

Accounting for a Factor Structure in High-Dimensional Data to Improve Multiple Testing Procedures

C. Friguet, M. Kloareg and D. Causeur

IRMAR, UMR CNRS 6625
Laboratoire de Mathématiques Appliquées
Agrocampus Rennes
65 rue de Saint Briec
CS 84215 - 35042 Rennes cedex
chloe.friguet@agrocampus-rennes.fr

Abstract. Among the topics that are widely discussed in the literature on statistical methodologies for microarray data, the impact of dependence between the genes on the properties of multiple testing procedures has known an increasing interest recently. This can be explained by the need for geneticists to control strongly the actual rate of false discoveries, which is not guaranteed by the current methods of simultaneous testing: although they are shown to control mean type-I error rates, they are also known to suffer from high instability in the presence of correlation between the gene expressions. Moreover, getting knowledge on the dependency structure in microarray data has also become a challenging task to give more insight into the complex gene regulation network involved in the biological system that yield the observed expressions. However, in many papers, the high dimensionality of the data reduces the scope of investigation for possible structures of dependencies to simple ones assuming for instance independence between blocks of gene expressions and simple within-block correlation patterns.

In our presentation, we assume a factor-analysis model for covariance, which is mostly used by social scientists or psychometricians as a dimension reduction technique. The $m \times m$ - covariance matrix Σ can be expressed as $\Sigma = B.B' + \Psi$ where B is a $m \times q$ - matrix of loadings associated to the common variability and Ψ a $m \times m$ - diagonal matrix for variable specific variance also called "uniqueness". In this context, the data Y can be decomposed as $Y = \mu + F.B' + E$. F can be interpreted as a small number q of latent factors $\{F_1, \dots, F_q\}$ that sum up the common information shared by all the variables, as in the Factor Analysis scheme. Within this framework, we study the stability of multiple testing procedures on high-dimensional data considering conditional independence structures. Close form expressions for the variance of error rates are derived. A new test statistic is also defined, taking into account the factor structure. It is shown to decrease distinctly the correlation between the test statistics and reduce the variance of error rates. The proportion of non discoveries is noticeably lessened with respect to usual methods.