Kernel-based testing and their applications to single-cell data analysis

Anthony Ozier-Lafontaine*[1,2], Bertrand Michel , and Franck Picard

[1]Laboratoire de Mathématiques Jean Leray – Centre National de la Recherche Scientifique : UMR6629, Nantes université - UFR des Sciences et des Techniques – France
[2]Nantes Université - École Centrale de Nantes – Nantes Université – France

Résumé

Single-cell RNA sequencing (scRNAseq) is a high-throughput technology quantifying gene expression at the single-cell level, for hundreds to thousands of observations (i.e. cells) and tens of thousands of variables (i.e. genes). New methodological challenges arose to fully exploit the potentialities of these complex data. A major statistical challenge in scRNAseq data analysis is to distinguish biological information from technical noise in order to compare conditions or tissues. The principal approach to do this is Differential Expression Analysis (DEA), which is basically gene-wise univariate two-sample tests. However, DEA misses the multivariate aspects of scRNAseq data, which carries information about gene dependencies and gene regulatory networks, and does not inform about the global similarity of the compared datasets. Thus there is a need to develop specific multivariate two-sample tests to test for any global difference between two scRNAseq datasets.
I will present a kernel based two-sample test called truncated Kernel Fisher Discriminant Analysis (tKFDA) test that can be used for DEA as well as for multivariate two-sample tests.
The tKFDA test have been introduced in the seminal work of Harchaoui et al. and we propose its first ready-to-use implementation dedicated to scRNAseq data. The tKFDA test can be interpreted as a regularized version of the famous Maximum Mean Discrepancy (MMD) test. This regularization is based on a Kernel Principal Component Analysis (KPCA) like dimension reduction, allowing to distinguish insightful information from technical noise. Moreover, the regularization is designed such that the associated empirical statistic has an asymptotic chi-square distribution, which allows for asymptotic testing.

Besides reaching state of the art performances in DEA, our approach has a geometrical interpretation of finding the optimal nonlinear transformation of the data that discriminates between the two compared conditions. Visualization tools dedicated to highlighting the main differences can be derived from this interpretation and be the basis of a cell-wise investigation to identify the cells or populations of cells that are the more different between the two conditions.

Mots-Clés: kernel testing, single, cell

*Intervenant