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`,

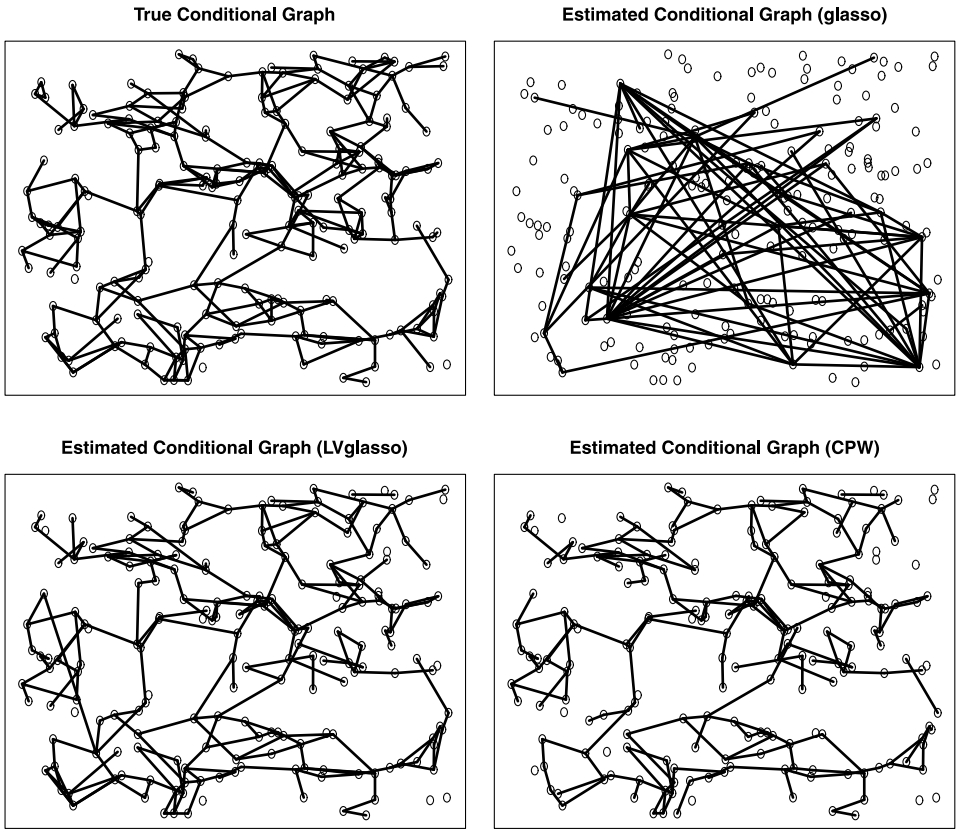

FIG. 1. *True graphical model and its estimates.*

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

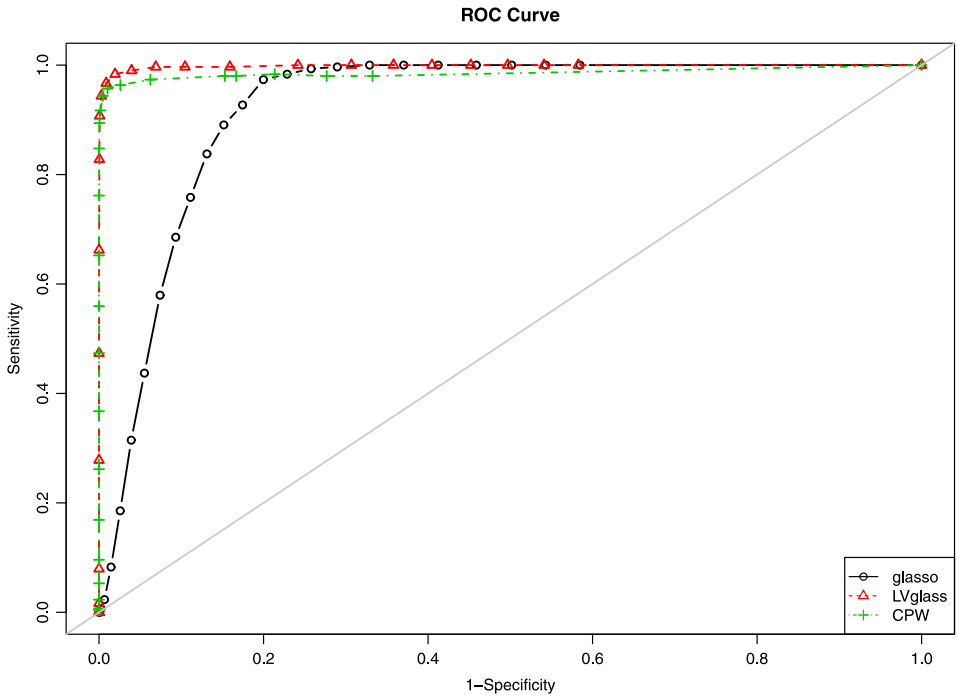

FIG. 2. *Accuracy of reconstructed conditional graphical model.*

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