)}, K\succ 0}{\arg\min} \{L(K) + \lambda \|K_O^\dagger\|_1\},$$

where

$$L(K) = -\ln\det(K) + \operatorname{trace}\bigl(\Sigma_{(OH)}^n K\bigr)$$

and $\Sigma_{(OH)}^n$ is the sample covariance matrix of the full data. Now that X_H is unobservable, we can use an EM algorithm which iteratively applies the following two steps:

EXPECTATION STEP (E STEP). Calculate the expected value of the penalized negative log-likelihood function, with respect to the conditional distribution of X_H given X_O under the current estimate $K^{(t)}$ of K, leading to the so-called Q function:

$$\begin{aligned} Q\bigl(K|K^{(t)}\bigr) &= \mathbb{E}_{X_H|X_O,K^{(t)}}[L(K) + \lambda\|K_O^\dagger\|_1] \\ &= -\ln\det(K) + \operatorname{trace}\bigl\{\mathbb{E}_{X_H|X_O,K^{(t)}}\bigl(\Sigma_{(OH)}^n\bigr)K\bigr\} + \lambda\|K_O^\dagger\|_1. \end{aligned}$$

Recall that $X_H|X_O, K^{(t)}$ follows a normal distribution with

$$\mathbb{E}(X_H|X_O, K^{(t)}) = \Sigma_{HO}^{(t)}(\Sigma_O^{(t)})^{-1}X_O$$

and

$$\mathrm{Var}(X_H|X_O, K^{(t)}) = \Sigma_H^{(t)} - \Sigma_{HO}^{(t)}(\Sigma_O^{(t)})^{-1}\Sigma_{OH}^{(t)},$$

where $\Sigma^{(t)} = (K^{(t)})^{-1}$. Therefore,

$$\mathbb{E}_{X_H|X_O,K^{(t)}}(\Sigma_{OH}^n) = \Sigma_O^n(\Sigma_O^{(t)})^{-1}\Sigma_{oH}^{(t)}$$

and

$$\mathbb{E}_{X_H|X_O,K^{(t)}}(\Sigma_H^n) = \Sigma_H^{(t)} - \Sigma_{HO}^{(t)}(\Sigma_O^{(t)})^{-1}\Sigma_{OH}^{(t)} + \Sigma_{HO}^{(t)}(\Sigma_O^{(t)})^{-1}\Sigma_O^n(\Sigma_O^{(t)})^{-1}\Sigma_{OH}^{(t)}.$$

MAXIMIZATION STEP (M STEP). Maximize $Q(\cdot|K^{(t)})$ over all $(p+r)\times(p+r)$ positive definite matrices. We first note that if we replace the penalty term $\|K_O^\dagger\|_1$ with $\|K^\dagger\|_1$, then maximizing $Q(\cdot|K^{(t)})$ becomes a `glasso` problem:

$$\max_{K\in\mathbb{R}^{(p+r)\times(p+r)}, K\succ 0}\{-\ln\det(K) + \mathrm{trace}\{WK\} + \lambda\|K^\dagger\|_1\},$$

where $W = \mathbb{E}_{X_H|X_O,K^{(t)}}(\Sigma_{(OH)}^n)$. As shown in Banerjee, El Ghaoui and d'Aspremont (2008), Friedman, Hastie and Tibshirani (2008) and Yuan (2008), this problem can be solved iteratively. At each iteration, one row and, correspondingly, one column of K, due to symmetry, are updated by solving a Lasso problem. The same idea can be applied here to maximize $Q(\cdot|K^{(t)})$. The only difference is that in each of the Lasso problems, we leave the coordinates corresponding to the latent variables unpenalized. This extension has been implemented in the `R` package `glasso` [Friedman, Hastie and Tibshirani (2008)].

3. Example. For illustration purposes, I conducted a simple numerical experiment. In this experiment the interest was in recovering a $p = 198$ dimensional graphical model with $h = 2$ missing variables. The graphical model was generated in a similar fashion as that from Meinshausen and Bühlmann (2006). I first simulated 198 locations uniformly over a square. Between each pair of locations, I put an edge with probability $2\phi(d\sqrt{p})$, where $\phi(\cdot)$ is the density function of the standard normal distribution and d is the distance between the two locations, unless one of the locations is already connected with four other locations. The two hidden variables were connected with all p observed variables. The entries of the inverse covariance matrix corresponding to the edges between the observables were assigned with value 0.2, between the observables and the latent variables were assigned with a uniform random value between 0 and 0.12, to ensure the positive definiteness. A typical simulated graphical model among the 198 observed variables conditional on the two latent variables is given in the top left panel of Figure 1. We apply both the method of CPW and `LVglasso`, along with `glasso`,

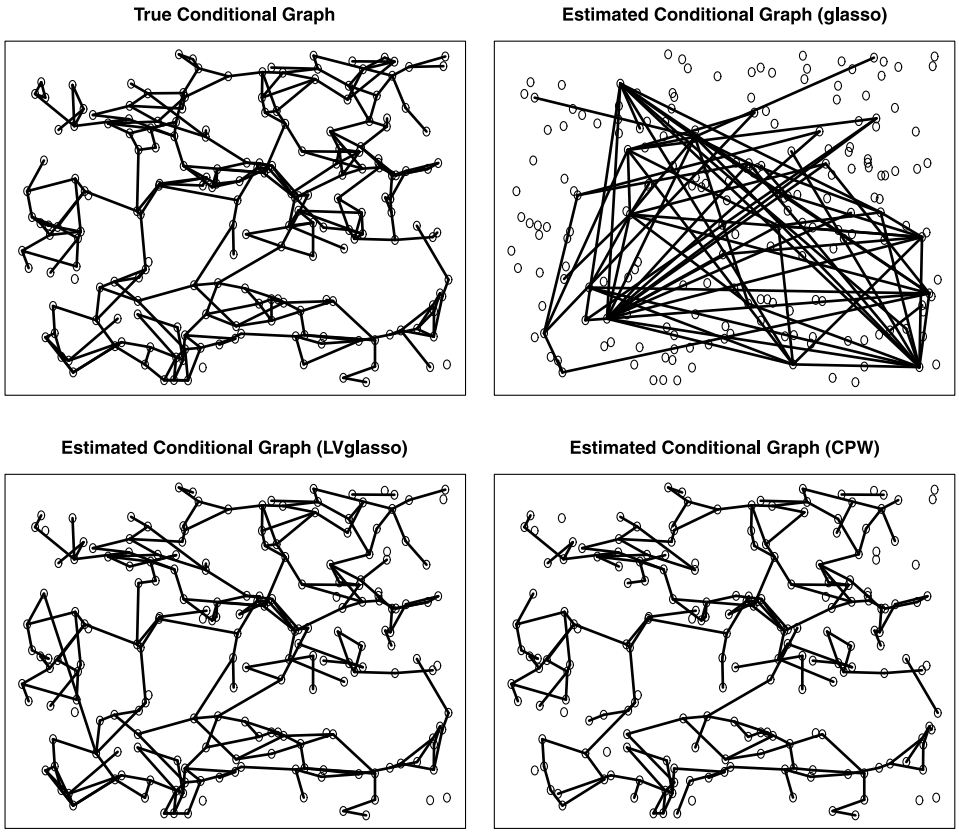


FIG. 1. *True graphical model and its estimates.*

to the data. We used the MATLAB code provided by CPW to compute their estimates. As observed by CPW, their estimate typically is insensitive to a wide range of values of γ, and we report here the results with the default choice of $\gamma = 5$ without loss of generality. Similarly, for LVglasso, little variation was observed for $r = 2, \ldots, 10$, and we shall focus on $r = 2$ for brevity. The choice of λ plays a critical role for both methods. We compute both estimators for a fine grid of λ. With the main focus on recovering the conditional graphical model, that is, the sparsity pattern of S, we report in Figure 2 the ROC curve for both methods. For contrast, we also reported the result for glasso which neglects the missingness. In Figure 1, we also presented the estimated graphical model for each method that is closest to the truth. These results clearly demonstrate the necessity of accounting for the latent variables. It is also interesting to note that the rank constrained estimator performs slightly better in this example over the trace norm penalization method of CPW.

The preliminary results presented here suggest that direct rank constraint may provide a competitive alternative to the trace norm penalization for recovering

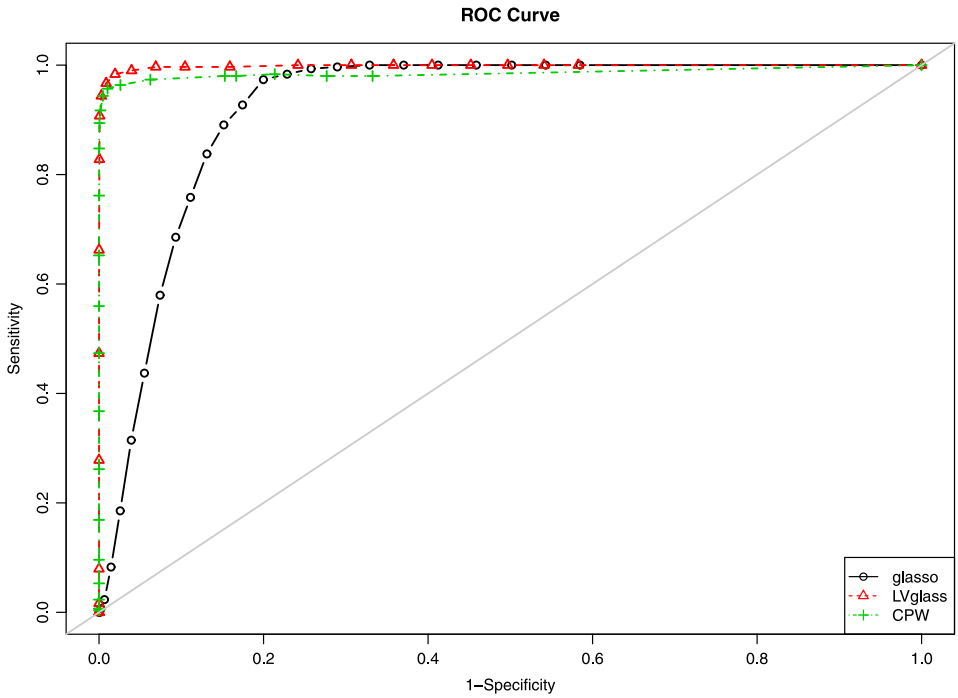


FIG. 2. *Accuracy of reconstructed conditional graphical model.*

graphical models with latent variables. It is of interest to investigate more rigorously how the two methods compare with each other.

REFERENCES

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, T. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

YUAN, M. (2008). Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models. *J. Comput. Graph. Statist.* **17** 809–826. MR2649068

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

H. MILTON STEWART SCHOOL OF INDUSTRIAL
AND SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332
USA

The Annals of Statistics
2012, Vol. 40, No. 4, 1973–1977
DOI: 10.1214/12-AOS980
Main article DOI: 10.1214/11-AOS949

DISCUSSION: LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION

BY STEFFEN LAURITZEN AND NICOLAI MEINSHAUSEN

University of Oxford

We want to congratulate the authors for a thought-provoking and very interesting paper. Sparse modeling of the concentration matrix has enjoyed popularity in recent years. It has been framed as a computationally convenient convex ℓ_1-constrained estimation problem in Yuan and Lin (2007) and can be applied readily to higher-dimensional problems. The authors argue—we think correctly—that the sparsity of the concentration matrix is for many applications more plausible after the effects of a few latent variables have been removed. The most attractive point about their method is surely that it is formulated as a convex optimization problem. Latent variable fitting and sparse graphical modeling of the conditional distribution of the observed variables can then be obtained through a single fitting procedure.

Practical aspects. The method deserves wide adoption, but this will only be realistic if software is made available, for example, as an R-package. Not many users will go to the trouble of implementing the method on their own, so we will strongly urge the authors to do so.

An imputation method. In the absence of readily available software, it is worth thinking whether the proposed fitting procedure can be approximated by methods involving known and well-tested computational techniques. The concentration matrix of observed and hidden variables is

$$K = \begin{pmatrix} K_O & K_{OH} \\ K_{HO} & K_H \end{pmatrix},$$

where we have deviated from the notation in the paper by omitting the asterisk. The proposed estimator $\hat{S}_n = \hat{K}_O$ of K_O was defined as

$$(\hat{K}_O, \hat{L}_n) = \operatorname{argmin}_{S,L} -\ell(S - L; \Sigma_O^n) + \lambda_n\big(\gamma \|S\|_1 + \operatorname{tr}(L)\big) \tag{1}$$

$$\text{such that } S - L \succ 0, L \succ 0, \tag{2}$$

where Σ_O^n is the empirical covariance matrix of the observed variables.

An alternative would be to replace the nuclear-norm penalization with a fixed constraint κ on the rank of the hidden variables, replacing problem (1) with

$$(\hat{K}_O, \hat{L}_n) = \operatorname{argmin}_{S,L} -\ell(S - L; \Sigma_O^n) + \lambda_n \|S\|_1$$
$$\text{such that } S - L \succ 0 \text{ and } L \succ 0 \text{ and } \operatorname{rank}(L) \le \kappa. \tag{3}$$

Received February 2012.

This can be achieved by a missing-value formulation in combination with use of the EM algorithm, which also applies in a penalized likelihood setting [Green (1990)]. Let the hidden variables be of a fixed dimensionality κ and assume for a moment these are observed so one would find the concentration matrix $\hat{K}$ of the joint distribution of the observed variables X_O and hidden variables X_H based on the complete data penalized likelihood as

$$\text{argmin}_K - \log f_K(X_O, X_H) + \lambda \|K_O\|_1, \tag{4}$$

where f_K is the joint density of (X_O, X_H). This formulation is very similar to the missing-value problem treated in Städler (2012), except for the fact that we only penalize the concentration matrix K_O of the observed variables, in analogy with the proposed latent-variable approach. The EM algorithm iteratively replaces the likelihood in (4) for $t = 1, \ldots, T$ by its conditional expectation and thus finds $\hat{K}^{t+1}$ as

$$\hat{K}^{t+1} = \text{argmin}_K - E_{\hat{K}^t}\{\log f_K(X_O, X_H)|X_O\} + \lambda \|K_O\|_1. \tag{5}$$

The iteration is guaranteed not to increase the negative marginal penalized likelihood at every stage and will, save for unidentifiability, converge to the minimizer in (3) for most starting values. Without loss of generality, one can fix the conditional concentration matrix K_H of the hidden variables to be the identity so that these are conditionally independent with variance 1, given the observed variables. Then $-K_{OH}$ is equal to the regression coefficients of the observed variables on the hidden variables. As starting value we have let $-\hat{K}^0_{OH}$ be equal to these with hidden variables determined by a principal component analysis.

The expectation in (5) can be written as the log-likelihood of a Gaussian distribution with concentration matrix K and empirical covariance matrix W^t, where

$$W^t = \begin{pmatrix} \Sigma^n_O & -\Sigma^n_O \hat{K}^t_{OH} \\ -\hat{K}^t_{HO}\Sigma^n_O & \mathbf{I} + \hat{K}^t_{HO}\Sigma^n_O \hat{K}^t_{OH} \end{pmatrix}.$$

The sufficient statistics involving the missing data are thus "imputed" in W^t. Each of the updates (5) can now be computed with the *graphical lasso* [Friedman, Hastie and Tibshirani (2008)].

We thought it would be interesting to compare the two methods on the data example given in the paper. Figure 1 shows the solution $\hat{K}_O$ for the stock-return example when using the proposed method (1) and the imputation method (4) with 4 iterations. The number κ of latent variables and the number of nonzero edges in $\hat{K}_O$ is adjusted to be the same as in the original estimator.

The three pairs with the highest absolute entries in the fitted conditional concentration matrix are identical (AT&T—Verizon, Schlumberger—Baker Hughes and Merrill Lynch—Morgan Stanley) for the two methods and the 15 pairs with highest absolute entries in the off-diagonal concentration matrix have an overlap of size 12. The resulting graphs are slightly different although they share many

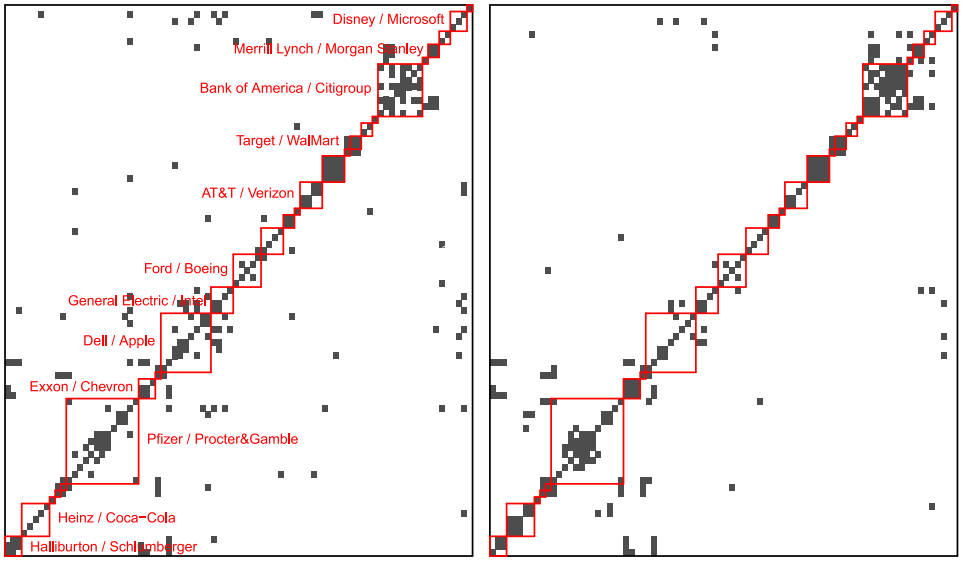


FIG. 1. *The nonzero entries of the concentration matrix* $\hat{K}_O$, *using the proposed procedure* (1) (*left*) *and the imputation method in* (4) (*right*). *Two representative companies are shown for some of the sectors*.

features. Our graph has 136 edges, one more than that in the procedure described in the paper, and 77 of the edges are shared. Our graph has more isolated vertices (15 vs. 9), slightly fewer cliques (62 vs. 81) and the largest clique in our graph has six variables rather than four. The graph is displayed to the left in Figure 2 and features some clearly identified clusters of variables.

The selected graph is very unstable under bootstrap simulations. In the spirit of Meinshausen and Bühlmann (2010), we fit the graph on 2000 bootstrap samples. Only 28 edges are selected in more than half of these samples. The resulting graph is shown in Figure 2. As many as 25 of these edges appear also as edges of the estimator proposed in (1). It would have been interesting to be able to compare with the same "stability graph" of the proposed procedure but we suspect that they will match closely.

Latent directed structures. In a sense the procedure described in this paper can be seen as a modification of, or an alternative to, factor analysis, in which independent latent variables are sought to explain all the correlations, corresponding to the graph for the observed variables being completely empty.

