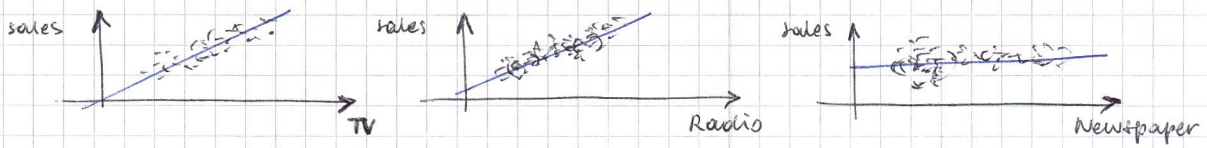
EXPLORATING A MULTIVARIATE DATASET

Descriptive statistics and graphical display. The geometry of multivariate sample. Sample mean, covariance and correlation. Generalized variance and total variance. The metric induced by the covariance matrix.

What is statistical learning?
Consider for instance a dataset of sales:



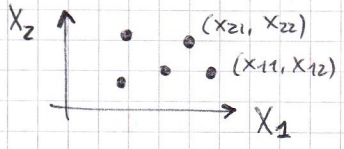
A linear regression line fit separately each one. Can we predict sales using all the informations? (All the features together?)

We'll always assume that we observed n statistical units, each one with p variables ($\equiv p$ features).

$$\Longrightarrow \quad \begin{aligned} \underline{x}_1 &= [x_{11}, x_{12}, \dots, x_{1p}]^T \in \mathbb{R}^p \\ \underline{x}_2 &= [x_{21}, x_{22}, \dots, x_{2p}]^T \\ &\dots \\ \underline{x}_n &= [x_{n1}, x_{n2}, \dots, x_{np}]^T \end{aligned} \quad \Bigg\} : \underset{\text{DATA MATRIX}}{\mathbb{X}} = \begin{array}{c} \begin{matrix} X_1 & X_2 & & X_p \end{matrix} \leftarrow \text{features} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \dots & \dots & & \dots \\ x_{n1} & x_{n2} & & x_{np} \end{bmatrix} \end{array}$$

$\underline{x}_n$: n-th statistical unit

We can read this data it two ways: by columns and by rows. Each row talks about a unit, each column about a variable. In the past statistical courses we saw how to determine the distribution of one feature, so we developed tools to analyze data by columns (mean, variance, quantile, ...), but the usual way to look at the data is by rows (by units). In this case we obtain a set of points in a $\mathbb{R}^p$ space.



: these data live in $\mathbb{R}^p$ ($p=2$).
(note that usually $n \gg p$)

Often it happens that we got a "special" variable:

$$\begin{bmatrix} X_1 & X_2 & \dots & X_p & Y \\ x_{11} & x_{12} & \dots & x_{1p} & y_1 \\ x_{21} & x_{22} & & x_{2p} & y_2 \\ \dots & \dots & & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} & y_p \end{bmatrix}$$

→ what we want to predict / explain in terms of the other features

(Y can also be categorical (a label): in this case we use information of $X_1, \dots, X_p$ to say something about membership of a group $\Longrightarrow$ classification problem)

So the problem is: observing p features we have to predict Y.
How can we translate this problem mathematically?

- $\underline{X} \in \mathbb{R}^p$: random vector of features ($\equiv$ what we observe)
- $Y \in \mathbb{R}$: what we want to predict

$\Longrightarrow$ **Problem**: use $\underline{X}$ to predict Y. What is the <u>best</u> function $f: \mathbb{R}^p \to \mathbb{R}$ to predict Y? (f in terms of $\underline{X}$)

What do we mean for "best"? We need a criterium for optimization. We want a function f such that:

$$\underbrace{\mathbb{E}[(Y - f(\underline{X}))^2]}_{\text{MEAN SQUARE ERROR}} \quad \text{is minimum.}$$

MEAN SQUARE ERROR: average of all the possible realizations of $\underline{X}$ and Y. We use $(\cdot)^2$ because we don't want the negative errors to compense the positive ones.

if f was a constant we would have: $k = \arg\min_k \mathbb{E}[(Y-k)^2] = \mathbb{E}[Y]$, so we have sort of an intuition of what the function should be, except for the

fact that we don't want f to be a constant because it's useless for accurate predictions (in fact we wouldn't even be using $\underline{X}$).
The important fact is: whatever f will be, it should be sort of a mean: $f(\underline{X}) = \mathbb{E}[\cdot]$.

