UNSUPERVISED LEARNING (h 10)

3/5/21

So far : supervised learning (regression, classification)

$$X_1, \ldots, X_p \rightarrow Y$$

Unsupervised : $X_1, \ldots, X_p$

- Finding patterns / structure
- Visualizing the data
- exploratory data analysis

① Principal Component Analysis (PCA)

→ reducing the dimension of the data

② Clustering.

→ Discover unknown subgroups of data

PCA

$X_1, \ldots, X_p$ features

n observations

We would like to do some visualisation:

- We could do it pairwise (Junction pairs in R)
  - $\binom{p}{2} = \frac{p(p-1)}{2}$ possible pairs of features
  - $p = 10 \rightarrow 45$ plots
  - informative up to a certain point

PCA tries to find a low-dimensional representation of the data

Two components/ingredients of PCA:

① Mathematical component: We want to rewrite the data in a different basis

② Statistical: We want to choose a basis so that most of the information in the data is contained in the first "components"

MATHEMATICAL VIEW OF PCA

It consists of:

- Eigendecomposition of Covariance Matrix
- Singular Value Decomposition of Data Matrix

The two methods are very much related but they lead to different software implementations.

COVARIANCE MATRIX

$(X_1, \dots, X_p) = \underline{X}$ p random variables

$$\Sigma = (\sigma_{ij})_{\substack{i=1,\dots,p \\ j=1,\dots,p}} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \\ \sigma_{p1} & & \sigma_{pp} \end{pmatrix}, \quad \sigma_{ij} = Cov(X_i, X_j) = E\left[(X_i - E(X_i)) \cdot (X_j - E(X_j)) \right]$$

The matrix is symmetric and positive definite:

① $\sigma_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i) = \sigma_{ji}$ (symmetric)

② For any $\underline{v} \in \mathbb{R}^p$,

$\underline{v}^t \Sigma \underline{v} = Var(\underbrace{\underline{v}^t \underline{X}}_{v_1 X_1 + \cdots + v_p X_p}) > 0$ (positive definite, unless there is perfect multicollinearity - $\underline{v}^t \underline{X}$ constant)

(Spectral Theorem)

Σ symmetric $\rightarrow$ p real eigenvalues $\rightarrow$

$$\Sigma = U D U^t$$

EIGENDECOMPOSITION OF Σ

$$= (\underline{u}_1, \dots, \underline{u}_p) \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_p \end{pmatrix} U^t$$

$\underline{u}_1, \dots, \underline{u}_p$ are eigenvectors of length 1 and orthogonal

$\lambda_1, \dots, \lambda_p$ eigenvalues

Some remarks:

① U is orthogonal : $U^t U = U U^t = I_p$

Orthogonal matrices have special properties in the context of linear transformations:

① Preserve scalar product : $(U\underline{w}) \cdot (U\underline{v}) = \underline{w} \cdot (\underbrace{U^t U}_{I} \underline{v}) = \underline{w} \cdot \underline{v}$

② $\det(U^t U) = 1 \Rightarrow \det(U) = \pm 1$

$U = (\underline{w}, \underline{v})$

So U is associated to rotations or reflections

② Σ positive definite $\Rightarrow \lambda_1, \dots, \lambda_p > 0$

③ In practice Σ is unknown and needs to be estimated from data

Data $\quad X_{n\times p} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$ n observations of p variables

SAMPLE COVARIANCE $\quad S = (s_{kj})_{kj}$

$$s_{kj} = \frac{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j)}{n} \quad \text{with} \quad \bar{x}_k = \frac{\sum_{i=1}^{n} x_{ik}}{n} \text{ sample means}$$

Consider centered data : $\bar{x}_k = 0 \ \forall j$

$$\Rightarrow s_{kj} = \frac{\sum_{i=1}^{n} x_{ik} x_{ij}}{n} \ (*) \quad k,j = 1, \dots, p$$

$$\Rightarrow \underset{p\times p}{S} = \frac{\sum_{i=1}^{n} \underset{p\times 1}{\underline{x}_i} \cdot \underset{1\times p}{\underline{x}_i^t}}{n} \quad \text{with} \quad \underline{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \ i^{th} \text{ "row" of } X$$

