Sparse PCA and LSA

**Latent Semantic Analytics (LSA)**

LSA is a technique in NLP of analysing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

It assumes that words that are close in meaning will occur in similar pieces of text.

It uses SVD to condense a large matrix of word by context data into a much smaller matrix for analysis.

Cosine similarity is used for the resulting vectors.

**Sparse PCA**

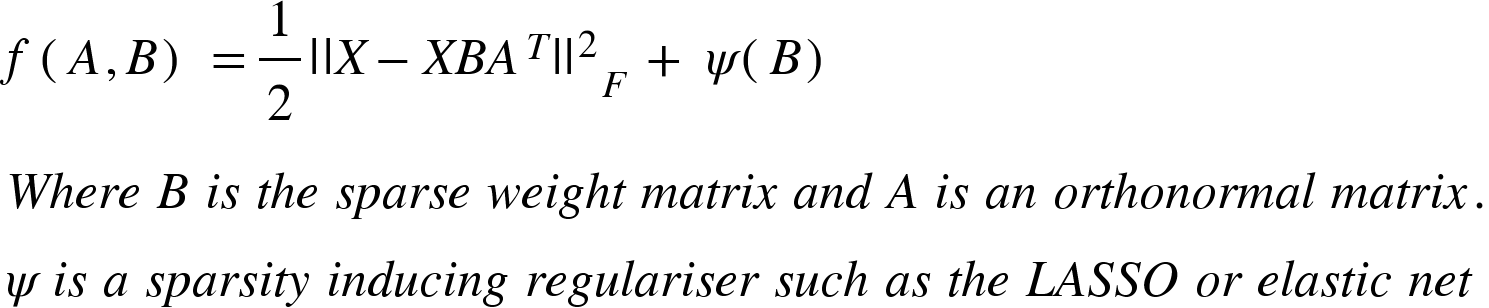
Sparse Machine learning refers to a collection of methods to learning that seek a tradeoff between some goodness-of-fit measure and sparsity of the result for better interpretability.

The sparsity of a sparse matrix is given by the total number of zero elements divided by the total number of elements. A matrix is sparse if its sparsity is greater than 0.5.

SPCA is a modern variant of PCA, where it attempts to find sparse vectors (loadings), or a vector with a few non zero values, causing better interpretability of the model.

It avoids overfitting in a high-dimensional data setting where number of variables is greater than the number of observations.

Given an n x p data matrix X, SPCA minimises



The principal components are formed as Z = XB.

The data can be approximately rotated back as X- = ZAT.