Multivariate Statistics – PCA (data sample view)

Data are typically stored in a data matrix $X \in \mathbb{R}^{n \times p}$, keeping n samples of a p-dimensional random vector. Just a few examples:

(i) Gene expression data: each row of X holds the data from one patient, where the data for a patient consist of the expression levels of p genes of a cell (or cells) of a patient. Rows correspond to patients and columns to genes.

(ii) Images: each image consists of p pixels. For a grey-scale, image, these p pixels can form one vector and a data-matrix can store n images (in rows), where each column corresponds to one pixel in the image.

(iii) Financial data: the daily returns of p financial instruments for n days can be stored in a data matrix. Rows are days (or time) and columns correspond to different instruments.

(iv) Movie reviews: a matrix that stores the movie ratings of n users in the rows, where each of the p columns corresponds to a user. Many entries might be not available if a movie has not yet been rated by a given user.

(v) Text data: a matrix can store a bag-of-words summary of a text in a matrix, where the n rows correspond to n texts/websites and the columns correspond to the presence of p different words.

Principal component analysis tries again to find a linear subspace of dimension $q < p$in the data that preserves most of the variance. It is used as a dimension reduction tool; either to make subsequent analysis more computationally efficient or reduce noise in the data or both.

Assume for the following that the columns are mean-centered, that is

$$\sum_{i=1}^{n} X_{ik} = 0 \quad \forall k \in \{1, \ldots, p\}.$$

The empirical covariance matrix of the data can the be written as [1]

$$\hat{\Sigma} = n^{-1} X^t X,$$

and we denote the EVD of $X^t X$ by

$$X^t X = V \Lambda V^t,$$

[1] sometimes a factor $1/(n-1)$ is used instead of $1/n$ to make the variances unbiased but for simplicity just work with n here

again with orthogonal V (where the first eigenvector is in the first column etc.) and diagonal Λ with decreasing (or rather non increasing) eigenvalues on the diagonal

$$\Lambda_{11} \geq \Lambda_{22} \geq \ldots.. \geq \Lambda_{pp}.$$

Lets discuss first the case of the first principal component ($q = 1$). There are again two views on this, just as in the population case:

(i) Find a direction $w_1 \in \mathbb{R}^p$ that maximizes variance.

(ii) Find a direction that minimizes error

Again these two views are identical and will use the first one in the following.

The first principal component is defined as follows

$$w_1 = \text{argmax}_{w:\|w\|_2=1} \underbrace{\text{Var}(Xw)}_{\|Xw\|_2^2}$$

and this can be reformulated as

$$\begin{aligned} w_1 &= \text{argmax}_{w:\|w\|_2=1} \underbrace{\text{Var}(Xw)}_{\|Xw\|_2^2} \\ &= \text{argmax}_{w:\|w\|_2=1} \; w^t \underbrace{X^t X}_{n\hat{\Sigma}=nV\Lambda V^t} w \\ &= \text{argmax}_{w:\|w\|_2=1} \; w^t V \Lambda V^t w \end{aligned}$$

Let $b_1 = V^t w_1 \in \mathbb{R}^p$ be coefficients of w_1 in basis $v_1, \ldots, v_p$:

$$w_1 = \sum_{k=1}^{p} (b_1)_k v_k.$$

Then $\|w_1\|_2 = 1$ is identical to $\|b_1\|_2 = 1$ since V is orthonormal and hence $\|Vx\|_2 = \|x\|_2$ and also $\|V^t x\| = \|V^{-1}x\| = \|x\|_2$ for all $x \in \mathbb{R}^p$.

