Kernel-based perturbation testing for single-cell data

Franck Picard[1][0000−0001−8084−5481]

Laboratory of Biology and Modelling of the Cell,
Université de Lyon, Ecole Normale Supérieure de Lyon, CNRS, UMR5239,
Université Claude Bernard Lyon 1,Lyon, France
franck.picard@ens-lyon.fr

The convergence of single-cell biology and high-throughput sequencing technologies has made it possible to generate high-dimensional molecular portraits of individual cells. As a result, the field of single-cell data science has emerged, focusing on the development of statistical and computational tools to analyze the complexity of these datasets. A central challenge in single-cell analysis is the statistical comparison of datasets across conditions or tissues. Comparative analysis is crucial for distinguishing biological variation from technical noise and for detecting meaningful expression or regulatory changes. Traditionally, differential expression analysis (DEA) has been addressed through gene-wise two-sample testing frameworks, but standard approaches often fail to capture the full distributional complexity of single-cell data. Single-cell perturbation experiments add another layer of complexity and opportunity. These experiments compare perturbed and control cells, allowing for fine-grained analysis of population responses to stimuli or drugs. However, the resulting data often exhibit complex, non-linear structures that standard linear methods cannot adequately capture. This calls for methods capable of performing non-linear comparisons and modeling intricate data geometries.

To address these challenges, we propose a kernel-based framework for the differential analysis of single-cell data [1], including perturbation studies [2]. Kernel methods, widely used in machine learning, allow for non-linear operations by embedding data into a reproducing kernel Hilbert space (RKHS)—an infinite-dimensional space in which linear operations correspond to non-linear operations in the input space. This enables powerful and flexible statistical comparisons that are robust to the high dimensionality, sparsity, and noise typical of single-cell data. Our approach captures the distance between distributions by measuring the separation of mean embeddings with respect to biological variability. For perturbation experiments, we further develop a linear model in RKHS and introduce a formal hypothesis testing of high-dimensional data structures in complex experimental designs. We implement our methodology in a software package, kaov, available in Python. The package includes tools for visualization and interpretation, and offers a user-friendly interface modeled after popular analysis libraries.

As single-cell datasets continue to grow in size and complexity, traditional statistical methods struggle to uncover subtle but biologically important patterns. Kernel-based testing offers a robust, distribution-free, and non-linear alternative that is well suited to the unique challenges of single-cell data. By accounting for complex dependencies and offering interpretable results, it enables deeper insights into gene regulation, cell fate, and disease progression, paving the way for biomarker discovery and precision medicine.

References

1. Ozier-Lafontaine, A.: Kernel-based testing for single-cell differential analysis. Genome Biology **25**(1), 114 (2024)
2. Ozier-Lafontaine, A.: Extending Kernel Testing to general designs. Arxiv **2405.13799**, (2024)