$$\mathbb{E}[(Y - f(\underline{X}))^2] = \mathbb{E}[\underbrace{(Y - \mathbb{E}[Y|\underline{X}]}_{a} + \underbrace{\mathbb{E}[Y|\underline{X}] - f(\underline{X}))^2}_{b}]$$

$\mathbb{E}[Y|\underline{X}]$ = predictional expectation of Y given $\underline{X}$ (= the best guess of Y once you know $\underline{X}$).
From a geometrical point of view this is simply a projection of Y on the σ-field generated by $\underline{X}$ (it's a Radon-Nikodym derivative)

$$= \mathbb{E}[(Y - \mathbb{E}[Y|\underline{X}])^2] + \mathbb{E}[(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))^2] + 2\underbrace{\mathbb{E}[(Y - \mathbb{E}[Y|\underline{X}])(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))]}_{(*)}$$

We remember that: $\mathbb{E}[W] = \mathbb{E}[\mathbb{E}[W|Z]]$

$$\Longrightarrow (*) = 2\,\mathbb{E}[\mathbb{E}[(\cdot)(\cdot)|\underline{X}]]$$
$$= 2\,\mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y|\underline{X}])(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))|\underline{X}]]$$

$\mathbb{E}[g(X)|X] = \mathbb{E}[g(X)]$

$$= 2\,\mathbb{E}[(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))\,\mathbb{E}[(Y - \mathbb{E}[Y|\underline{X}))|\underline{X}]]$$
$$= 2\,\mathbb{E}[(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))\,(\mathbb{E}[Y|\underline{X}] - \mathbb{E}[Y|\underline{X}])]$$
$$= 0$$

$$\Longrightarrow \mathbb{E}[(Y - f(\underline{X}))^2] = \underbrace{\mathbb{E}[(Y - \mathbb{E}[Y|\underline{X}])^2]}_{(1)} + \underbrace{\mathbb{E}[(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))^2]}_{(2)}$$

So the problem is to find f s.t. (1) + (2) is minimum. For doing that, (1) is useless, it's like a constant, and for minimizing (2) we take (looking at the const case):

$$f(\underline{X}) = \mathbb{E}[Y|\underline{X}] \quad \leftarrow \text{solution of the optimization problem}$$

Of course there always be an error ε:

$$\varepsilon = Y - f(\underline{X})$$

(what we cannot capture from $\underline{X}$ to predict Y, it's **IRREDUCIBLE** using $\underline{X}$)

$\Longrightarrow$ We want to explain Y through a function f of $\underline{X}$ and a residual ε.
What can we say (statically) about ε?

$$\mathbb{E}[Y] = \mathbb{E}[f(\underline{X}) + \varepsilon] = \mathbb{E}[f(\underline{X})] + \mathbb{E}[\varepsilon] = \mathbb{E}[\mathbb{E}[Y|\underline{X}]] + \mathbb{E}[\varepsilon]$$
$$\downarrow$$
$$\mathbb{E}[Y] + \mathbb{E}[\varepsilon]$$

$$\Longrightarrow \mathbb{E}[\varepsilon] = 0$$

(this mean that if f is the best predictor then f has the same average that Y)

$\Longrightarrow$ The **model** is:

$$\begin{cases} Y = f(\underline{X}) + \varepsilon \\ f(\underline{X}) = \mathbb{E}[Y|\underline{X}] \\ \mathbb{E}[\varepsilon] = 0 \end{cases}$$

What do we know about f? We should know the joint distribution $Y|\underline{X}$ but usually we don't know it
$\Longrightarrow$ we need to use the data to estimate f

Suppose now that we have an estimation of f: $\hat{f}$.
How good is this estimation? Note that we need $\hat{f}$ not only to fitting the data, but for predictions too. We want to know what are the uncertainties of $\hat{f}$.
We want a criteria not only to check if the model is good for the data, we want to see how good it is when we try to predict something new.

