Eugene Katsevich
Assistant Professor
Department of Statistics and Data Science
The Wharton School, University of Pennsylvania
ekatsevi@wharton.upenn.edu
https://ekatsevi.github.io/

August 26, 2024

Dear Drs. Doss, Witten and Yao,

We are very pleased to submit the attached manuscript titled "Computationally efficient and statistically accurate conditional independence testing with spaCRT" to the *Journal of the Royal Statistical Society: Series B.*

Our motivation stems from the analysis of the data produced by *single-cell CRISPR screens,* a cutting-edge biological assay. A central task in analyzing these data is to test for associations between CRISPR perturbations and the expressions of genes, while accounting for confounding technical variation. In an earlier work, we formulated this as a *conditional independence testing* problem and noted that the data involved are highly sparse (Barry et al., 2021 and 2024). This leads to the following dilemma: Asymptotic conditional independence tests like the GCM test (Shah and Peters, 2020) are statistically inaccurate (e.g. have inflated Type-I error), whereas resampling-based tests like the dCRT (Candès et al. 2018, Liu et al., 2022) are more accurate but more computationally expensive (especially since hundreds of thousands of tests can be carried out in a single experiment). Our state-of-the-art sceptre software for single-cell CRISPR screen analysis takes the latter approach.

To address these challenges, we propose a new conditional independence testing methodology called the *spaCRT,* which enjoys the statistical accuracy of resampling-based procedures and the computational speed of asymptotic procedures. This methodology is based on a *saddlepoint approximation (SPA)* to the tail probability of the resampling distribution of the dCRT test statistic, which gives an accurate approximation to the dCRT p-value. While SPAs have been applied to simpler resampling-based procedures like permutation tests, the spaCRT is the first application of the SPA to any conditional independence test. To justify this methodology, we rigorously prove that the relative error in our approximation vanishes asymptotically, so that approximation accuracy is maintained even for very small p-values. We confirm the excellent statistical and computational performance of the spaCRT on both simulated and real single-cell CRISPR screen data. On the real data, we found that the spaCRT matches the statistical performance of the dCRT while accelerating it by a factor of about 250.

We view the significance of our work as two-fold. First, it offers an immediately actionable statistical solution to a real computational problem in a high-impact genomics application area. We plan to implement the spaCRT in the sceptre software to offer users excellent statistical performance at a fraction of the computational cost. Second, by integrating classical SPA techniques with modern CRT methodology, our work dispels the impression that SPAs are a classical tool suitable only simple inferential tasks, showcasing instead how this powerful approach can be applied in a much broader variety of settings. Given the novelty and significance our contributions, we believe this manuscript would be a great fit at the *Journal of the Royal Statistical Society: Series B.* We thank you for your time and consideration.

Sincerely,

Eugene Katsevich

Ziang Niu, Jyotishka Ray Choudhury, and Eugene Katsevich