ALL SPARSE PCA MODELS ARE WRONG, BUT SOME ARE USEFUL

J. Camacho[(1)], A. K. Smilde[(2)], E. Saccenti[(3)], J. A. Westerhuis[(2)]

(1) School of Computer Science and Telecommunications, University of Granada, Spain
(2) Biosystems Data Analysis, University of Amsterdam, the Netherlands
(3) Laboratory of Systems and Synthetic Biology, Wageningen University & Research, the Netherlands

josecamacho@ugr.es, a.k.smilde@uva.nl, edoardo.saccenti@wur.nl, j.a.westerhuis@uva.nl

Topic: Sparse Principal Component Analysis (sPCA) is a popular matrix factorization approach based on Principal Component Analysis (PCA) that combines variance maximization and sparsity with the ultimate goal of improving data interpretation.

Focus: When moving from PCA to sPCA, there are a number of implications that the practitioner needs to be aware of. We study some of these implications both theoretically and numerically using simulations for several state-of-the-art sPCA algorithms.

Results: We show that all sPCA methods considered have significant drawbacks that make models either **lie** (model spurious variance) or **hide information** (leave variance unmodeled).

Computation of Scores, Residuals and Explained Variance

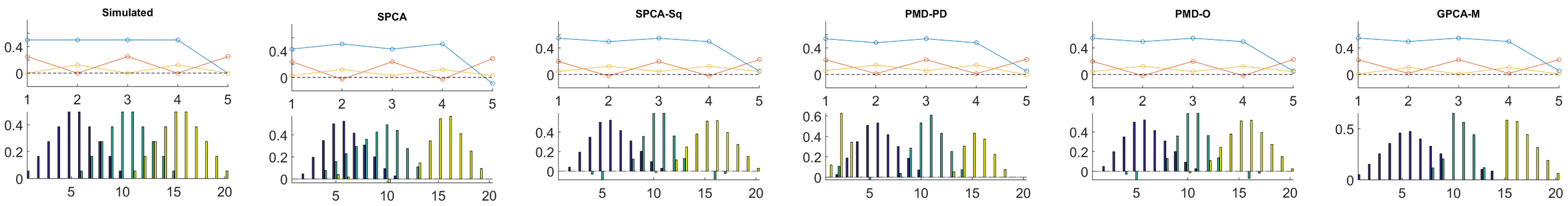
Disagreement in the literature on how to compute scores, residuals and explained variance. These quantities are relevant: scores and residuals are central for data visualization and interpretation. Explained variance is useful for comparison among sPCA variants, and with other modeling approaches.

Given that loadings and scores in sPCA can be correlated, estimates should follow these expressions:

$$\hat{\mathbf{T}} = \mathbf{X}\hat{\mathbf{P}}(\hat{\mathbf{P}}^\top\hat{\mathbf{P}})^+ \qquad \mathbf{E} = \mathbf{X} - \hat{\mathbf{T}}\hat{\mathbf{P}}^\top \qquad \mathrm{Var(sPCA)} = \mathrm{tr}(\hat{\mathbf{P}}\hat{\mathbf{T}}^\top\hat{\mathbf{T}}\hat{\mathbf{P}}^\top)$$

Simulation: Noise-free spectra

Algorithms: The sPCA algorithm by Zou et al. [1] (SPCA) and the sequential implementation in the SPASM toolbox [2] (SPCA-Sq), the PMD algorithm by Witten et al. [3] with projection (PMD-PD) and orthogonalized deflation (PMDO) and the GPCA algorithm by [4] Camacho et al.(GPCA-M)



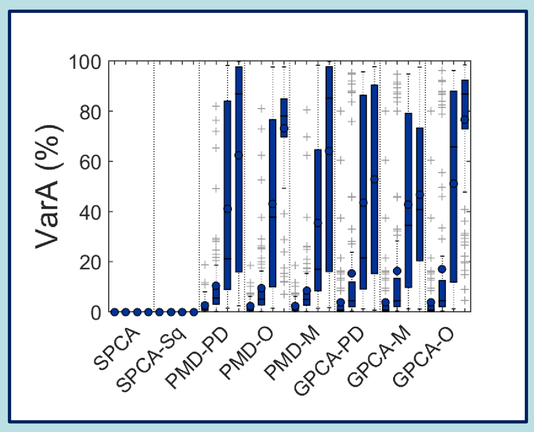
None of the algorithms provides an accurate approximation of noise-free sparse data: Why?

sPCA can lie: sPCA loadings are outside the data row-space. As a result, residuals of deflation-based sPCA (PMD and GPCA above) are also outside this space: this can introduce artifacts in higher order components.

