Multivariate Statistics – Introduction and notation

Main goals of multivariate statistics (as for this course) are dimensionality reduction, visualization, deconvolution and in general concise summaries of multivariate data.

Topics covered might/will include:

1. Principal component analysis (PCA)
2. Non-negative matrix factorization
3. Sparse dictionary learning
4. Spectral clustering
5. Autoencoders as nonlinear PCA
6. Visualization (multidimensional scaling and t-SNE)
7. Linear unmixing and independent component analysis (ICA)
8. Optimal Transport

1. Notation

Objects of interest are either a random vector $X \in \mathbb{R}^p$ or n independent realizations of such a vector (i.e. a data sample).

Population view. The random vector is given by

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \in \mathbb{R}^p.$$

Its first and second moments are given by

$$\mu := E(X) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} \in \mathbb{R}^p \text{ and } \Sigma := \mathrm{Cov}(X) = \begin{pmatrix} \Sigma_{11} & \dots & \Sigma_{1p} \\ \vdots & \ddots & \vdots \\ \Sigma_{p1} & \dots & \Sigma_{pp} \end{pmatrix} \in \mathbb{R}^{p \times p},$$

where for all $k, l \in \{1, \dots, p\}$,

$$\begin{aligned} \Sigma_{kk} &= \mathrm{Var}(X_k) = E[(X_k - \mu_k)^2] \\ \Sigma_{kl} &= \mathrm{Cov}(X_k, X_l) = E[(X_k - \mu_k)(X_l - \mu_l)]. \end{aligned}$$

The correlation ρ is given by the normalized covariance

$$\rho := \mathrm{Cor}(X) = \begin{pmatrix} \rho_{11} & \dots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \dots & \rho_{pp} \end{pmatrix} \in \mathbb{R}^{p \times p},$$

with

$$\rho_{kl} = \frac{\Sigma_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}} \text{ for } k, l \in \{1, \dots, p\}.$$

Data sample. The data matrix is given by

$$X = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix} \in \mathbb{R}^{n \times p},$$

where a row of the matrix corresponds to the p-dimensional observation of the random variable for all rows or samples $1, \dots, n$. Since we mostly work with the data matrix, we will not distinguish between the random vector and the data matrix notationally but this could be done by setting one of the two in bold script for example.

The first and second moments are given by

$$\hat{\mu} := \overline{X} = \begin{pmatrix} \overline{X_1} \\ \vdots \\ \overline{X_p} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} X_{ip} \end{pmatrix} \in \mathbb{R}^p$$

and

$$\hat{\Sigma} := \begin{pmatrix} \hat{\Sigma}_{11} & \dots & \hat{\Sigma}_{1p} \\ \vdots & \ddots & \vdots \\ \hat{\Sigma}_{p1} & \dots & \hat{\Sigma}_{pp} \end{pmatrix} \in \mathbb{R}^{p \times p},$$

where for all $k, l \in \{1, \dots, p\}$,

$$\hat{\Sigma}_{kl} = \frac{1}{n}\sum_{i=1}^{n}(X_{ik} - \hat{\mu}_k)(X_{il} - \hat{\mu}_l).$$

Let $\tilde{X} \in \mathbb{R}^{n \times p}$ be a mean-zero version of X, in which the mean of each of the p columns of X is set to zero. Then

$$\hat{\Sigma} = \frac{1}{n}\tilde{X}^t\tilde{X}.$$

Without removing the mean, the $p \times p$ -dimensional matrix

$$X^t X$$

is often referred to as the Gram matrix.

The empirical correlation $\hat{\rho}$ is given by the normalized empirical covariance

$$\hat{\rho} := \begin{pmatrix} \hat{\rho}_{11} & \cdots & \hat{\rho}_{1p} \\ \vdots & \ddots & \vdots \\ \hat{\rho}_{p1} & \cdots & \hat{\rho}_{pp} \end{pmatrix} \in \mathbb{R}^{p \times p},$$

with

$$\hat{\rho}_{kl} = \frac{\hat{\Sigma}_{kl}}{\sqrt{\hat{\Sigma}_{kk}\hat{\Sigma}_{ll}}} \text{ for } k, l \in \{1, \ldots, p\}.$$

The n-dimensional identity matrix is denoted by

$$\mathbf{1}_{n \times n} = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}.$$

2. Recap: singular value decomposition (SVD) and eigenvalue decomposition (EVD)

Singular value decomposition of a data matrix. The data matrix $X \in \mathbb{R}^{n \times p}$ can be decomposed in a singular-value decomposition (SVD) as

$$X = UDV^t$$

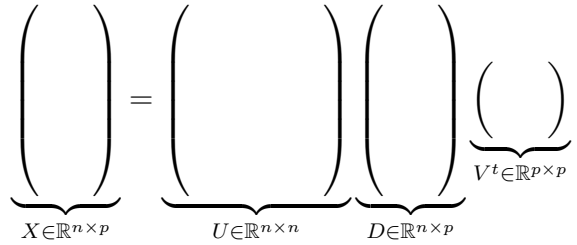
where

(a) U is orthogonal in $\mathbb{R}^n$: $U^tU = UU^t = \mathbf{1}_{n \times n}$.

