Modeling complex phenotypes with latent variables - PCA

Paulina Kucharewicz

March 11, 2023

1 Brief definitions

1. **Complex phenotype**

   Complex phenotype is a set of observable characteristics that are a result of many genetic (or not) factors that interact with each other. As such, these traits are difficult to model.

2. **Latent variable**

   Latent variable is a variable that is not directly observed in the experiment but can be inferred from observed variables. In the context of modeling complex phenotypes we might want to find latent variables that are underlying cause of variables observed in the experiment. Identifying such variables hopefully makes modeling easier (especially if their number is smaller than that of observable variables).

3. **Principal Component Analysis**

   PCA is a method used in the statistical data analysis that allows for a representation of data in a lower dimension in the form of principal components (PCs). PCs are uncorrelated variables that preserve as much information about the variance in the original data set as possible.

2 Computing PCA

PCA is mathematically related to the Singular Value Decomposition (SVD). SVD is a matrix factorization method that results in expression of a given matrix X of dimensions $m \times n$ as a product of matrices: USV^T, such that U is $m \times n$, S is $n \times n$ and diagonal (with values decreasing with column index) and V is $n \times n$ matrix. If mean of every column in X is equal to zero, PCs can be computed using matrices from SVD. Detailed explanation of SVD and PCA calculations are included in [WRR02].

In the context of gene expression analysis it would be more beneficial to understand how SVD and PCA relate to gene expression arrays.

$$X = U \times S \times V^T$$

- **X**: Every i-th row of X (g_i) is a transcriptional response of i-th gene. j-th column of X represents expression of every gene in a j-th sample and is an expression profile. An x_{ij} is expression of a i-th gene in a j-th sample.

- $\mathbf{V}$ is orthonormal and its rows (*eigengenes*) span the space of transcriptional responses.
- rows of $Z = SV^T$ are the PCs, corresponding columns of U are their loadings [CS14].

3 Pros and cons of the PCA

PCs preserve as much variability in columns of $\mathbf{X}$ as possible. First PC is explains the biggest portion of the variance and each next PC represents less of it. By selecting first k PCs such that most of the variance is explained, but k is still much smaller than m, we can achieve big reduction in data dimension while still representing most of the information from the data set. Additionally principal components are uncorrelated which makes theirs analysis easier. PCs are a great tool for phenotypes modeling as they can represent the influence of latent variables on observed data while reducing the dimension.

However PCs are linear combinations of usually n variables and it can make their interpretation difficult. In response to this problem Sparse Principal Component Analysis was developed [ZHT06]. As a result, PCs can be expressed using less variables, a huge improvent for analysis of multivariate data sets such as gene expression data.

4 Applications in gene expression analysis

PCA is widely used in data analysis as a dimension reduction tool which makes visualising data sets with large number of variables easier but also can help identify underling patterns in data.

PCA is suitable for gene expression data analysis as it deals with noise in the data. PCA is used for detecting genes crucial for observed phenotype and expression patterns. It is also used in population structure analysis, a crucial step in genome-wide association studies.

SVD also allows for dimension reduction depending on the rank of the matrix and is a usefull tool in detecting signals in data.

References

[CS14] Neo Christopher Chung and John D. Storey. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554, 10 2014.

[WRR02] Michael E. Wall, Andreas Rechtsteiner, and Luis M. Rocha. Singular value decomposition and principal component analysis. 2002.

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.