Submitted to the Annals of Statistics

DISCUSSION: LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION

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 had 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. (2008). Cai, Liu & 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^*, i.e., $(\Sigma^*_O)^{-1} = S^* - L^*$. 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^*,

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

where Σ^n_O is the sample covariance matrix, $\left\|S\right\|_1 = \sum_{i,j}|s_{ij}|$, and γ and χ_n are regularization tuning parameters. Here $\operatorname{tr}(L)$ is the trace of L. The

*The research was supported in part by NSF Career Award DMS-0645676 and NSF FRG Grant DMS-0854975.

notation $A \succ 0$ means A is positive definite, and $A \succeq 0$ denotes that A is non-negative.

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 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 $\|\cdot\|_{1\to 1}$ denotes the matrix l_1 norm, i.e., $\|S\|_{1\to 1} = \max_{1\le i\le p}\sum_{j=1}^{p}|s_{ij}|$. We assume that S^* is in the following uniformity class,

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

where we allow $s_0\left(p\right)$ and M_p to grow as p and n increase. This uniformity class was considered in Ravikumar et al. (2008) 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, i.e. 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 (2008). We further assume that

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

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, i.e., $\theta = \min_{i,j} |s_{ij}| \mathbf{1}\{s_{ij} \neq 0\}$, and the minimum nonzero eigenvalue of L^* by σ, i.e., $\sigma = \min_{1 \leq i \leq r_0} \lambda_i$. 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) $n \gtrsim s_0^4(p)\sqrt{p/n}$.

However, for sparse graphical model selection without latent variables, either l_1-regularized maximum likelihood approach (see Ravikumar et al. (2008)) 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 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 sub-optimality 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,

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

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 $\hat{S}_1 = \left(\hat{s}_{ij}^1\right)$ be the solution. The CLIME-like estimator $\hat{S} = (\hat{s}_{ij})$ is obtained by symmetrizing $\hat{S}_1$ as follows,

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

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

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

to estimate the support of S^*.

Theorem 1 *Suppose that* $S^* \in \mathcal{U}\left(s_0(p), M_p\right)$,

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

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

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

Hence if the minimum magnitude of nonzero entries $\theta > 18 M_p \tau_n$, *we obtain the sign consistency* $sign\left(\tilde{S}\right) = sign\left(S^*\right)$. *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 the Theorem 7 in Cai, Liu and Luo (2011). The subgaussian condition with spectral norm upper bound M implies that each empirical covariance $\widetilde{\sigma}_{ij}$ satisfies the following large deviation result

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

where C_s, C_2 and ϕ only depends 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

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

Let

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

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

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

which immediately yield

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

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

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

By the matrix l_1 norm assumption, we could obtain that

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

From (8) and (9) we have

$$\left\|\Sigma_O^n (S^* - L^*) - I\right\|_\infty = \left\|\left(\Sigma_O^n - \Sigma_O^*\right)(\Sigma_O^*)^{-1}\right\|_\infty \leq \left\|\Sigma_O^n - \Sigma_O^*\right\|_\infty \left\|(\Sigma_O^*)^{-1}\right\|_{1\to 1} \leq \tau_n,$$

which implies

$$\left\|\Sigma_O^n (S^* - L^*) - \Sigma_O^n \hat{S}_1\right\|_\infty \leq \left\|\Sigma_O^n (S^* - L^*) - I\right\|_\infty + \left\|\Sigma_O^n \hat{S}_1 - I\right\|_\infty \leq 2\tau_n. \tag{10}$$

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

$$\left\|\hat{S}_1\right\|_{1\to 1} \leq \|S^* - L^*\|_{1\to 1} \leq 2M_p, \tag{11}$$

