Analysis of Pharmaceutical revenue drivers

Maxim Khomiakov

02582 – Computational Data Analysis

Data pre-whitening

The data has been pre-whitened by dividing each axes of the components by the standard deviation, and the mean has been subtracted to centralize the component axis around 0. Furthermore the SVD is restricted by strictly positive singular values, as revenue figures cannot be negative.

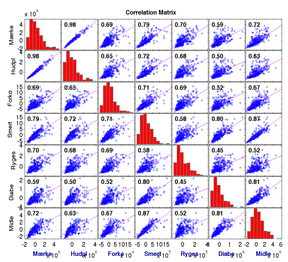
Unsupervised learning of pharmaceutical data

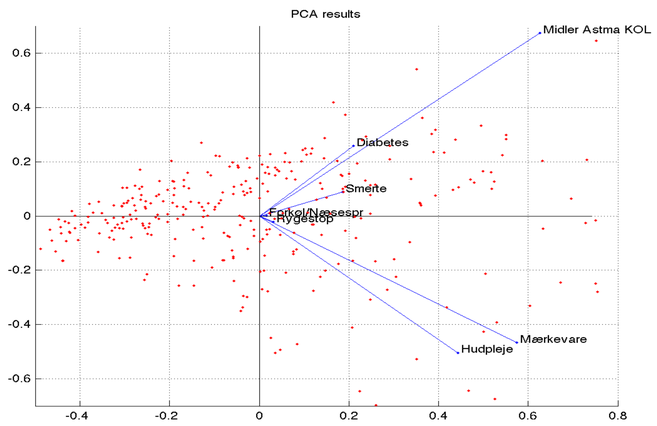
The dataset consists of two matrices, one which contains main revenue information, and another matrix with gender and age distributions. Both matrices are linked to a demographic region.

The main question of analysis is whether there is any basis for suggesting complementary purchases in different revenue categories. I.e. Would a customer purchasing good X have a certain propensity to purchase good Y as well? And would such a relationship be affected by gender, age or specific demography?

We will utilize the Principal Component Analysis and the PARAFAC algorithm to quantify, if possible, the above.

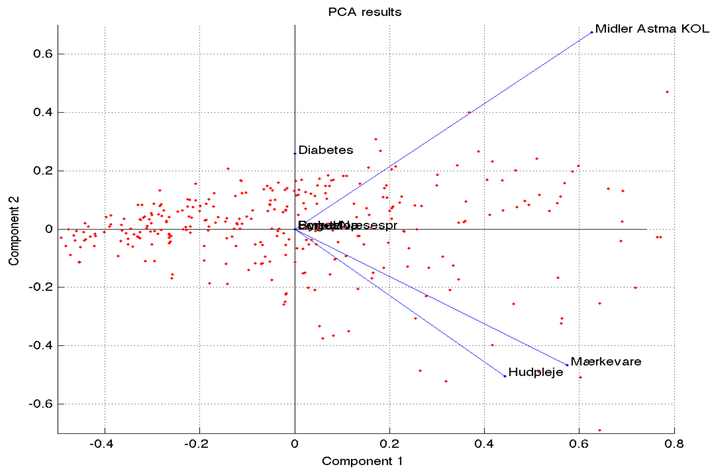
Results of the PCA.





Clockwise from top: Correlation of variables, Screeplot, Bi-plot of PCA 1 and 2

Results of the S-PCA. *(Conditioned on loadings, L < |0.25| = 0)*



Bi-plot of S-PCA components 1 and 2, with restricted loadings, *L <|.25|=0*

The PCA and S-PCA

Principal Component Analysis is an excellent and efficient way to discover underlying relationships in data. The PCA is calculated by singular value decomposition of the data matrix.

By SVD we can decompose the data matrix, $X \in \mathbb{R}^{mxn}$ into

$$X = U\Sigma V^T$$
$$S = U\Sigma$$
$$L = V$$
$$\sigma^2 = diag(\Sigma)^2/n$$

Where the loadings $L \in \mathbb{R}^{mxm}$ and the scores $S \in \mathbb{R}^{nxm}$ and the variances $\sigma^2 \in \mathbb{R}^{mx1}$. The scores form the axes that maximizes the variance from the components. The loadings measure the amount of information described by a particular component.

In Sparse-PCA we restrict a particular portion of the loadings to be zero, which in effect will make for more explicit interpretation, as the independence between the variables becomes larger.

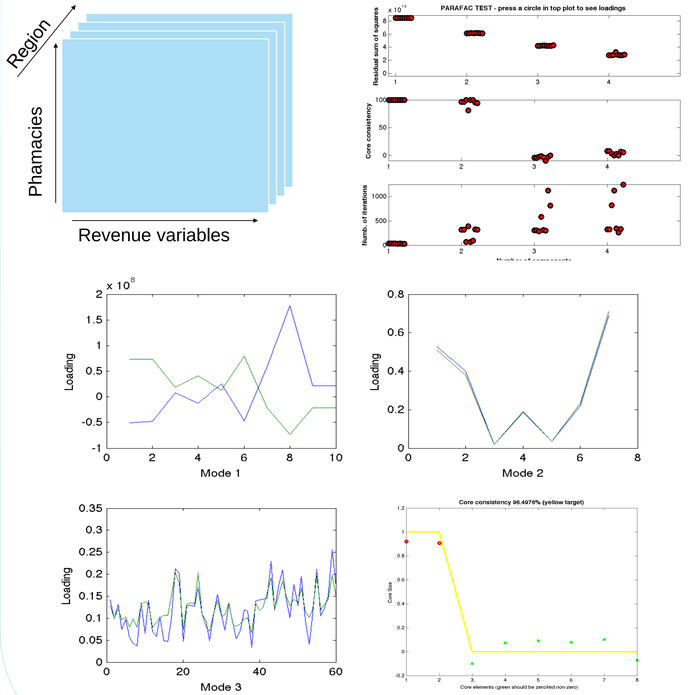
The PARAFAC/Canonical Decomposition

The PARAFAC/CP Algorithm is analogous to the PCA, while quite different. It could be described as the PCA equivalent for tensor based data problems i.e. $X \in \mathbb{R}^{nxmxj}$ .The PARAFAC can be represented as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + ei_{jk}$$

Results of the PARAFAC algorithm

Data representation in the case of the PARAFAC is key, since our data is not naturally adjusted to such an analysis. Through many iterations a choice of variables have been made, and the tensor of dimensions $\mathbb{R}^{10x7x60}$. Representing pharmacies, revenue variables and demographic region consequtively, has been chosen.



Clockwise from top: Tensor, PARAFAC 10 est. 4 comp. diagnostics, loadings of a 2-component PARAFAC-model, core diagnostics for PARAFAC 2-component model

Conclusion

PCA & S-PCA
Customers who purchase medicine against Astma, may also buy diabetes medicine. Likewise, skincare products are usually sold together with branded products.
Component 1 describes the amount of revenue the variable attributes towards. Component 2 describes the type of product/medicine. I.e. Required prescription or not.

PARAFAC
Revenues have opposite movements in the loadings. Confirmed by the PCA, Astma vs. Branded products. The second mode given by X amount of pharmacies in a region seems indifferent. Lastly, regional movements seem more or less equal as well.