Modeling complex phenotypes with latent variables - PCA (class notes)

Roksana Malinowska

February 2023

1 Some definitions

1.1 Phenotype

Phenotype is described by features of an organism that are observable. It's the result of two main factors which are: expression of genetic code and influence of the environment.

1.2 Latent variables

Latent variables are unobserved variables that underlie the observed data. Here, we can define them as underlying factors that contribute to a complex phenotype.

1.3 PCA

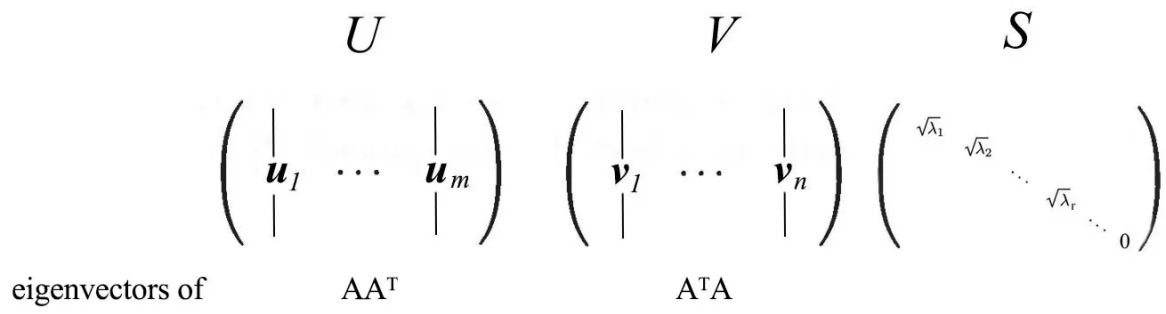
Principal Component Analysis (PCA) is a statistical technique used to identify patterns and reduce the dimensionality of complex data sets. It transformes a large set of variables into a smaller one that still containes most of the information.
Some variables are correlated with each other and PCA transforms data linearly into new properties that are not correlated with each other - **principal components**. Hence, the dimensionality reduction.

1.4 Eigenvalue and eigenvector

Scalar b and vector v are the eigenvalue and eigenvector of matrix A if Av=bv

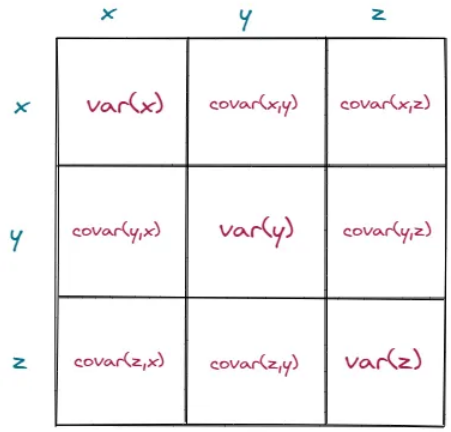
1.5 SVD

Singular Value Decomposition of a matrix is a factorization of that matrix into three matrices that are easy to manipulate and analyze. It's the foundation for unraveling data into independent components. SVD states that any matrix can be presented as $A = USV^T$.



1.6 Covariance matrix

Matrix with values that correspond to variance between a set of features. If the covariance is positive, it means that two features (variables) are correlated (in the same direction), and negative, when they are correlated but inversely (in the opposite direction).



2 What's the relationship between SVD and PCA?

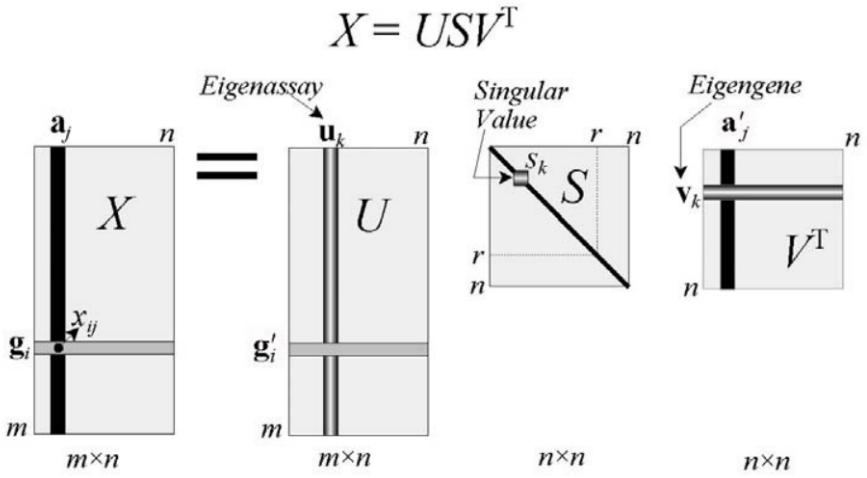
The relationship between SVD and PCA is direct in the case, where principal components are calculated from the covariance matrix. Let's say that X is a matrix with some real life data, for example gene expression matrix, where rows are samples and columns are genes. If we assume it's centered, which means column means are zero, we can get covariance matrix C as follows:
$C = X^T X/(n-1)$,
where n is the number of rows. Then we can diagonalize it:
$C = VLV^T$,
where V is a matrix of eigenvectors, L is a diagonal matrix with eigenvalues. Those eigenvectors are called **principal directions** of the data. When we multiply X by V (XV) we get principal components aka PC scores. From applying SVD on X we get $X = USV^T$. LET's substitute X to its SVD to the formula for covariance matrix:
$C = X^T X/(n-1) = \frac{VSU^T USV^T}{n-1} = V\frac{S^2}{n-1}V^T$. V are here principal directions of the data. **Principal components** are given by $XV = USV^T V = US$.

3 Example!

Let's look at the example of using SVD/PCA on gene expression data. Such data is well suited for those techniques.

3.1 Interpretation of SVD on gene expression data

There is no specific answear for the question of biological significance of the SVD. Let's now look at the X matrix where c columns are assays and m rows are genes. Here, columns of U are called the left singular vectors.They form an orthonormal basis for the assay expression profiles. The rows of V^T contain the elements of the right singular vectors. They form an orthonormal basis for the gene transcriptional responses. Left singular vectors (u vectors from U matrix after applying SVD) are eigenassays. Right singular vectors (v from V matrix) are eigengenes. We can sometimes refer to them as a components (consideering analogy to PCA).



For example, is diagnostic applications we might want to clasiffy samples of tissue from patients with and without a disease. Here, signal of interest would be the assay expression profile a, which we can define as: $a_j = \sum v_{jk} s_k u_k$.
j^th column of V^T, a'_j (look at the figure above) containes the coordinates of the j^th assay, which is kind of smaller representation of the expression profile of the assay. We get fewer variables to analyse. Such gene profile can be refereed to a phenotype. Having lesser number of variables it is easier to see some correlations between gene expression and the phenotype.

3.2 Summary

PCA (principal components analysis) is a a common unsupervised method for the analysis of gene expression data. It provides information about general structure of the dataset that we are studying. It reduces the dimensionality of the data, hence we can look at a kind of global map of gene expression, not on some huge sample/gene table. PCA presents information about directions where the data varies the most. We take a set of individual samples and through PCA we calculate similarity between them. This procedure can be sufficient to cluster them into some gene profiles which might be biologically significant (they might correspond to the presence of some disease). This is very valuable, because by looking at this huge table of gene expression we are not able to detect previously unknown relationships or characteristics of samples that don't have annotated phenotype.