Thesis topic

Sparse Principal Component Analysis in Word Vector Spaces: Can there be Guarantees?

Principal component analysis (PCA) is a well-established technique on its own for data analysis and data compression, but also an important ingredient in complexer algorithms. Especially in high-dimensional situations but also for reasons of interpretability in lower dimensions, it is helpful if a PCA is sparse, i.e., the coefficients are mostly zero. Several algorithms exist that (heuristically) enforce sparsity (e.g., Guerra-Urzola et al., 2021). Singular value decomposition (SVD) is a related problem, that is often used to find the PCA.

In this thesis, the candidate will study the potential of approximation algorithms as alternatives to the established heuristic or computationally demanding methods. The key idea is to randomly sample rows and columns of the original data and compute a sparse approximation to the SVD from these. With high probability, the sparse approximation is similar to the rang k approximation of the SVD. An efficient implementation is an intermediate step and it is helpful to be or become familiar with matrix algebra in R, Python or C++ for this purpose. Next to a comparison to other algorithms, an application to word vector spaces (latent semantic analysis) could result as a part of this thesis.

This thesis will be supervised by Prof. Dr. Philipp Doebler and Dr. Alexander Munteanu.

Contact: doebler@statistik.tu-dortmund and alexander.munteanu@tu-dortmund.de

Date: 27.5.2022

References

Guerra-Urzola, R., Van Deun, K., Vera, J. C., & Sijtsma, K. (2021). A guide for sparse PCA: Model comparison and applications. Psychometrika, *86*(4), 893-919.