$\hat{f}$: estimation of f through data $\mathbb{X}$
$\underline{x}_0 \in \mathbb{R}^p$: new observation ($\notin \mathbb{X}$) for which we want to predict y_0

We want to know about the error in predicting y_0:

$$\mathbb{E}_{\mathbb{X}}[(y_0 - \hat{f}(\underline{x}_0))^2] = \mathbb{E}[(y_0 - \hat{f}(\underline{x}_0))^2 \mid \mathbb{X} = x]$$

$$= \mathbb{E}_{\mathbb{X}}[(f(\underline{x}_0) + \varepsilon - \hat{f}(\underline{x}_0))^2]$$

$y_0 = f(\underline{x}_0) + \varepsilon_0$ — error referring to y_0

they're constant (we've conditioning on $\mathbb{X}$, the only random is ε)

$$= \mathbb{E}_{\mathbb{X}}[(f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2] + \mathbb{E}_{\mathbb{X}}[\varepsilon_0^2] + 2\,\mathbb{E}_{\mathbb{X}}[(f(\underline{x}_0) - \hat{f}(\underline{x}_0))\,\varepsilon_0]$$

$$= (f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2 + \mathrm{Var}(\varepsilon_0) + 2(f(\underline{x}_0) - \hat{f}(\underline{x}_0))\,\mathbb{E}_{\mathbb{X}}[\varepsilon_0]$$

since $\mathbb{E}_{\mathbb{X}}[\varepsilon_0] = \mathbb{E}[\varepsilon_0] = 0$

NOTE: $\varepsilon_0 \perp\!\!\!\perp \mathbb{X} \Rightarrow \mathbb{E}_{\mathbb{X}}[\varepsilon_0] = \mathbb{E}[\varepsilon_0]$

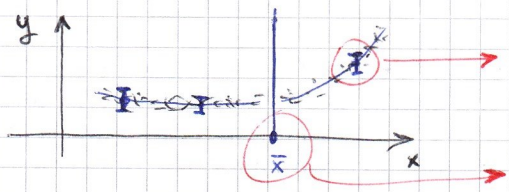
$$\Longrightarrow \boxed{\mathbb{E}_{\mathbb{X}}[(y_0 - \hat{f}(\underline{x}_0))^2] = (f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2 + \mathrm{Var}(\varepsilon_0)}$$

we can try to make this error as small as possible.
NOTE: $\underline{x}_0$ is a new observation, $\underline{x}_0 \notin \mathbb{X}$, it didn't influence the model, this is the right way to evaluate the model

we're not able to reduce this.

this is what we are not able to catch with any $\hat{f}$ (or any f), it's the part of Y which is not explainable by $\underline{X}$
It's **IRREDUCIBLE!**

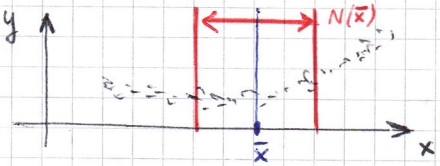
How can we get an estimation $\hat{f}$?
We wanted the **regression function** $\hat{f}(\underline{x}) = \mathbb{E}[Y \mid \underline{X} = \underline{x}]$ but there is a problem: sometimes we have few data for a particular $\underline{x}$ so we can't estimate $\hat{f}(\underline{x})$.



in each point x we do the mean of all the observations and we state $\hat{f}(x)$ equal to that mean.

in this point we don't have any observation so we can't find an estimation

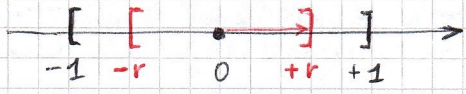
A possible solution is to relax the definition: $\hat{f}(\underline{x}) = \mathrm{Ave}(Y \mid \underline{X} \in N(\underline{x}))$.
Now the estimation of $\hat{f}(\underline{x})$ is the average in a neighborhood ($N(\underline{x})$):



: for the estimation of $\hat{f}(\underline{x})$ we use all the observations in a neighborhood.

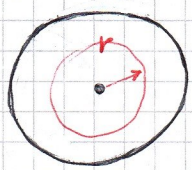
But there is a problem: **CURSE OF DIMENSIONALITY.**
This method fails when p becomes large. Why? To be efficient, the average must be taken in a neighborhood, more precisely in a local neigh. When p becomes large the points result to be far away from each other, so we lose the local property.

- **$p = 1$** : Let's consider $X \sim U([-1,1]) = U(S^1(1))$ (sphere of radius 1)
How far do we have to go from 0 to capture 10% of the data?