Then

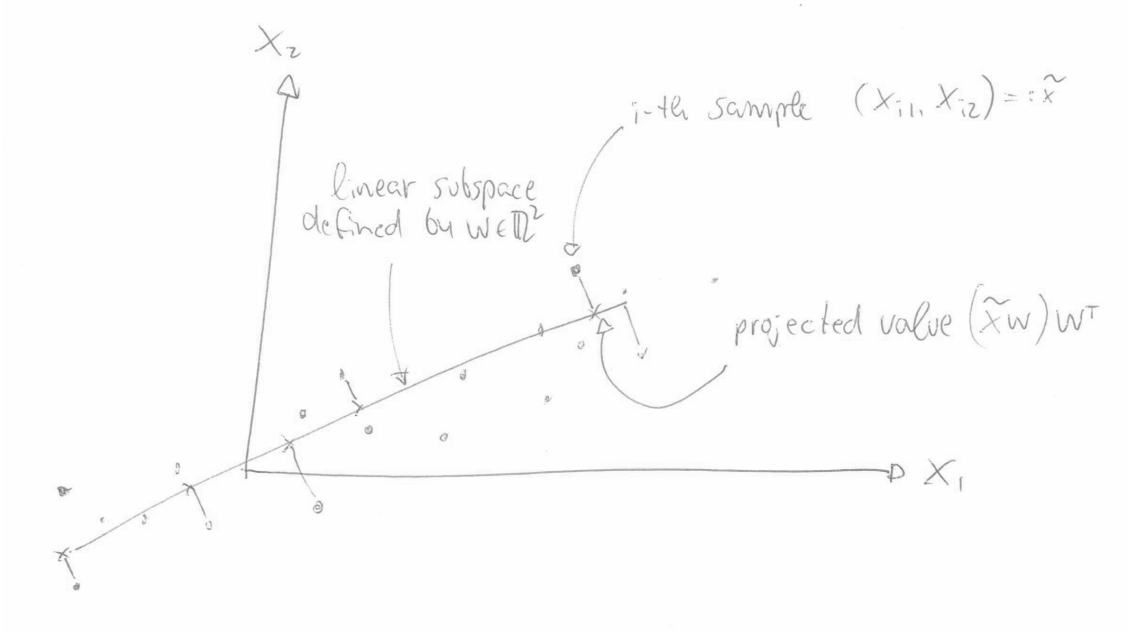
$$b_1 = \text{argmax}_{b:\|b\|_2=1} b^t \Lambda b = \sum_{k=1}^{p} \Lambda_{kk} b_k^2$$

is solved by

$$(b_1)_k = \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } k = 1, \ldots, p$$

since $\Lambda_{11} \geq \Lambda_{22} \geq \ldots..$ The first principal components is thus given by

$$w_1 = V b_1 = v_1,$$



An example of a projection for two dimensions ($p = 2$) and projection dimension $q = 1$ under the constraint that $\|w\|_2 = 1$.

the first eigenvector of $\hat{\Sigma}$ (associated with the largest eigenvalue). Solution is unique iff $\Lambda_{11} > \Lambda_{22}$.

Second principal component is defined as the direction that maximizes the variances among all directions orthogonal to the first principal component:

$$w_2 = \text{argmax}_{w:\|w\|_2=1 \text{ and } w^t w_1=0} \text{Var}(Xw).$$

We get, using the same approach, that the solution is $w_2 = v_2$, that is the second eigenvector of $X^t X$ (associated with the second largest eigenvalue) and so on for larger values of $q > 2$.

Projected data-points in a rank-q approximation can be computed directly by a truncated SVD. For $n > p$, SVD is given by

$$X = \underbrace{\begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}}_{X \in \mathbb{R}^{n \times p}} = \underbrace{\begin{pmatrix} | & & | \\ | & & | \\ u_1 & & u_n \\ | & & | \\ | & & | \end{pmatrix}}_{U \in \mathbb{R}^{n \times n}} \underbrace{\begin{pmatrix} D_{11} & & & \\ & D_{22} & & \\ & & \ddots & \\ & & & D_{pp} \\ & & 0 & \end{pmatrix}}_{D \in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} - & v_1 & - & - \\ & & & \\ & & & \\ - & v_p & - & - \end{pmatrix}}_{V^t \in \mathbb{R}^{p \times p}}.$$

Truncating to q dimensions while keeping the q largest singular values (that correspond to the q largest eigenvalues of $X^t X$) is achieved by setting the singular values equal to zero

after the q-th diagonal element of D:

$$\tilde{X} = \underbrace{\begin{pmatrix} | & & | \\ | & & | \\ u_1 & & u_n \\ | & & | \\ | & & | \end{pmatrix}}_{U \in \mathbb{R}^{n \times n}} \underbrace{\begin{pmatrix} D_{11} & & & \\ & \cdots & & \\ & & D_{qq} & \\ & & & 0 \\ & 0 & & \end{pmatrix}}_{\text{a truncated } D} \underbrace{\begin{pmatrix} - & v_1 & - & - \\ & & & \\ - & v_p & - & - \end{pmatrix}}_{V^t \in \mathbb{R}^{p \times p}}.$$

