Gene expression

dslice: an R package for nonparametric testing of associations with application in QTL and gene set analysis

Chao Ye¹, Bo Jiang², Xuegong Zhang¹,³,* and Jun S. Liu²,⁴,*

¹MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIS, Department of Automation, Tsinghua University, Beijing 100084, China, ²Department of Statistics, Harvard University, Cambridge, MA 02138, USA, ³School of Life Sciences, Tsinghua University, Beijing 100084, China and ⁴Center of Statistics, Tsinghua University, Beijing 100084, China

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on July 3, 2014; revised on December 28, 2014; accepted on January 12, 2015

Abstract

Summary: Many statistical problems in bioinformatics and genetics can be formulated as the testing of associations between a categorical variable and a continuous variable. A dynamic slicing method was proposed for non-parametric dependence testing, which has been demonstrated to have higher powers compared with traditional methods such as Kolmogorov–Smirnov test. We introduce an R package dslice to facilitate the use of dynamic slicing method in bioinformatic applications such as quantitative trait loci study and gene set enrichment analysis.

Availability and implementation: dslice is implemented in Rcpp and available in the Comprehensive R Archive Network. The package is distributed under the GNU General Public License (version 2 or later).

Contact: zhangxg@tsinghua.edu.cn or jliu@stat.harvard.edu.

1 Introduction

Often in biological studies, we need to test whether the underlying distributions of two or more populations are different from each other on the basis of independent samples from these populations. The K-sample test problem is equivalent to testing whether the value of an observation is independent of its label indicator (i.e. which sample it comes from). Traditional methods such as the t-test and the rank-sum test can only test differences in means or medians. On the other hand, classic omnibus testing methods such as Kolmogorov–Smirnov test have limited power in detecting differences between distributions of different types. Jiang et al. (2014) proposed an omnibus K-sample testing method based on regularized likelihood-ratio test and dynamic slicing (here we refer to the test statistic as ‘DS-statistic’) and demonstrated its statistical power compared with some existing well-known methods through extensive simulation studies. Here, we introduce an R package dslice that implements this novel dynamic slicing (‘DS’) method for nonparametric tests and makes it a versatile statistical tool for applications in quantitative trait loci (QTL) study and gene set enrichment analysis (GSEA).

2 Implementation

Core functions for testing dependence in dslice are implemented in Cpp and are integrated in R through the Rcpp package (Eddelbuettel and François, 2011). R package dslice contains functions for omnibus K-sample hypothesis testing (ds_test), illustrations of optimal slicing scheme (slice_show) and gene set analysis (ds_gsa) on dataset downloaded from GSEA website (http://www.broadinstitute.org/gsea/index.jsp).

3 Examples

3.1 QTL study

The QTLs analysis attempts to detect relationship between genetic variation (single-nucleotide polymorphism) and continuous variable
of phenotypes (human height or gene expression levels). Traditional analyzing methods are mainly based on the researcher’s foresight about the agnostic mechanism of QTLs, such as linear marginal effect or interaction effect (Aschard et al., 2013). Dynamic slicing method views the QTL problem as dependence test between a categorical variable (genotype) and a continuous variable (phenotype). It has the advantage of revealing nonlinear associations free of assumptions on underlying mechanism.

In view of power study on DS method in detecting non-linear effects (Jiang et al., 2014), we apply the ds.test function in dslice package on a mouse QTL dataset by Burke et al. (2012) with 558 binary genetic markers and two phenotypes. We removed observations with missing phenotype values and set the penalty parameter $\lambda = 1.0$ in ds.test. Then, we randomly shuffled the phenotypes of individuals to generate empirical $P$ values of DS-statistic. By controlling a false discovery rate (FDR, Benjamini and Hochberg, 1995) of 0.05, we identified 6 QTLs and 15 QTLs associated with mouse femur length and vertebra length, respectively (Table 1). Figure 1 illustrates the optimal slicing schemes generated by dslice for mouse vertebra length QTLs rs4222738 and rs13477864. Each panel shows the relative proportion of genotypes in the corresponding slice.

### 3.2 Gene Set Analysis

Subramanian et al. (2005) introduced GSEA to the aggregate effect of genes in unit of ‘gene set’. Specifically, GSEA attempts to determine whether the distribution of biological phenotypes are different between genes in a gene set and the other genes, which can be formulated as a non-parametric two-sample testing problem. However, Goeman and Bühlmann (2007) discussed two different null hypotheses in gene set analysis: competitive null, which is a two-sample test comparing genes in the gene set and genes not in the set, and self-contained null, which is a one-sample test of differential expression on a set of genes. Here, we focus on the application of dynamic slicing method on a well-studied dataset P53 NCI-60 (Ackermann and Strimmer, 2009; Efron and Tibshirani, 2007; Subramanian et al., 2005) in testing competitive null but note that our method can also be used in one-sample test for self-contained null (see function ds_.1 in dslice package for details).

This dataset assays 10 100 gene expression levels and consists of 17 normal samples and 33 samples with mutated p53. The C2 gene set contains 308 predefined functional gene sets (with gene set size between 15 and 500). The dataset is available on the GSEA website. We set the penalty parameter $\lambda = 1.0$ and obtained the empirical $P$ values by randomly assigning sample labels. Table 2 lists significant gene sets reported by DS under a FDR cutoff of 0.05. In comparison, the GSA method by Efron and Tibshirani (2007) missed the gene set P53 signalling under the same cutoff, and all the significant gene sets reported by GSA have FDR values larger than 0.01.

### 4 Discussion

Testing of associations between a categorical variable and a continuous response is a frequently encountered statistical problem in bioinformatics. R package dslice implements a recently proposed nonparametric dependency detection method. Although we use QTL and gene set analysis as examples to demonstrate the use of dslice, it can be applied to other biological problems such as protein.
binding inference from nucleosome occupancy (Meyer et al., 2011) and binding profile differences between protein mutants (Gisselbrecht et al., 2013), and we anticipate it to be of great use in future bioinformatic studies.

**Funding**

This work was supported, in part, by the National Basic Research Program of China (2012CB316304), Shenzhen Special Fund for Strategic Emerging Industry grant ZD201111080127A and the National Science Foundation Grants DMS-1120368 and IIS-1017967.

**Conflict of Interest**: none declared.

**References**