$\underline{x}_i \cdot \underline{x}_i^t$ has rank 1
so S does not have rank larger than n

$\begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} (x_{i1} \cdots x_{ip})$

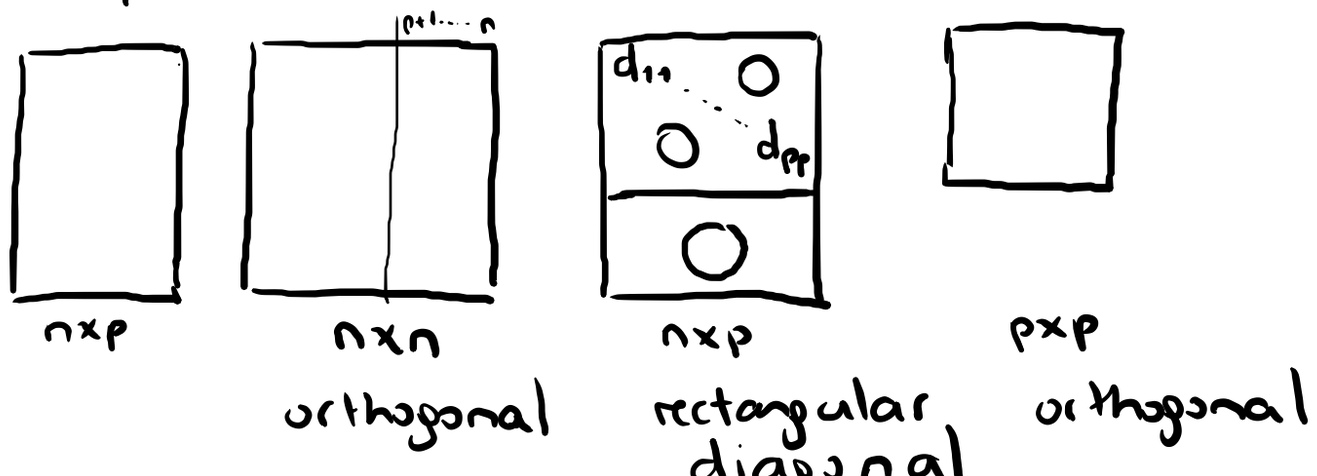
So if $p > n$, S is not full rank $\rightarrow$ ill-conditioned

$$\Rightarrow S = \frac{X^t X}{n} \qquad s_{kj} = \frac{1}{n}(x_{1k} \cdots x_{nk}) \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} = \frac{\sum_{i=1}^{n} x_{ik} \cdot x_{ij}}{n}$$

SINGULAR VALUE DECOMPOSITION ($of\ X_{n \times p}$)

Any matrix X can be factorized as

$$X_{n \times p} = V \cdot D \cdot U^t$$



$d_{11}, \dots, d_{pp} \geq 0$
singular values

In order to show that any matrix X admits at least one SVD, we can construct one in this way:

1. Set V as the matrix of eigenvectors of $\underbrace{XX^t}_{n\times n}$ and U the matrix of unit eigenvectors of $\underbrace{X^tX}_{p\times p}$.
2. The singular values in D are the square root of the eigenvalues of X^tX.

