Robust and sparse Gaussian graphical modelling under cell-wise contamination

Tokyo Institute of Technology Shota Katayama
The Institute of Statistical Mathematics Hironori Fujisawa
University of Washington Mathias Drton

Gaussian graphical modelling explores dependences among a collection of variables by inferring a graph that encodes pairwise conditional independences through the support of precision matrix. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$ be a p-dimensional random vector representing a multivariate observation. The conditional independence graph of $\boldsymbol{Y}$ is the undirected graph $G = (V, E)$ whose vertex set $V = \{1, \ldots, p\}$ indexes the individual variables and whose edge set E indicates conditional dependences among them. More precisely, $(i, j) \notin E$ if and only if Y_i and Y_j are conditionally independent given $Y_{V\setminus\{i,j\}} = \{Y_k : k \neq i, j\}$. It is well known that if $\boldsymbol{Y}$ follows a multivariate Gaussian distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(i, j) \notin E$ if and only if $\boldsymbol{\Omega}_{ij} = 0$, where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix. The goal is to detect the support of $\boldsymbol{\Omega}$.

Many modern applications such as bioinformatics and economics feature high dimensional and contaminated data that complicate this task. Traditional robust techniques that down-weight entire observation vectors would often be inappropriate as high dimensional data may feature partial contamination in many observations, i.e., cell-wise contamination. This talk develops a robust estimation method for the large conditional independence graph G from cell-wise contaminated observations.

The node-wise regression proposed by Meinshausen and Bühlmann (2006), graphical lasso (Glasso) by Friedman et al. (2008) and constrained ℓ_1 minimization for inverse matrix estimation (CLIME) by Cai et al. (2011) are commonly used techniques to infer the support of precision matrix. These methods process an estimate of the covariance matrix $\boldsymbol{\Sigma}$. Our strategy is thus simply to apply these procedure using a covariance matrix estimator that is robust against cell-wise contamination via γ-divergence by Fujisawa and Eguchi (2008). The γ-divergence can automatically reduce the impact of contaminations, and it is known to be robust even for highly contaminated data.

References

[1] Cai, T., Liu, W. and Luo, X. (2011), A constrained ℓ_1 minimization approach to sparse precision matrix estimation, *J. Am. Stat. Assoc.*, **106**, 594–607.
[2] Friedman, H., Hastie, T. and Tibshirani, R. (2008), Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, **9**, 432–441.
[3] Fujisawa, H. and Eguchi, S. (2008), Robust parameter estimation with a small bias against heavy contamination, *Journal of Multivariate Analysis*, **99**, 2053–2081.
[4] Meinshausen, N. and Bühlmann, P. (2006), High-dimensional graphs and variable selection with the lasso, *The Annals of Statistics*, **34**, 1436–1462.