Estimation of covariance structure in high-dimensional setting using graphical lasso with application to gene expression data

Annika Tillander

November 19, 2016

Background

A high dimensional setting is when the number of features/variables (p) are larger than the number of observations (n). In this setting the standard maximum likelihood estimated covariance matrix ($\hat{\Sigma}$) is singular and considering the expected value for the inverse of $\hat{\Sigma}$, using the properties of the Gaussian distribution [6]

$$E\left[\hat{\Sigma}^{-1}\right] = \psi\left(p, n\right)\Sigma^{-1}, \quad \psi\left(p, n\right) = \frac{n}{n-p-1} = \frac{1}{1-\frac{p-1}{n}}$$

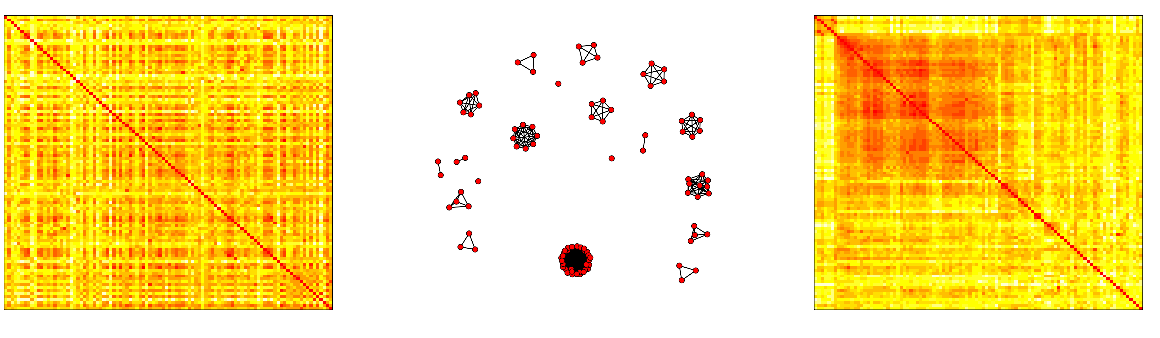
the effect of the relation between p and n for $\hat{\Sigma}^{-1}$ can be clearly seen. As Σ^{-1} is used in many statistical methods e.g. discriminant analysis and regression analysis it is essential to get reliable and non-biased estimates. Hence in recent years many methods for estimating Σ^{-1} in the high dimensional setting have been suggested, one of the most popular methods being the graphical lasso proposed by [10]. This method have been extended and improved by e.g. [4, 5, 3, 8]

Aim

The aim is to present existing graphical lasso methods for estimating covariance structure and to explore and compare these methods using both simulations and application of gene expression data.

Method

To explore the graphical lasso methods different covariance structures should generated according algorithms presented in e.g. [9, 2, 8]. Data will be simulated based on the different structures and then permuted. The performance of the graphical lasso methods to recover and identify true structure will be evaluated using true positive rates and false positive rates with regard to the estimated edges. Plotting the true positive rates against the false positive rates gives the receiver operating characteristic (ROC curve). Application to gene expression data will be to publicly available data such as [1, 7]



References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96(12):6745–6750, June 1999.

[2] M. Aoshima and K. Yata. A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, 66(5):983–1010, 2014.

[3] P. Danaher, P Wang, and D.M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

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

[5] J. Friedman, T. Hastie, and R. Tibshirani. *Graphical lasso- estimation of Gaussian graphical models*, February 2009. Manual to the R-package glasso.

[6] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley-Interscience, 2004.

[7] Y. Pawitan, J. Bjohle, L. Amler, A. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E.T. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P. M. Shaw, J. Smeds, L. Skoog, S. Wedren, and J. Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7(6):953–964, 2005.

[8] K.M. Tan, D. Witten, and A. Shojaie. The cluster graphical lasso for improved estimation of gaussian graphical models. *Computational Statistics Data Analysis*, 85(C):23–36, 2015.

[9] A.S. Wagaman and E. Levina. Discovering sparse covariance structures with the isomap. *Journal of Computational and Graphical Statistics*, 18(3):551–572, September 2009.

[10] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94:19–35, 2007.