Proof Firstly, X^tX is symmetric and positive semi-definite, since

1. $(X^tX)^t = X^t(X^t)^t = X^tX$
2. $\underline{v}^t(X^tX)\underline{v} = (\underline{v}^tX^t)(X\underline{v}) = (X\underline{v})^t(X\underline{v}) = \|X\underline{v}\|^2 \geq 0$ for any $\underline{v}$

Therefore, I can consider the eigendecomposition of X^tX:

$$X^tX = U\Lambda U^t \quad \text{with} \quad \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}, \ \lambda_1,\dots,\lambda_p \geq 0$$

$U = (\underline{u}_1,\dots,\underline{u}_p)$ unit eigenvectors of X^tX

$\frac{X^tX}{n}$ sample covariance

$$\boxed{(X^tX)\underline{u}_i = \lambda_i \underline{u}_i}$$

Let $\boxed{d_{ii} = \sqrt{\lambda_i}}$ and suppose that X^tX is full rank (proof can be generalised), i.e.

X^tX positive definite $\Rightarrow \lambda_1, \dots, \lambda_p > 0$

Then * $\boxed{\underline{v}_i = \frac{X\underline{u}_i}{d_{ii}}}$, $i = 1, \dots, p$

and create a matrix $V = (\underline{v}_1, \dots, \underline{v}_p, \underline{v}_{p+1}, \dots, \underline{v}_n)$ with $\underline{v}_{p+1}, \dots, \underline{v}_n$ n-p orthogonal vectors (not so important as they are multiplied with the "zero" part of D)

By construction, $\underline{v}_1, \dots, \underline{v}_p$ are unit eigenvectors of XX^t:

① $(XX^t)\underline{v}_i = (XX^t)\frac{X\underline{u}_i}{d_{ii}} = \frac{X(X^tX\underline{u}_i)}{d_{ii}} = \frac{X\lambda_i\underline{u}_i}{d_{ii}} = d_{ii}X\underline{u}_i$

$= (d_{ii}^2)\underline{v}_i$ $\quad \underline{v}_i$ is an eigenvector of XX^t with eigenvalue $d_{ii}^2 = \lambda_i$

② $\|\underline{v}_i\|^2 = \left(\frac{X\underline{u}_i}{d_{ii}}\right)^t\left(\frac{X\underline{u}_i}{d_{ii}}\right) = \frac{\underline{u}_i^t(X^tX\underline{u}_i)}{d_{ii}^2} = \frac{\underline{u}_i^t \cancel{d_{ii}^2}\underline{u}_i}{\cancel{d_{ii}^2}} = \|\underline{u}_i\|^2 = 1$

Let $D = \begin{pmatrix} d_{11} & & 0 \\ & \ddots & \\ 0 & & d_{pp} \\ \hline & 0 & \end{pmatrix}$; $\underline{v}_i := \frac{X \underline{u}_i}{d_{ii}}$

In matrix form:

$$V = X U D^{-1} \qquad \text{with} \qquad D^{-1} := \begin{pmatrix} \frac{1}{d_{11}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{d_{pp}} \\ \hline & 0 & \end{pmatrix}$$

Equivalently:

$$VD = XU \Rightarrow VDU^t = X \underbrace{UU^t}_{I} = X$$

$$\Rightarrow X = V \cdot D \cdot U^t$$

Interpretation of SVD: $X = V D U^t$

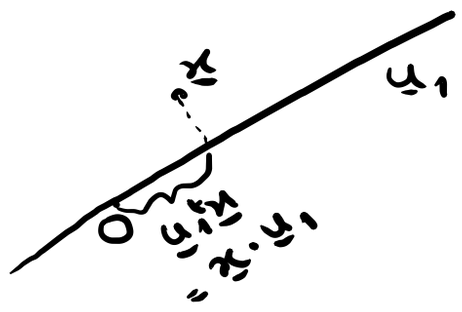
① Since $XU = VD$, the matrix VD provides the representation of the data X in the new basis U.

