Handling dependence or not in statistical learning for high-dimensional data

D. Causeur

[a]Irmar, UMR 6625 CNRS, Agrocampus Ouest
65 rue de Saint-Brieuc, CS84215
35042 Rennes cedex, France
david.causeur@agrocampus-ouest.fr

Keywords : Dependence, Genome-Wide Association Studies, Global Testing, Functional Analysis of Variance, High dimension, Statistical learning.

The proper way to handle dependence across features in high-throughput data has raised fundamental discussions with unclear general conclusions or final recommendations. One of the most obvious illustration of this point is the tremendous effort of the statistics research community to address the impact of dependence on the False Discovery Rate (FDR)-controlling method by Benjamini and Hochberg (1995), which was initially designed under an independence assumption. Another famous questioning example is provided by the strikingly good performance of a naïve Bayes procedure ignoring dependence in a comparative study of machine learning methods by Dudoit *et al.* (2002) to predict classes from gene expression data.
Addressing the dependence issue has often consisted in assessing its detrimental impact on the performance of standard methods designed to be optimal under independence, and deduce patches. To be valid for arbitrarily complex dependence patterns, such approaches in which dependence is viewed as a curse can lead to poorly powerful procedures. Therefore, both for machine learning and testing issues, a new generation of methods have emerged, advocating for an ad-hoc handling of dependence consisting in a preliminary whitening of the data (see Ahdesmäki and Strimmer, 2010, Hall and Jin, 2010). However, disentangling the dependent noise and the true association signal is very challenging and decorrelation can then lead to an alteration of the true association signal.
For the purpose of global testing, where the objective is to test for the significance of an association signal between a set of features and a covariate, Arias-Castro *el al.* (2011) suggests that the optimal handling of dependence shall be specific of the pattern of the true association signal, especially through its sparsity rate. The former global testing framework covers a wide scope of applications, such as functional Analysis of Variance (fANOVA) and association tests between a region of the genome formed by contiguous Single Nucleotide Polymorphisms (SNP) and a case/control response variable in Genome Wide Association Studies. Interestingly, in the two former fields of applications, many popular methods are just based on simple aggregation of pointwise test statistics ignoring their dependence.
The talk will start by a selection of short stories with confusing conclusions about the proper way to handle dependence. After a tentative clarification, I will introduce two methods for global testing in which whitening is adapted both to the pattern of dependence across pointwise test statistics and to the pattern of the true signal. The performance of the two testing methods will be illustrated by applications to significance analysis of ElectroEncephalogram curves in Event-Related Potentials (ERP) designs and by SNPset approaches of Genome Wide Association Studies for genetic epidemiology issues. We also discuss the applications of the former general principles to prediction in high-dimension.

joint work with Florian Hébert and Mathieu Emily

References

[1] Arias-Castro, E., Candès, E.J. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics.* **39**, 5, 2533–2556.

[2] Ahdesmäki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Annals of Applied Statistics.* **4**. 503–519.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, **57** (1). 289–300.

[4] Dudoit, S., Fridlyand, J. and Speed, T.P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data *Journal of the American Statistical Association.* **97** (457), 77–87

[5] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics.* **38**, 3, 1686–1732.