Methods for identifying such models can, for example, be developed using tetrad constraints [Spirtes, Glymour and Scheines (1993), Drton, Sturmfels and Sullivant (2007)]. Another generalization of factor analysis is to look for sparse *directed* graphical models, which have now been rather well established through, for example, the FCI algorithm [Spirtes, Glymour and Scheines (1993), Richardson

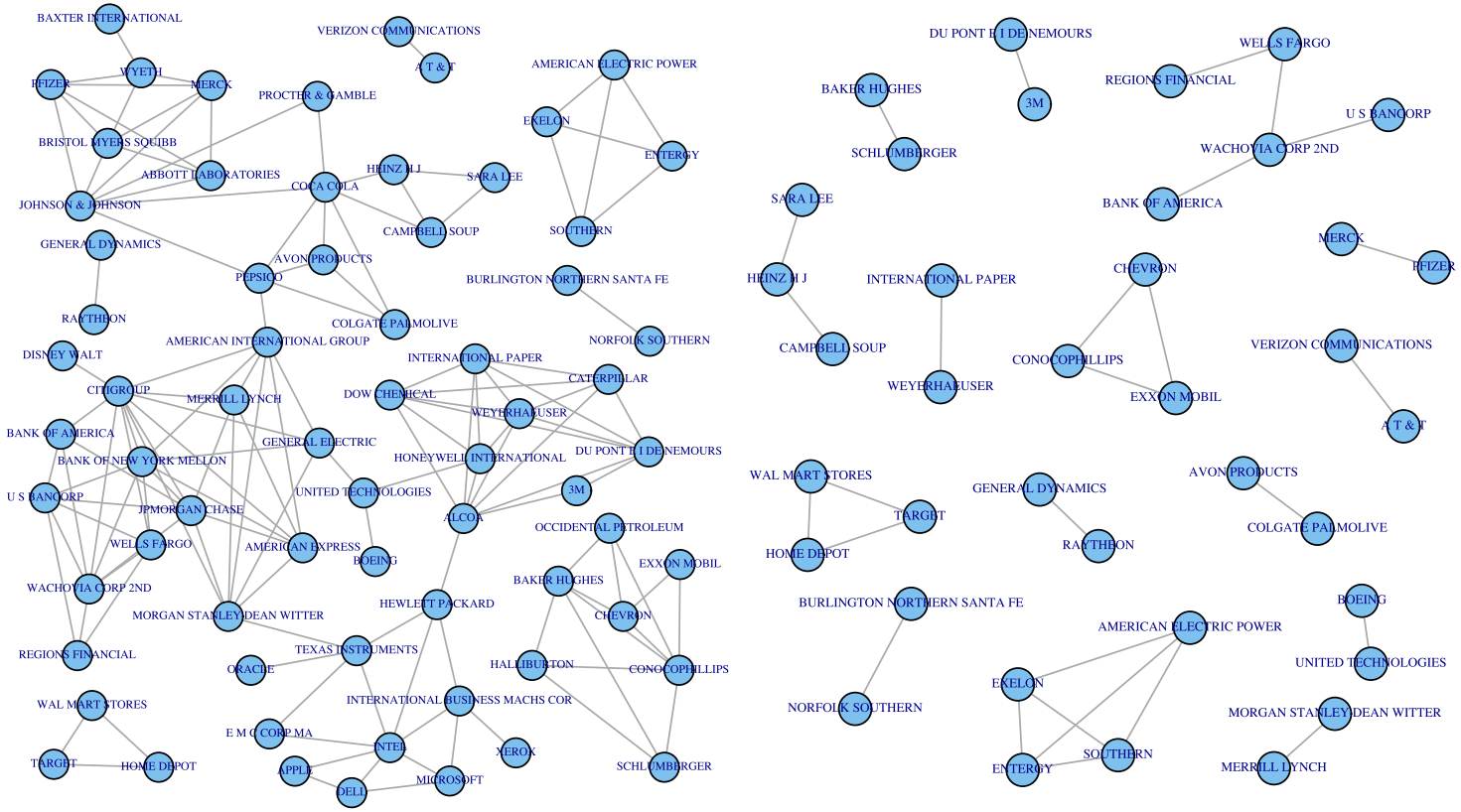


FIG. 2. *Left*: *the graph of the imputation method as in* (4). *Right*: *the graph of the stable edges*. *In both cases*, *isolated vertices have been removed from the display.*

and Spirtes (2002)] with an algebraic underpinning in Sullivant (2008). Again this could be an alternative to the procedure described in this interesting paper.

Summary. We effectively replaced the nuclear norm penalization of L in the paper by a fixed constraint on the rank. This might be easier to do than choosing a reasonable value for the penalty on the trace of L. Using this formulation, we could combine the EM algorithm with the graphical lasso, enabling us to compute the solution with readily available software. It would be interesting to see whether our procedure can be shown to recover the correct sparsity structure under similar assumptions to those in the paper. We want to congratulate the authors again for a very interesting discussion paper.

REFERENCES

DRTON, M., STURMFELS, B. and SULLIVANT, S. (2007). Algebraic factor analysis: Tetrads, pentads and beyond. *Probab. Theory Related Fields* **138** 463–493. MR2299716

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432.

GREEN, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52** 443–452. MR1086796

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

RICHARDSON, T. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search. Lecture Notes in Statistics* **81**. Springer, New York. MR1227558

STÄDLER, N. and BÜHLMANN, P. (2012). Missing values: Sparse inverse covariance estimation and an extension to sparse regression. *Statist. Comput.* **22** 219–235.

SULLIVANT, S. (2008). Algebraic geometry of Gaussian Bayesian networks. *Adv. in Appl. Math.* **40** 482–513. MR2412156

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
1 SOUTH PARKS ROAD
OXFORD, OX1 3TG
UNITED KINGDOM
E-MAIL: meinshausen@stats.ox.ac.uk

The Annals of Statistics
2012, Vol. 40, No. 4, 1978–1983
DOI: 10.1214/12-AOS981
Main article DOI: 10.1214/11-AOS949

DISCUSSION: LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION

BY MARTIN J. WAINWRIGHT

University of California at Berkeley

1. Introduction. It is my pleasure to congratulate the authors for an innovative and inspiring piece of work. Chandrasekaran, Parrilo and Willsky (hereafter CPW) have come up with a novel approach, combining ideas from convex optimization and algebraic geometry, to the long-standing problem of Gaussian graphical model selection with latent variables. Their method is intuitive and simple to implement, based on solving a convex log-determinant program with suitable choices of regularization. In addition, they establish a number of attractive theoretical guarantees that hold under high-dimensional scaling, meaning that the graph size p and sample size n are allowed to grow simultaneously.

1.1. *Background.* Recall that an undirected graphical model (also known as a Markov random field) consists of a family of probability distributions that factorize according to the structure of undirected graph $G = (V, E)$. In the multivariate Gaussian case, the factorization translates into a sparsity assumption on the inverse covariance or precision matrix [9]. In particular, given a multivariate Gaussian random vector $(X_1, \ldots, X_p)$ with covariance matrix Σ, it is said to be Markov with respect to the graph G if its precision matrix $K = \Sigma^{-1}$ has zeroes for each distinct pair of indices (j, k) *not* in the edge set E of the graph. Consequently, the sparsity pattern of the inverse covariance K encodes the edge structure of the graph. The goal of Gaussian graphical model selection is to determine this unknown edge structure, and hence the sparsity pattern of the inverse covariance matrix. It can also be of interest to estimate the matrices K or Σ, for instance, in the Frobenius or ℓ_2-operator norm sense. In recent years, under the assumption that all entries of X are fully observed, a number of practical methods have been proposed and shown to perform well under high-dimensional scaling (e.g., [2, 5–7]).

Chandrasekaran et al. tackle a challenging extension of this problem, in which one observes only p coordinates of a larger $p + h$ dimensional Gaussian random vector. In this case, the $p \times p$ precision matrix K of the observed components need not be sparse, but rather, by an application of the Schur complement formula, can be written as the difference $K = S^* - L^*$. The first matrix S^* is sparse, whereas the second matrix L^* is not sparse (at least in general), but has rank at most h, corresponding to the number of latent or hidden variables. Consequently, the problem

Received February 2012.

of latent Gaussian graphical model selection can be cast as a form of *matrix decomposition*, involving a splitting of the precision matrix into sparse and low-rank components. Based on this nice insight, CPW propose a natural M-estimator for this problem, based on minimizing a regularized form of the (negative) log likelihood for a multivariate Gaussian, where the elementwise ℓ_1-norm is used as a proxy for sparsity, and the nuclear or trace norm as a proxy for rank. Overall, the method is based on the convex program

$$(\widehat{S}, \widehat{L}) \in \arg\min\bigl\{-\ell(S-L; \widehat{\Sigma}^n) + \lambda_n\bigl(\gamma\|S\|_1 + \operatorname{trace}(L)\bigr)\bigr\} \quad \text{such that } S \succeq L \succeq 0, \tag{1}$$

where $\ell(S-L; \widehat{\Sigma}^n)$ is the Gaussian log-likelihood as a function of the precision matrix $S-L$ and the empirical covariance matrix $\widehat{\Sigma}^n$ of the observed variables.

1.2. *Sharpness of rates*. On one hand, the paper provides attractive guarantees on the procedure (1)—namely, that under suitable incoherence conditions (to be discussed below) and a sample size $n \gtrsim p$, the method is guaranteed with high probability: (a) to correctly recover the signed support of the sparse matrix S^*, and hence the full graph structure; (b) to correctly recover the rank of the component L^*, and hence the number of latent variables; and (c) to yield operator norm consistency of the order $\sqrt{\frac{p}{n}}$. The proof itself involves a clever use of the primal-dual witness method [6], in which one analyzes an M-estimator by constructing a primal solution and an associated dual pair, and uses the construction to show that the optimum has desired properties (in this case, support and rank recovery) with high probability. A major challenge, not present in the simpler problem without latent variables, is dealing with the potential nonidentifiability of the matrix decomposition problem (see below for further discussion); the authors overcome this challenge via a delicate analysis of the tangent spaces associated with the sparse and low-rank components.

On the other hand, the scaling $n \gtrsim p$ is quite restrictive, at least in comparison to related results without latent variables. To provide a concrete example, consider a Gaussian graphical model with maximum degree d. For any such graph, again under a set of so-called incoherence or irrepresentability conditions, the neighborhood-based selection of approach of Meinshausen and Bühlmann [5] can be shown to correctly specify the graph structure with high probability based on $n \gtrsim d \log p$ samples. Moreover, under a similar set of assumptions, Ravikumar et al. [6] show that the ℓ_1-regularized Gaussian MLE returns an estimate of the precision matrix with operator norm error of the order $\sqrt{\frac{d^2 \log p}{n}}$. Consequently, whenever the maximum degree d is significantly smaller than the dimension, results of this type allow for the sample size n to be much smaller than p. This discrepancy—as to whether or not the sample size can be smaller than the dimension—thus raises some interesting directions for future work. More precisely, one wonders

whether or not the CPW analysis might be sharpened so as to reduce the sample size requirements. Possibly this might require introducing additional structure in the low-rank matrix. From the other direction, an alternative approach would be to develop minimax lower bounds on latent Gaussian model selection, for instance, by using information-theoretic techniques that have been exploited in related work on model/graph selection and covariance estimation (e.g., [2, 8, 10]).

1.3. *Relaxing assumptions*. The CPW analysis also imposes lower bounds on the minimum absolute values of the nonzero entries in S^*, as well as the minimum nonzero singular values of L^*—both must scale as $\Omega(\sqrt{\frac{p}{n}})$. Clearly, some sort of lower bound on these quantities is necessary in order to establish exact recovery guarantees, as in the results (a) and (b) paraphrased above. It is less clear whether lower bounds of this order are the weakest possible, and if not, to what extent they can be relaxed. For instance, again in the setting of Gaussian graph selection without latent variables [5, 6], the minimum values are typically allowed to be as small as $\Omega(\sqrt{\frac{\log p}{n}})$. More broadly, in many applications, it might be more natural to assume that the data is not actually drawn from a sparse graphical model, but rather can be well-approximated by such a model. In such settings, although exact recovery guarantees would no longer be feasible, one would like to guarantee that a given method, either the M-estimator (1) or some variant thereof, can recover all entries of S^* with absolute value above a given threshold, and/or estimate the number of eigenvalues of L^* above a (possibly different) threshold. Such guarantees are possible for ordinary Gaussian graph selection, where it is known that ℓ_1-based methods will recover all entries with absolute values above the regularization parameter [5, 6].

The CPW analysis also involves various types of incoherence conditions on the matrix decomposition. As noted by the authors, some of these assumptions are related to the incoherence or irrepresentability conditions imposed in past work on ordinary Gaussian graph selection [5, 6, 11]; others are unique to the latent problem, since they are required to ensure identifiability (see discussion below). It seems worthwhile to explore which of these incoherence conditions are artifacts of a particular methodology and which are intrinsic to the problem. For instance, in the case of ordinary Gaussian graph selection, there are problems for which the neighborhood-based Lasso [5] can correctly recover the graph while the ℓ_1-regularized log-determinant approach [4, 6] cannot. Moreover, there are problems for which, with the same order of sample size, the neighborhood-based Lasso will fail whereas an oracle method will succeed [10]. Such differences demonstrate that certain aspects of the incoherence conditions are artifacts of ℓ_1-relaxations. In the context of latent Gaussian graph selection, these same issues remain to be explored. For instance, are there alternative polynomial-time methods that can perform latent graph selection under milder incoherence conditions? What conditions are required by an oracle-type approach—that is, involving exact cardinality and rank constraints?

1.4. *Toward partial identifiability.* On the other hand, certain types of incoherence conditions are clearly intrinsic to the problem. Even at the population level, it is clearly not possible in general to identify the components (S^*, L^*) based on observing only the sum $K = S^* - L^*$. A major contribution of the CPW paper, building from their own pioneering work on matrix decompositions [3], is to provide sufficient conditions on the pair (S^*, L^*) that ensure identifiability. These sufficient conditions are based on a detailed analysis of the algebraic structure of the spaces of sparse and low-rank matrices, respectively.

In a statistical setting, however, most models are viewed as approximations to reality. With this mindset, it could be interesting to consider matrix decompositions that satisfy a weaker notion of partial identifiability. To provide a concrete illustration, suppose that we begin with a matrix pair (S^*, L^*) that is identifiable based on observing the difference $K = S^* - L^*$. Now imagine that we perturb K by a matrix that is both sparse and low-rank—for instance, a matrix of the form $E = zz^T$ where z is a sparse vector. If we then consider the perturbed matrix $\widetilde{K} := K + \delta E = S^* - L^* + \delta E$ for some suitably small parameter δ, the matrix decomposition is longer identifiable. In particular, at the two extremes, we can choose between the decompositions $\widetilde{K} = (S^* + \delta E) - L^*$, where the matrix $(S^* + \delta E)$ is sparse, or the decomposition $\widetilde{K} = S^* - (L^* - \delta E)$, where the matrix $L^* - \delta E$ is low-rank. Note that this nonidentifiability holds regardless of how small we choose the scalar δ. However, from a more practical perspective, if we relax our requirement of exact identification, then such a perturbation need not be a concern as long as δ is relatively small. Indeed, one might expect that it should be possible to recover estimates of the pair (S^*, L^*) that are accurate up to an error proportional to δ.

In some of our own recent work [1], we have provided such guarantees for a related class of noisy matrix decomposition problems. In particular, we consider the observation model[1]

$$Y = \mathfrak{X}(S^* - L^*) + W, \tag{2}$$

where $\mathfrak{X} : \mathbb{R}^{p \times p} \to \mathbb{R}^{n_1 \times n_2}$ is a known linear operator and $W \in \mathbb{R}^{n_1 \times n_2}$ is a noise matrix. In the simplest case, $\mathfrak{X}$ is simply the identity operator. Observation models of this form (2) arise in robust PCA, sparse factor analysis, multivariate regression and robust covariance estimation.

Instead of enforcing incoherence conditions sufficient for identifiability, the analysis is performed under related but milder conditions on the interaction between S^* and L^*. For instance, one way of controlling the radius of nonidentifiability is via control on the "spikiness" of the low-rank component, as measured by the ratio $\alpha(L^*) := \frac{p\|L^*\|_\infty}{|||L^*|||_F}$, where $\|\cdot\|_\infty$ denotes the elementwise absolute maximum and $|||\cdot|||_F$ denotes the Frobenius norm. For any nonzero p-dimensional matrix, this spikiness ratio ranges between 1 and p:

[1] Here we follow the notation of the CPW paper for the sparse and low-rank components.

- On one hand, it achieves its minimum value by a matrix that has all its entries equal to the same nonzero constant (e.g., $L^* = 11^T$, where $1 \in \mathbb{R}^p$ is a vector of all ones).
- On the other hand, the maximum is achieved by a matrix that concentrates all its mass in a single position (e.g., $L^* = e_1 e_1^T$, where $e_1 \in \mathbb{R}^p$ is the first canonical basis vector).

Note that it is precisely this latter type of matrix that is troublesome in sparse plus low-rank matrix decomposition, since it is simultaneously sparse *and* low-rank. In this way, the spikiness ratio limits the effect of such troublesome instances, thereby bounding the radius of nonidentifiability of the model. The paper [1] analyzes an M-estimator, also based on elementwise ℓ_1 and nuclear norm regularization, for estimating the pair (S^*, L^*) from the noisy observation model (2). The resulting error bounds involve both terms arising from the (possibly stochastic) noise matrix W and additional terms associated with the radius of nonidentifiability.

The same notion of partial identifiability is applicable to latent Gaussian graph selection. Accordingly, it seems worthwhile to explore whether similar techniques can be used to obtain error bounds with a similar form—one component associated with the stochastic noise (induced by sampling), and a second deterministic component. Interestingly, under the scaling $n \gtrsim p$ assumed in the CPW paper, the empirical covariance matrix $\widehat{\Sigma}^n$ will be invertible with high probability and, hence, it can be cast as an observation model of the form (2)—namely, we can write $(\widehat{\Sigma}^n)^{-1} = S^* - L^* + W$, where the noise matrix W is induced by sampling.

1.5. *Extensions to non-Gaussian variables*. A final more speculative yet intriguing question is whether the techniques of CPW can be extended to graphical models involving non-Gaussian variables, for instance, those with binary or multinomial variables for a start. The main complication here is that factorization and conditional independence properties for non-Gaussian variables do not translate directly into sparsity of the inverse covariance matrix. Nonetheless, it might be possible to reveal aspects of this factorization by some type of spectral analysis, in which context related matrix-theoretic approaches could be brought to bear. Overall, we should all be thankful to Chandrasekaran, Parillo and Willsky for their innovative work and the exciting line of questions and possibilities that it has raised for future research.

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.* **40** 1171–1197.

[2] CAI, T. and ZHOU, H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statistica Sinica*. To appear.

[3] CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479

[4] MEINSHAUSEN, N. (2008). A note on the Lasso for Gaussian graphical model selection. *Statist. Probab. Lett.* **78** 880–884. MR2398362

[5] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[6] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

[7] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[8] SANTHANAM, N. P. and WAINWRIGHT, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Inform. Theory* **58** 4117–4134.

[9] SPEED, T. P. and KIIVERI, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14** 138–150. MR0829559

[10] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** 5728–5741. MR2597190

[11] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA AT BERKELEY
421 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wainwrig@stat.berkeley.edu

The Annals of Statistics
2012, Vol. 40, No. 4, 1984–1988
DOI: 10.1214/12-AOS984
Main article DOI: 10.1214/11-AOS949

DISCUSSION: LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION

BY CHRISTOPHE GIRAUD AND ALEXANDRE TSYBAKOV

Ecole Polytechnique and CREST-ENSAE

Recently there has been an increasing interest in the problem of estimating a high-dimensional matrix K that can be decomposed in a sum of a sparse matrix S^* (i.e., a matrix having only a small number of nonzero entries) and a low rank matrix L^*. This is motivated by applications in computer vision, video segmentation, computational biology, semantic indexing, etc. The main contribution and novelty of the Chandrasekaran, Parrilo and Willsky paper (CPW in what follows) is to propose and study a method of inference about such decomposable matrices for a particular setting where K is the precision (concentration) matrix of a partially observed sparse Gaussian graphical model (GGM). In this case, K is the inverse of the covariance matrix of a Gaussian vector X_O extracted from a larger Gaussian vector (X_O, X_H) with sparse inverse covariance matrix. Then it is easy to see that K can be represented as a sum of a sparse precision matrix S^* corresponding to the observed variables X_O and a matrix L^* with rank at most h, where h is the dimension of the latent variables X_H. If h is small, which is a typical situation in practice, then L^* has low rank. The GGM with latent variables is of major interest for applications in biology or in social networks where one often does not observe all the variables relevant for depicting sparsely the conditional dependencies. Note that formally this is just one possible motivation and mathematically the problem is dealt with in more generality, namely, postulating that the precision matrix satisfies

$$K = S^* + L^* \tag{1}$$

with sparse S^* and low-rank L^*, both symmetric matrices. A small amendment to that inherited from the latent variables motivation is that L^* is assumed negative definite (in our notation, L^* corresponds to $-L^*$ in the paper). We believe that this is not crucial and all the results remain valid without this assumption.

CPW propose to estimate the pair (S^*, L^*) from a n-sample of X_O by the pair $(\widehat{S}, \widehat{L})$ obtained by minimizing the negative log-likelihood with mixed ℓ^1 and nuclear norm penalties; cf. (1.2) of the paper. The key issue in this context is identifiability. Under what conditions can we identify S^* and L^* separately? CPW provide geometric conditions of identifiability based on transversality of tangent spaces to the varieties of sparse and low-rank matrices. They show that, under these conditions, with probability close to 1, it is possible to recover the support of S^*, the rank

Received February 2012.

of L^* and to get a bound of order $O(\sqrt{p/n})$ on the estimation errors $|\widehat{S} - S^*|_{\ell^\infty}$ and $\|\widehat{L} - L^*\|_2$. Here, p is the dimension of X_O and $|\cdot|_{\ell^q}$ and $\|\cdot\|_2$ stand for the componentwise ℓ^q-norm and the spectral norm of a matrix, respectively.

