A085 · Stacked SVD or SVD stacked? A Random Matrix Theory perspective on data integration

Presenter: Tavor Baharav — Broad Institute

Authors: Tavor Z. Baharav, Phillip B. Nicol, Rafael A. Irizarry, Rong Ma

Session: Day 1 · Thu Oct 1, 2026 · 2:15–4:15 PM

Modern data analysis increasingly requires identifying shared latent structure across multiple high-dimensional datasets. A commonly used model assumes that the data matrices are noisy observations of low-rank matrices with a shared singular subspace. In this case, two primary methods have emerged for estimating this shared structure, which vary in how they integrate information across datasets. The first approach, termed , concatenates all the datasets, and then performs a singular value decomposition (SVD). The second approach, termed , first performs an SVD separately for each dataset, then aggregates the top singular vectors across these datasets, and finally computes a consensus amongst them. While these methods are widely used, they have not been rigorously studied in the proportional asymptotic regime, which is of great practical relevance in today's world of increasing data size and dimensionality. Consequently, it remains unclear when one method should be preferred over another. In this work, we derive exact expressions for the asymptotic performance and phase transitions of these two methods and develop optimal weighting schemes to further improve both methods. Our analysis reveals that while neither method uniformly dominates the other in the unweighted case, optimally weighted dominates optimally weighted when the low rank signal is fully shared across the datasets. We then extend our analysis to handle multiple, partially shared components per dataset and demonstrate that can yield improved performance without requiring estimation of subspace alignment. Finally, we provide practical algorithms for estimating optimal weights from data, offering theoretical guidance for method selection in practical data integration problems. Extensive numerical simulations and semi-synthetic experiments on genomic data corroborate our theoretical findings.

STACKED SVD OR SVD STACKED? A RANDOM MATRIX THEORY PERSPECTIVE ON DATA INTEGRATION

By Tavor Z. Baharav[1,2,*,a], Phillip B. Nicol[2,3,*,c], Rafael A. Irizarry[2,3,b] and Rong Ma[1,2,3,†,d]

[1]Eric and Wendy Schmidt Center, Broad Institute, Cambridge, MA, 02142, USA, [a]baharav@broadinstitute.org

[2]Department of Data Science, Dana Farber Cancer Institute, Boston, MA, 02115, USA, [b]rafael__irizarry@dfci.harvard.edu

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA, [c]phillipnicol@g.harvard.edu; [d]rongma@hsph.harvard.edu

Modern data analysis increasingly requires identifying shared latent structure across multiple high-dimensional datasets. A commonly used model assumes that the data matrices are noisy observations of low-rank matrices with a shared singular subspace. In this case, two primary methods have emerged for estimating this shared structure, which vary in how they integrate information across datasets. The first approach, termed `Stack-SVD`, concatenates all the datasets, and then performs a singular value decomposition (SVD). The second approach, termed `SVD-Stack`, first performs an SVD separately for each dataset, then aggregates the top singular vectors across these datasets, and finally computes a consensus amongst them. While these methods are widely used, they have not been rigorously studied in the proportional asymptotic regime, which is of great practical relevance in today's world of increasing data size and dimensionality. Consequently, it remains unclear when one method should be preferred over another. In this work, we derive exact expressions for the asymptotic performance and phase transitions of these two methods and develop optimal weighting schemes to further improve both methods. Our analysis reveals that while neither method uniformly dominates the other in the unweighted case, optimally weighted `Stack-SVD` dominates optimally weighted `SVD-Stack` when the low rank signal is fully shared across the datasets. We then extend our analysis to handle multiple, partially shared components per dataset and demonstrate that `SVD-Stack` can yield improved performance without requiring estimation of subspace alignment. Finally, we provide practical algorithms for estimating optimal weights from data, offering theoretical guidance for method selection in practical data integration problems. Extensive numerical simulations and semi-synthetic experiments on genomic data corroborate our theoretical findings.

*Equal contributions, listed alphabetically.

†Corresponding author.

MSC2020 subject classifications: Primary 15A18, 62H25.

Keywords and phrases: Data integration, Random matrix theory, Spectral methods.