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

Abstract:

An important task in modern data analysis is to determine the rank of a corrupted data matrix. Random matrix theory provides useful insights into this task 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 admits the celebrated Marchenko-Pastur (MP) law, providing a simple approach for rank estimation. However, in many practical situations, such as in single-cell RNA sequencing (scRNA-seq), the noise is far from being homoskedastic. In this talk, focusing on a Poisson data model, I will present a simple procedure termed biwhitening, which enforces the MP law to hold by appropriate diagonal scaling. Aside from the Poisson distribution, this procedure is extended to families of distributions with quadratic variance functions. I will demonstrate this procedure on both simulated and experimental data, showcasing accurate rank estimation in simulations and excellent fits to the MP law for real scRNA-seq datasets.