ISSN 1973-2201 STATISTICA, anno LXXXIV, n. 2, 2024
DOI: https://doi.org/10.60923/issn.1973-2201/22034

DISCUSSION OF THE PAPER "CONNECTING MODEL-BASED AND MODEL-FREE APPROACHES TO LINEAR LEAST SQUARES REGRESSION" BY LUTZ DÜMBGEN AND LAURIE DAVIES (2024)

Jan Hannig [1]
Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill

1. Discussion

Dümbgen and Davies (2024) propose a fascinating new framework that leverages random rotations and exact Beta distributions to tackle the problem of separating signal from noise in matrix data. The presentation highlights several key aspects of their methodology and offers intriguing connections to both classical and modern statistical tools. A central theme of the paper is the notion of using random rotations as am alternative to the more traditional permutation-based tests. These random rotations can reveal how subspace alignments (or angles) between an estimated signal and the truth behave under the null hypothesis, leading to exact or near-exact Beta distributions for certain traditional test statistics. This is particularly remarkable because permutation tests are themselves widely respected for their exactness under exchangeability assumptions and yet the proposed tests based on rotations often provide a compelling, theoretically elegant alternative.

Beyond that, the emphasis on angles — e.g., the use of principal angles or directional angles between estimated and true signals — underscores an under-appreciated aspect of statistical inference. While singular value decomposition (SVD), principal components, and related methods are mainstays, explicit angle-based inference has not received commensurate attention. The paper argues that systematically studying these angles can offer clearer insights into how accurately estimated subspaces align with the true underlying structure of data. This is especially relevant in large-scale, high-dimensional problems in bioinformatics and machine learning, where one seeks to distinguish genuine patterns (low-rank signal) from noise.

[1] Corresponding Author. E-mail: jan.hannig@unc.edu

My question is whether the work can be extended to improve on common approaches in matrix denoising and low-rank approximation (e.g., via shrinkage estimators of singular values). I can see at least two current approaches to this problem that may benefit from having a closed form null distribution of a test statistics. The first is the *Jackstraw* procedure Chung and Storey (2014), which tests the significance of traits (rows) in the SVD using permutations of selected raws of data. Jackstraw can be computationally expensive especially when the analyzed dataset is large, as is the case in genomics applications. However, there is no inherent need to favor permutation over rotation and rotation based version of Jackstraw could be very useful.

The second problem in this space is estimating the perturbation angle. Assume that we have $\mathbf{X} = \mathbf{A} + \mathbf{E}$, where $\mathbf{X}$ is observed data matrix, $\mathbf{A}$ is a low rank signal matrix and $\mathbf{E}$ is a full rank noise matrix. Let $\hat{\mathbf{A}}$ be estimator of $\mathbf{A}$ obtained by truncated SVD. While the problem of estimating singular values of $\mathbf{A}$ is well studied and solved by shrinkage Gavish and Donoho (2014) the angular deviation of $\hat{\mathbf{A}}$ from $\mathbf{A}$ is less understood. The main theoretical contributions are due to Cai and Zhang (2018); Wedin (1972). However these results can be quite conservative for use in statistical uncertainty quantification. An angle bootstrap approach was proposed by Prothero *et al.* (2024) where one simulates new data by randomly rotating the estimated signal components while preserving the structure of the residuals. While this somewhat mitigates the heavy computational demands of full permutation tests and does not rely on normality assumptions, it still can be quite computationally costly. The question is, whether there is a closed formed formula similar to the magic beta formula in this case.

2. Conclusion

Dümbgen and Davies (2024) provide a fresh perspective on signal extraction and uncertainty quantification in high-dimensional data matrices. Their emphasis on geometry — specifically angles and subspaces — offers both theoretical depth and computational practicality. The interplay between exactness (through Beta distributions) and resampling (via random rotations) opens up a rich avenue of research. Not only do these ideas complement classical hypothesis testing methods, they also suggest new frontiers in statistical inference where geometry, randomization, and computational efficiency can jointly yield more transparent insights.

I look forward to seeing how this line of work evolves. In particular, it will be interesting to track whether the angle-based bootstrap machinery spurs further methodological developments, or finds immediate traction in applied fields like bioinformatics, image processing, and latent space modeling. In any case, this paper underscores just how pivotal angles and rotations can be in statistical problems — a notion that will likely spark significant and sustained interest in the years to come.

References

T. Cai, A. Zhang (2018). *Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics*. The Annals of Statistics, 46, no. 1, pp. 60–80.

N. Chung, J. Storey (2014). *Statistical significance of variables driving systematic variation in high-dimensional data*. Bioinformatics, 31, no. 4, pp. 545–554.

L. Dümbgen, L. Davies (2024). *Connecting model-based and model-free approaches to linear least squares regression*. Statistica, 84, no. 2, pp. 65–81.

M. Gavish, D. Donoho (2014). *The optimal hard threshold for singular values is* $4/\sqrt{3}$. IEEE Transactions on Information Theory, 60, no. 8, pp. 5040–5053.

J. Prothero, M. Jiang, J. Hannig, Q. Tran-Dinh, A. Ackerman, J. Marron (2024). *Data integration via analysis of subspaces (divas)*. TEST, 33, pp. 633–674.

P.-A. Wedin (1972). *Perturbation bounds in connection with singular value decomposition*. BIT Numerical Mathematics, 12, pp. 99–111.