Overall, CPW pioneer a hard and important problem of high-dimensional statistics and provide an original solution both in the theory and in numerically implementable realization. While being the first work to shed light on the problem, the paper does not completely raise the curtain and several aspects still remain to be understood and elucidated.

The nature of the results. The most important problem for current applications appears to be the estimation of S^* or the recovery of its support. Indeed, the main interest is in the conditional dependencies of the coordinates of X_O in the complete model (X_O, X_H) and this information is carried by the matrix S^*. In this context, L^* is essentially a nuisance, so that bounds on the estimation error of L^* and the recovery of the rank of L^* are of relatively moderate interest. However, mathematically, the most sacrifice comes from the desire to have precise estimates of L^*. Indeed, if $\widehat{\Sigma}_n$ and Σ denote the empirical and population covariance matrices, the slow rate $O(\sqrt{p/n})$ comes from the bound on $\|\widehat{\Sigma}_n - \Sigma\|_2$ in Lemma 5.4, that is, from the stochastic error corresponding to L^*. Since the sup-norm error $|\widehat{\Sigma}_n - \Sigma|_{\ell^\infty}$ is of order $\sqrt{(\log p)/n}$, can we get a better rate when solely focusing on $|\widehat{S} - S^*|_{\ell^\infty}$?

Extension to high dimensions. The results of the paper are valid and meaningful only when $p < n$. However, for the applications of GGM, the case $p \gg n$ is the most common. A key question is whether the restriction $p < n$ is intrinsic, that is, whether it is possible to have results on S^* in model (1) when $p \gg n$. Since the traditional model with sparse component S^* alone is still tractable when $p \gg n$, a related question is whether introducing the model (1) with two components and estimating both S^* and L^* gives any improvement in the $p \gg n$ setting as compared to estimation in the model with a sparse component alone. A small simulation study that we provide below suggests that already for $p = n$, including the low-rank component *in the estimator* may yield no improvement as compared to traditional sparse estimation without the low-rank component, although this low-rank component is effectively present *in the model*.

Optimal rates. The paper obtains bounds of order $O(\sqrt{p/n})$ on the estimation errors $|\widehat{S} - S^*|_{\ell^\infty}$ and $\|\widehat{L} - L^*\|_2$ with probability $1 - 2\exp(-p)$. Can we achieve a better rate than $\sqrt{p/n}$ when solely focusing on the recovery of S^* with the usual probability $1 - p^{-a}$ for some $a > 0$? Is the rate $\sqrt{p/n}$ optimal in a minimax sense on some class of matrices? Note that one should be careful in defining the class of matrices because in reality the rate is not $O(\sqrt{p/n})$ but rather $O(\psi\sqrt{p/n})$, where ψ is the spectral norm of Σ depending on p. It can be large for large p. Surprisingly, not much is known about the optimal rates even in the

simpler case of purely sparse precision matrices, without the low-rank component. In this case, [1, 7] and [8] provide some analysis of the upper bounds on the estimation error of different estimators and under different sets of assumptions on the precision matrix. All these bounds are of "order" $O(\sqrt{(\log p)/n})$, but again one should be very careful here because of the factors depending on p that multiply this rate. In [1], the factor is the squared $\ell^1 \to \ell^1$ norm of the precision matrix while in [7], it is the squared degree of the graphical model multiplied by some combinations of powers of matrix norms that are not easy to interpret. The most recent paper [8] obtains the rate $O(d\sqrt{(\log p)/n})$, where d is the degree of the graph for ℓ^∞-bounded precision matrices. An open problem is to find optimal rates of convergence on classes of precision matrices defined via sparsity and low rank characteristics. The same problem makes sense for covariance matrices. Here, some advances have been achieved very recently. In particular, some optimal rates of estimation of low-rank covariance matrices are provided by [5].

The assumptions of the paper are stated in terms of some inaccessible characteristics such as $\xi(T)$ and $\mu(\Omega)$ and seem to be very strong. They are in the spirit of the irrepresentability condition for the vector case used to prove model selection consistency of the Lasso. For a given set of data, there is no means to check whether these assumptions are satisfied. What happens when they do not hold? Can we still have some convergence properties under no assumption at all or under weaker assumptions akin to the restricted eigenvalue condition in the vector case?

Choice of the tuning parameters. The choice of parameters (γ, λ_n) ensuring algebraic consistency in Theorem 4.1 depends on various unknown quantities. Proposing a reasonable data-driven selector for (γ, λ_n) (e.g., similarly to [4] for the pure sparse setting) would be very helpful for the practice.

Alternative methods of estimation. Constructively, the method of CPW is obtained from the GLasso of [2] by adding a penalization by the nuclear norm of the low-rank component. Similar low-rank extensions can be readily derived from other methods, such as the Dantzig type approach of [1] and the regression approach of [3, 6]. Consider a Gaussian random vector $X \in \mathbb{R}^p$ with mean 0 and nonsingular covariance matrix Σ. Let $K = \Sigma^{-1}$ be the precision matrix. We assume that K is of the form (1) where S^* is sparse and L^* has low rank.

(a) *Dantzig type approach.* In the spirit of [1], we may define our estimator as a solution of the following convex program:

$$(\widehat{S}, \widehat{L}) = \operatorname*{argmin}_{(S,L)\in\mathcal{G}} \{|S|_{\ell^1} + \mu\|L\|_*\}, \tag{2}$$

where $\|\cdot\|_*$ is the nuclear norm, $\mathcal{G} = \{(S, L) : |\widehat{\Sigma}_n(S+L) - I|_{\ell^\infty} \le \lambda\}$ and $\mu, \lambda > 0$ are tuning constants. Here, the nuclear norm $\|L\|_*$ is a convex relaxation of the rank of L^*.

(b) *Regression approach.* The regression approach [3, 6] is an alternative point of view for estimating the structure of a GGM. In the pure sparse setting, some numerical experiments [9] suggest that it may be more reliable than the ℓ^1-penalized log-likelihood approach. Let $\operatorname{diag}(A)$ denote the diagonal of square matrix A and $\|A\|_F$ its Frobenius norm. Defining

$$\Theta = \operatorname*{argmin}_{A:\operatorname{diag}(A)=0} \|\Sigma^{1/2}(I-A)\|_F^2,$$

we have $\Theta = K\Delta + I$, where I is the identity matrix and Δ is the diagonal matrix with diagonal elements $\Delta_{jj} = -1/K_{jj}$ for $j = 1, \ldots, p$. Thus, we have the decomposition

$$\Theta = \bar{S} + \bar{L}, \qquad \text{where } \bar{S} = S^*\Delta + I \text{ and } \bar{L} = L^*\Delta.$$

Note that $\operatorname{rank}(\bar{L}) = \operatorname{rank}(L^*)$ and the nondiagonal elements $\bar{S}_{ij}$ of matrix $\bar{S}$ are nonzero only if S^*_{ij} is nonzero. Therefore, recovering the support of S^* and $\operatorname{rank}(L^*)$ is equivalent to recovering the support of $\bar{S}$ and $\operatorname{rank}(\bar{L})$.

Now, we estimate $(\bar{S}, \bar{L})$ from an n-sample of X represented as an $n \times p$ matrix $\mathbf{X}$. Noticing that the sample analog of $\|\Sigma^{1/2}(I-A)\|_F^2$ is $\|\mathbf{X}(I-A)\|_F^2/n$ and using the decomposition $\Theta = \bar{S} + \bar{L}$, we arrive at the following estimator:

$$(\widehat{S}, \widehat{L}) = \operatorname*{argmin}_{(S,L):\operatorname{diag}(S+L)=0} \left\{ \frac{1}{2}\|\mathbf{X}(I-S-L)\|_F^2 + \lambda|S|_{\ell^1,\text{off}} + \mu\|\mathbf{X}L\|_* \right\}, \tag{3}$$

where μ, λ are positive tuning constants and $|S|_{\ell^1,\text{off}} = \sum_{i\neq j}|S_{ij}|$. Note that here the low-rank shrinkage is driven by the nuclear norm $\|\mathbf{X}L\|_*$ rather than by $\|L\|_*$. The convex minimization in (3) can be performed efficiently by alternating block descents on the off-diagonal elements of S, the matrix L and the diagonal of S. The off-diagonal support of S^* is finally estimated by the off-diagonal support of $\widehat{S}$.

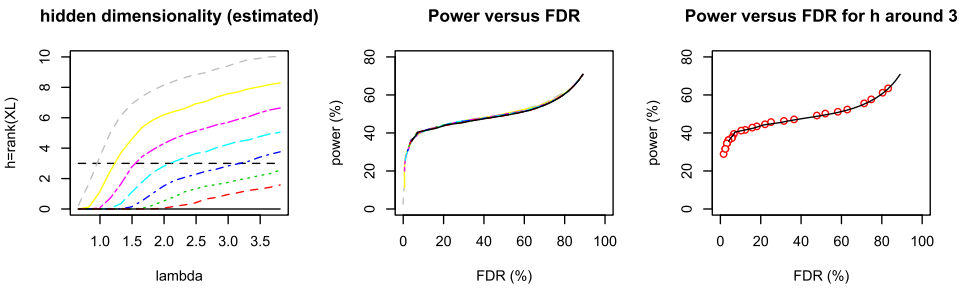


FIG. 1. *Each color corresponds to a fixed value of μ, the solid-black color being for $\mu = +\infty$. For each choice of μ, different quantities are plotted for a series of values of λ. Left*: *Mean rank of $\mathbf{X}\widehat{L}$. Middle*: *The curve of estimated power versus estimated FDR. Right*: *The power versus FDR for the estimators fulfilling $\mathbb{E}[\operatorname{rank}(\mathbf{X}\widehat{L})] \approx h = 3$ (red dots), superposed with the Power versus the FDR for $\mu = +\infty$ (in solid-black).*

Numerical experiment. A sparse Gaussian graphical model in $\mathbb{R}^{30}$ is generated randomly according to the procedure described in Section 4 of [4]. A sample of size $n = 30$ is drawn from this distribution and $\mathbf{X}$ is obtained by hiding the values of 3 variables. These 3 hidden variables are chosen randomly among the connected variables. The estimators $(\widehat{S}, \widehat{L})$ defined in (3) are then computed for a grid of values of λ and μ. The results are summarized in Figure 1 (average over 100 simulations).

Strikingly, there is no significative difference in these examples between the procedure of [6] (corresponding to $\mu = +\infty$, in solid-black) and the procedure (3) that includes the low-rank component (corresponding to finite μ).

REFERENCES

[1] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[2] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.

[3] GIRAUD, C. (2008). Estimation of Gaussian graphs by model selection. *Electron. J. Stat.* **2** 542–563. MR2417393

[4] GIRAUD, C., HUET, S. and VERZELEN, N. (2012). Graph selection with GGMselect. *Stat. Appl. Genet. Mol. Biol.* **11** 1–50.

[5] LOUNICI, K. (2012). High-dimensional covariance matrix estimation with missing observations. Available at arXiv:1201.2577.

[6] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[7] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

[8] SUN, T. and ZHANG, C. H. (2012). Sparse matrix inversion with scaled lasso. Available at arXiv:1202.2723.

[9] VILLERS, F., SCHAEFFER, B., BERTIN, C. and HUET, S. (2008). Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems. *Stat. Appl. Genet. Mol. Biol.* **7** Art. 14, 36. MR2465331

CMAP, UMR CNRS 7641
ECOLE POLYTECHNIQUE
ROUTE DE SACLAY
F-91128 PALAISEAU CEDEX
FRANCE
E-MAIL: Christophe.Giraud@polytechnique.edu

LABORATOIRE DE STATISTIQUE
CREST-ENSAE
3, AV. PIERRE LAROUSSE
F-92240 MALAKOFF CEDEX
FRANCE
E-MAIL: Alexandre.Tsybakov@ensae.fr

The Annals of Statistics
2012, Vol. 40, No. 4, 1989–1996
DOI: 10.1214/12-AOS985
Main article DOI: 10.1214/11-AOS949

DISCUSSION: LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION[1]

BY ZHAO REN AND HARRISON H. ZHOU

Yale University

1. Introduction. We would like to congratulate the authors for their refreshing contribution to this high-dimensional latent variables graphical model selection problem. The problem of covariance and concentration matrices is fundamentally important in several classical statistical methodologies and many applications. Recently, sparse concentration matrices estimation has received considerable attention, partly due to its connection to sparse structure learning for Gaussian graphical models. See, for example, Meinshausen and Bühlmann (2006) and Ravikumar et al. (2011). Cai, Liu and Zhou (2012) considered rate-optimal estimation.

The authors extended the current scope to include latent variables. They assume that the fully observed Gaussian graphical model has a naturally sparse dependence graph. However, there are only partial observations available for which the graph is usually no longer sparse. Let X be $(p+r)$-variate Gaussian with a sparse concentration matrix $S^*_{(O,H)}$. We only observe X_O, p out of the whole $p+r$ variables, and denote its covariance matrix by Σ^*_O. In this case, usually the $p \times p$ concentration matrix $(\Sigma^*_O)^{-1}$ are not sparse. Let S^* be the concentration matrix of observed variables conditioned on latent variables, which is a submatrix of $S^*_{(O,H)}$ and hence has a sparse structure, and let L^* be the summary of the marginalization over the latent variables and its rank corresponds to the number of latent variables r for which we usually assume it is small. The authors observed $(\Sigma^*_O)^{-1}$ can be decomposed as the difference of the sparse matrix S^* and the rank r matrix L^*, that is, $(\Sigma^*_O)^{-1} = S^* - L^*$. Then following traditional wisdoms, the authors naturally proposed a *regularized maximum likelihood approach* to estimate both the sparse structure S^* and the low-rank part L^*,

$$\min_{(S,L):S-L\succ 0, L\succeq 0} \operatorname{tr}\big((S-L)\Sigma^n_O\big) - \log\det(S-L) + \chi_n\big(\gamma\|S\|_1 + \operatorname{tr}(L)\big),$$

where Σ^n_O is the sample covariance matrix, $\|S\|_1 = \sum_{i,j}|s_{ij}|$, and γ and χ_n are regularization tuning parameters. Here $\operatorname{tr}(L)$ is the trace of L. The notation $A \succ 0$ means A is positive definite, and $A \succeq 0$ denotes that A is nonnegative.

There is an obvious identifiability problem if we want to estimate both the sparse and low-rank components. A matrix can be both sparse and low rank. By exploring the geometric properties of the tangent spaces for sparse and low-rank components, the authors gave a beautiful sufficient condition for identifiability, and then

Received February 2012.
[1]Supported in part by NSF Career Award DMS-06-45676 and NSF FRG Grant DMS-08-54975.

provided very much involved theoretical justifications based on the sufficient condition, which is beyond our ability to digest them in a short period of time in the sense that we don't fully understand why those technical assumptions were needed in the analysis of their approach. Thus, we decided to look at a relatively simple but potentially practical model, with the hope to still capture the essence of the problem, and see how well their regularized procedure works. Let $\|\cdot\|_{1\to 1}$ denote the matrix l_1 norm, that is, $\|S\|_{1\to 1} = \max_{1\le i\le p}\sum_{j=1}^{p}|s_{ij}|$. We assume that S^* is in the following uniformity class:

$$\mathcal{U}(s_0(p), M_p) = \Bigg\{ S = (s_{ij}) : S \succ 0,\ \|S\|_{1\to 1} \le M_p, \max_{1\le i\le p}\sum_{j=1}^{p}\mathbf{1}\{s_{ij}\neq 0\} \le s_0(p)\Bigg\}, \tag{1}$$

where we allow $s_0(p)$ and M_p to grow as p and n increase. This uniformity class was considered in Ravikumar et al. (2011) and Cai, Liu and Luo (2011). For the low-rank matrix L^*, we assume that the effect of marginalization over the latent variables spreads out, that is, the low-rank matrix L^* has row/column spaces that are not closely aligned with the coordinate axes to resolve the identifiability problem. Let the eigen-decomposition of L^* be as follows:

$$L^* = \sum_{i=1}^{r_0(p)} \lambda_i u_i u_i^T, \tag{2}$$

where $r_0(p)$ is the rank of L^*. We assume that there exists a universal constant c_0 such that $\|u_i\|_\infty \le \sqrt{\frac{c_0}{p}}$ for all i, and $\|L^*\|_{1\to 1}$ is bounded by M_p which can be shown to be bounded by $c_0 r_0$. A similar incoherence assumption on u_i was used in Candès and Recht (2009). We further assume that

$$\lambda_{\max}(\Sigma_O^*) \le M \quad \text{and} \quad \lambda_{\min}(\Sigma_O^*) \ge 1/M \tag{3}$$

for some universal constant M.

As discussed in the paper, the goals in latent variable model selection are to obtain the sign consistency for the sparse matrix S^* as well as the rank consistency for the low-rank semi-positive definite matrix L^*. Denote the minimum magnitude of nonzero entries of S^* by θ, that is, $\theta = \min_{i,j}|s_{ij}|\mathbf{1}\{s_{ij}\neq 0\}$, and the minimum nonzero eigenvalue of L^* by σ, that is, $\sigma = \min_{1\le i\le r_0}\lambda_i$. To obtain theoretical guarantees of consistency results for the model described in (1), (2) and (3), in addition to the strong irrepresentability condition which seems to be difficult to check in practice, the authors require the following assumptions (by a translation of the conditions in the paper to this model) for θ, σ and n:

(1) $\theta \gtrsim \sqrt{p/n}$, which is needed even when $s_0(p)$ is constant;

(2) $\sigma \gtrsim s_0^3(p)\sqrt{p/n}$ under the additional strong assumptions on the Fisher information matrix $\Sigma_O^* \otimes \Sigma_O^*$ (see the footnote for Corollary 4.2);

(3) $n \gtrsim s_0^4(p)p$.

However, for sparse graphical model selection without latent variables, either the l_1-regularized maximum likelihood approach [see Ravikumar et al. (2011)] or CLIME [see Cai, Liu and Luo (2011)] can be shown to be sign consistent if the minimum magnitude nonzero entry of concentration matrix θ is at the order of $\sqrt{(\log p)/n}$ when M_p is bounded, which inspires us to study rate-optimalites for this latent variables graphical model selection problem. In this discussion, we propose a procedure to obtain an algebraically consistent estimate of the latent variable Gaussian graphical model under a much weaker condition on both θ and σ. For example, for a wide range of $s_0(p)$, we only require θ is at the order of $\sqrt{(\log p)/n}$ and σ is at the order of $\sqrt{p/n}$ to consistently estimate the support of S^* and the rank of L^*. That means the *regularized maximum likelihood approach* could be far from being optimal, but we don't know yet whether the suboptimality is due to the procedure or their theoretical analysis.

2. Latent variable model selection consistency. In this section we propose a procedure to obtain an algebraically consistent estimate of the latent variable Gaussian graphical model. The condition on θ to recover the support of S^* is reduced to that in Cai, Liu and Luo (2011) which studied sparse graphical model selection without latent variables, and the condition on σ is just at an order of $\sqrt{p/n}$, which is smaller than $s_0^3(p)\sqrt{p/n}$ assumed in the paper when $s_0(p) \to \infty$. When M_p is bounded, our results can be shown to be rate-optimal by lower bounds stated in Remarks 2 and 4 for which we are not giving proofs due to the limitation of the space.

2.1. *Sign consistency procedure of S^*.* We propose a CLIME-like estimator of S^* by solving the following linear optimization problem:

$$\min \|S\|_1 \qquad \text{subject to} \qquad \|\Sigma_O^n S - I\|_\infty \le \tau_n, \qquad S \in \mathbb{R}^{p\times p},$$

where $\Sigma_O^n = (\widetilde{\sigma}_{ij})$ is the sample covariance matrix. The tuning parameter τ_n is chosen as $\tau_n = C_1 M_p \sqrt{\frac{\log p}{n}}$ for some large constant C_1. Let $\hat{S}_1 = (\hat{s}_{ij}^1)$ be the solution. The CLIME-like estimator $\hat{S} = (\hat{s}_{ij})$ is obtained by symmetrizing $\hat{S}_1$ as follows:

$$\hat{s}_{ij} = \hat{s}_{ji} = \hat{s}_{ij}^1 \mathbf{1}\{|\hat{s}_{ij}^1| \le \hat{s}_{ji}^1\} + \hat{s}_{ji}^1 \mathbf{1}\{|\hat{s}_{ij}^1| > \hat{s}_{ji}^1\}.$$

In other words, we take the one with smaller magnitude between $\hat{s}_{ij}^1$ and $\hat{s}_{ji}^1$. We define a thresholding estimator $\tilde{S} = (\tilde{s}_{ij})$ with

$$\tilde{s}_{ij} = \tilde{s}_{ij} \mathbf{1}\{|\tilde{s}_{ij}| > 9M_p\tau_n\} \tag{4}$$

to estimate the support of S^*.