We now need to keep only the first q columns of U and V in the low-rank approximation:

$$\tilde{X} = \underbrace{\begin{pmatrix} | & & | \\ | & & | \\ u_1 & & u_q \\ | & & | \\ | & & | \end{pmatrix}}_{\tilde{U} \in \mathbb{R}^{n \times q}} \underbrace{\begin{pmatrix} D_{11} & & \\ & \cdots & \\ & & D_{qq} \end{pmatrix}}_{\tilde{D} \in \mathbb{R}^{q \times q}} \underbrace{\begin{pmatrix} - & v_1 & - & - \\ & & & \\ - & v_q & - & - \end{pmatrix}}_{\tilde{V}^t \in \mathbb{R}^{q \times p}}.$$

We can combine $\tilde{U}$ and $\tilde{D}$ to get

$$\begin{aligned} \tilde{X} &= \tilde{U}\tilde{D}\tilde{V}^t \\ &= AH, \qquad \text{where } A = \tilde{U}\tilde{D} \text{ and } H = \tilde{V}^t. \end{aligned}$$

In matrix notation,

$$\tilde{X} = \underbrace{\begin{pmatrix} & \\ & \\ & \end{pmatrix}}_{\tilde{X} \in \mathbb{R}^{n \times p}} = \underbrace{\begin{pmatrix} & \\ & \\ & \end{pmatrix}}_{A \in \mathbb{R}^{n \times q}} \underbrace{\begin{pmatrix} & & \end{pmatrix}}_{H \in \mathbb{R}^{q \times p}}.$$

The PCA decomposition is by construction optimal solution to

$$\text{argmin}_{A \in \mathbb{R}^{n \times q}, H \in \mathbb{R}^{q \times p} : HH^t = \mathbf{1}_{q \times q}} \|X - AH\|_2^2.$$

1. Entries in A are often called scores. Each of the n rows of A can be thought of as containing the "activity-coefficients" of the basis vectors in H.
2. Entries in H are often called loadings. Each of the q rows of H contains a basis vector (which in the PCA solution is just the corresponding eigenvector $v_1, \ldots, v_q$ of X^tX).

Solution is only determined up to rotations of H, even if all eigenvalues are distinct. To make this more precise, let $R \in \mathbb{R}^{q\times q}$ contain in its q columns an orthonormal basis of $\mathbb{R}^q$. The operation R^t can be seen as a rotation by transforming the coefficients in the basis $v_1, \ldots, v_1$ into a rotated basis. The operation R is the inverse and transforms the coefficients back into the original eigenvector basis. Then $R^tR = RR^t = \mathbf{1}_{q\times q}$ and

$$X - AH = X - ARR^tH = X - A'H',$$

where $A' = AR$ and $H' = R^tH$. and $H'(H')^t = HH^t = \mathbf{1}_{q\times q}$.

Some comments:

(i) Factor analysis (used a lot in psychology for example) often tries to find special rotations within the space of solutions by introducing different criteria to the objective of a low Frobenius norm error. The goal is to make the loadings (the rows in H) more interpretable. The most popular rotation is the varimax rotation which starts from the PCA solution typically and then chooses the rotation R so as to maximize the variance of the squared loadings (ie choose R as a rotation in q-dimensional space so as to maximize the variance of the squared loadings, ie the rows of the rotated factors R^tH).

(ii) The PCA solution/approximation is not invariant under scaling

$$X_k \leftarrow c \cdot X_k \text{ for some } c \in \mathbb{R}.$$

Typically data are scaled so that all columns have unit variance, ie

$$\sum_{i=1}^{k} X_{ik}^2 = n \text{ for all } k = 1, \ldots, p.$$

Exceptions might be sparse data or data where the units of the variables are comparable and meaningful (all variables are given in units of meters, for example).

(iii) The total variance "explained" is equal to the sum of the q largest squared values of X squared or the sum of the q largest eigenvalues of X^tX in that

$$\|\tilde{X}\|_2^2 = \sum_{\ell=1}^{q} D_{\ell\ell}^2 = \sum_{\ell=1}^{q} \Lambda_{\ell\ell}.$$

The relative error by keeping only the first q terms is

$$\text{err}(q) = \frac{\|X - \tilde{X}\|_2^2}{\|X\|_2^2} = 1 - \frac{\sum_{\ell=1}^{q} D_{\ell\ell}^2}{\sum_{\ell=1}^{p} D_{\ell\ell}^2} = 1 - \frac{\sum_{\ell=1}^{q} \Lambda_{\ell\ell}}{\sum_{\ell=1}^{p} \Lambda_{\ell\ell}}$$

and is monotonically decreasing (with $\text{err}(0) = 1$ and $\text{err}(p) = 0$).