Post-clustering inference: practical limitations for application to scRNA-seq data-analysis

Boris Hejblum*†[1]

[1]Inserm Bordeaux Population Health research center U1219 – Bordeaux Population Health U1219
Inserm - Université de Bordeaux – France

Abstract

Post-clustering inference in single-cell RNA sequencing (scRNA-seq) analysis presents significant challenges in controlling Type I error during differential expression analysis. Among the current state-of-the-art, data fission represents a promising approach that aims to split data into two independent parts. However, it relies on strong parametric assumptions of non-mixture distributions that are inherently violated in clustered data. To address this limitation, we introduced conditional data fission, an extension designed to decompose each mixture component into two independent parts. However, we demonstrate it requires prior knowledge of the clustering structure to ensure valid post-clustering inference. We theoretically quantify how biases in estimating component-specific scale parameters lead to deviations from independence, and thus to inflated Type I error rates. Given that mixture components are typically unknown in practice, our results underscore the fundamental difficulty of applying data fission in real-world settings for scRNA-seq data analysis.

Keywords: Next generation hypothesis testing

*Speaker
†Corresponding author: boris.hejblum@u-bordeaux.fr