Revisiting PCA with SPOQ norm-ratio sparsity penalties

1 General project information

- Supervisors:
  - ☐ Émilie Chouzenoux (INRIA OPIS, CentraleSupélec, University Paris-Saclay, emilie.chouzenoux@inria.fr)
  - ☐ Laurent Duval (Digital Science and Technology Dpt., IFP Energies nouvelles, laurent.duval@ifpen.fr)
- Type: Master's project
- Location: INRIA OPIS, University Paris-Saclay

2 Scientific context

Principal component analysis (PCA) is a workhorse in linear dimensionality reduction [Jol02]. It is widely applied in exploratory data analysis, visualization, data preprocessing). Principal components are usually linear combinations of all input variables. For high-dimension data, this may involve input variables that contribute very little to the understanding. Finding the few directions in space that explain best observations is desirable. Sparse PCA overcomes this disadvantage by finding linear combinations that contain just a few input variables, by adding sparsity constraints [ZX18, CR24]. One of such is formulated (cf. lasso) with the help of an absolute norm penalty/regularization. In [MBPS10], one designs this matrix factorization problem as:

$$\min \|X - D\alpha\|_F^2 + \lambda \|\alpha\|_{1,1}$$

where: $X = [x_1, \ldots, x_n]$ is the matrix of data vectors; D is a square matrix from a suitable basis set, $\|\cdots\|_F$ denotes the Frobenius norm; $\|\cdots\|_{1,1}$ denotes the sum of the magnitude of matrix coefficients.

A penalty such as $\|\cdots\|_{1,1}$ is 1-homogeneous ($\|a\alpha\|_{1,1} = |a|\|\alpha\|_{1,1}$). This may only weakly emulate the sheer count of non-zero entries of a matrix, that would be scale-invariant or 0-homogeneous. Recently, the SOOT/SPOQ family of penalties has been developed, as smooth emulations to the scale-invariant $\ell_p(\cdot)/\ell_q(\cdot)$ norm ratios. The latter had been used for a while, as stopping-criteria, penalties or "continuous" sparsity count estimators [HR09]. They have been used successfully for the restoration/deconvolution/source separation of sparse signals [RPD+15, CCDP20, ZCD23]. The goal of this subject is to:

- investigate potential derivations using SOOT/SPOT penalties,
- implement the algorithmic work-flow in a scientific toolkit (*eg* scikit-learn),
- benchmark it against competing methods.

References

[CCDP20] Afef Cherni, Emilie Chouzenoux, Laurent Duval, and Jean-Christophe Pesquet. SPOQ ℓ_p-over-ℓ_q regularization for sparse signal recovery applied to mass spectrometry. *IEEE Trans. Signal Process.*, 68:6070–6084, 2020.

[CR24] Fan Chen and Karl Rohe. A new basis for sparse principal component analysis. *J. Comp. Graph. Stat.)*, 33(2):421–434, 2024.

[HR09] N. Hurley and S. Rickard. Comparing measures of sparsity. *IEEE Trans. Inform. Theory*, 55(10):4723–4741, Oct. 2009.

[Jol02] I. T. Jolliffe. *Principal component analysis*. Springer Series in Statistics, 2nd edition, 2002.

[MBPS10] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, 2010.

[RPD+15] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet. Euclid in a taxicab: Sparse blind deconvolution with smoothed ℓ_1/ℓ_2 regularization. *IEEE Signal Process. Lett.*, 22(5):539–543, May 2015.

[ZCD23] Paul Zheng, Emilie Chouzenoux, and Laurent Duval. PENDANTSS: PEnalized Norm-ratios Disentangling Additive Noise, Trend and Sparse Spikes. *IEEE Signal Process. Lett.*, 30:215–219, 2023.

[ZX18] Hui Zou and Lingzhou Xue. A selective overview of sparse principal component analysis. *Proc. IEEE*, 106(8):1311–1320, August 2018.