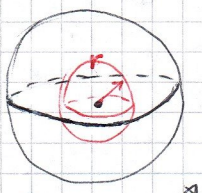
$$0,1 = \frac{S^1(r)}{S^1(1)} = \frac{2r}{2} = r \implies$$ to capture 10% of the data we have to move $r = 0,1$

- **$p = 2$** : $X \sim U(S^2(1))$



$$0,1 = \frac{S^2(r)}{S^2(1)} = \frac{\pi r^2}{\pi} = r^2 \implies$$ to capture 10% of the data we have to move $r = \sqrt{0.1} = 0,31$

- **$p = 3$** : $X \sim U(S^3(1))$



$$0,1 = \frac{S^3(r)}{S^3(1)} = \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi} = r^3 \implies$$ to capture 10% of the data we need to move $r = \sqrt[3]{0,1} = 0,46$

- **$p = 1000$** : $r = 0,99$: we have to travel the 99% of the universe to capture the 10% of the data. In this way we completely lose the sense of "local"

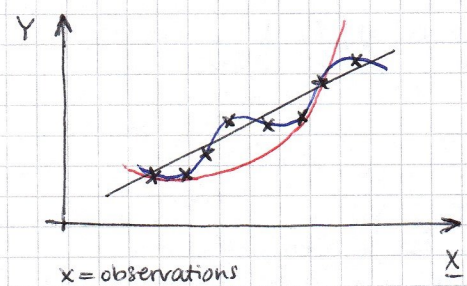
(every time we're solving : $r^p = 0,1$)

So, with p large we have the curse of dimensionality. How do we react?
We have two possibilities:

1. We reduce the dimensionality (we reduce p).
How? We forget part of the information. That's ok but we need a lot of knowledge of the phenomenal. We want a DATA DRIVEN REDUCTION so we project the data in a linear subspace (of dim < p) where the variability will be preserved.
(**PCA**: PRINCIPAL COMPONENT ANALYSIS)

2. We can use a **PARAMETRIC MODEL**: $Y = f(\underline{X}) + \varepsilon \rightarrow Y = \beta_0 + \beta_1 X_1 + .. + \beta_p X_p + \varepsilon$.
Everything is reduced to estimate $p+1$ parameters : we're choosing the form of the model. By parametrizing f we reduce the dimensionality of the problem.
It's not Data Driven (we don't know if it is better to consider, for instance, highs or highs^2), that's on us.

Problem of the error : **BIAS - VARIANCE TRADE OFF**



x = observations

$$\underline{f(x)} = \beta_0 + \beta_1 X$$

$$\underline{f(x)} = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$\underline{f(x)} = \beta_0 + \beta_1 X + \dots + \beta_k X^k$$

model that goes through all the points

FITTING EXACTLY the data is the worst thing for a predictor since given a x there is variability of y not captured by f. The more we use the data for the model (for the estimate of the parameters) the worst is the model in terms of variability.
(BIAS - VARIANCE trade-off)

We want to judge how good is a model for predictions: $x_0 \xrightarrow{\hat{f}} y_0$ ($\hat{f}$ based on $\mathbb{X}$):

$$\mathbb{E}_{\mathbb{X}}[(y_0 - f(x_0))^2] = (f(x_0) - \hat{f}(x_0))^2 + \mathrm{Var}(\varepsilon_0)$$

this estimate the prediction error given the data $\mathbb{X}$ but with another $\mathbb{X}$ this will change

We want our model to be that good that if we have $\hat{f}_1$ based on $\mathbb{X}_1$ and $\hat{f}_2$ based on $\mathbb{X}_2$ (where $\mathbb{X}_1$ and $\mathbb{X}_2$ are different observations of the same phenomenal) $\Longrightarrow \hat{f}_1 = \hat{f}_2$.

We need to do an average through all the possible $\mathbb{X}$:

$$\mathbb{E}[\mathbb{E}_{\mathbb{X}}[(y_0 - \hat{f}(x_0))^2]] = \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2 + \mathrm{Var}(\varepsilon_0)]$$

