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

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

$$(2) \qquad \qquad \text{such that } S - L \succ 0, L \succ 0,$$

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

$$(\hat{K}_O, \hat{L}_n) = \mathrm{argmin}_{S,L} -\ell(S - L; \Sigma_O^n) + \lambda_n \|S\|_1$$

$$(3) \qquad \text{such that } S - L \succ 0 \text{ and } L \succ 0 \text{ and } \mathrm{rank}(L) \leq \kappa.$$

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

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

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

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

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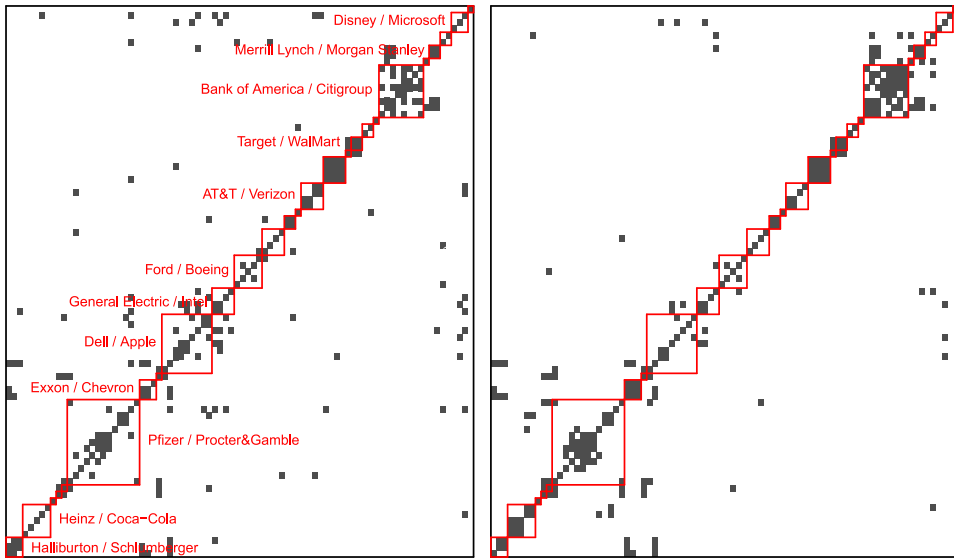
FIG. 1. *The nonzero entries of the concentration matrix* $\hat{K}_O$, *using the proposed procedure* (1) (*left*) *and the imputation method in* (4) (*right*). *Two representative companies are shown for some of the sectors.*

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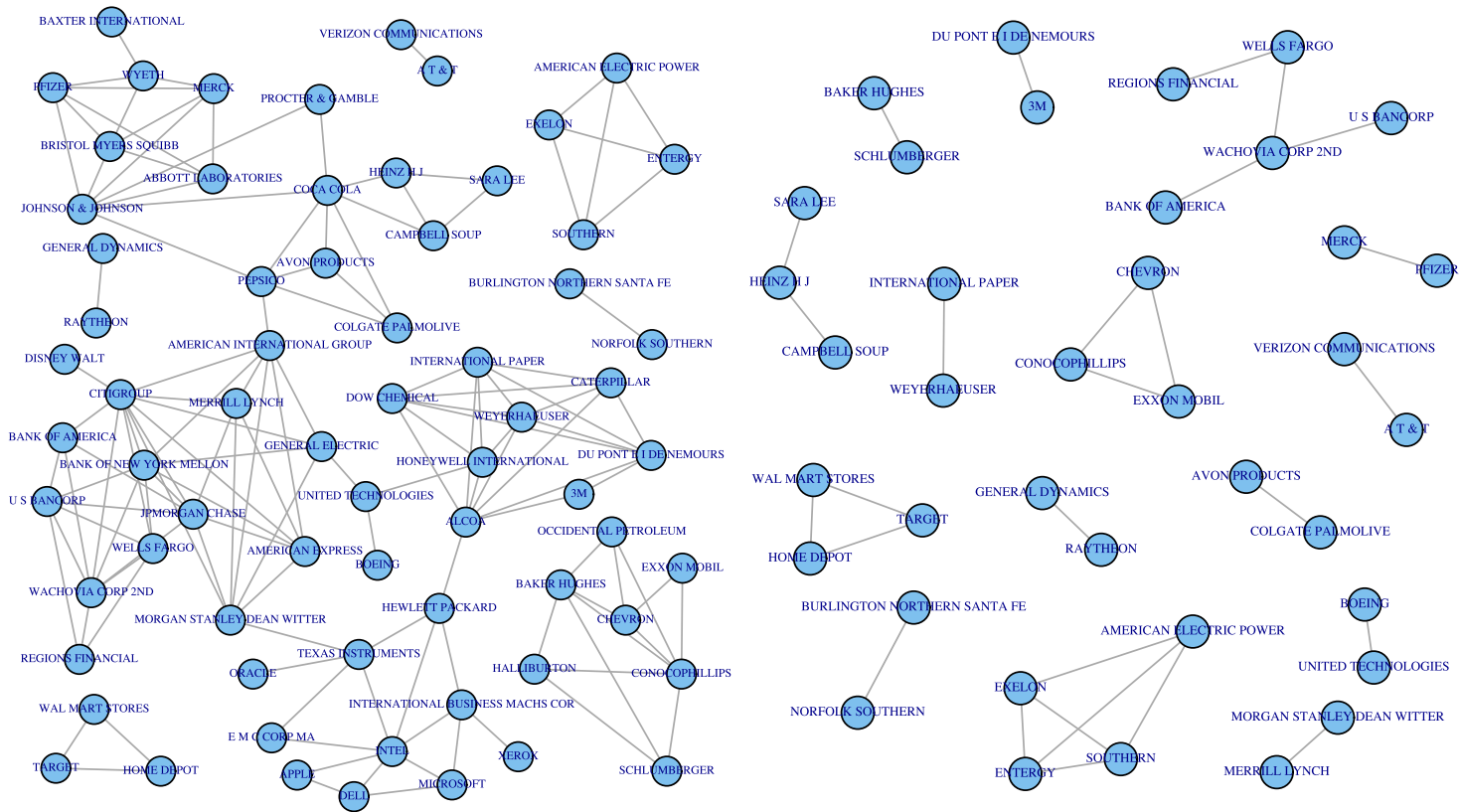
FIG. 2. *Left*: *the graph of the imputation method as in* (4). *Right*: *the graph of the stable edges. In both cases, isolated vertices have been removed from the display.*

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