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

$$\begin{aligned}
\left\|\Sigma_O^*\left((S^* - L^*) - \hat{S}_1\right)\right\|_\infty &\leq \left\|\Sigma_O^n (S^* - L^*) - \hat{S}_1\right\|_\infty + \left\|\left(\Sigma_O^* - \Sigma_O^n\right)\left((S^* - L^*) - \hat{S}_1\right)\right\|_\infty \\
&\leq 2\tau_n + \left\|\Sigma_O^n - \Sigma_O^*\right\|_\infty \left\|(S^* - L^*) - \hat{S}_1\right\|_{1\to 1} \\
&\leq 2\tau_n + 4M_p \left\|\Sigma_O^n - \Sigma_O^*\right\|_\infty \leq 4\tau_n.
\end{aligned}$$

Thus we have

$$\left\|(S^* - L^*) - \hat{S}_1\right\|_\infty \leq \left\|(\Sigma_O^*)^{-1}\right\|_{1\to 1} \left\|\Sigma_O^*\left((S^* - L^*) - \hat{S}_1\right)\right\|_\infty \leq 8M_p\tau_n.$$

■

Remark 1 *By the choice of our τ_n and the eigen-decomposition of L^*, the condition $\|L^*\|_\infty \leq M_p\tau_n$ holds when $r_0(p)C_0/p \leq C_1M_p^2\sqrt{(\log p)/n}$, i.e., $p^2\log p \gtrsim nr_0^2(p)M_p^{-4}$. If M_p is slowly increasing (for instance $p^{1/4-\tau}$ 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 the 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 $0 < c < c_1$,*

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

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 $\lambda_i(\tilde{L}) = \lambda_i(\hat{L})1\left\{\lambda_i(\hat{L}) > C_3\sqrt{\frac{p}{n}}\right\}$ 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 $rank\left(L^*\right) = rank\left(\tilde{L}\right)$.

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

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

Then there exists some constant C_3 such that

$$\left\|\hat{L} - L^*\right\| \leq C_3\sqrt{\frac{p}{n}}$$

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

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

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

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

(13)

$$\left\|(\Sigma_O^*)^{-1} - (\Sigma_O^n)^{-1}\right\| \leq \left\|(\Sigma_O^*)^{-1}\right\| \left\|(\Sigma_O^n)^{-1}\right\| \|\Sigma_O^* - \Sigma_O^n\| \leq 2M^2\sqrt{128}M\sqrt{\frac{p}{n}} = 16\sqrt{2}M^3\sqrt{\frac{p}{n}}.$$

From Theorem 1 we know

$$sign\left(\tilde{S}\right) = sign\left(S^*\right), \text{ and } \left\|\tilde{S} - S^*\right\|_\infty \leq 9M_p\tau_n$$

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

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

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

$$\left\|\hat{L} - L^*\right\| \leq \left\|(\Sigma_O^*)^{-1} - (\Sigma_O^n)^{-1}\right\| + \left\|\tilde{S} - S^*\right\| \leq 16\sqrt{2}M^3\sqrt{\frac{p}{n}} + 9C_1M_p^2 s_0\left(p\right)\sqrt{\frac{\log p}{n}} \leq C_3\sqrt{\frac{p}{n}}$$

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 $s_0^3\left(p\right)\sqrt{\frac{p}{n}}$ 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 $0 < c < c_2$,*

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

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 sub-optimal. 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 sub-optimality 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 lights on understanding optimalities in estimation.

REFERENCES

[1] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199-227.

[2] Cai, T. 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.

[3] Cai, T. T., Liu, W. and Zhou, H. H. (2012). Optimal estimation of large sparse precision matrices. Manuscript.

[4] Cai, T. T. (2011). Personal communication.

[5] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. of Comput. Math.* **9** 717-772.

[6] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

[7] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2008). High-dimensional covariance estimation by minimizing $\ell_1$penalized log-determinant divergence. Preprint.

Department of Statistics,
Yale University
New Haven, CT 06511
USA
E-mail: zhao.ren@yale.edu
E-mail: huibin.zhou@yale.edu