($\hat{f}$ depends on the dataset)

$$= \mathbb{E}[(f(x_0) - \hat{f}(x_0) \pm \mathbb{E}[\hat{f}(x_0)])^2] + \mathrm{Var}(\varepsilon_0)$$

$$= \underbrace{\mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] + \mathbb{E}[(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2]}_{(*)} + \underbrace{2\,\mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))]}_{(*1)} + \mathrm{Var}(\varepsilon_0)$$

$$(*) = 2\,\mathbb{E}[\underbrace{(f(x_0) - \mathbb{E}[\hat{f}(x_0)])}_{\text{number}}\underbrace{(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))}_{\text{number}}]$$

$$= 2\,(f(x_0) - \mathbb{E}[\hat{f}(x_0)])\,\cancel{(\mathbb{E}[\hat{f}(x_0)] - \mathbb{E}[\hat{f}(x_0)])}$$

$$= 0$$

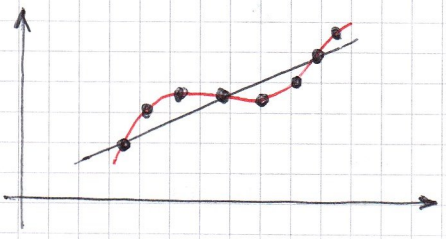
$$(*) = \mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] + \mathbb{E}[(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2]$$

$$= (f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2 + \mathrm{Var}(\hat{f}(x_0))$$

$$\Longrightarrow \boxed{\mathbb{E}[\mathbb{E}_{\mathbb{X}}[(y_0 - \hat{f}(x_0))^2]] = \underbrace{\mathrm{Var}(\varepsilon_0)}_{\substack{\text{variability} \\ \text{of the model}}} + \underbrace{\mathrm{Var}(\hat{f}(x_0))}_{\substack{\text{IRREDUCIBLE} \\ \text{variability}}} + \underbrace{(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2}_{:= \text{Bias}^2}}$$

(how far is the model from what we want to estimate (on average))

$\Longrightarrow$



—— : fits all the points, reduce the bias but the variability is pretty high (OVERFITTING). If we change 1 observ. then the model changes all

—— : linear model: the bias is high but the variability is very low

We want to minimize all $\Longrightarrow$ bias-variance trade off.

RECAP: Data: units (↕) features (↔) → predict data: regression funct. → estimate of the regression function

→ local average? No: curse of dim. → sol. → reduction of dimensionality (PCA)

→ parametric models

GEOMETRY OF THE DATA

We'll always assume to have:
- n statistical units
- p features / variables

$$\mathbb{X} = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & & x_{np} \end{bmatrix} \end{array} \in \mathbb{R}^{n \times p}$$

Data matrix (/ Data frame)

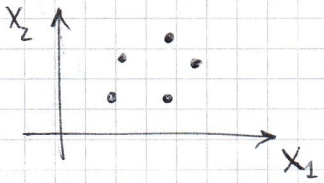
We can look at the data in two ways:

(1) BY ROWS:

$$\underline{x}_1^T = [\, x_{11} \;\; x_{12} \;\cdots\; x_{1p} \,] \in \mathbb{R}^p$$

$$\underline{x}_i^T = [\, x_{i1} \;\; x_{i2} \;\cdots\; x_{ip} \,] \in \mathbb{R}^p$$

i-th row



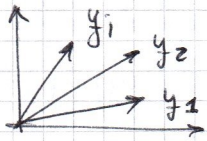
(2) BY COLUMNS:

$$y_1 = [\, x_{11} \;\; x_{21} \;\cdots\; x_{n1} \,]^T \in \mathbb{R}^n = \text{sample from } \underline{X}_1$$

$$y_j = [\, x_{1j} \;\; x_{2j} \;\cdots\; x_{nj} \,]^T \in \mathbb{R}^n = \text{sample from } \underline{X}_j$$

j-th column

(We can represent them as vectors, we'll see why:



in $\mathbb{R}^n$)

---

1. BY COLUMN

The way of summarize the informations coming from a sample is through the distribution of the sample, and we can summarize the distribution with mean and variance. Let's formally define it. Let's look at $\mathbb{X}$ by columns:

$$y_j = [\, x_{1j} \;\; x_{2j} \;\cdots\; x_{nj} \,]^T$$

$$\Longrightarrow \boxed{\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}} \longrightarrow \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \text{sample mean vector}$$

MEAN for variable j

The mean is the baricentrum of the sample:

This is only a partial information.
We want to know about the variability too (the dispersion around the mean), so we introduce the variance.

$$\Longrightarrow \boxed{S_{jj} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \longrightarrow \sqrt{S_{jj}} = \text{STANDARD DEVIATION of } \underline{X}_j$$

