Chandrasekaran, Venkat; Parrilo, Pablo A.; Willsky, Alan S.
Latent variable graphical model selection via convex optimization. (English) Zbl 1257.62061
Ann. Stat. 40, No. 4, 1935-1967 (2012).

Summary: 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 $\ell_1$ 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.

MSC:
62H12 Estimation in multivariate analysis
90C25 Convex programming
05C90 Applications of graph theory
62F12 Asymptotic properties of parametric estimators

Keywords:
Gaussian graphical models; covariance selection; latent variables; regularization; sparsity; low-rank; algebraic statistics; high-dimensional asymptotics

Software:
PMA; SDPT3; YALMIP

Full Text: DOI arXiv Euclid

References:
[1] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. Ann. Statist. 36 199-227. - Zbl 1132.62040 · doi:10.1214/009053607000000758
[2] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. Ann. Statist. 36 2577-2604. - Zbl 1196.62062 · doi:10.1214/08-AOS600
[3] Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? J. ACM 58 Art. 11, 37. - Zbl 1327.62369
[4] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. Found. Comput. Math. 9 717-772. - Zbl 1219.90124 · doi:10.1007/s10208-009-9045-5
[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. - Zbl 1231.94017 · doi:10.1109/TIT.2005.862083
[6] Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2011). Supplement to “Latent variable graphical model selection via convex optimization.” - Zbl 1288.62085
[7] Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decompo-
