Optimal classification of high-dimensional data with sparsity patterns

Tatjana Pavlenko[1] **and Annika Tillander**[2]
[1] KTH Royal Institute of Technology, Sweden
[2] Karolinska Institutet, Sweden

We consider a two-class classification problem where the number of feature variables, p greatly exceeds the number of observations n. The feature vectors are coming from populations with Gaussian distributions with the precision matrix $\Xi = \Sigma^{-1}$, that is unknown but assumed to be sparse. The discriminative feature variables, also unknown, are sparse and each contributes weakly (i.e. sparse and weak setting) towards the classification decision. It is shown that in this setting, a reliable classification can be achieved by simultaneously using the sparsity of the shift vector and the precision matrix, i.e. sparsity of the graph underlying the data.

By obtaining an accurate estimator of the graph topology, we then use this topology to construct a classifier with group-wise feature selection by adapting the Higher Criticism (HC) thresholding approach (see e.g. Donoho and Jin (2009)). One key component of our analysis the quantifying of a group separation strength; by extending the results by Kawasaki and Seo (2014) to the case of $p > n$, we propose an approach that naturally connected the measure of separation strength and group-wise selection to the classification accuracy. Another key component is the choice of the selection threshold based on the data and its limit behavior in a high-dimensional setting.

We formalize the asymptotic framework for analyzing the sparse and weak model in $p \gg n$ regime; we set up a sequence of classification problems where the sparsity and weakness are liked to p by fixed parameters, and the sample size n, depends on p according to condition $(\log n)/\log p \to \kappa, \quad 0 \leq \kappa < 1$. We then show that along this sequence, our suggested HC thresholding is optimally adaptive in a sense that the asymptotic performance of the classifier based on the selected features can be as good as the asymptotic performance with feature selection by exactly known threshold. Using the two-dimensional diagram with coordinates representing the sparsity and weakness of the model, we empirically determine the sharp separating boundary where this optimal property is attained.

Compared to recent work (e.g Jin (2009)) where Ξ is the identity matrix, our approach is more general since it allows to take into account the dependence structure underlying the data. We thoroughly investigate how the uncertainty in estimating Ξ, given the graph topology, may effect classification accuracy. We also show that by replacing the popular cross-validated threshold with computationally less expensive and simpler HC threshold, the misclassification rate is reduced.

The advantages of the proposed classification with appropriately chosen HC thresholding, are illustrated by using both simulated and real-life data.

Keywords: High-dimensional statistics, feature selection, separation strength, sparse and weak model

References:

Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367**, 44494470.

Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA,* **106**, 88598864.

Kawasaki, T. and Seo T. (2014). A Two Sample Test for Mean Vectors with Unequal Covariance Matrices. *Communication in Statistics–Simulation and Computation*, DOI: 10.1080/03610918.2013.824587.