Sparse inverse covariance estimation in the supervised classification of high-dimensional data

Tatjana Pavlenko and Anders Björkström

Stockholm University, Sweden

Abstract

The performance accuracy of the sample based supervised classifiers is known to be poor in a high-dimensional setting, i.e. when the data dimension p is comparable to or larger then the sample size n. To overcome this problem, we suggest to construct a sparse estimate of the class inverse covariance matrix, which in a Gaussian case can be obtained as a minimizer of its negative log-likelihood, subject to a Lasso-type penalty on its off-diagonal elements. Our procedure consists of two-stages; we first estimate the sparsity patterns in the class inverse covariance matrix and then form its block-diagonal approximation using Cuthill-McKee reordering algorithm. We then incorporate this technique in the supervised classification framework and investigate the effect of the suggested approximation on the classification accuracy in the growing dimension asymptotics, i.e. when both p and n are allowed to grow. Further, we show that our approach allows for substantial dimensionality reduction while maintaining the misclassification probability at a certain desired level.

Keywords

High-dimensional data, Sparse covariance structure, Lasso, Supervised classification, Misclassification probability.

References

Bühlmann, P. and Rütimann, P. (2009). High dimensional sparse covariance estimation via directed acyclic graphs. *Electronic J. Statist. 3*, 1133–1160.

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

Hall, P., Pittelkow, Y., and Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. R. Stat. Soc. Ser. B Stat. Methodol. 70*, 159–173.