(b) V is orthogonal in $\mathbb{R}^p$: $V^tV = VV^t = \mathbf{1}_{p \times p}$.

(c) D diagonal in $\mathbb{R}^{n \times p}$.

If $n > p$, the decomposition looks like



where

(a) the columns of U contain the left-singular vectors $u_1, \ldots, u_n$ that form an orthonormal basis of $\mathbb{R}^n$,

(b) the columns of V contain the right-singular vectors $v_1, \ldots, v_n$ that form an orthonormal basis of $\mathbb{R}^p$,

(c) and the diagonal elements of D (usually denoted by $d_1 \geq d_2 \geq \ldots d_{\min\{n,p\}}$ where $d_i = D_{ii}$ for $i = 1, \ldots, \min\{n, p\}$) contain the non-negative singular values which are ordered in decreasing magnitude.

The matrix V^t can be seen as a rotation (with $\|V^t x\|_2^2 = \|x\|_2^2$ for all $x \in \mathbb{R}^p$) that transform coefficients in standard basis

$$\begin{pmatrix} 1 \\ 0 \\ . \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ . \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ 0 \\ . \\ 1 \end{pmatrix}$$

into coefficients in a basis given by the right-singular vectors

$$v_1, \ldots, v_p.$$

The transformation V reverses this transformation, ie performs the transformation in the opposite direction (and $V^t V = V V^t = \mathbf{1}_{p \times p}$). Analogous for U in $\mathbb{R}^n$.

Eigenvalue decomposition. The second moment $X^t X \in \mathbb{R}^{p \times p}$ is sometimes called the Gram matrix and it is equal to the empirical covarince if the columns of X are mean-centered (modulo factor n). It can be decomposed, using the SVD of X, as

$$\begin{aligned} X^t X &= (UDV^t)^t (UDV^t) \\ &= VD^t \underbrace{(U^t U)}_{\mathbf{1}_{n \times n}} DV^t \\ &= VD^t DV^t, \end{aligned}$$

where

$$D^t D = \begin{pmatrix} D_{11}^2 & & & \\ & D_{22}^2 & & \\ & & \ddots & \\ & & & D_{pp}^2 \end{pmatrix}.$$

The eigenvalue decomposition of $X^t X \in \mathbb{R}^{p \times p}$ is hence given by

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

where the orthogonal matrix $V \in \mathbb{R}^{p \times p}$ is identical to the matrix in the SVD of X and $\Lambda \in \mathbb{R}^{p \times p}$ is the diagonal matrix containing the eigenvalues

$$\lambda_k := \Lambda_{kk} = D_{kk}^2 \text{ for } k = 1, \ldots, p$$

equal to the squared singular values in SVD of X. The EVD implies that

$$(X^tX)v_k = \lambda_k v_k,$$

where v_k is again k-th column of V. Eigenvalues ordered again such that

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0,$$

and by the connection to the SVD above it is clear that at most $\min\{n, p\}$ of these eigenvalues can be non-zero.

Rank and trace. The rank of a matrix X is defined as the number of non-zero singular values and is also equal to the number of non-zero eigenvalues of X^tX:

$$\text{rank}(X^tX) = \sum_{k=1}^{p} 1\{\lambda_k \neq 0\} \leq \min\{n, p\}$$

as

$$\sum_{k=1}^{p} 1\{\lambda_k \neq 0\} = \sum_{k=1}^{p} 1\{\Lambda_{kk} \neq 0\} = \sum_{k=1}^{\min\{n,p\}} 1\{D_{kk} \neq 0\}.$$

Note that the sum of the eigenvalues of X^tX is equal to the trace as

$$\text{trace}(X^tX) := \sum_{k=1}^{p} (X^tX)_{kk} = \sum_{k=1}^{p} \Lambda_{kk} = \sum_{k=1}^{p} \lambda_k$$

as

$$\text{trace}(X^tX) = \text{trace}(V\Lambda V^t) = \text{trace}(\underbrace{V^tV}_{\mathbf{1}_{p\times p}}\Lambda) = \text{trace}(\Lambda) = \sum_{k=1}^{p} \Lambda_{kk}.$$

Projection. A linear projection $P : \mathbb{R}^p \mapsto \mathbb{R}^p$ has defining property $P^2x = Px$ for all $x \in \mathbb{R}^p$. Projecting into the space spanned by the vectors

$$v_1, \ldots, v_q \in \mathbb{R}^p \text{ with } q < p$$

can be achieved (in Euclidean space) by

$$P = V(V^tV)^{-1}V^t,$$

where the k-th column of $V \in \mathbb{R}^{p\times q}$ is given by v_k for all $k = 1, \ldots, q$. Projection P now has rank $q < p$. More specifically, P has

1. q eigenvalues with value 1 since $Pv_k = v_k$ for all $k = 1, \ldots, q$
2. $(p - q)$ eigenvalues with value 0 since $Pw \equiv 0$ for all w orthogonal to space spanned by $v_1, \ldots, v_q$.