THE UNIVERSITY OF CHICAGO

COMPUTATIONAL AND APPLIED MATHEMATICS COLLOQUIUM

BORIS LANDA

Department of Mathematics

Yale University

When Sinkhorn-Knopp Meets Marchenko-Pastur: Simple Diagonal Scaling Reveals the Rank of a Count Matrix

THURSDAY, April 29, 2021 at 4:15pm (Central)

Via ZOOM

A ubiquitous step in many data analysis pipelines is to project the data on a small number of principal components. Yet, a longstanding question is how to choose the number of these components. Random matrix theory provides useful insights into this question by assuming a "signal+noise" model, where the goal is to estimate the rank of the underlying signal matrix. If the noise is homoskedastic, i.e., the noise variances are identical across all entries, the spectrum of the noise matrix can be explained by the celebrated Marchenko-Pastur (MP) law, which provides a natural threshold on the eigenvalues for rank estimation. However, in many practical situations the noise is not homoskedastic. One such example is single-cell RNA sequencing (scRNA-seq), where the entries in the data matrix are count random variables, e.g., Poisson, in which case the noise variances can be arbitrary. In this case, the MP law does not generally hold, and the spectrum of the noise is not available in advance, posing a major challenge for rank estimation.

In this talk, I will present a procedure termed biwhitening for a Poisson data model, which enforces the MP law to hold regardless of the Poisson parameters. This procedure operates by scaling the rows and columns of the data matrix simultaneously so that the average noise variance in each row and each column is precisely 1. Even though the noise variances are unknown (as they depend on the underlying Poisson parameters), the scaling factors required for biwhitening can be estimated consistently by applying the well-known Sinkhorn-Knopp algorithm directly to the observed matrix. I will also discuss generalizations of our approach to other discrete distributions, and demonstrate biwhitening on both simulated and experimental count data. In particular, when applying biwhitening to real scRNA-seq data we get an almost perfect fit to the MP law with a variance slightly larger than 1, which is explained by a generalized Poisson model with a small overdispersion.

Organizer:

Daniel Sanz-Alonso, Department of Statistics, sanzalonso@uchicago.edu

CAM Colloquium URL: https://cam.uchicago.edu/events/cam-colloquium/

For further information and inquiries about building access for persons with disabilities, please contact Zellencia Harris, zellenciah@uchicago.edu. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/subscribe/cam_colloquium/.