Statistic: *VarA* represents the percentage of artifacts in a component, and ideally should be as close as possible to 0.

$$\mathbf{O}_i = (\mathbf{I} - \hat{\mathbf{X}}_0^\top(\hat{\mathbf{X}}_0^\top)^+)\hat{\mathbf{X}}_i^\top$$

$$\mathrm{VarA} = 100 \cdot \mathrm{tr}(\mathbf{O}_i^\top\mathbf{O}_i)/\mathrm{tr}(\hat{\mathbf{X}}_i^\top\hat{\mathbf{X}}_i)$$



Monte Carlo Simulation

PMD and GPCA components can contain a high percentage of artifacts in a component when they share variables

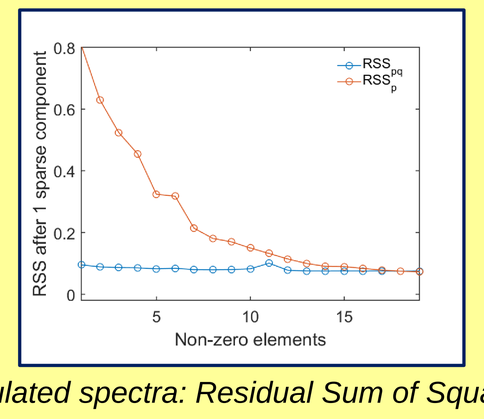
sPCA can hide information: in sPCA based on weights (**P**) and loadings (**Q**), like SPCA and SPCA-Sq above, we can compute the residuals using **P** or **P** and **Q**, and the result is quite different.

$$\mathbf{X} = \mathbf{X}\hat{\mathbf{P}}(\hat{\mathbf{P}}^\top\hat{\mathbf{P}})^+\hat{\mathbf{P}}^\top + \mathbf{E}_p$$

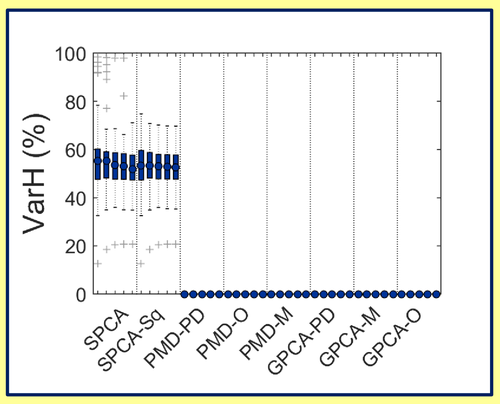
$$\mathbf{X} = \mathbf{X}\hat{\mathbf{P}}(\hat{\mathbf{Q}}^\top\hat{\mathbf{P}})^+\hat{\mathbf{Q}}^\top + \mathbf{E}_{pq}$$

Statistic: *VarH* represents the percentage of hidden variance in a multi-component model, and ideally should be as close as possible to 0.

$$\mathrm{VarH} = 100 \cdot (\mathrm{tr}(\mathbf{E}_p^\top\mathbf{E}_p) - \mathrm{tr}(\mathbf{E}_{pq}^\top\mathbf{E}_{pq}))/\mathrm{tr}(\mathbf{X}^\top\mathbf{X})$$



Simulated spectra: Residual Sum of Squares



Monte Carlo Simulation

SPCA components can hide a large percentage of variance when they go very sparse

Different sPCA algorithms can have different performance. We provide statistics to detect problems (VarH or VarA).
Future research is in due to solve the limitations found.

[1] Zou Hui, Hastie Trevor, Tibshirani Robert. Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics. 2006;15:265-286.
[2] Sjöstrand Karl, Clemmensen Line, Larsen Rasmus, Einarsson Gudmundur, Ersbøll Bjarne. SpaSM: A MATLAB Toolbox for Sparse Statistical Modeling Journal of Statistical Software, Articles. 2018;84:1-37.
[3] Witten Daniela M., Tibshirani Robert, Hastie Trevor. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics. 2009;10:515-534.
[4] Camacho Jose, Rodríguez-Gómez Rafael A., Saccenti Edoardo. Group-wise Principal Component Analysis for Exploratory Data Analysis, Journal of Computational and Graphical Statistics. 2017;26:501-512.

ACKNOWLEDGEMENTS: This work is partly supported by the Spanish Ministry of Economy and Competitiveness and FEDER funds, project TIN2017-83494-R and the “Plan Propio” of University of Granada.