THEOREM 1. *Suppose that* $S^* \in \mathcal{U}(s_0(p), M_p)$,

$$\sqrt{(\log p)/n} = o(1) \quad \text{and} \quad \|L^*\|_\infty \le M_p \tau_n. \tag{5}$$

With probability greater than $1 - C_s p^{-6}$ *for some constant* C_s *depending on* M *only, we have*

$$\|\hat{S} - S^*\|_\infty \le 9 M_p \tau_n.$$

Hence, if the minimum magnitude of nonzero entries $\theta > 18 M_p \tau_n$, *we obtain the sign consistency* $\operatorname{sign}(\tilde{S}) = \operatorname{sign}(S^*)$. *In particular, if* M_p *is in the constant level, then to consistently recover the support of* S^*, *we only need that* $\theta \asymp \sqrt{(\log p)/n}$.

PROOF. The proof is similar to Theorem 7 in Cai, Liu and Luo (2011). The sub-Gaussian condition with spectral norm upper bound M implies that each empirical covariance $\widetilde{\sigma}_{ij}$ satisfies the following large deviation result:

$$\mathbb{P}(|\widetilde{\sigma}_{ij} - \sigma_{ij}| > t) \le C_s \exp\left(-\frac{8}{C_2^2} n t^2\right) \qquad \text{for } |t| \le \phi,$$

where C_s, C_2 and ϕ only depend on M. See, for example, Bickel and Levina (2008). In particular, for $t = C_2\sqrt{(\log p)/n}$ which is less than ϕ by our assumption, we have

$$\mathbb{P}(\|\Sigma_O^* - \Sigma_O^n\|_\infty > t) \le \sum_{i,j} \mathbb{P}(|\widetilde{\sigma}_{ij} - \sigma_{ij}| > t) \le p^2 \cdot C_s p^{-8}. \tag{6}$$

Let

$$A = \{\|\Sigma_O^* - \Sigma_O^n\|_\infty \le C_2\sqrt{(\log p)/n}\}.$$

Equation (6) implies $\mathbb{P}(A) \ge 1 - C_s p^{-6}$. On event A, we will show

$$\|(S^* - L^*) - \hat{S}_1\|_\infty \le 8 M_p \tau_n, \tag{7}$$

which immediately yields

$$\|S^* - \hat{S}\|_\infty \le \|(S^* - L^*) - \hat{S}_1\|_\infty + \|L^*\|_\infty \le 8 M_p \tau_n + M_p \tau_n = 9 M_p \tau_n.$$

Now we establish equation (7). On event A, for some large constant $C_1 \ge 2C_2$, the choice of τ_n yields

$$2 M_p \|\Sigma_O^* - \Sigma_O^n\|_\infty \le \tau_n. \tag{8}$$

By the matrix l_1 norm assumption, we could obtain that

$$\|(\Sigma_O^*)^{-1}\|_{1\to 1} \le \|S^*\|_{1\to 1} + \|L^*\|_{1\to 1} \le 2 M_p. \tag{9}$$

From (8) and (9) we have

$$\begin{aligned} \|\Sigma_O^n (S^* - L^*) - I\|_\infty &= \|(\Sigma_O^n - \Sigma_O^*)(\Sigma_O^*)^{-1}\|_\infty \\ &\le \|\Sigma_O^n - \Sigma_O^*\|_\infty \|(\Sigma_O^*)^{-1}\|_{1\to 1} \le \tau_n, \end{aligned}$$

which implies

$$\begin{aligned}(10)\qquad &\|\Sigma_O^n(S^*-L^*)-\Sigma_O^n\hat{S}_1\|_\infty\\ &\quad\le\|\Sigma_O^n(S^*-L^*)-I\|_\infty+\|\Sigma_O^n\hat{S}_1-I\|_\infty\le 2\tau_n.\end{aligned}$$

From the definition of $\hat{S}_1$ we obtain that

$$(11)\qquad \|\hat{S}_1\|_{1\to1}\le\|S^*-L^*\|_{1\to1}\le 2M_p,$$

which, together with equations (8) and (10), implies

$$\begin{aligned}&\big\|\Sigma_O^*\big((S^*-L^*)-\hat{S}_1\big)\big\|_\infty\\ &\quad\le\|\Sigma_O^n(S^*-L^*)-\hat{S}_1\|_\infty+\big\|(\Sigma_O^*-\Sigma_O^n)\big((S^*-L^*)-\hat{S}_1\big)\big\|_\infty\\ &\quad\le 2\tau_n+\|\Sigma_O^n-\Sigma_O^*\|_\infty\|(S^*-L^*)-\hat{S}_1\|_{1\to1}\\ &\quad\le 2\tau_n+4M_p\|\Sigma_O^n-\Sigma_O^*\|_\infty\le 4\tau_n.\end{aligned}$$

Thus, we have

$$\|(S^*-L^*)-\hat{S}_1\|_\infty\le\|(\Sigma_O^*)^{-1}\|_{1\to1}\big\|\Sigma_O^*\big((S^*-L^*)-\hat{S}_1\big)\big\|_\infty\le 8M_p\tau_n.\quad\square$$

REMARK 1. By the choice of our τ_n and the eigen-decomposition of L^*, the condition $\|L^*\|_\infty\le M_p\tau_n$ holds when $r_0(p)C_0/p\le C_1M_p^2\sqrt{(\log p)/n}$, that is, $p^2\log p\gtrsim nr_0^2(p)M_p^{-4}$. If M_p is slowly increasing (e.g., $p^{1/4-\tau}$ for any small $\tau>0$), the minimum requirement $\theta\asymp M_p^2\sqrt{(\log p)/n}$ is weaker than $\theta\gtrsim\sqrt{p/n}$ required in Corollary 4.2. Furthermore, it can be shown that the optimal rate of minimum magnitude of nonzero entries for sign consistency is $\theta\asymp M_p\sqrt{(\log p)/n}$ as in Cai, Liu and Zhou (2012).

REMARK 2. Cai, Liu and Zhou (2012) showed the minimum requirement for θ, $\theta\asymp M_p\sqrt{(\log p)/n}$ is necessary for sign consistency for sparse concentration matrices. Let $\mathcal{U}_S(c)$ denote the class of concentration matrices defined in (1) and (2), satisfying assumption (5) and $\theta>cM_p\sqrt{(\log p)/n}$. We can show that there exists some constant $c_1>0$ such that for all $0<c<c_1$,

$$\lim_{n\to\infty}\inf_{(\hat{S},\hat{L})}\sup_{\mathcal{U}_S(c)}\mathbb{P}\big(\operatorname{sign}(\hat{S})\ne\operatorname{sign}(S^*)\big)>0,$$

similar to Cai, Liu and Zhou (2012).

2.2. *Rank Consistency Procedure of L^*.* In this section we propose a procedure to estimate L^* and its rank. We note that with high probability Σ_O^n is invertible, then define $\hat{L}=(\Sigma_O^n)^{-1}-\tilde{S}$, where $\tilde{S}$ is defined in (4). Denote the eigen-decomposition of $\hat{L}$ by $\sum_{i=1}^p\lambda_i(\hat{L})v_iv_i^T$, and let $\lambda_i(\tilde{L})=\lambda_i(\hat{L})1\{\lambda_i(\hat{L})>C_3\sqrt{\frac{p}{n}}\}$, where constant C_3 will be specified later. Define $\tilde{L}=\sum_{i=1}^p\lambda_i(\tilde{L})v_iv_i^T$. The following theorem shows that estimator $\tilde{L}$ is a consistent estimator of L^* under the spectral norm and with high probability $\operatorname{rank}(L^*)=\operatorname{rank}(\tilde{L})$.

THEOREM 2. *Under the conditions in Theorem 1, we assume that*

$$\sqrt{\frac{p}{n}} \le \frac{1}{16\sqrt{2}M^2} \quad \text{and} \quad M_p^2 s_0(p) \le \sqrt{\frac{p}{\log p}}. \tag{12}$$

Then there exists some constant C_3 such that

$$\|\hat{L} - L^*\| \le C_3\sqrt{\frac{p}{n}}$$

with probability greater than $1 - 2e^{-p} - C_s p^{-6}$. Hence, if $\sigma > 2C_3\sqrt{\frac{p}{n}}$, we have $\operatorname{rank}(L^*) = \operatorname{rank}(\tilde{L})$ *with high probability.*

PROOF. From Corollary 5.5 of the paper and our assumption on the sample size, we have

$$\mathbb{P}\left(\|\Sigma_O^* - \Sigma_O^n\| \ge \sqrt{128}M\sqrt{\frac{p}{n}}\right) \le 2\exp(-p).$$

Note that $\lambda_{\min}(\Sigma_O^*) \ge 1/M$, and $\sqrt{128}M\sqrt{\frac{p}{n}} \le 1/(2M)$ under the assumption (12), then $\lambda_{\min}(\Sigma_O^n) \ge 1/(2M)$ with high probability, which yields the same rate of convergence for the concentration matrix, since

$$\begin{aligned}\|(\Sigma_O^*)^{-1} - (\Sigma_O^n)^{-1}\| &\le \|(\Sigma_O^*)^{-1}\|\|(\Sigma_O^n)^{-1}\|\|\Sigma_O^* - \Sigma_O^n\| \\ &\le 2M^2\sqrt{128}M\sqrt{\frac{p}{n}} = 16\sqrt{2}M^3\sqrt{\frac{p}{n}}.\end{aligned} \tag{13}$$

From Theorem 1 we know

$$\operatorname{sign}(\tilde{S}) = \operatorname{sign}(S^*) \quad \text{and} \quad \|\tilde{S} - S^*\|_\infty \le 9M_p\tau_n$$

with probability greater than $1 - C_s p^{-6}$. Since $\|B\| \le \|B\|_{1\to 1}$ for any symmetric matrix B, we then have

$$\|\tilde{S} - S^*\| \le \|\tilde{S} - S^*\|_{1\to 1} \le s_0(p)9M_p\tau_n = 9C_1M_p^2 s_0(p)\sqrt{\frac{\log p}{n}}. \tag{14}$$

Equations (13) and (14), together with the assumption $M_p^2 s_0(p) \le \sqrt{\frac{p}{\log p}}$, imply

$$\begin{aligned}\|\hat{L} - L^*\| &\le \|(\Sigma_O^*)^{-1} - (\Sigma_O^n)^{-1}\| + \|\tilde{S} - S^*\| \\ &\le 16\sqrt{2}M^3\sqrt{\frac{p}{n}} + 9C_1M_p^2 s_0(p)\sqrt{\frac{\log p}{n}} \le C_3\sqrt{\frac{p}{n}}\end{aligned}$$

with probability greater than $1 - 2e^{-p} - C_s p^{-6}$. □

REMARK 3. We should emphasize the fact that in order to consistently estimate the rank of L^* we need only that $\sigma > 2C_3\sqrt{\frac{p}{n}}$, which is smaller than $s_0^3(p)\sqrt{\frac{p}{n}}$ required in the paper (see the footnote for Corollary 4.2), as long as $M_p^2 s_0(p) \leq \sqrt{\frac{p}{\log p}}$. In particular, we don't explicitly constrain the rank $r_0(p)$. One special case is that M_p is constant and $s_0(p) \asymp p^{1/2-\tau}$ for some small $\tau > 0$, for which our requirement is $\sqrt{\frac{p}{n}}$ but the assumption in the paper is at an order of $p^{3(1/2-\tau)}\sqrt{\frac{p}{n}}$.

REMARK 4. Let $\mathcal{U}_L(c)$ denote the class of concentration matrices defined in (1), (2) and (3), satisfying assumptions (12), (5) and $\sigma > c\sqrt{\frac{p}{n}}$. We can show that there exists some constant $c_2 > 0$ such that for all $0 < c < c_2$,

$$\lim_{n\to\infty} \inf_{(\hat{S},\hat{L})} \sup_{\mathcal{U}_L(c)} \mathbb{P}\big(\operatorname{rank}(\hat{L}) \neq \operatorname{rank}(L^*)\big) > 0.$$

The proof of this lower bound is based on a modification of a lower bound argument in a personal communication of T. Tony Cai (2011).

3. Concluding remarks and further questions. In this discussion we attempt to understand optimalities of results in the present paper by studying a relatively simple model. Our preliminary analysis seems to indicate that their results in this paper are suboptimal. In particular, we tend to conclude that assumptions on θ and σ in the paper can be potentially very much weakened. However, it is not clear to us whether the suboptimality is due to the methodology or just its theoretical analysis. We want to emphasize that the preliminary results in this discussion can be strengthened, but for the purpose of simplicity of the discussion we choose to present weaker but simpler results to hopefully shed some light on understanding optimalities in estimation.

REFERENCES

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

CAI, T. T. (2011). Personal communication.

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CAI, T. T., LIU, W. and ZHOU, H. H. (2012). Optimal estimation of large sparse precision matrices. Unpublished manuscript.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: zhao.ren@yale.edu
huibin.zhou@yale.edu

The Annals of Statistics
2012, Vol. 40, No. 4, 1997–2004
DOI: 10.1214/12-AOS1001
Main article DOI: 10.1214/11-AOS949

DISCUSSION: LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION

BY EMMANUEL J. CANDÉS AND MAHDI SOLTANOLKOTABI

Stanford University

We wish to congratulate the authors for their innovative contribution, which is bound to inspire much further research. We find latent variable model selection to be a fantastic application of matrix decomposition methods, namely, the superposition of low-rank and sparse elements. Clearly, the methodology introduced in this paper is of potential interest across many disciplines. In the following, we will first discuss this paper in more detail and then reflect on the versatility of the low-rank + sparse decomposition.

Latent variable model selection. The proposed scheme is an extension of the *graphical lasso* of Yuan and Lin [15] (see also [1, 6]), which is a popular approach for learning the structure in an undirected Gaussian graphical model. In this setup, we assume we have independent samples $X \sim \mathcal{N}(0, \Sigma)$ with a covariance matrix Σ exhibiting a sparse dependence structure but otherwise unknown; that is to say, most pairs of variables are conditionally independent given all the others. Formally, the concentration matrix Σ^{-1} is assumed to be sparse. A natural fitting procedure is then to regularize the likelihood by adding a term proportional to the ℓ_1 norm of the estimated inverse covariance matrix S:

$$\text{minimize } -\ell(S, \Sigma_0^n) + \lambda \|S\|_1 \tag{1}$$

under the constraint $\mathbf{S} \succeq 0$, where Σ_0^n is the empirical covariance matrix and $\|S\|_1 = \sum_{ij} |S_{ij}|$. (Variants are possible depending upon whether or not one would want to penalize the diagonal elements.) This problem is convex.

When some variables are unobserved—the observed and hidden variables are still jointly Gaussian—the model above may not be appropriate because the hidden variables can have a confounding effect. An example is this: we observe stock prices of companies and would like to infer conditional (in)dependence. Suppose, however, that all these companies rely on a commodity, a source of energy, for instance, which is not observed. Then the stock prices might appear dependent even though they may not be once we condition on the price of this commodity. In fact, the marginal inverse covariance of the observed variables decomposes into two terms. The first is the concentration matrix of the observed variables in the full model conditioned on the latent variables. The second term is the effect of

Received April 2012.

marginalization over the hidden variables. Assuming a sparse graphical model, the first term is sparse, whereas the second term may have low rank; in particular, the rank is at most the number of hidden variables. The authors then penalize the negative log-likelihood with a term proportional to

$$\gamma \|S\|_1 + \operatorname{trace}(L) \tag{2}$$

since the trace functional is the usual convex surrogate for the rank over the cone of positive semidefinite matrices. The constraints are $S \succ L \succeq 0$.

Adaptivity. The penalty (2) is simple and flexible since it does not really make special parametric assumptions. To be truly appealing, it would also need to be adaptive in the following sense: suppose there is no hidden variable, then does the low-rank + sparse model (L + S) behave as well or nearly as well as the graphical lasso? When there are few hidden variables, does it behave nearly as well? Are there such theoretical guarantees? If this is the case, it would say that using the L + S model would protect against the danger of not having accounted for all possible covariates. At the same time, if there were no hidden variable, one would not suffer any loss of performance. Thus, we would get the best of both worlds.

At first sight, the analysis presented in this paper does not allow us to reach this conclusion. If X is p-dimensional, the number of samples needed to show that one can obtain accurate estimates scales like $\Omega(p/\xi^4)$, where ξ is a modulus of continuity introduced in the paper that is typically much smaller than 1. We can think of $1/\xi$ as being related to the maximum degree d of the graph so that the condition may be interpreted as having a number of observations very roughly scaling like $d^4 p$. In addition, accurate estimation holds with the proviso that the signal is strong enough; here, both the minimum nonzero singular value of the low-rank component and the minimum nonzero entry of the sparse component scale like $\Omega(\sqrt{p/n})$. On the other hand, when there are no hidden variables, a line of work [11, 13, 14] has established that we could estimate the concentration matrix with essentially the same accuracy if $n = \Omega(d^2 \log p)$ and the magnitude of the minimum nonvanishing value of the concentration matrix scales like $\Omega(\sqrt{n^{-1} \log p})$. As before, d is the maximum degree of the graphical model. In the high-dimensional regime, the results offered by this literature seem considerably better. It would be interesting to know whether this could be bridged, and if so, under what types of conditions—if any.

Interestingly, such adaptivity properties have been established for related problems. For instance, the L + S model has been used to suggest the possibility of a principled approach to robust principal component analysis [2]. Suppose we have incomplete and corrupted information about an $n_1 \times n_2$ low-rank matrix L^0. More precisely, we observe $M_{ij} = L^0_{ij} + S^0_{ij}$, where $(i, j) \in \Omega_{\text{obs}} \subset \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$. We think of S^0 as a corruption pattern so that some

entries are totally unreliable but we do not know which ones. Then [2] shows that under rather broad conditions, the solution to

$$\begin{aligned} &\text{minimize } \|L\|_* + \lambda\|S\|_1 \\ &\text{subject to } M_{ij} = L_{ij} + S_{ij},\ (i,j) \in \Omega_{\text{obs}}, \end{aligned} \tag{3}$$

where $\|L\|_*$ is the nuclear norm, recovers L^0 exactly. Now suppose there are no corruptions. Then we are facing a matrix completion problem and, instead, one would want to minimize the nuclear norm of L under data constraints. In other words, there is no need for S in (3). The point is that there is a fairly precise understanding of the minimal number of samples needed for this strategy to work; for incoherent matrices [3], $|\Omega_{\text{obs}}|$ must scale like $(n_1 \vee n_2) r \log^2 n$, where r is the rank of L^0. Now some recent work [10] establishes the adaptivity in question. In details, (3) recovers L^0 from a minimal number of samples, in the sense defined above, even though a positive fraction may be corrupted. That is, the number of reliable samples one needs, regardless of whether corruption occurs, is essentially the same. Results of this kind extend to other settings as well. For instance, in sparse regression or compressive sensing we seek a sparse solution to $y = Xb$ by minimizing the ℓ_1 norm of b. Again, we may be worried that some equations are unreliable because of gross errors and would solve, instead,

$$\begin{aligned} &\text{minimize } \|b\|_1 + \lambda\|e\|_1 \\ &\text{subject to } y = Xb + e \end{aligned} \tag{4}$$

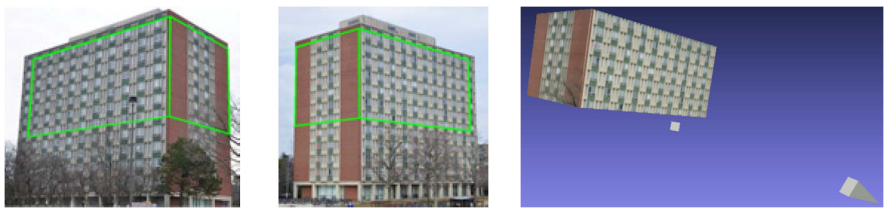
to achieve robustness. Here, [10] shows that the minimal number of reliable samples/equations required, regardless of whether the data is clean or corrupted, is essentially the same.

