Master Internship offer
Sensitivity Analysis for kernel testing with applications to Single Cell Genomics

Keywords: kernel-based testing, sensitivity analysis, high dimensional statistics, single cell genomics

Context and challenges

Single Cell Genomics. Single-Cell transcriptomics now allows the quantification of gene expression at the scale of individual cells, encoded in count matrices countaining thousands observations (cells) and tens of thousands features (gene expression values). The analysis of such data requires new methodological frameworks, dedicated to their complexity and size. A major challenge consists in comparing the distribution of gene expression between conditions (ex: control vs treatment).

Kernel testing. In recent years, there have been significant advancements in statistical hypothesis testing and one important breakthrough was the combination of kernel methods with statistical testing [4]. Kernel methods, widely used in machine learning, are based on embedding data into a feature space, enabling non-linear data analysis in the original input space. Kernel testing has been introduced by the seminal work of Gretton et al. [3], who proposed to combine the power of kernel methods with the framework of statistical testing. The pioneering work of Gretton et al. was further extended by Moulines et al. [7] to account for covariance structures in the feature space, both within and between groups in two-sample testing. In a recent contribution [8], we demonstrated the power and versatility of this kernel-based approach for the statistical comparison of gene expression distributions. Our method is available as a Python/R package, `ktest` [1], and has already been used to differentiate cell populations based on single-cell expression data and to explore cancer cell heterogeneity using single-cell epigenomic data.

Sensitivity Analysis for kernel testing. The next step now, which is the subject of this master internship, is to enhance the interpretability of our kernel-based testing method, improving our understanding of the complex gene expression contributions that drive differences between cell populations. Since our test relies on a kernel-based non-linear classifier, it lacks interpretability—specifically, identifying the genes contributing to transcriptomic differences is challenging. This highlights a contemporary challenge in using non-linear machine learning methods for scientific discovery: both kernel testing and state-of-the-art AI techniques improve data representation but at the cost of increased complexity and reduced interpretability. To address this challenge, the fields of Explainable AI [6] and sensitivity analysis [2] seek to develop methodologies that can identify the key factors contributing to machine learning models' decisions.

[1] `https://github.com/LMJL-Alea/ktest`

Project description

The objective of this internship is to exploit sensitivity analysis techniques to assess the influence of individual genes and their interactions on the classifier, which serves as a proxy for hypothesis testing. Based on sensitivity analysis and explainable AI, we will then develop a specific interpretability framework dedicated to non-linear testing based on classification.

One first approach to describe and visualize some characteristics of the kernel-based test is to use the feature map's derivative [1]. This strategy aligns with Derivation-Based Global Sensitivity measures [5], also known as sensitivity maps [9]. The aim is to develop a sensitivity map for the kernel-based differential analysis test. Our approach involves a local mathematical perturbation of the gene expression data, allowing us to study the derivatives of the classification rules and the impact of small changes in the expression of individual genes. As a result, we expect that influential genes identified in this context will be more biologically relevant than differentially expressed genes tested marginally. This approach can also be extended to gene sets, allowing for the study of the influence of specific gene regulatory networks on transcriptomic variations between conditions. The statistical developments will focus on the sampling and convergence properties of these sensitivity maps to provide statistical confidence in the measurement of the influence of individual genes or gene sets

A more advanced approach is developing a sensitivity analysis method tailored to non-linear testing of single-cell transcriptomic data. While most sensitivity analysis methods are designed for black-box models producing continuous real-valued outputs, there is no established framework for assessing feature influence in statistical testing. We will explore a novel approach by treating the p-value from differential transcriptome analysis as the key quantity of interest in SA. We will study how perturbations in the input gene expression distributions affect the risk of false positive rejection.

The methods developed in this project will be applied to single-cell data from our collaborators in biology.

Relevance of the identified gene set signatures and differentially expressed genes obtained through our new methodological approaches will be tested on biological samples obtained from the training data sets.

The candidate will work at **Laboratoire de Mathématiques Jean Leray** in **Nantes** and will be supervised by **Bertrand Michel** (EC Nantes) and **Franck Picard** (CNRS, ENS Lyon).

The candidate will benefit from the **Ai4scmed** project that gathers an interdisciplinary consortium dedicated to single cell genomics, with experts in machine learning, statistics and biostatiscs.

The ideal candidate holds a Master's degree in **Statistics** with a strong background in mathematics and an **interest in pursuing a PhD** on this project.

Contact: Bertrand Michel (Email: bertrand.michel@ec-nantes.fr) and Franck Picard (Email: franck.picard@ens-lyon.fr).

References

[1] M. Briscik, M.-A. Dillies, and S. Déjean. Improvement of variables interpretability in kernel PCA. arXiv preprint arXiv:2303.16682, 2023.

[2] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. Basics and trends in sensitivity analysis: Theory and practice in R. SIAM, 2021.

[3] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. Journal of Machine Learning Research, 13(25):723–773, 2012.

[4] Z. Harchaoui, F. Bach, O. Cappe, and E. Moulines. Kernel-Based Methods for Hypothesis Testing: A Unified View. IEEE Signal Processing Magazine, 30(4):87–97, July 2013.

[5] S. Kucherenko and B. Iooss. Derivative-Based Global Sensitivity Measures, pages 1–24. Springer International Publishing, Cham, 2016.

[6] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1):18, 2020.

[7] E. Moulines, F. Bach, and Z. Harchaoui. Testing for homogeneity with kernel fisher discriminant analysis. In Advances in Neural Information Processing Systems, volume 20, 2007.

[8] A. Ozier-Lafontaine, C. Fourneaux, G. Durif, P. Arsenteva, C. Vallot, O. Gandrillon, S. Gonin-Giraud, B. Michel, and F. Picard. Kernel-based testing for single-cell differential analysis. Genome Biology, 25(1), May 2024.

[9] P. M. Rasmussen, K. H. Madsen, T. E. Lund, and L. K. Hansen. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. NeuroImage, 55(3):1120–1131, 2011.