Take the generic row of X, $\underline{x}$. Then

$$U^t \underline{x} = \begin{pmatrix} \underline{u}_1^t \underline{x} \\ \vdots \\ \underline{u}_p^t \underline{x} \end{pmatrix} = \underline{t}$$

↑

p "principal components", projections of the point $\underline{x}$ on the new p directions $\underline{u}_1, \dots, \underline{u}_p$



② SVD is not unique:
- D can be arranged from the largest singular value to the smallest (unique)
- V and U^t are not unique (think of the case of repeated eigenvalues)

③ $X = V D U^t = \underbrace{\underset{n\times 1}{\underline{v}_1}\ \underset{1\times 1}{d_{11}}\ \underset{1\times p}{\underline{u}_1^t}}_{n\times p} + \cdots + \underline{v}_p d_{pp} \underline{u}_p^t$

(X: $n\times p$; largest → d_{11}; smallest → d_{pp})

Proof:

$$DU^t = \begin{pmatrix} d_{11} & & 0 \\ & \ddots & \\ 0 & & d_{pp} \\ \hline 0 & & \end{pmatrix} \begin{pmatrix} \underline{u}_1^t \\ \vdots \\ \underline{u}_p^t \end{pmatrix} = \begin{pmatrix} d_{11}\,\underline{u}_1^t \\ \vdots \\ d_{pp}\,\underline{u}_p^t \\ 0 \end{pmatrix}$$

So

$$(X)_{ij} = \underbrace{V_{[i,\cdot]}}_{\text{ith row of } V} \cdot (DU^t)_{[\cdot,j]} = \sum_{k=1}^{p} v_{ik} (DU^t)_{kj} = \sum_{k=1}^{p} v_{ik}\, d_{kk} (u_k^t)_j$$

$$= \left(\sum_{k=1}^{p} \underline{v}_k d_{kk} u_k^t \right)_{ij} \quad \Rightarrow$$

$$X = \underline{v}_1 d_{11} \underline{u}_1^t + \cdots + \underline{v}_p d_{pp} \underline{u}_p^t$$

Consider $\tilde{X} = \underline{v}_1 d_{11} \underline{u}_1^t + \cdots + \underline{v}_k d_{kk} \underline{u}_k^t$. Is $\tilde{X} \simeq X$?

Lower-dimensional representation of X

PCA- STATISTICAL VIEW

The choice of U orthonormal basis leads to p <u>UNCORRELATED</u> variables

$$X_1, \dots, X_p \xrightarrow[U^t \underline{x} = \underline{t}]{} T_1, \dots, T_p$$

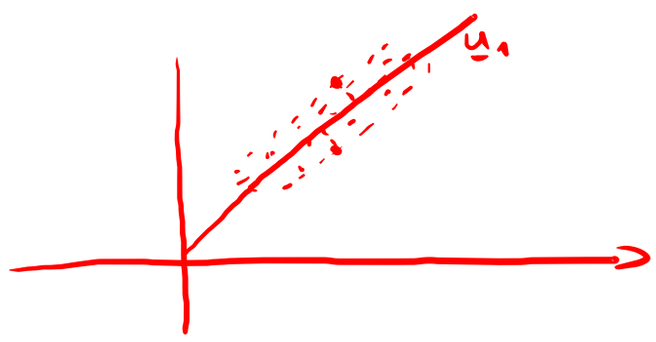
$$Var(\underline{t}) = Var(U^t \underline{x}) = U^t \, Var(\underline{x}) \, U = U^t \Sigma U$$

$$\widehat{Var(\underline{t})} = U^t \frac{X^t X}{n} U \overset{X = VDU^t}{=} \frac{U^t (VDU^t)^t (VDU^t) U}{n} =$$

$$(ABC)^t = C^t B^t A^t$$

$$= \frac{U^t U D^t \overbrace{V^t V}^{I} D \overbrace{U^t U}^{I}}{n} = \frac{D^t D}{n} = \frac{1}{n} \begin{pmatrix} d_{11}^2 & & 0 \\ & \ddots & \\ 0 & & d_{pp}^2 \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

$\lambda_1 = Var(t_1)$, $0 = Cov(t_i, t_j)$, $\lambda_p = Var(t_p)$



So the components $t_1, \dots, t_p$ are uncorrelated and "sorted" according to their variance.