The versatility of the L + S model. We now move to discuss the L + S model more generally and survey a set of circumstances where it has proven useful and powerful. To begin with, methods which simply minimize an ℓ_1 norm, or a nuclear norm, or a combination thereof are seductive because they are flexible and apply to a rich class of problems. The L + S model is nonparametric and does not make many assumptions. As a result, it is widely applicable to problems ranging from latent variable model selection [4] (arguably one of the most subtle and beautiful applications of this method) to video surveillance in computer vision and document classification in machine learning [2]. In any given application, when much is known about the problem, it may not return the best possible answer, but our experience is that it is always fairly competitive. That is, the little performance loss we might encounter is more than accounted for by the robustness we gain vis a vis various modeling assumptions, which may or may not hold in real applications. A few recent applications of the L + S model demonstrate its flexibility and robustness.

Applications in computer vision. The L + S model has been applied to address several problems in computer vision, most notably by the group of Yi Ma and colleagues. Although the low-rank + sparse model may not hold precisely, the nuclear + ℓ_1 relaxation appears practically robust. This may be in contrast with algorithms which use detailed modeling assumptions and may not perform well under slight model mismatch or variation.

Video surveillance. An important task in computer vision is to separate background from foreground. Suppose we stack a sequence of video frames as columns of a matrix (rows are pixels and columns time points), then it is not hard to imagine that the background will have low-rank since it is not changing very much over time, while the foreground objects, such as cars, pedestrians and so on, can be seen as a sparse disturbance. Hence, finding an L + S decomposition offers a new way of modeling the background (and foreground). This method has been applied with some success [2]; see also the online videos Video 1 and Video 2.

From textures to 3D. One of the most fundamental steps in computer vision consists of extracting relevant features that are subsequently used for high-level vision applications such as 3D reconstruction, object recognition and scene understanding. There has been limited success in extracting stable features across variations in lightening, rotations and viewpoints. Partial occlusions further complicate matters. For certain classes of 3D objects such as images with regular symmetric patterns/textures, one can bypass the extraction of local features to recover 3D structure from 2D views. To fix ideas, a vertical or horizontal strip can be regarded as a rank-1 texture and a corner as a rank-2 texture. Generally speaking, surfaces may exhibit a low-rank texture when seen from a suitable viewpoint; see Figure 1. However, their 2D projections as captured by a camera will typically not be low rank. To see why, imagine there is a low-rank texture $L^0(x, y)$ on a planar surface. The image we observe is a transformed version of this texture, namely, $L^0 \circ \tau^{-1}(x, y)$. A technique named TILT [16] recovers τ simply by seeking a low-rank and sparse superposition. In spite of idealized assumptions, Figures 1 and 2 show that the L + S model works well in practice.



(a) (b)

FIG. 1. (a) *Pair of images from distinct viewpoints.* (b) *3D reconstruction (TILT) from photographs in* (a) *using the L + S model. The geometry is recovered from two images.*

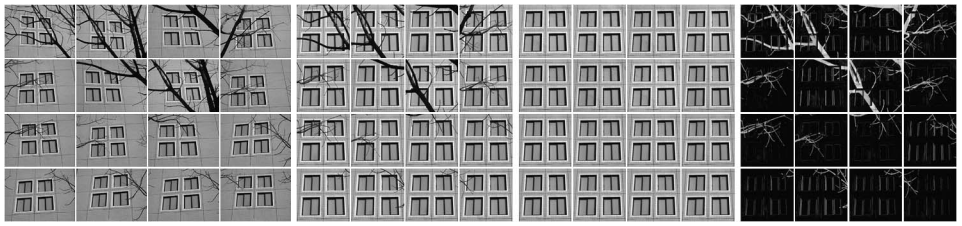


FIG. 2. *We are given the* 16 *images on the right. The task is to remove the clutter and align the images. Stacking each image as a column of a matrix, we look for planar homeographies that reveal a low-rank plus sparse structure* [12]. *From left to right*: *original data set*, *aligned images*, *low-rank component* (*columns of* L), *sparse component* (*columns of* S).

Compressive acquisition. In the spirit of compressive sensing, the L + S model can also be used to speed up the acquisition of large data sets or lower the sampling rate. At the moment, the theory of compressive sensing relies on the sparsity of the object we wish to acquire, however, in some setups the L + S model may be more appropriate. To explain our ideas, it might be best to start with two concrete examples. Suppose we are interested in the efficient acquisition of either (1) a hyper-spectral image or (2) a video sequence. In both cases, the object of interest is a data matrix M which is $N \times d$, where each column is an N-pixel image and each of the d columns corresponds to a specific wavelength (as in the hyper-spectral example) or frame (or time point as in the video example). In the first case, the data matrix may be thought of as $M(x, \lambda)$, where x indexes position and λ wavelength, whereas in the second example, we have $M(x, t)$ where t is a time index. We would like to obtain a sequence of highly resolved images from just a few measurements; an important application concerns dynamic magnetic resonance imaging where it is only possible to acquire a few samples in k-space per time interval.

Clearly, frames in a video sequence are highly correlated in time. And in just the same way, two images of the same scene at nearby wavelengths are also highly correlated. Obviously, images are correlated in space as well. Suppose that $W \otimes F$ is a tensor basis, where W sparsifies images and F time traces (W might be a wavelet transform and F a Fourier transform). Then we would expect WMF to be a nearly sparse matrix. With undersampled data of the form $y = \mathcal{A}(M) + z$, where $\mathcal{A}$ is the operator supplying information about M and z is a noise term, this leads to the low-rank + sparse decomposition problem

$$\begin{aligned} &\text{minimize } \|X\|_* + \lambda \|WXF\|_1 \\ &\text{subject to } \|\mathcal{A}(X) - y\|_2 \le \varepsilon, \end{aligned} \tag{5}$$

where ε^2 is the noise power. A variation, which is more in line with the discussion paper is a model in which L is a low-rank matrix modeling the static background, and S is a sparse matrix roughly modeling the innovation from one frame to the

next; for instance, S might encode the moving objects in the foreground. This would give

$$\begin{aligned} &\text{minimize } \lambda\|L\|_* + \|WSF\|_1 \\ &\text{subject to } \|\mathcal{A}(L+S) - y\|_2 \le \varepsilon. \end{aligned} \tag{6}$$

One could imagine that these models might be useful in alleviating the tremendous burden on system resources in the acquisition of ever larger 3D, 4D and 5D data sets.

We note that proposals of this kind have begun to emerge. As we were preparing this commentary, we became aware of [8], which suggests a model similar to (5) for hyperspectral imaging. The difference is that the second term in (5) is of the form $\sum_i \|X_i\|_{\mathrm{TV}}$ in which X_i is the ith column of X, the image at wavelength λ_i; that is, we minimize the total variation of each image, instead of looking for sparsity simultaneously in space and wavelength/frequency. The results in [8] show that dramatic undersampling ratios are possible. In medical imaging, movement due to respiration can degrade the image quality of Computed Tomography (CT), which can lead to incorrect dosage in radiation therapy. Using time-stamped data, 4D CT has more potential for precise imaging. Here, one can think of the object as a matrix with rows labeling spatial variables and columns time. In this context, we have a low-rank (static) background and a sparse disturbance corresponding to the dynamics, for example, of the heart in cardiac imaging. The recent work [7] shows how one can use the L + S model in a fashion similar to (6). This has interesting potential for dose reduction since the approach also supports substantial undersampling.

Connections with theoretical computer science and future directions. A class of problems where further study is required concerns situations in which the low-rank and sparse components have a particular structure. One such problem is the *planted clique problem*. It is well known that finding the largest clique in a graph is NP hard; in fact, it is even NP-hard to approximate the size of the largest clique in an n vertex graph to within a factor $n^{1-\varepsilon}$. Therefore, much research has focused on an "easier" problem. Consider a random graph $G(n, 1/2)$ on n vertices where each edge is selected independently with probability $1/2$. The expected size of its largest clique is known to be $(2 - o(1)) \log n$. The planted clique problem adds a clique of size k to G. One hopes that it is possible to find the planted clique in polynomial time whenever $k \gg \log n$. At this time, this is only known to be possible if k is on the order of $\sqrt{n}$ or larger. In spite of its seemingly simple formulation, this problem has eluded theoretical computer scientists since 1998, and is regarded as a notoriously difficult problem in modern combinatorics. It is also fundamental to many areas in machine learning and pattern recognition. To emphasize its wide applicability, we mention a new connection with game theory. Roughly speaking, the recent work [9] shows that finding a near-optimal

Nash equilibrium in two-player games is as hard as finding hidden cliques of size $k = C_0 \log n$, where C_0 is some universal constant.

One can think about the planted clique as a low rank + sparse decomposition problem. To be sure, the adjacency matrix of the graph can be written as the sum of two matrices: the low-rank component is of rank 1 and represents the clique of size k (a submatrix with all entries equal to 1); the sparse component stands for the random edges (and with -1 on the diagonal if and only if that vertex belongs to the hidden clique). Interestingly, low-rank + sparse regularization based on nuclear and ℓ_1 norms have been applied to this problem [5]. (Here the clique is both low-rank and sparse and is the object of interest so that we minimize $\|X\|_* + \lambda\|X\|_1$ subject to data constraints.) These proofs show that these methods find cliques of size $\Omega(\sqrt{n})$, thus recovering the best known results, but they may not be able to break this barrier. It is interesting to investigate whether tighter relaxations, taking into account the specific structure of the low-rank and sparse components, can do better.

REFERENCES

[1] BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[2] CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000

[3] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

[4] CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479

[5] DOAN, X. V. and VAVASIS, S. A. (2010). Finding approximately rank-one submatrices with the nuclear norm and ℓ_1 norm. Available at http://arxiv.org/abs/1011.1839.

[6] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[7] GAO, H., CAI, J., SHEN, Z. and ZHAO, H. (2011). Robust principal component analysis-based four-dimensional computed tomography. *Phys. Med. Biol.* **56** 3181.

[8] GOLBABAEE, M. and VANDERGHEYNST, P. (2012). Joint trace/TV norm minimization: A new efficient approach for spectral compressive imaging. In *IEEE International Conference on Image Processing (ICIP), Orlando, Florida*.

[9] HAZAN, E. and KRAUTHGAMER, R. (2011). How hard is it to approximate the best Nash equilibrium? *SIAM J. Comput.* **40** 79–91. MR2765712

[10] LI, X. (2011). Compressed sensing and matrix completion with constant proportion of corruptions. Available at http://arxiv.org/abs/1104.1041.

[11] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[12] PENG, Y., GANESH, A., WRIGHT, J., XU, W. and MA, Y. (2010). RASL: Robust alignment by sparse and low-rank de-composition for linearly correlated images. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA*.

[13] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

[14] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[15] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

[16] ZHANG, Z., GANESH, A., LIANG, X. and MA, Y. (2012). TILT: Transform-invariant low-rank textures. *International Journal of Computer Vision (IJCV)*. To appear.

DEPARTMENT OF STATISITCS
390 SERRA MALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: candes@stanford.edu
mahdisol@stanford.edu

The Annals of Statistics
2012, Vol. 40, No. 4, 2005–2013
DOI: 10.1214/12-AOS1020
Main article DOI: 10.1214/11-AOS949

REJOINDER: LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION

BY VENKAT CHANDRASEKARAN, PABLO A. PARRILO AND ALAN S. WILLSKY

California Institute of Technology, Massachusetts Institute of Technology and Massachusetts Institute of Technology

1. Introduction. We thank all the discussants for their careful reading of our paper, and for their insightful critiques. We would also like to thank the editors for organizing this discussion. Our paper contributes to the area of high-dimensional statistics which has received much attention over the past several years across the statistics, machine learning and signal processing communities. In this rejoinder we clarify and comment on some of the points raised in the discussions. Finally, we also remark on some interesting challenges that lie ahead in latent variable modeling.

Briefly, we considered the problem of latent variable graphical model selection in the Gaussian setting. Specifically, let X be a zero-mean Gaussian random vector in $\mathbb{R}^{p+h}$ with O and H representing disjoint subsets of indices in $\{1, \ldots, p+h\}$ with $|O| = p$ and $|H| = h$. Here the subvector X_O represents the observed variables and the subvector X_H represents the latent variables. Given samples of only the variables X_O, is it possible to consistently perform model selection? We noted that if the number of latent variables h is small relative to p and if the conditional statistics of the observed variables X_O conditioned on the latent variables X_H are given by a sparse graphical model, then the marginal concentration matrix of the observed variables X_O is given as the sum of a sparse matrix and a low-rank matrix. As a first step we investigated the identifiability of latent variable Gaussian graphical models—effectively, this question boils down to one of uniquely decomposing the sum of a sparse matrix and a low-rank matrix into the individual components. By studying the geometric properties of the algebraic varieties of sparse and low-rank matrices, we provided natural sufficient conditions for identifiability and gave statistical interpretations of these conditions. Second, we proposed the following regularized maximum-likelihood estimator to decompose the concentration matrix into sparse and low-rank components:

$$\begin{aligned}(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} &-\ell(S - L; \Sigma_O^n) + \lambda_n\big(\gamma\|S\|_1 + \operatorname{tr}(L)\big)\\ &\text{s.t. } S - L \succ 0, L \succeq 0.\end{aligned} \tag{1.1}$$

Received May 2012.

Here Σ_O^n represents the sample covariance formed from n samples of the observed variables, ℓ is the Gaussian log-likelihood function, $\hat{S}_n$ represents the estimate of the conditional graphical model of the observed variables conditioned on the latent variables, and $\hat{L}_n$ represents the extra correlations induced due to marginalization over the latent variables. The ℓ_1 norm penalty induces sparsity in $\hat{S}_n$ and the trace norm penalty induces low-rank structure in $\hat{L}_n$. An important feature of this estimator is that it is given by a convex program that can be solved efficiently. Our final contribution was to establish the high-dimensional consistency of this estimator under suitable assumptions on the Fisher information underlying the true model (in the same spirit as irrepresentability conditions for sparse model selection [11, 16]).

2. Alternative estimators. A number of the commentaries described alternative formulations for estimators in the latent variable setting.

2.1. *EM-based methods*. The discussions by Yuan and by Lauritzen and Meinshausen describe an EM-based alternative in which the rank of the matrix L is explicitly constrained:

$$(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} -\ell(S - L; \Sigma_O^n) + \lambda_n \|S\|_1 \quad \text{s.t. } S - L \succ 0, L \succeq 0, \operatorname{rank}(L) \leq r. \tag{2.1}$$

The experimental results based on this approach seem quite promising, and certainly deserve further investigation. On the one hand, we should reiterate that the principal motivation for our convex optimization based formulation was to develop a method for latent variable modeling with provable statistical *and* computational guarantees. One of the main drawbacks of EM-based methods is the existence of local optima in the associated variational formulations, thus leading to potentially different solutions depending on the initial point. On the other hand, one of the reasons for the positive empirical behavior observed by Yuan and by Lauritzen and Meinshausen may be that all the local optima in the experimental settings considered by the authors may be "good" models. Such behavior has in fact been rigorously characterized recently for certain nonconvex estimators in some missing data problems [7].

One of the motivations for the EM proposal of Yuan and of Lauritzen and Meinshausen seems to be that there are fairly mature and efficient solvers for the graphical lasso. As our estimator is relatively newer and as its properties are better understood going forward, we expect that more efficient solvers will be developed for (1.1) as well. Indeed, the LogdetPPA solver [15] that we cite in our paper already scales to instances involving several hundred variables, while more recent efforts [8] have resulted in algorithms that scale to instances with several thousand variables.

2.2. *Thresholding estimators*. Ren and Zhou propose and analyze an interesting thresholding based estimator for decomposing a concentration matrix into sparse and low-rank components. They apply a two-step procedure—ℓ_1 norm thresholding followed by trace norm thresholding—to obtain the sparse component followed by the low-rank component. Roughly speaking, this two-step estimator can be viewed as the application of the first cycle of a block coordinate descent procedure to compute our estimator that alternately updates the sparse and low-rank pieces (we also refer the reader to the remarks in [1]).

However, in Theorem 1 in the discussion by Ren and Zhou, a quite stringent assumption requires that in some scaling regimes the true low-rank component L^* must vanish, that is, $\|L^*\|_{\ell_\infty} \lesssim \sqrt{\frac{\log p}{n}} \to 0$. The reason for this condition is effectively to ensure sign consistency in recovering the sparse component. In a pure sparse model selection problem (with no low-rank component in the population), the deviation away from the sparse component is given only by noise due to finite samples and this deviation is on the order of $\sqrt{\frac{\log p}{n}}$ in the Gaussian setting—consequently, sparse model selection via ℓ_1 norm thresholding is sign-consistent when the minimum magnitude nonzero entry in the true model is larger than $\sqrt{\frac{\log p}{n}}$. In contrast, if the true model consists of both a sparse component and a low-rank component, the total deviation away from the sparse component in the finite sample regime is given by both sample noise as well as the low-rank component. This seems to be the reason for the stringent assumption on the vanishing of the low-rank component in Theorem 1 of Ren and Zhou.

More broadly, one of the motivations of Ren and Zhou in proposing and analyzing their estimator is that it may be possible to weaken the assumptions on the minimum magnitude nonzero entry θ of the true sparse component S^* and the minimum nonzero singular value σ of the true low-rank component L^*—whether this is possible under less stringent assumptions on L^* is an interesting question, and we comment on this point in Section 3 in the more general context of potentially improving the rates in our paper.

2.3. *Other proposals*. Giraud and Tsybakov propose two alternative estimators for decomposing a concentration matrix into sparse and low-rank components. While our approach (1.1) builds on the graphical lasso, their proposed approaches build on the Dantzig selector of Candès and Tao [2] and the neighborhood selection approach of Meinshausen and Bühlmann [9]. Several comments are in order here.

First, we note that the extension of neighborhood selection proposed by Giraud and Tsybakov to deal with the low-rank component begins by reformulating the neighborhood selection procedure to obtain a "global" estimator that simultaneously estimates all the neighborhoods. This reformulation touches upon a fundamental aspect of latent variable modeling. In many applications marginalization

over the latent variables typically induces correlations between most pairs of observed variables—consequently, local procedures that learn model structure one node at a time are ill-suited for latent variable modeling. Stated differently, requiring that a matrix be sparse with few nonzeros per row or column (e.g., expressing preference for a graphical model with bounded degree) can be done by imposing column-wise constraints. On the other hand, the constraint that a matrix be low-rank is really a global constraint expressed by requiring all minors of a certain size to vanish. Thus, any estimator for latent variable modeling (in the absence of additional conditions on the latent structure) must necessarily be global in nature.

Second, we believe that the reformulation based on the Dantzig selector perhaps ought to have an additional constraint. Recall that the Dantzig selector [2] constrains the ℓ_∞ norm (the dual norm of the ℓ_1 norm) of the correlated residuals rather than the ℓ_2 norm of the residuals as in the lasso. As the dual norm of our combined ℓ_1/trace norm regularizer involves both an ℓ_∞ norm and a spectral norm, the following constraint set may be more appropriate in the Dantzig selector based reformulation of Giraud and Tsybakov:

$$\mathcal{G} = \{(S, L) : \|\Sigma_O^n(S + L) - I\|_{\ell_\infty} \leq \gamma\lambda_n, \|\Sigma_O^n(S + L) - I\|_2 \leq \lambda_n\}.$$

Finally, we note that the Dantzig selector of [2] has the property that its constraint set contains the lasso solution (with the same choice of regularization/relaxation parameters). In contrast, this property is not shared in general by the Dantzig selector reformulation of Giraud and Tsybakov in relation to our regularized maximum-likelihood estimator (1.1). It is unclear how one might achieve this property via suitable convex constraints in a Dantzig selector type reformulation of our estimator.

In sum, both of these alternative estimators deserve further study.

3. Comments on rates. Several of the commentaries (Wainwright, Giraud and Tsybakov, Ren and Zhou and Candès and Soltanolkotabi) bring up the possibility of improving the rates given in our paper. At the outset we believe that $n \gtrsim p$ samples is inherent to the latent variable modeling problem if spectral norm consistency is desired in the low-rank component. This is to be expected since the spectral norm of the deviation of a sample covariance from the underlying population covariance is on the order of $\sqrt{\frac{p}{n}}$. However, some more subtle issues remain.

Giraud and Tsybakov point out that one may be concerned purely with estimation of the sparse component, and that the low-rank component may be a "nuisance" parameter. While this is not appropriate in every application, in problem domains where the conditional graphical model structure of the observed variables is the main quantity of interest one can imagine quantifying deviations in the low-rank component via "weaker" norms than the spectral norm—this may lead to consistent estimates for the sparse component with $n \ll p$ samples. The analysis in our paper does not rule out this possibility, and a more careful investigation is needed to establish such results.

