Nonlinear Data Integration through Joint Embedding and Data Alignment

Prof. Rong Ma
(Harvard T.H. Chan School of Public Health, Boston, USA)

December 11, 2026

Modern data analysis increasingly requires identifying shared latent structure across multiple high-dimensional datasets. A primary approach for learning such a shared structure is to jointly embed different datasets into a common low-dimensional space. In this talk, I will present two recent and closely related kernel methods for nonlinear joint embedding of high-dimensional datasets. The first method builds on ideas from entropic optimal transport, while the second is based on duo-landmark integral operators. Both are principled approaches for aligning and jointly embedding multiple datasets, supported by rigorous theoretical guarantees. We show that for a pair of noisy, high-dimensional datasets, these methods consistently recover the shared underlying manifold structure while mitigating dataset-specific nuisance structures. I will provide an intuitive geometric explanation of each methodology, along with the theoretical foundations that justify their performance. I will demonstrate their effectiveness in analyzing a single-cell multiomic dataset for human brain cells, which uncovers interesting cell-type-specific interactions between transcription and epigenomic regulation. This talk is based on recent work in collaboration with Xiucai Ding, Boris Landa, and Yuval Kluger.