Clusterpath Gaussian Graphical Modeling

Daniel J.W. Touw[1], Patrick J.F. Groenen[1], Ines Wilms[2], and Andreas Alfons[1]

[1] Erasmus School of Economics, Erasmus University Rotterdam, (e-mail: touw@ese.eur.nl, groenen@ese.eur.nl, alfons@ese.eur.nl)

[2] Department of Quantitative Economics, Maastricht University, (e-mail: i.wilms@maastrichtuniversity.nl)

Abstract: Gaussian graphical models (GGMs) serve as a means of summarizing conditional dependencies among a set of p variables. Such models are structured as networks, in which nodes represent individual variables and edges denote the presence of conditional dependence between two variables. Estimating GGMs in cases where the sample size n is smaller than the number of variables ($n < p$) can present a challenge. To address this issue, existing estimation methods frequently rely on applying regularization techniques to the edges within the network, with the aim of obtaining a sparse network where many variables are represented as conditionally independent (see, e.g., Cai *et al.*, 2011; Friedman *et al.*, 2008; Meinshausen & Bühlmann, 2006; Peng *et al.*, 2009; Rothman *et al.*, 2008; Yuan, 2010).

Nevertheless, relying solely on edge sparsity does have limitations. First, when the number of variables is substantially larger than the sample size ($n \ll p$), the conditional dependencies between variables may become too weak to detect (Eisenach *et al.*, 2020). Second, sparse GGMs that include many variables can still contain a substantial number of edges, making interpretation difficult (Grechkin *et al.*, 2015). Last, real-world networks often exhibit more complex structures than mere edge sparsity (Heinävaara *et al.*, 2016; Hosseini & Lee, 2016).

To overcome these challenges, node aggregation has emerged as a means to perform dimension reduction in GGMs (see, e.g., Hosseini & Lee, 2016; Pircalabelu & Claeskens, 2020; Tarzanagh & Michailidis, 2018; Wilms & Bien, 2022). For example, instead of estimating the conditional dependencies between all observed variables, one may be interested in identifying the dependencies among a smaller number of clusters that share the same behavior. To achieve this, we propose the clusterpath GGM (CGGM), a model-based convex clustering Gaussian graphical model that automatically clusters groups of variables by means of the penalty structure used in the convex clustering literature (Hocking *et al.*, 2011; Lindsten *et al.*, 2011; Pelckmans *et al.*, 2005).

Keywords: convex clustering, dimension reduction, graphical modeling, regulatization

References

Cai, T., Liu, W., & Luo, X. 2011. A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, **106**(494), 594–607.

Eisenach, C., Bunea, F., Ning, Y., & Dinicu, C. 2020. High-Dimensional Inference for Cluster-Based Graphical Models. *Journal of Machine Learning Research*, **21**(53), 1–55.

Friedman, J., Hastie, T., & Tibshirani, R. 2008. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, **9**(3), 432–441.

Grechkin, M., Fazel, M., Witten, D., & Lee, S.-I. 2015. Pathway Graphical Lasso. *In: Proceedings of the AAAI conference on artificial intelligence*, vol. 29.

Heinävaara, O., Leppä-Aho, J., Corander, J., & Honkela, A. 2016. On the Inconsistency of ℓ_1-penalised Sparse Precision Matrix Estimation. *BMC bioinformatics*, **17**(16), 99–107.

Hocking, T.D., Joulin, A., Bach, D., & Vert, J.-P. 2011. Clusterpath: An Algorithm for Clustering Using Convex Fusion Penalties. *In: The 28th International Conference on Machine Learning*.

Hosseini, M.J., & Lee, S.-I. 2016. Learning Sparse Gaussian Graphical Models with Overlapping Blocks. *Advances in Neural Information Processing Systems*, **29**.

Lindsten, F., Ohlsson, H., & Ljung, L. 2011. *Just Relax and Come Clustering!: A Convexification of K-Means Clustering*. Tech. rept. Department of Electrical Engineering, Linköping University, Linköping, Sweden.

Meinshausen, N., & Bühlmann, P. 2006. High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, **34**(3), 1436–1462.

Pelckmans, K., de Brabanter, J., Suykens, J.A.K., & de Moor, B. 2005. Convex Clustering Shrinkage. *In: PASCAL Workshop on Statistics and Optimization of Clustering Workshop*.

Peng, J., Wang, P., Zhou, N., & Zhu, J. 2009. Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association*, **104**(486), 735–746.

Pircalabelu, E., & Claeskens, G. 2020. Community-Based Group Graphical Lasso. *Journal of Machine Learning Research*, **21**(1), 2406–2437.

Rothman, A.J., Bickel, P.J., Levina, E., & Zhu, J. 2008. Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics*,

2, 494–515.

TARZANAGH, D.A., & MICHAILIDIS, G. 2018. Estimation of Graphical Models through Structured Norm Minimization. *Journal of Machine Learning Research*, **18**(1), 1–48.

WILMS, I., & BIEN, J. 2022. Tree-based Node Aggregation in Sparse Graphical Models. *Journal of Machine Learning Research*, **23**(243), 1–36.

YUAN, M. 2010. High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research*, **11**(79), 2261–2286.