Electronic Journal of Statistics > Vol. 3 (2009) open journal systems

Inferring sparse Gaussian graphical models with latent structure

Christophe Ambroise, *Université d'Évry*
Julien Chiquet, *Université d'Évry*
Catherine Matias, *CNRS*

Abstract

Our concern is selecting the concentration matrix's nonzero coefficients for a sparse Gaussian graphical model in a high-dimensional setting. This corresponds to estimating the graph of conditional dependencies between the variables. We describe a novel framework taking into account a latent structure on the concentration matrix. This latent structure is used to drive a penalty matrix and thus to recover a graphical model with a constrained topology. Our method uses an ℓ_1 penalized likelihood criterion. Inference of the graph of conditional dependencies between the variates and of the hidden variables is performed simultaneously in an iterative em-like algorithm named SIMoNe (Statistical Inference for Modular Networks). Performances are illustrated on synthetic as well as real data, the latter concerning breast cancer. For gene regulation networks, our method can provide a useful insight both on the mutual influence existing between genes, and on the modules existing in the network.

AMS 2000 subject classifications: Primary 62H20, 62J07; secondary 62H30.

Keywords: Gaussian graphical model, Mixture model, ℓ_1-penalization, Model selection, Variational inference, EM algorithm.

Full Text: PDF

Ambroise, Christophe, Chiquet, Julien, Matias, Catherine, Inferring sparse Gaussian graphical models with latent structure, Electronic Journal of Statistics, 3, (2009), 205-238 (electronic). DOI: 10.1214/08-EJS314.

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res., 9:485–516, 2008. MR2417243

Biernacki, C., Celeux, G., and Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell., 22(7):719–725, 2000.

Castelo, R. and Roverato, A. A robust procedure for Gaussian graphical model search from microarray data with p larger than n. J. Mach. Learn. Res., 7:2621–2650, 2006. MR2274453

Chen, S.S., Donoho, D.L., and Saunders, M.A. Atomic decomposition by basis pursuit. SIAM Rev., 43(1):129–159, 2001. MR1854649

Chiquet, J., Smith, A., Grasseau, G., Matias, C., and Ambroise, C. Simone: Statistical inference for modular networks. Bioinformatics, 25(3):417–418, 2009. doi:10.1093/bioinformatics/btn637.

Daudin, J.-J., Picard, F., and Robin, S. A mixture model for random graphs. Stat. Comput., 18(2):173–183, 2008.

Dempster, A.P. Covariance selection. Biometrics, Special Multivariate Issue, 28:157–175, 1972.

Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B, 39(1):1–38, 1977. MR0501537

Dobra, A., Hans, C., Jones, B., Nevins, J.R., Yao, G., and West, M. Sparse graphical models for exploring gene expression data. J. Multivariate Anal., 90(1):196–212, 2004. MR2064941

Donoho, D.L. and Johnstone, I.M. Adapting to unknown smoothness via wavelet shrinkage. J. Amer. Statist. Assoc., 90(432):1200–1224, 1995. MR1379464

Drton, M. and Perlman, M.D. Multiple testing and error control in Gaussian graphical model selection. Statist. Sci., 22:430, 2007. MR2416818

Drton, M. and Perlman, M.D. A SINful approach to Gaussian graphical model selection. J. Statist. Plann. Inference, 138(4):1179–1200, 2008. MR2416875

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. Ann. Statist., 32(2):407–499, 2004. MR2060166

Frank, O. and Harary, F. Cluster inference by using transitivity indices in empirical graphs. J. Amer. Statist. Assoc., 77(380):835–840, 1982. MR0686407

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. Pathwise coordinate optimization. Ann. Appl. Stat., 1(2):302–332, 2007. MR2415737

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432–441, 2008.

Fu, W.J. Penalized regressions: the bridge versus the lasso. J. Comput. Graph. Statist., 7(3):397–416, 1998. MR1646710

Hess, K.R., Anderson, K., Symmans, W.F., Valero, V., Ibrahim, N., Mejia, J.A., Booser, D., Theriault, R.L., Buzdar, U., Dempsey, P.J., Rouzier, R., Sneige, N., Ross, J.S., Vidaurre, T., Gómez, H.L., Hortobagyi, G.N., and Pustzai, L. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. Journal of Clinical Oncology, 24(26):4236–4244, 2006.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. Revealing modular organization in the yeast transcriptional network. Nature Genetics, pages 370–377, July 2002.

Jaakkola, T. Advanced mean field methods: theory and practice, chapter Tutorial on variational approximation methods. Neural Information Processing Series. MIT Press, Cambridge, MA, 2001. MR1863214

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. Experiments in stochastic computation for high-dimensional graphical models. Statist. Sci., 20(4):388–400, 2005. MR2210226

Lauritzen, S.L. Graphical models, volume 17 of Oxford Statistical Science Series. The Clarendon Press Oxford University Press, New York, 1996. MR1419991

Mariadassou, M. and Robin, S. Uncovering latent structure in valued graphs: a variational approach. Technical Report 10, Statistics for Systems Biology, 2007.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. Ann. Statist., 34(3):1436–1462, 2006. MR2278363

Natowicz, R., Incitti, R., Horta, E.G., Charles, B., Guinot, P., Yan, K., Coutant, C., André, F., Pusztai, R., and Rouzier, L. Prediction of the outcome of a preoperative chemotherapy in breast cancer using dna probes that provide information on both complete and incomplete response. BMC Bioinformatics, 9(149), 2008.

Ng, A.Y., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In NIPS 14, 2002.

Nowicki, K. and Snijders, T.A.B. Estimation and prediction for stochastic blockstructures. J. Amer. Statist. Assoc., 96(455):1077–1087, 2001. MR1947255

Osborne, M.R., Presnell, B., and Turlach, B.A. On the LASSO and its dual. J. Comput. Graph. Statist., 9(2):319–337, 2000. MR1822089

Schäfer, J. and Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4(1), 2005.

Snijders, T.A.B. and Nowicki, K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J. Classification, 14(1):75–100, 1997. MR1449742

Tallberg, C. A Bayesian approach to modeling stochastic blockstructures with covariates. Journal of Mathematical Sociology, 29(1):1–23, 2005.

Tibshirani, R. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B, 58(1):267–288, 1996. MR1379242

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl., 109(3):475–494, 2001. MR1835069

Wille, A. and Bühlmann, P. Low-order conditional independence graphs for inferring genetic networks. Statistical Applications in Genetics and Molecular Biology, 5(1), 2006. MR2221304

Wu, T.T. and Lange, K. Coordinate descent algorithms for lasso penalized regression. Ann. Appl. Stat., 2(1):224–244, 2008.

Yuan, M. and Lin, Y. Model selection and estimation in the Gaussian graphical model. Biometrika, 94(1):19–35, 2007. MR2367824

Zanghi, H., Ambroise, C., and Miele, V. Fast online graph clustering via Erdös Rényi mixture. Pattern Recognition, 41(12):3592–3599, 2008.

Zou, H. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc., 101 (476):1418–1429, 2006. MR2279469