Extreme Value Analysis of Huge Datasets

Tail Estimation Methods in High-Throughput
Screening and Bioinformatics

Dmitrii Zholud

Thesis for the degree of Doctor of Philosophy, to be defended in public on
Thursday, November 3, 2011 at 10:15, Room Pascal,
Department of Mathematical Sciences, Chalmers Tvärgata 3, Göteborg.
Faculty opponent is Professor Johan Segers, Institut de statistique,
biostatistique et sciences actuarielles, Université catholique de Louvain.
The defence will be conducted in English.

CHALMERS | GÖTEBORGS UNIVERSITET

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology
and Göteborg University
Göteborg, Sweden 2011

Abstract

This thesis presents results in Extreme Value Theory with applications to High-Throughput Screening and Bioinformatics. The methods described here, however, are applicable to statistical analysis of huge datasets in general. The main results are covered in four papers.

The first paper develops novel methods to handle false rejections in High-Throughput Screening experiments where testing is done at extreme significance levels, with low degrees of freedom, and when the true null distribution may differ from the theoretical one. We introduce efficient and accurate estimators of False Discovery Rate and related quantities, and provide methods of estimation of the true null distribution resulting from data preprocessing, as well as techniques to compare it with the theoretical null distribution. Extreme Value Statistics provides a natural analysis tool: a simple polynomial model for the tail of the distribution of p-values. We exhibit the properties of the estimators of the parameters of the model, and point to model checking tools, both for independent and dependent data. The methods are tried out on two large scale genomic studies and on an fMRI brain scan experiment.

The second paper gives a strict mathematical basis for the above methods. We present asymptotic formulas for the distribution tails of probably the most commonly used statistical tests under non-normality, dependence, and non-homogeneity, and derive bounds on the absolute and relative errors of the approximations.

In papers three and four we study high-level excursions of the Shepp statistic for the Wiener process and for a Gaussian random walk. The application areas include finance and insurance, and sequence alignment scoring and database searches in Bioinformatics.

Keywords: Extreme Value Statistics, High Throughput Screening, HTS, Bioinformatics, analysis of huge datasets, quality control, correction of theoretical p-values, comparison of pre-processing methods, SmartTail, estimation of False Discovery Rates, test power, distribution tail, high level excursions, quantile estimation, multiple testing, Student $t-$test, Welch statistic, small sample sizes, $F-$test, Wiener process, Gaussian random walk, Shepp statistic, limit theorems, exotic options.