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 $(X_1, \dots, X_p)$ 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

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

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