Ren and Zhou suggest that while $n \gtrsim p$ may be required for consistent estimation, one may be able to weaken the assumptions on θ and σ (the minimum magnitude nonzero entry of the sparse component and the minimum nonzero singular value of the low-rank component, respectively). From the literature on sparse model selection, a natural lower bound on the minimum magnitude nonzero entry for consistent model selection is typically given by the size of the noise measured in the ℓ_∞ norm (the dual of the ℓ_1 regularizer). Building on this intuition, a natural lower bound that one can expect in our setting on θ is $\frac{1}{\gamma}\|\Sigma_O^n - \Sigma\|_{\ell_\infty}$, while a natural bound on σ would be $\|\Sigma_O^n - \Sigma\|_2$. The reason for this suggestion is that $\max\{\frac{\|S\|_{\ell_\infty}}{\gamma}, \|L\|_2\}$ is the dual norm of the regularizer used in our paper. Therefore, it may be possible to only require $\theta \sim \frac{1}{\gamma}\sqrt{\frac{\log p}{n}}$ and $\sigma \sim \sqrt{\frac{p}{n}}$. However, one issue here is that the ℓ_∞ norm bound kicks in when $n \gtrsim \log p$ with probability approaching one polynomially fast, while the spectral norm bound only kicks in when $n \geq p$ but holds with probability approaching one exponentially fast. Thus (as also noted by Giraud and Tsybakov), it may be possible that $n \gtrsim p$ is required for overall consistent estimation, but that the assumption on θ could be weakened by only requiring that the probability of consistent estimation approach one polynomially fast.

Candès and Soltanolkotabi comment that it would be of interest to establish an "adaptivity" property whereby if no low-rank component were present, the number of samples required for consistent estimation would boil down to just the rate for sparse graphical model selection, that is, $n \sim \log p$. While such a feature would clearly be desirable to establish for our estimator, one potential roadblock may be that our estimator (1.1) "searches" over a larger classes of models than just those given by sparse graphical models; consequently, rejecting the hypothesis that the observed variables are affected by any latent variables may require that $n \gg \log p$. This question deserves further investigation and, as suggested by Candès and Soltanolkotabi, recent results on adaptivity could inform a more refined analysis of our estimator.

Finally, Wainwright suggests the intriguing possibility that faster rates may be possible if the low-rank component has additional structure. For example, there may exist a sparse factorization of the low-rank component due to special structure between the latent and observed variables. In such settings the trace norm regularizer applied to the low-rank component is not necessarily the tightest convex penalty. In recent joint work by the authors and Recht [4], a general framework for constructing convex penalty functions based on some desired structure is presented. The trace norm penalty for inducing low-rank structure is motivated from the viewpoint that a low-rank matrix is the sum of a small number of rank-one matrices and, therefore, the norm induced by the convex hull of the rank-one matrices (suitably scaled) is a natural convex regularizer as this convex hull (the trace norm ball) carries precisely the kind of facial structure required for inducing low-rank structure in matrices. In this spirit, one can imagine constructing convex penalty

functions by taking the convex hull of *sparse* rank-one matrices. While this convex hull is in general intractable to represent, relaxations of this set that are tighter than the trace norm ball could provide faster rates than can be obtained by using the trace norm.

4. Weakening of irrepresentability conditions. Wainwright asks a number of insightful questions regarding the potential for weakening our Fisher information based conditions. Giraud and Tsybakov also bring up connections between our conditions and irrepresentability conditions in previous papers on sparse model selection [11, 16].

In order to better understand if the Fisher information based conditions stated in our paper are necessary, Wainwright raises the question of obtaining a converse result by comparing to an oracle method that directly minimizes the rank and the cardinality of the support of the components. A difficulty with this approach is that we don't have a good handle on the set of matrices that are expressible as the sum of a sparse matrix and a low-rank matrix. The properties of this set remain poorly understood, and developing a better picture has been the focus of research efforts in algebraic geometry [6] and in complexity theory [14]. Nonetheless, a comparison to oracle estimators that have side information about the support of the sparse component and the row/column spaces of the low-rank component (in effect, side information about the tangent spaces at the two components) appears to be more tractable. This is closer to the viewpoint we have taken in our paper in which we consider the question of identifiability of the components given information about the underlying tangent spaces. Essentially, our Fisher information conditions state that these tangent spaces must be sufficiently transverse with respect to certain natural norms and in a space in which the Fisher information is the underlying inner-product. More generally, as also pointed out by Giraud and Tsybakov, the necessity of Fisher information based conditions is an open question even in the sparse graphical model selection setting considered in [11]. The experimental studies in [11] describing comparisons to neighborhood selection in some simple cases provide a good starting point.

Wainwright raises the broader question of consistent model selection when transversality of the underlying tangent spaces does not hold. One approach [1] is to quantify the level of identifiability based on a "spikiness" condition. A more geometric viewpoint may be that only those pieces of the sparse and low-rank components that do not lie in the intersection of their underlying tangent spaces are fundamentally identifiable and, therefore, consistency should be quantified with respect to these identifiable pieces.

Giraud and Tsybakov ask about the interpretability of our conditions $\xi(T)$ and $\mu(\Omega)$. These quantities are geometric in nature and relate to the tangent space conditions for identifiability. In particular, they are closely related to (and bounded by) the incoherence of the row/column spaces of the low-rank component and the maximum number of nonzeros per row/column [5]. These latter quantities have

appeared in many papers on sparse graphical model selection (e.g., [9, 11]) as well as on low-rank matrix completion [3], and computing them is straightforward. In our previous work on matrix decomposition [5], we note that these quantities are bounded for natural random families of sparse and low-rank matrices based on results in [3].

5. Experimental issues and applications. Lauritzen and Meinshausen as well as Giraud and Tsybakov raise several points about the choice of the regularization parameters. Choosing these parameters in a data-driven manner (e.g., using the methods described in [10]) is clearly desirable. We do wish to emphasize that the sensitivities of the solution with respect to the parameters λ_n and γ are qualitatively different. As described in our main theorem and in our experimental section, the solution of our estimator (1.1) is stable for a range of values of γ (see also [5])—this point is observed by Yuan as well in his experiments. Further, the choice of γ ideally should not depend on n, while the choice of λ_n clearly should.

On a different point regarding experimental results, Giraud and Tsybakov suggest at the end of their discussion that latent variable models don't seem to provide significantly more expressive power than a sparse graphical model. In contrast, Yuan's synthetic experiment seems to provide compelling evidence that our approach (1.1) provides better performance relative to models learned by the graphical lasso. The reason for these different observations may be tied to the manner in which their synthetic models were generated. Specifically, latent variable model selection using (1.1) is likely to be most useful when the latent variables affect many observed variables upon marginalization (e.g., latent variables are connected to many observed variables), while the conditional graphical model among the observed variables conditioned on the latent variables is sparse and has bounded degree. This intuition is based on the theoretical analysis in our paper and is also the setting considered in the experiment in Yuan's discussion (as well as in the synthetic experiments in our paper). On the other hand, the experimental setup followed by Giraud and Tsybakov seems to generate a graphical model with large maximum degree and low average degree, and randomly selects a subset of the variables as latent variables. It is not clear if these latent variables are the ones with large degree, which may explain their remarks.

Finally, we note that sparse and low-rank matrix decomposition is relevant in applications beyond the one described in our paper. As observed by Lauritzen and Meinshausen, a natural matrix decomposition problem involving *covariance* matrices may arise if one considers directed latent variable models in the spirit of factor analysis. In such a context the covariance matrix may be expressed as the sum of a low-rank matrix and a sparse (rather than just diagonal) matrix, corresponding to the setting in which the distribution of the observed variables conditioned on the latent variables is given by a sparse covariance matrix. More broadly, similar matrix decomposition problems arise in domains beyond statistical estimation such as optical system decomposition, matrix rigidity and system identification in control [5], as well as others as noted by Candès and Soltanolkotabi.

6. Future questions. Our paper and the subsequent discussions raise a number of research and computational challenges in latent variable modeling that we wish to highlight briefly.

6.1. *Convex optimization in R.* As mentioned by Lauritzen and Meinshausen, R remains the software of choice for practitioners in statistics. However, some of the recent advances in high-dimensional statistical estimation have been driven by sophisticated convex optimization based procedures that are typically prototyped using packages such as SDPT3 [13] and others in Matlab and Python. It would be of general interest to develop packages to invoke SDPT3 routines directly from R.

6.2. *Sparse/low-rank decomposition as infimal convolution.* Given a matrix $M \succ 0$, consider the following function:

$$\|M\|_{S/L,\gamma} = \min_{S,L} \gamma \|S\|_{\ell_1} + \operatorname{tr}(L), \qquad \text{s.t. } M = S - L,\ L \succeq 0. \tag{6.1}$$

It is clear that $\|\cdot\|_{S/L,\gamma}$ is a norm, and it can be viewed as the infimal convolution [12] of the (scaled) ℓ_1 norm and the trace norm. In essence, it is a norm whose minimization induces matrices expressible as the sum of sparse and low-rank components (see also the atomic norm viewpoint of [4]). We could then effectively restate (1.1) as

$$\hat{M}_n = \arg\min_{M \succ 0} -\ell(M; \Sigma_O^n) + \lambda_n \|M\|_{S/L,\gamma}$$

and then decompose $\hat{M}_n$ by solving (6.1). This two-step approach suggests the possibility of decoupling the decomposition problem from the conditions fundamentally required for consistency via regularized maximum-likelihood, as the latter only ought to depend on the composite norm $\|\cdot\|_{S/L,\gamma}$. This decoupling also highlights the different roles played by the parameters λ_n and γ (as discussed in Section 5). More broadly, such an approach may be useful as one analyzes general regularizers, for example, convex penalties other than the trace norm as described in Section 3.

6.3. *Non-Gaussian latent variable modeling.* As described in our paper and as raised by Wainwright, latent variable modeling with non-Gaussian variables is of interest in many applications. Both the computational and algebraic aspects present major challenges in this setting. Specifically, the secant varieties arising due to marginalization in non-Gaussian models (e.g., in models with categorical variables) are poorly understood, and computing the likelihood is also intractable. An approach based on matrix decomposition as described in our paper may be appropriate, although one would have to quantify the effects of the Gaussianity assumption.

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Preprint.

[2] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[3] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

[4] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2010). The convex geometry of linear inverse problems. Preprint.

[5] CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479

[6] DRTON, M., STURMFELS, B. and SULLIVANT, S. (2007). Algebraic factor analysis: Tetrads, pentads and beyond. *Probab. Theory Related Fields* **138** 463–493. MR2299716

[7] LOH, P. and WAINWRIGHT, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. Preprint.

[8] MA, S., XUE, L. and ZOU, H. (2012). Alternating direction methods for latent variable Gaussian graphical model selection. Technical report, IMA and School of Statistics, Univ. Minnesota.

[9] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[10] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

[11] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

[12] ROCKAFELLAR, R. T. (1997). *Convex Analysis*. Princeton Univ. Press, Princeton, NJ. MR1451876

[13] TOH, K. C., TODD, M. J. and TUTUNCU, R. H. SDPT3—A MATLAB software package for semidefinite-quadratic-linear programming. Available at http://www.math.nus.edu.sg/~mattohkc/sdpt3.html.

[14] VALIANT, L. G. (1977). Graph-theoretic arguments in low-level complexity. In *Mathematical Foundations of Computer Science* (*Proc. Sixth Sympos., Tatranská Lomnica*, 1977). *Lecture Notes in Comput. Sci.* **53** 162–176. Springer, Berlin. MR0660702

[15] WANG, C., SUN, D. and TOH, K.-C. (2010). Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *SIAM J. Optim.* **20** 2994–3013. MR2735941

[16] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

V. CHANDRASEKARAN
DEPARTMENT OF COMPUTING
AND MATHEMATICAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91106
USA
E-MAIL: venkatc@caltech.edu

P. A. PARRILO
A. S. WILLSKY
LABORATORY FOR INFORMATION
AND DECISION SYSTEMS
DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: parrilo@mit.edu
willsky@mit.edu

The Annals of Statistics

LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION – SUPPLEMENTARY MATERIAL

By Venkat Chandrasekaran, Pablo A. Parrilo and Alan S. Willsky

California Institute of Technology, Massachusetts Institute of Technology, and Massachusetts Institute of Technology

1. Matrix perturbation bounds. Given a low-rank matrix we consider what happens to the invariant subspaces when the matrix is perturbed by a small amount. We assume without loss of generality that the matrix under consideration is square and symmetric, and our methods can be extended to the general non-symmetric non-square case. We refer the interested reader to [1, 3] for more details, as the results presented here are only a brief summary of what is relevant for this paper. In particular the arguments presented here are along the lines of those presented in [1]. The appendices in [1] also provide a more refined analysis of second-order perturbation errors.

The resolvent of a matrix M is given by $(M - \zeta I)^{-1}$ [3], and it is well-defined for all $\zeta \in \mathbb{C}$ that do not coincide with an eigenvalue of M. If M has no eigenvalue with magnitude equal to η, then we have by the Cauchy residue formula that the projector onto the invariant subspace of a matrix M corresponding to all singular values smaller than η is given by

$$P_{M,\eta} = \frac{-1}{2\pi i} \oint_{\mathcal{C}_\eta} (M - \zeta I)^{-1} d\zeta, \tag{1.1}$$

where $\mathcal{C}_\eta$ denotes the positively-oriented circle of radius η centered at the origin. Similarly, we have that the weighted projection onto the invariant subspace corresponding to the smallest singular values is given by

$$P^w_{M,\eta} = M P_{M,\eta} = \frac{-1}{2\pi i} \oint_{\mathcal{C}_\eta} \zeta \ (M - \zeta I)^{-1} d\zeta, \tag{1.2}$$

Suppose that M is a low-rank matrix with smallest nonzero singular value σ, and let Δ be a perturbation of M such that $\|\Delta\|_2 \leq \kappa < \frac{\sigma}{2}$. We have the following identity for any $|\zeta| = \kappa$, which will be used repeatedly:

$$[(M + \Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} = -[M - \zeta I]^{-1} \Delta [(M + \Delta) - \zeta I]^{-1}. \tag{1.3}$$

We then have that

$$\begin{aligned} P_{M+\Delta,\kappa} - P_{M,\kappa} &= \frac{-1}{2\pi i} \oint_{\mathcal{C}_\kappa} [(M+\Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} d\zeta. \end{aligned} \tag{1.4}$$

Similarly, we have the following for $P^w_{M,\kappa}$:

$$\begin{aligned} P^w_{M+\Delta,\kappa} &- P^w_{M,\kappa} \\ &= \frac{-1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ \left\{ [(M+\Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} \right\} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ \left\{ [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} \right\} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ [M - \zeta I]^{-1} \Delta [M - \zeta I]^{-1} d\zeta \\ &\quad - \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ [M - \zeta I]^{-1} \Delta [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} d\zeta. \end{aligned} \tag{1.5}$$

Given these expressions, we have the following two results.

PROPOSITION 1.1. *Let $M \in \mathbb{R}^{p \times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\kappa}{2}$ with $\kappa < \frac{\sigma}{2}$. Then we have that*

$$\|P_{M+\Delta,\kappa} - P_{M,\kappa}\|_2 \leq \frac{\kappa}{(\sigma - \kappa)(\sigma - \frac{3\kappa}{2})} \ \|\Delta\|_2.$$

Proof: This result follows directly from the expression (1.4), and the sub-multiplicative property of the spectral norm:

$$\begin{aligned} \|P_{M+\Delta,\kappa} - P_{M,\kappa}\|_2 &\leq \frac{1}{2\pi} \ 2\pi \ \kappa \ \frac{1}{\sigma - \kappa} \ \|\Delta\|_2 \ \frac{1}{\sigma - \frac{3\kappa}{2}} \\ &= \frac{\kappa}{(\sigma - \kappa)(\sigma - \frac{3\kappa}{2})} \ \|\Delta\|_2. \end{aligned}$$

Here, we used the fact that $\|[M - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \kappa}$ and $\|[(M+\Delta) - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \frac{3\kappa}{2}}$ for $|\zeta| = \kappa$. $\square$

Next, we develop a similar bound for $P^w_{M,\kappa}$. Let $U(M)$ denote the invariant subspace of M corresponding to the nonzero singular values, and let $P_{U(M)}$ denote the projector onto this subspace.

PROPOSITION 1.2. *Let $M \in \mathbb{R}^{p\times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\kappa}{2}$ with $\kappa < \frac{\sigma}{2}$. Then we have that*

$$\|P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} - (I - P_{U(M)})\Delta(I - P_{U(M)})\|_2 \leq \frac{\kappa^2}{(\sigma-\kappa)^2(\sigma - \frac{3\kappa}{2})}\ \|\Delta\|_2^2.$$

Proof: One can check that

$$\frac{1}{2\pi i}\oint_{\mathcal{C}_\kappa} \zeta\ [M - \zeta I]^{-1}\Delta[M - \zeta I]^{-1}d\zeta = (I - P_{U(M)})\Delta(I - P_{U(M)}).$$

Next we use the expression (1.5), and the sub-multiplicative property of the spectral norm:

$$\begin{aligned}\|P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} &- (I - P_{U(M)})\Delta(I - P_{U(M)})\|_2 \\ &\leq \frac{1}{2\pi}\ 2\pi\ \kappa\ \kappa\ \frac{1}{\sigma-\kappa}\ \|\Delta\|_2\ \frac{1}{\sigma-\kappa}\ \|\Delta\|_2\ \frac{1}{\sigma - \frac{3\kappa}{2}} \\ &= \frac{\kappa^2}{(\sigma-\kappa)^2(\sigma - \frac{3\kappa}{2})}\ \|\Delta\|_2^2.\end{aligned}$$

As with the previous proof, we used the fact that $\|[M - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma-\kappa}$ and $\|[(M+\Delta) - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \frac{3\kappa}{2}}$ for $|\zeta| = \kappa$. □

We will use these expressions to derive bounds on the "twisting" between the tangent spaces at M and at $M + \Delta$ with respect to the rank variety.

2. Curvature of rank variety. For a symmetric rank-r matrix M, the projection onto the tangent space $T(M)$ (restricted to the variety of symmetric matrices with rank less than or equal to r) can be written in terms of the projection $P_{U(M)}$ onto the row space $U(M)$. For any matrix N

$$\mathcal{P}_{T(M)}(N) = P_{U(M)}N + NP_{U(M)} - P_{U(M)}NP_{U(M)}.$$

One can then check that the projection onto the normal space $T(M)^\perp$

$$\mathcal{P}_{T(M)^\perp}(N) = [I - \mathcal{P}_{T(M)}](N) = (I - P_{U(M)})\ N\ (I - P_{U(M)}).$$

PROPOSITION 2.1. *Let $M \in \mathbb{R}^{p\times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\sigma}{8}$. Further, let $M + \Delta$ be a rank-r matrix. Then we have that*

$$\rho(T(M+\Delta), T(M)) \leq \frac{2}{\sigma}\ \|\Delta\|_2.$$

Proof: For any matrix N, we have that

$$\begin{aligned}[\mathcal{P}_{T(M+\Delta)} - \mathcal{P}_{T(M)}](N) &= [P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}] \\ &\quad + [I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}].\end{aligned}$$

Further, we note that for $\kappa < \frac{\sigma}{2}$

$$\begin{aligned}P_{U(M+\Delta)} - P_{U(M)} &= [I - P_{U(M)}] - [I - P_{U(M+\Delta)}] \\ &= P_{M,\kappa} - P_{M+\Delta,\kappa},\end{aligned}$$

where $P_{M,\kappa}$ is defined in the previous section. Thus, we have the following sequence of inequalities for $\kappa = \frac{\sigma}{4}$:

$$\begin{aligned}\rho(T(M+\Delta), T(M)) &= \max_{\|N\|_2 \leq 1} \|[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}] \\ &\quad + [I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]\|_2 \\ &\leq \max_{\|N\|_2 \leq 1} \|[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}]\|_2 \\ &\quad + \max_{\|N\|_2 \leq 1} \|[I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]\|_2 \\ &\leq 2\ \|P_{M+\Delta,\frac{\sigma}{4}} - P_{M,\frac{\sigma}{4}}\|_2 \\ &\leq \frac{2}{\sigma}\ \|\Delta\|_2,\end{aligned}$$

where we obtain the last inequality from Proposition 1.1. □

PROPOSITION 2.2. *Let $M \in \mathbb{R}^{p \times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\| \leq \frac{\sigma}{8}$. Further, let $M + \Delta$ be a rank-r matrix. Then we have that*

$$\|\mathcal{P}_{T(M)^\perp}(\Delta)\|_2 \leq \frac{\|\Delta\|_2^2}{\sigma}.$$

Proof: Since both M and $M+\Delta$ are rank-r matrices, we have that $\mathcal{P}^w_{M+\Delta,\kappa} = \mathcal{P}^w_{M,\kappa} = 0$ for $\kappa = \frac{\sigma}{4}$. Consequently,

$$\begin{aligned}\|\mathcal{P}_{T(M)^\perp}(\Delta)\|_2 &= \|(I - P_{U(M)})\ \Delta\ (I - P_{U(M)})\|_2 \\ &\leq \frac{\|\Delta\|_2^2}{\sigma},\end{aligned}$$

where we obtain the last inequality from Proposition 1.2 with $\kappa = \frac{\sigma}{4}$. □

3. Proof of supplementary results of main theorem. Throughout this section we denote $m = \max\{1, \frac{1}{\gamma}\}$. Further $\Omega = \Omega(K_O^*)$ and $T = T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ denote the tangent spaces at the true sparse matrix $S^* = K_O^*$ and low-rank matrix $L^* = K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$. We assume that

$$\gamma \in \left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right] \tag{3.1}$$

We also let $E_n = \Sigma_O^n - \Sigma_O^*$ denote the difference between the true marginal covariance and the sample covariance. Finally we let $D = \max\{1, \frac{\nu\alpha}{3\beta(2-\nu)}\}$ throughout this section. For γ in the above range we note that

$$m \leq \frac{D}{\xi(T)}. \tag{3.2}$$

Standard facts that we use throughout this section are that $\xi(T) \leq 1$ and that $\|M\|_\infty \leq \|M\|_2$ for any matrix M.

