Robust and Sparse Canonical Correlation Analysis

Christophe Croux[1] **and Ines Wilms**[1]
[1] Faculty of Economics and Business, KU Leuven, Belgium

Canonical correlation analysis (CCA) describes the associations between two sets of variables by maximizing the correlation between linear combinations of the variables in each data set. However, in high-dimensional settings, where the number of variables exceeds the sample size, or when the variables are highly correlated, traditional CCA is no longer appropriate. This talk discusses a method for Robust Sparse CCA. Sparse estimation produces linear combinations of only a subset of variables from each data set. More precisely, some of the elements of the canonical vectors will be estimated as exactly zero. As such, the interpretability of the canonical variates is increased. We also robustify the method such that it can cope with outliers in the data.

We convert the CCA problem into an alternating regression framework. To obtain sparse canonical vectors, we add an L_1 penalty on the coefficient estimates to the Least Squares estimator. The lasso, however, is not robust to outliers. The method can be easily robustified by using the sparse Least Trimmed Squares estimator.

We illustrate the good performance of the Robust Sparse CCA method in several simulation studies. In addition, the Robust Sparse CCA method is applied to a genomic data set.

Keywords: Canonical Correlation Analysis, Robust regression, Sparsity.

References:

Branco, J.A., Croux, C., Filzmoser, P., Oliveira, M.R. (2005), Robust canonical correlations: A comparative study *Computational Statistics. 20*, 203–229.

Alfons,A., Croux, C., Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets *The Annals of Applied Statistics. 7*, 226–248.

Witten, D.M., Tibshirani, R., Hastie T. (2009). A penalized matrix decomposition, with applications to sparse principal component and canonical correlation analysis *Biostatistics. 10*, 515–534.