Multivariate data analysis

Theoretical properties and applications in genetics

Piotr Szulc

Abstract

We consider the so-called "large p, small n" problem, when we want to find a connection between a variable Y and a vector of variables $\mathbf{X} = (X_1, \ldots, X_p)$, and p can be much larger than number of observation n. Furthermore, we assume that only small number of variables from X, p_0, is truly connected with Y (sparsity). This assumption partially results from the fact that the construction of the model for $p_0 > n$ is usually impossible (and we assume that n is relatively small), and also because models with large number of variables not necessarily can be useful. That's why this assumption does not have to be treated as limiting.

One of the ways to combine Y and $\mathbf{X}$, is to fit the right regression model. In our situation it is connected with the necessity to choose a certain small subset of explaining variables. Well known criteria used to select such subset include the Akaike Information Criterion (Akaike 1974) or the Bayesian Information Criterion (Schwarz 1978), which are examples of classic statistical tools that were not designed for the "large p, small n" problem, and actually are not suitable in such situations. However, it turns out that analyzing the origin of the Bayesian informational criterion, we are able to take into account our new demand and get modified version of that criterion, which can be successfully used in the situation under consideration.

The "large p, small n" problem is especially present in genetics, where we have information about a very large number of genotypes, much larger than the number of individuals. Therefore we begin this dissertation by providing basic genetic facts and the simplest methods used in localizing genes. We also present the idea of so-called effective number of tests, which is the basis for the construction of statistical tools described in further part. We give theorems about consistency of mBIC and mBIC2 under the assumption that both n and p as well as p_0 go to infinity (with proofs) supported by simulation results, as well as we describe the package for R environment to choose regression model for very large sets of data (Szulc 2016), which use modi-

fied Bayesian criteria among others. That part of the dissertation include collected and expanded results coming from the papers: Szulc (2012, 2014, 2015) and Szulc, Bogdan (2012). We also present modification of mBIC2, thanks to which it can be used in a situation when we have additional predictors (of different nature than basic ones). This situation occurs for example in so-called admixture mapping. That part constitutes organized results from the papers: Szulc (2013) and Szulc, Bogdan, Frommlet, Tang (2016).

Piotr Szulc