We study the following convex program:

$$\begin{aligned}(\bar{S}_n, \bar{L}_n) = \arg\min_{S,L} \ & \operatorname{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \ + \ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.} \quad & S - L \succ 0.\end{aligned} \tag{3.3}$$

Comparing (3.3) with the convex program (1.2) (main paper), the main difference is that we do not constrain the variable L to be positive semidefinite in (3.3) (recall that the nuclear norm of a positive semidefinite matrix is equal to its trace). However we show that the unique optimum $(\bar{S}_n, \bar{L}_n)$ of (3.3) under the hypotheses of Theorem 4.1 (main paper) is such that $\bar{L}_n \succeq 0$ (with high probability). Therefore we conclude that $(\bar{S}_n, \bar{L}_n)$ is also the unique optimum of (1.2) (main paper). The subdifferential with respect to the nuclear norm at a matrix M with (reduced) SVD given by $M = UDV^T$ is as follows:

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

The proof of this theorem consists of a number of steps, each of which is analyzed in separate sections below. We explicitly keep track of the constants α, β, ν, ψ. The key ideas are as follows:

1. We show that if we solve the convex program (3.3) subject to the additional constraints that $S \in \Omega$ and $L \in T'$ for some T' "close to" T (measured by $\rho(T', T)$), then the error between the optimal solution $(\bar{S}_n, \bar{L}_n)$ and the underlying matrices (S^*, L^*) is small. This result is discussed in Appendix 3.2.

2. We analyze the optimization problem (3.3) with the additional constraint that the variables S and L belong to the algebraic varieties of sparse and low-rank matrices respectively, and that the corresponding tangent spaces are close to the tangent spaces at (S^*, L^*). We show that under suitable conditions on the minimum nonzero singular value of the true low-rank matrix L^* and on the minimum magnitude nonzero entry of the true sparse matrix S^*, the optimum of this modified program is achieved at a *smooth* point of the underlying varieties. In particular the bound on the minimum nonzero singular value of L^* helps bound the curvature of the low-rank matrix variety locally around L^* (we use the results described in Appendix 2). These results are described in Appendix 3.3.
3. The next step is to show that the variety constraint can be linearized and changed to a tangent-space constraint (see Appendix 3.4), thus giving us a *convex program*. Under suitable conditions this tangent-space constrained program also has an optimum that has the same support/rank as the true (S^*, L^*). Based on the previous step these tangent spaces in the constraints are close to the tangent spaces at the true (S^*, L^*). Therefore we use the first step to conclude that the resulting error in the estimate is small.
4. Finally we show that under suitable identifiability conditions these tangent-space constraints are inactive at the optimum. Therefore we conclude with the statement that the optimum of the convex program (3.3) without any variety constraints is achieved at a pair of matrices that have the same support/rank as the true (S^*, L^*) (with high probability). Further the low-rank component of the solution is positive semidefinite, thus allowing us to conclude that the original convex program (1.2) (main paper) also provides estimates that are algebraically correct.

3.1. *Proof of main paper Proposition 5.1 – Bounded curvature of matrix inverse.* Consider the Taylor series of the inverse of a matrix:

$$(M + \Delta)^{-1} = M^{-1} - M^{-1}\Delta M^{-1} + R_{M^{-1}}(\Delta),$$

where

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

This infinite sum converges for Δ sufficiently small. The following proposition provides a bound on the second-order term specialized to our setting:

PROPOSITION 3.1. *Suppose that γ is in the range given by (3.1). Let $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$ for $C_1 = \psi(1 + \frac{\alpha}{6\beta})$, and for any (Δ_S, Δ_L) with $\Delta_S \in \Omega$. Then we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))) \leq \frac{2D\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2}{\xi(T)}.$$

Proof: We have that

$$\begin{aligned}
\|\mathcal{A}(\Delta_S, \Delta_L)\|_2 &\leq \|\Delta_S\|_2 + \|\Delta_L\|_2 \\
&\leq \gamma\mu(\Omega)\frac{\|\Delta_S\|_\infty}{\gamma} + \|\Delta_L\|_2 \\
&\leq (1 + \gamma\mu(\Omega))g_\gamma(\Delta_S, \Delta_L) \\
&\leq (1 + \frac{\alpha}{6\beta})g_\gamma(\Delta_S, \Delta_L) \\
&\leq \frac{1}{2\psi},
\end{aligned}$$

where the second-to-last inequality follows from the range for γ (3.1) and that $\nu \in (0, \frac{1}{2}]$, and the final inequality follows from the bound on $g_\gamma(\Delta_S, \Delta_L)$. Therefore,

$$\begin{aligned}
\|R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2 &\leq \psi\sum_{k=2}^{\infty}(\|\Delta_S + \Delta_L\|_2\psi)^k \\
&\leq \psi^3\|\Delta_S + \Delta_L\|_2^2\frac{1}{1 - \|\Delta_S + \Delta_L\|_2\psi} \\
&\leq 2\psi^3(1 + \frac{\alpha}{6\beta})^2 g_\gamma(\Delta_S, \Delta_L)^2 \\
&= 2\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2.
\end{aligned}$$

Here we apply the last two inequalities from above. Since the $\|\cdot\|_\infty$-norm is bounded above by the spectral norm $\|\cdot\|_2$, we have the desired result. □

3.2. *Proof of main paper Proposition 5.2 – Bounded errors.* Next we analyze the following convex program subject to certain additional tangent-space constraints:

(3.4)

$$\begin{aligned}
(\hat{S}_\Omega, \hat{L}_{T'}) = \arg\min_{S,L} \operatorname{tr}[(S - L)\ \Sigma_O^n] - \log\det(S - L)\ +\ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\
\text{s.t.} \quad S - L \succ 0, \quad S \in \Omega, \quad L \in T',
\end{aligned}$$

for some subspace T'. We show that if T' is any tangent space to the low-rank matrix variety such that $\rho(T, T') \leq \frac{\xi(T)}{2}$, then we can bound the error

$(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'})$. Let $\mathcal{C}_{T'} = \mathcal{P}_{T'^\perp}(L^*)$ denote the normal component of the true low-rank matrix at T', and recall that $E_n = \Sigma_O^n - \Sigma_O^*$ denotes the difference between the true marginal covariance and the sample covariance. The proof of the following result uses Brouwer's fixed-point theorem [4], and is inspired by the proof of a similar result in [5] for standard sparse graphical model recovery without latent variables.

PROPOSITION 3.2. *Let the error (Δ_S, Δ_L) in the solution of the convex program (3.4) (with T' such that $\rho(T', T) \leq \frac{\xi(T)}{2}$) be as defined above. Further let $C_1 = \psi(1 + \frac{\alpha}{6\beta})$, and define*

$$r = \max\left\{\frac{8}{\alpha}\left[g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T'}) + \lambda_n\right], \|\mathcal{C}_{T'}\|_2\right\}.$$

If we have that

$$r \leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\},$$

for γ in the range given by (3.1), then

$$g_\gamma(\Delta_S, \Delta_L) \leq 2r.$$

Proof: Based on Proposition 3.3 (main paper) we note that the convex program (3.4) is strictly convex (because the negative log-likelihood term has a strictly positive-definite Hessian due to the constraints involving transverse tangent spaces), and therefore the optimum is unique. Applying the optimality conditions of the convex program (3.4) at the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$, we have that there exist Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T'^\perp} \in T'^\perp$ such that

$$\begin{aligned} \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} + Q_{\Omega^\perp} &\in -\lambda_n\gamma\partial\|\hat{S}_\Omega\|_1, \\ \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} + Q_{T'^\perp} &\in \lambda_n\partial\|\hat{L}_{T'}\|_*. \end{aligned}$$

Restricting these conditions to the space $\mathcal{Y} = \Omega \times T'$, one can check that

$$\mathcal{P}_\Omega[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z_\Omega, \quad \mathcal{P}_{T'}[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z_{T'},$$

where $Z_\Omega \in \Omega, Z_{T'} \in T'$ and $\|Z_\Omega\|_\infty = \lambda_n\gamma, \|Z_{T'}\|_2 \leq 2\lambda_n$ (we use here the fact that projecting onto a tangent space T' increases the spectral norm by at most a factor of two). Denoting $Z = [Z_\Omega, Z_{T'}]$, we conclude that

$$\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z, \tag{3.5}$$

with $g_\gamma(Z) \leq 2\lambda_n$. Since the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$ is unique, one can check using Lagrangian duality theory [6] that $(\hat{S}_\Omega, \hat{L}_{T'})$ is the unique solution

of the equation (3.5). Rewriting $\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}$ in terms of the errors (Δ_S, Δ_L), we have using the Taylor series of the matrix inverse that

$$\begin{aligned}\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} &= \Sigma_O^n - [\mathcal{A}(\Delta_S, \Delta_L) + (\Sigma_O^*)^{-1}]^{-1} \\ &= E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^*\mathcal{A}(\Delta_S, \Delta_L) \\ &= E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) \\ &\quad + \mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L) + \mathcal{I}^*\mathcal{C}_{T'}. \end{aligned} \tag{3.6}$$

Since T' is a tangent space such that $\rho(T', T) \leq \frac{\xi(T)}{2}$, we have from Proposition 3.3 (main paper) that the operator $\mathcal{B} = \left(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}\right)^{-1}$ from $\mathcal{Y}$ to $\mathcal{Y}$ is bijective and is well-defined. Now consider the following matrix-valued function from $(\delta_S, \delta_L) \in \mathcal{Y}$ to $\mathcal{Y}$:

$$\begin{aligned}F(\delta_S, \delta_L) = (\delta_S, \delta_L) \quad - \quad \mathcal{B}\Big\{&\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) \\ &+ \mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\delta_S, \delta_L) + \mathcal{I}^*\mathcal{C}_{T'}] - Z\Big\}.\end{aligned}$$

A point $(\delta_S, \delta_L) \in \mathcal{Y}$ is a fixed-point of F if and only if $\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\delta_S, \delta_L) + \mathcal{I}^*\mathcal{C}_{T'}] = Z$. Applying equations (3.5) and (3.6) above, we then see that the only fixed-point of F by construction is the "true" error $\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)$ restricted to $\mathcal{Y}$. The reason for this is that, as discussed above, $(\hat{S}_\Omega, \hat{L}_{T'})$ is the unique optimum of (3.4) and therefore is the *unique solution* of (3.5). Next we show that this unique fixed-point of F lies in the ball $\mathbb{B}_r = \{(\delta_S, \delta_L) \mid g_\gamma(\delta_S, \delta_L) \leq r, \quad (\delta_S, \delta_L) \in \mathcal{Y}\}$.

In order to prove this step, we resort to Brouwer's fixed point theorem [4]. In particular we show that the function F maps the ball $\mathbb{B}_r$ onto itself. Since F is a continuous function and $\mathbb{B}_r$ is a compact set, we can conclude the proof of this proposition. Simplifying the function F, we have that

$$F(\delta_S, \delta_L) = \mathcal{B}\left\{\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) - \mathcal{I}^*\mathcal{C}_{T'}] + Z\right\}.$$

Consequently, we have from Proposition 3.3 (main paper) that

$$\begin{aligned}g_\gamma(F(\delta_S, \delta_L)) &\leq \frac{2}{\alpha}\, g_\gamma\left(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{C}_{T'}] - Z\right) \\ &\leq \frac{4}{\alpha}\, \left\{g_\gamma(\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{C}_{T'}]) + \lambda_n\right\} \\ &\leq \frac{r}{2} \quad + \quad \frac{4}{\alpha}\, g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'}))),\end{aligned}$$

where in the second inequality we use the fact that $g_\gamma(\mathcal{P}_{\mathcal{Y}}(\cdot, \cdot)) \leq 2g_\gamma(\cdot, \cdot)$ and that $g_\gamma(Z) \leq 2\lambda_n$, and in the final inequality we use the assumption on r.

We now bound the term $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L)))$ using Proposition 3.1 as $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$:

$$\begin{aligned}
\frac{4}{\alpha}\, g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'}))) &\leq \frac{8D\psi C_1^2 (g_\gamma(\delta_S, \delta_L) + \|\mathcal{C}_{T'}\|_2)^2}{\xi(T)\alpha} \\
&\leq \frac{32 D\psi C_1^2 r^2}{\xi(T)\alpha} \\
&\leq \frac{32 D\psi C_1^2 r}{\xi(T)\alpha}\, \frac{\alpha\xi(T)}{64 D\psi C_1^2} \\
&\leq \frac{r}{2},
\end{aligned}$$

where we have used the fact that $r \leq \frac{\alpha\xi(T)}{64D\psi C_1^2}$. Hence $g_\gamma(\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) \leq r$ by Brouwer's fixed-point theorem. Finally we observe that

$$\begin{aligned}
g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) + \|\mathcal{C}_{T'}\|_2 \\
&\leq 2r.
\end{aligned}$$

□

3.3. *Solving a variety-constrained problem.* In order to prove that the solution $(\bar{S}_n, \bar{L}_n)$ of (3.3) has the same sparsity pattern/rank as (S^*, L^*), we will study an optimization problem that explicitly enforces these constraints. Specifically, we consider the following *non-convex* constraint set:

$$\begin{aligned}
\mathcal{M} \;=\; \{(S, L) \mid\; & S \in \Omega(S^*),\ \operatorname{rank}(L) \leq \operatorname{rank}(L^*), \\
& \|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi^2},\ g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n\}
\end{aligned}$$

Recall that $S^* = K_O^*$ and $L^* = K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$. The first constraint ensures that the tangent space at S is the same as the tangent space at S^*; therefore the support of S is contained in the support of S^*. The second and third constraints ensure that L lives in the appropriate low-rank variety, but has a tangent space "close" to the tangent space T. The final constraint roughly bounds the sum of the errors $(S - S^*) + (L^* - L)$; note that this does not necessarily bound the individual errors. Notice that the only non-convex constraint is that $\operatorname{rank}(L) \leq \operatorname{rank}(L^*)$. We then have the following nonlinear program:

(3.7)
$$\begin{aligned}
(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M}) = \arg\min_{S,L}\ & \operatorname{tr}[(S - L)\ \Sigma_O^n] - \log\det(S - L)\ +\ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\
\text{s.t.}\quad & S - L \succ 0,\quad (S, L) \in \mathcal{M}.
\end{aligned}$$

Under suitable conditions this nonlinear program is shown to have a unique solution. Each of the constraints in $\mathcal{M}$ is useful for proving the consistency of the solution of the convex program (3.3). We show that under suitable conditions the constraints in $\mathcal{M}$ are actually inactive at the optimal $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$, thus allowing us to conclude that the solution of (3.3) is also equal to $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$; hence the solution of (3.3) shares the consistency properties of $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$. A number of interesting properties can be derived simply by studying the constraint set $\mathcal{M}$.

PROPOSITION 3.3. *Consider* any $(S, L) \in \mathcal{M}$, *and let* $\Delta_S = S - S^*, \Delta_L = L^* - L$. *For* γ *in the range specified by* (3.1) *and letting* $C_2 = \frac{48}{\alpha} + \frac{1}{\psi^2}$, *we have that* $g_\gamma(\Delta_S, \Delta_L) \leq C_2 \lambda_n$.

Proof: We have by the triangle inequality that

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(\mathcal{P}_\Omega(\Delta_S), \mathcal{P}_T(\Delta_L))) &\leq 11\lambda_n + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(\mathcal{P}_{\Omega^\perp}(\Delta_S), \mathcal{P}_{T^\perp}(\Delta_L))) \\ &\leq 11\lambda_n + m\psi^2 \|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \\ &\leq 12\lambda_n, \end{aligned}$$

as $m \leq \frac{D}{\xi(T)}$. Therefore, we have that $g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq 24\lambda_n$, where $\mathcal{Y} = \Omega \times T$. Consequently, we can apply Proposition 3.3 (main paper) to conclude that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq \frac{48\lambda_n}{\alpha}.$$

Finally, we use the triangle inequality again to conclude that

$$\begin{aligned} g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) + g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}(\Delta_S, \Delta_L)) \\ &\leq \frac{48\lambda_n}{\alpha} + m\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \\ &\leq C_2 \lambda_n. \end{aligned}$$

□

This simple result immediately leads to a number of useful corollaries. For example we have that under a suitable bound on the minimum nonzero singular value of $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, the constraint in $\mathcal{M}$ along the normal direction $T^\perp$ is locally inactive. Next we list several useful consequences of Proposition 3.3.

COROLLARY 3.4. *Consider* any $(S, L) \in \mathcal{M}$, *and let* $\Delta_S = S - S^*, \Delta_L = L^* - L$. *Suppose that* γ *is in the range specified by* (3.1), *and let* $C_3 = \left(\frac{6(2-\nu)}{\nu} + 1\right) C_2^2 \psi^2 D$ *and* $C_4 = C_2 + \frac{3\alpha C_2^2 (2-\nu)}{16(3-\nu)}$ *(where* C_2 *is as defined*

in Proposition 3.3). Let the minimum nonzero singular value σ of $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ be such that $\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2}$ for $C_5 = \max\{C_3, C_4\}$, and suppose that the smallest magnitude nonzero entry of S^* is greater than $\frac{C_6\lambda_n}{\mu(\Omega)}$ for $C_6 = \frac{C_2\nu\alpha}{\beta(2-\nu)}$. Setting $T' = T(L)$ and $\mathcal{C}_{T'} = \mathcal{P}_{T'^{\perp}}(L^*)$, we then have that:

1. *L has rank equal to $\mathrm{rank}(L^*)$, i.e., L is a smooth point of the variety of matrices with rank less than or equal to $\mathrm{rank}(L^*)$. In particular L has the same inertia as L^*.*
2. $\|\mathcal{P}_{T^{\perp}}(\Delta_L)\|_2 \leq \frac{\xi(T)\lambda_n}{19D\psi^2}$.
3. $\rho(T, T') \leq \frac{\xi(T)}{4}$.
4. $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T'}) \leq \frac{\lambda_n\nu}{6(2-\nu)}$.
5. $\|\mathcal{C}_{T'}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$.
6. $\mathrm{sign}(S) = \mathrm{sign}(S^*)$.

Proof: We note the following facts before proving each step. First $C_2 \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2} \geq \frac{\xi(T)}{D\psi^2}$. Second $\xi(T) \leq 1$. Third we have from Proposition 3.3 that $\|\Delta_L\|_2 \leq C_2\lambda_n$. Finally $\frac{6(2-\nu)}{\nu} \geq 18$ for $\nu \in (0, \frac{1}{2}]$. We prove each step separately.

For the first step, we note that

$$\sigma \geq \frac{C_3\lambda_n}{\xi(T)^2} \geq \frac{19C_2^2\psi^2 D\lambda_n}{\xi(T)^2} \geq \frac{19C_2\lambda_n}{\xi(T)} \geq 8C_2\lambda_n \geq 8\|\Delta_L\|_2.$$

Hence L is a smooth point with rank equal to $\mathrm{rank}(L^*)$, and specifically has the same inertia as L^*.

For the second step, we use the fact that $\sigma \geq 8\|\Delta_L\|_2$ to apply Proposition 2.2:

$$\|\mathcal{P}_{T^{\perp}}(\Delta_L)\| \leq \frac{\|\Delta_L\|_2^2}{\sigma} \leq \frac{C_2^2\xi(T)^2\lambda_n^2}{C_3\lambda_n} \leq \frac{\xi(T)\lambda_n}{19D\psi^2}.$$

For the third step we apply Proposition 2.1 (by using the conclusion from above that $\sigma \geq 8\|\Delta_L\|_2$) so that

$$\rho(T, T') \leq \frac{2\|\Delta_L\|_2}{\sigma} \leq \frac{2C_2\xi(T)^2}{C_3} \leq \frac{2\xi(T)^2}{19C_2D\psi^2} \leq \frac{\xi(T)}{4}.$$

For the fourth step let σ' denote the minimum singular value of L. Consequently,

$$\sigma' \geq \frac{C_3\lambda_n}{\xi(T)^2} - C_2\lambda_n \geq C_2\lambda_n\left[\frac{19C_2D\psi^2}{\xi(T)^2} - 1\right] \geq 8\|\Delta_L\|_2.$$

Using the same reasoning as in the proof of the second step, we have that

$$\begin{aligned}\|\mathcal{C}_{T'}\|_2 &\leq \frac{\|\Delta_L\|_2^2}{\sigma'} \leq \frac{C_2^2\lambda_n^2}{(\frac{C_3}{\xi(T)^2} - C_2)\lambda_n} \\ &= \frac{C_2^2\xi(T)^2\lambda_n}{C_2^2 D\psi^2(\frac{6(2-\nu)}{\nu}) + C_2^2 D\psi^2 - C_2\xi(T)^2} \\ &\leq \frac{C_2^2\xi(T)^2\lambda_n}{C_2^2 D\psi^2(\frac{6(2-\nu)}{\nu})} \leq \frac{\nu\xi(T)\lambda_n}{6(2-\nu)D\psi^2}.\end{aligned}$$

Hence

$$g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T'}) \leq m\psi^2\|\mathcal{C}_{T'}\|_2 \leq \frac{\lambda_n\nu}{6(2-\nu)}.$$

For the fifth step the bound on σ' implies that

$$\sigma' \geq \frac{C_4\lambda_n}{\xi(T)^2} - C_2\lambda_n \geq \frac{3C_2^2\alpha(2-\nu)}{16(3-\nu)}\lambda_n$$

Since $\sigma' \geq 8\|\Delta_L\|_2$, we have from Proposition 2.2 and some algebra that

$$\|\mathcal{C}_{T'}\|_2 \leq \frac{C_2^2\lambda_n^2}{\sigma'} \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}.$$

For the final step since $\|\Delta_S\|_\infty \leq \gamma C_2\lambda_n$, the assumed lower bound on the minimum magnitude nonzero entry of S^* guarantees that $\text{sign}(S) = \text{sign}(S^*)$. □

Notice that this corollary applies to *any* $(S, L) \in \mathcal{M}$, and is hence applicable to *any solution* $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ of the $\mathcal{M}$-constrained program (3.7). For now we choose an arbitrary solution $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ and proceed. In the next steps we show that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ is *the unique* solution to the convex program (3.3), thus showing that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ is also the unique solution to (3.7).

3.4. *From variety constraint to tangent-space constraint.* Given the solution $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$, we show that the solution to the convex program (3.4) with the tangent space constraint $L \in T_\mathcal{M} \triangleq T(\hat{L}_\mathcal{M})$ is the same as $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ under suitable conditions:

(3.8)
$$\begin{aligned}(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}}) = \arg\min_{S,L}\ &\text{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.}\quad & S - L \succ 0, \quad S \in \Omega, \quad L \in T_\mathcal{M}.\end{aligned}$$

Assuming the bound of Corollary 3.4 on the minimum singular value of L^* the uniqueness of the solution $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ is assured. This is because we

have from Proposition 3.3 (main paper) and from Corollary 3.4 that $\mathcal{I}^*$ is injective on $\Omega \oplus T_{\mathcal{M}}$. Therefore the Hessian of the convex objective function of (3.8) is strictly positive-definite at $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$.

We let $\mathcal{C}_{\mathcal{M}} = \mathcal{P}_{T_{\mathcal{M}}^\perp}(L^*)$. Recall that $E_n = \Sigma_O^n - \Sigma_O^*$ denotes the difference between the sample covariance matrix and the marginal covariance matrix of the observed variables.

PROPOSITION 3.5. *Let γ be in the range specified by (3.1). Suppose that the minimum nonzero singular value σ of $L^* = K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ is such that $\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2}$ (C_5 is defined in Corollary 3.4). Suppose also that the minimum magnitude nonzero entry of S^* is greater than or equal to $\frac{C_6\lambda_n}{\mu(\Omega)}$ (C_6 is defined in Corollary 3.4). Let $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$. Further suppose that*

$$\lambda_n \leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.$$

Then we have that

$$(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}}) = (\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}}).$$

Proof: Note first that the condition on the minimum singular value of L^* in Corollary 3.4 is satisfied. Therefore we proceed with the following two steps:

1. First we can change the non-convex constraint $\text{rank}(L) \leq \text{rank}(L^*)$ to the linear constraint $L \in T(\hat{L}_{\mathcal{M}})$. This is because the lower bound assumed for σ implies that $\hat{L}_{\mathcal{M}}$ is a smooth point of the algebraic variety of matrices with rank less than or equal to $\text{rank}(L^*)$ (from Corollary 3.4). Due to the convexity of all the other constraints and the objective, the optimum of this "linearized" convex program will still be $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$.
2. Next we can again apply Corollary 3.4 (based on the bound on σ) to conclude that the constraint $\|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi^2}$ is *locally inactive* at the point $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$.

Consequently, we have that $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$ can be written as the solution of a *convex program*:

(3.9)
$$(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}}) = \arg\min_{S,L} \text{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \|L\|_*]$$
$$\text{s.t.} \quad S - L \succ 0, \quad S \in \Omega, \quad L \in T_{\mathcal{M}},$$
$$g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n.$$

We now need to argue that the constraint $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(S-S^*,L^*-L)) \le 11\lambda_n$ is also inactive in the convex program (3.9). We proceed by showing that the solution $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ of the convex program (3.8) has the property that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_\mathcal{M}})) < 11\lambda_n$, which concludes the proof of this proposition. We have from Corollary 3.4 that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) \le \frac{\lambda_n\nu}{6(2-\nu)}$. Since $g_\gamma(\mathcal{A}^\dagger E_n) \le \frac{\lambda_n\nu}{6(2-\nu)}$ by assumption, one can verify that

$$\begin{aligned}\frac{8}{\alpha}\left[\lambda_n + g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})\right] &\le \frac{8\lambda_n}{\alpha}\left[1+\frac{\nu}{3(2-\nu)}\right]\\ &= \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}\\ &\le \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.\end{aligned}$$

The last line follows from the assumption on λ_n. We also note that $\|C_{T_\mathcal{M}}\|_2 \le \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$ from Corollary 3.4, which in turn implies that $\|C_{T_\mathcal{M}}\|_2 \le \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}$. Letting $(\Delta_S, \Delta_L) = (S_\Omega - S^*, L^* - \hat{L}_{T_\mathcal{M}})$, we can conclude from Proposition 3.2 that $g_\gamma(\Delta_L, \Delta_S) \le \frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)}$. Next we apply Proposition 3.1 (as $g_\gamma(\Delta_L,\Delta_S) \le \frac{1}{2C_1}$) to conclude that

$$\begin{aligned}g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S+\Delta_L)) &\le \frac{2D\psi C_1^2 g_\gamma(\Delta_S,\Delta_L)^2}{\xi(T)}\\ &\le \frac{2D\psi C_1^2}{\xi(T)}\frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)}\frac{\alpha\xi(T)}{32D\psi C_1^2}\\ &\le \frac{2(3-\nu)\lambda_n}{3(2-\nu)}. \qquad (3.10)\end{aligned}$$

From the optimality conditions of (3.8) one can also check that for $\mathcal{Y} = \Omega \times T_\mathcal{M}$,

$$\begin{aligned}g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S,\Delta_L)) &\le 2\lambda_n + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S+\Delta_L))\\ &\quad + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger E_n)\\ &\le 2[\lambda_n + g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S+\Delta_L))\\ &\quad + g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})]\\ &\le 4\left[\frac{2(3-\nu)\lambda_n}{3(2-\nu)}\right].\end{aligned}$$

Here we used (3.10) in the last inequality, and also that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) \le \frac{\lambda_n\nu}{6(2-\nu)}$ (as noted above from Corollary 3.4) and that $g_\gamma(\mathcal{A}^\dagger E_n) \le \frac{\lambda_n\nu}{6(2-\nu)}$.

Therefore,

$$g_\gamma(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) \quad \leq \quad \frac{16\lambda_n}{3}, \tag{3.11}$$

because $\nu \in (0, \frac{1}{2}]$. Based on Proposition 3.3 in the main paper (the second part), we also have that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) \leq (1-\nu)\frac{16\lambda_n}{3} \leq \frac{16\lambda_n}{3}. \tag{3.12}$$

Summarizing steps (3.11) and (3.12),

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\Delta_S,\Delta_L)) &\leq g_\gamma(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) \\ &\quad + g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_{\mathcal{M}}}) \\ &\leq \frac{16\lambda_n}{3} + \frac{16\lambda_n}{3} + \frac{\lambda_n\nu}{6(2-\nu)} \\ &\leq \frac{32\lambda_n}{3} + \frac{\lambda_n}{18} \\ &< 11\lambda_n. \end{aligned}$$

This concludes the proof of the proposition. □

This proposition has the following important consequence.

COROLLARY 3.6. *Under the assumptions of Proposition 3.5 we have that* $\mathrm{rank}(\hat{L}_{T_{\mathcal{M}}}) = \mathrm{rank}(L^*)$ *and that* $T(\hat{L}_{T_{\mathcal{M}}}) = T_{\mathcal{M}}$. *Moreover,* $\hat{L}_{T_{\mathcal{M}}}$ *actually has the same inertia as* L^*. *We also have that* $\mathrm{sign}(\hat{S}_\Omega) = \mathrm{sign}(S^*)$.

3.5. *Proof of main paper Proposition 5.3 – Removing the tangent-space constraints.* The following lemma provides a simple set of sufficient conditions under which the optimal solution $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ of (3.8) satisfies the optimality conditions of the convex program (3.3) (without the tangent space constraints). This lemma, along with Corollary 3.4 and Corollary 3.6, proves Proposition 5.3 of the main paper.

LEMMA 3.7. *Let* $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ *be the solution to the tangent-space constrained convex program* (3.8). *Suppose that the assumptions of Proposition 3.5 hold. If in addition we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S,\Delta_L))) \leq \frac{\lambda_n\nu}{6(2-\nu)},$$

then $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ *is also the unique optimum of the convex program* (3.3).

Proof: Recall from Corollary 3.6 that the tangent space at $\hat{L}_{T_{\mathcal{M}}}$ is equal to $T_{\mathcal{M}}$. Applying the optimality conditions of the convex program (3.8) at the optimum $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$, we have that there exist Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T_{\mathcal{M}}^\perp} \in T_{\mathcal{M}}^\perp$ such that

$$\begin{aligned} \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} + Q_{\Omega^\perp} &\in -\lambda_n \gamma \partial \|\hat{S}_\Omega\|_1, \\ \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} + Q_{T_{\mathcal{M}}^\perp} &\in \lambda_n \partial \|\hat{L}_{T_{\mathcal{M}}}\|_*. \end{aligned}$$

Restricting these conditions to the space $\mathcal{Y} = \Omega \times T_{\mathcal{M}}$, one can check that

$$\begin{aligned} \mathcal{P}_\Omega[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] &= -\lambda_n \gamma \operatorname{sign}(S^*), \\ \mathcal{P}_{T_{\mathcal{M}}}[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] &= \lambda_n U V^T, \end{aligned}$$

where $\hat{L}_{T_{\mathcal{M}}} = UDV^T$ is a reduced SVD of $\hat{L}_{T_{\mathcal{M}}}$. Setting $Z = [-\lambda_n \gamma \operatorname{sign}(S^*), \lambda_n U V^T]$, we conclude that

$$\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger [\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] = Z, \tag{3.13}$$

with $g_\gamma(Z) = \lambda_n$. It is clear that the optimality condition of the convex program (3.3) (without the tangent-space constraints) on $\mathcal{Y}$ is satisfied. All we need to show is that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger [\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}]) < \lambda_n. \tag{3.14}$$

Rewriting $\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}$ in terms of the error $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_{\mathcal{M}}})$, we have that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} = E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^* \mathcal{A}(\Delta_S, \Delta_L).$$

Restating the condition (3.13) on $\mathcal{Y}$, we have that

$$\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L) = Z + \mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}]. \tag{3.15}$$

(Recall that $\mathcal{C}_{T_{\mathcal{M}}} = \mathcal{P}_{T_{\mathcal{M}}^\perp}(L^*)$.) A sufficient condition to show (3.14) and complete the proof of this lemma is that

$$\begin{aligned} g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) < \lambda_n - g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) \\ - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}]). \end{aligned}$$

We prove this inequality next. Recall from Corollary 3.4 that $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}) \leq$

$\frac{\lambda_n \nu}{6(2-\nu)}$. Therefore, from equation (3.15) we can conclude that

$$\begin{aligned} g_\gamma(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) &\leq \lambda_n + 2(g_\gamma(\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S,\Delta_L)) \\ &\qquad -\mathcal{I}^*\mathcal{C}_{T_{\mathcal{M}}}])) \\ &\leq \lambda_n + 2\left[\frac{3\lambda_n\nu}{6(2-\nu)}\right] \\ &= \frac{2\lambda_n}{2-\nu}. \end{aligned}$$

Here we used the bounds on $g_\gamma(\mathcal{A}^\dagger E_n)$ and on $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S,\Delta_L)))$.

Applying the second part of Proposition 3.3 (main paper), we have that

$$\begin{aligned} g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) &\leq \frac{2\lambda_n(1-\nu)}{2-\nu} \\ &= \lambda_n - \frac{\nu\lambda_n}{2-\nu} \\ &< \lambda_n - \frac{\nu\lambda_n}{2(2-\nu)} \\ &\leq \lambda_n - g_\gamma(\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S,\Delta_L)) \\ &\qquad -\mathcal{I}^*\mathcal{C}_{T_{\mathcal{M}}}]) \\ &\leq \lambda_n - g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S,\Delta_L)) \\ &\qquad -\mathcal{I}^*\mathcal{C}_{T_{\mathcal{M}}}]). \end{aligned}$$

Here the second-to-last inequality follows from the bounds on $g_\gamma(\mathcal{A}^\dagger E_n)$, $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S,\Delta_L)))$, and $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_{\mathcal{M}}})$. This concludes the proof of the lemma. □

3.6. *Proof of main paper Lemma 5.4 – Probabilistic analysis.* All the analysis described so far in this section has been completely deterministic in nature. Here we present the probabilistic component of our proof. Specifically, we study the rate at which the sample covariance matrix converges to the true covariance matrix. The following result from [2] plays a key role in our analysis:

THEOREM 3.8. *Given natural numbers n, p with $p \leq n$, let Γ be a $p \times n$ matrix with i.i.d. Gaussian entries that have zero-mean and variance $\frac{1}{n}$. Then the largest and smallest singular values $s_1(\Gamma)$ and $s_p(\Gamma)$ of Γ are such that*

$$\max\left\{\Pr\left[s_1(\Gamma) \geq 1 + \sqrt{\tfrac{p}{n}} + t\right], \Pr\left[s_p(\Gamma) \leq 1 - \sqrt{\tfrac{p}{n}} - t\right]\right\} \leq \exp\left\{-\tfrac{nt^2}{2}\right\},$$

for any $t > 0$.

Using this result the next lemma provides a probabilistic bound between the sample covariance Σ_O^n formed using n samples and the true covariance Σ_O^* in spectral norm. This result is well-known, and we mainly discuss it here for completeness and also to show explicitly the dependence on $\psi = \|\Sigma_O^*\|_2$.

LEMMA 3.9. *Let $\psi = \|\Sigma_O^*\|_2$. Given any $\delta > 0$ with $\delta \leq 8\psi$, let the number of samples n be such that $n \geq \frac{64p\psi^2}{\delta^2}$. Then we have that*

$$\Pr\left[\|\Sigma_O^n - \Sigma_O^*\|_2 \geq \delta\right] \leq 2\exp\left\{-\tfrac{n\delta^2}{128\psi^2}\right\}.$$

Proof: Since the spectral norm is unitarily invariant, we can assume that Σ_O^* is diagonal without loss of generality. Let $\bar{\Sigma}^n = (\Sigma_O^*)^{-\frac{1}{2}}\Sigma_O^n(\Sigma_O^*)^{-\frac{1}{2}}$, and let $s_1(\bar{\Sigma}^n), s_p(\bar{\Sigma}^n)$ denote the largest/smallest singular values of $\bar{\Sigma}^n$. Note that $\bar{\Sigma}^n$ can be viewed as the sample covariance matrix formed from n independent samples drawn from a model with identity covariance, i.e., $\bar{\Sigma}^n = \Gamma\Gamma^T$ where Γ denotes a $p \times n$ matrix with i.i.d. Gaussian entries that have zero-mean and variance $\frac{1}{n}$. We then have that

$$\begin{aligned}
\Pr\left[\|\Sigma_O^n - \Sigma_O^*\|_2 \geq \delta\right] &\leq \Pr\left[\|\bar{\Sigma}^n - I\|_2 \geq \tfrac{\delta}{\psi}\right] \\
&\leq \Pr\left[s_1(\bar{\Sigma}^n) \geq 1 + \tfrac{\delta}{\psi}\right] + \Pr\left[s_p(\bar{\Sigma}^n) \leq 1 - \tfrac{\delta}{\psi}\right] \\
&= \Pr\left[s_1(\Gamma)^2 \geq 1 + \tfrac{\delta}{\psi}\right] + \Pr\left[s_p(\Gamma)^2 \leq 1 - \tfrac{\delta}{\psi}\right] \\
&\leq \Pr\left[s_1(\Gamma) \geq 1 + \tfrac{\delta}{4\psi}\right] + \Pr\left[s_p(\Gamma) \leq 1 - \tfrac{\delta}{4\psi}\right] \\
&\leq \Pr\left[s_1(\Gamma) \geq 1 + \sqrt{\tfrac{p}{n}} + \tfrac{\delta}{8\psi}\right] \\
&\qquad + \Pr\left[s_p(\Gamma) \leq 1 - \sqrt{\tfrac{p}{n}} - \tfrac{\delta}{8\psi}\right] \\
&\leq 2\exp\left\{-\tfrac{n\delta^2}{128\psi^2}\right\}.
\end{aligned}$$

Here we used the fact that $n \geq \frac{64p\psi^2}{\delta^2}$ in the fourth inequality, and we applied Theorem 3.8 to obtain the final inequality by setting $t = \frac{\delta}{8\psi}$. □

References.

[1] BACH, F. (2008). Consistency of trace norm minimization. *J. Mach. Lear. Res.*. **9** 1019–1048.

[2] DAVIDSON, K. R. AND SZAREK, S.J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces*. **I** 317–366.

[3] KATO, T. (1995). *Perturbation theory for linear operators*. Springer.

[4] ORTEGA, J. M. AND RHEINBOLDT, W. G. (1970). *Iterative solution of nonlinear equations in several variables*. Academic Press.

[5] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Elec. Jour. of Stat.* **4** 935–980.
[6] Rockafellar, R. T. (1996). *Convex Analysis*. Princeton University Press.

V. Chandrasekaran
Department of Computing
and Mathematical Sciences
California Institute of Technology
Pasadena, California 91125
USA
E-mail: venkatc@caltech.edu

P. A. Parrilo
A. S. Willsky
Laboratory for Information
and Decision Systems
Department of Electrical Engineering
and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
USA
E-mail: parrilo@mit.edu
willsky@mit.edu