VARIANCE for the variable $\underline{X}_j$

(For the moment we want to divide by n and we will consider dividing by $n-1$ later. When we want the variability of the sample we divide by n, when we consider S_{jj} as an estimator of the variance of the entire population then we divide by $n-1$)

What about the covariability between two variables?

$$\Longrightarrow \boxed{\text{Cov}(X_j, X_k) = \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j) := S_{kj}} \qquad k, j = 1, \dots, p$$

COVARIANCE between X_j and X_k

$$Cov(X_j, X_j) = S_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

$\Longrightarrow$ we can introduce a matrix: $S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & & S_{2p} \\ \cdots & & & \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix} \in \mathbb{R}^{p\times p}$, symmetric

COVARIANCE MATRIX
of the sample

$\Longrightarrow$ $r_{kj} = \dfrac{S_{kj}}{\sqrt{S_{kk}\,S_{jj}}} = Cor(X_k, X_j)$ $\longrightarrow$ $r = \begin{bmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & & \\ r_{p1} & \cdots & r_{pp} \end{bmatrix}$, symmetric

CORRELATION
between X_j and X_k
($r_{kj} \in [-1, 1]$)

CORRELATION MATRIX
(covariance matrix normalized by the product of std deviation)

These are the basic tools for # dataset.

Example: X_1 = high of italians, $\bar{X}_1 = 1{,}80$ m, $\sqrt{S_{11}} = 0{,}03$ m

$\Longrightarrow$ $[\bar{x}_1 - 3\sqrt{S_{11}},\ \bar{x}_1 + 3\sqrt{S_{11}}] = [1.71,\ 1.86]$

Whatever is the distribution of X_1, we can say that (almost) 90 % of the population is in the interval (thanks to Chebyshev).

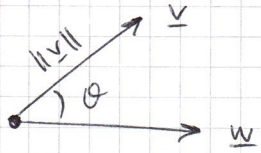
$\Longrightarrow$ $Fr(\bar{x}_1 - k\sqrt{S_{11}} \le X_1 \le \bar{x}_1 + k\sqrt{S_{11}}) \ge 1 - \frac{1}{k^2}$

CHEBYSHEV
alternatively: $P(|X - \mu| \le k\sigma) \ge 1 - \frac{1}{k^2}$

$\Longrightarrow$ Thanks to Chebyshev, whenever we have the mean and the standard deviation we can construct an interval. Note that, if we know that X is gaussian then the Chebyshev inequality changes a little (s.t. the frequency of $(|x_i - \mu| \le k\sqrt{S_{11}})$ is higher than $1 - 1/k^2$).

GEOMETRICAL INTERPRETATION

$\underline{v}, \underline{w} \in \mathbb{R}^n$

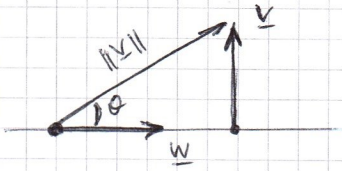


$\langle \underline{v}, \underline{w} \rangle = \underline{v}^T \underline{w}$

$\|\underline{v}\| = \sqrt{\langle \underline{v}, \underline{v} \rangle} = \sqrt{\sum_i v_i^2}$

$$\cos(\theta) = \frac{\langle \underline{v}, \underline{w} \rangle}{\|\underline{v}\| \cdot \|\underline{w}\|} = \frac{\underline{v}^T \underline{w}}{\sqrt{\underline{v}^T\underline{v}}\sqrt{\underline{w}^T\underline{w}}} = \frac{\underline{v}^T \underline{w}}{\sqrt{(\underline{v}^T\underline{v})(\underline{w}^T\underline{w})}} = \frac{\sum_i v_i w_i}{\sqrt{(\sum_i v_i^2)(\sum_i w_i^2)}}$$

Orthogonal projector:



$\Pi_{\underline{v}|\underline{w}} = \Pi_{\underline{v}|\mathcal{L}(\underline{w})}$ = projection of $\underline{v}$ on $\underline{w}$ (on the linear space generated by $\underline{w}$)

$\mathcal{L}(\underline{w}) = \{\underline{z} : \underline{z} = c\underline{w},\ c \in \mathbb{R}\}$
linear space generated by $\underline{w}$

$$\Pi_{\underline{v}|\mathcal{L}(\underline{w})} = \underbrace{\|\underline{v}\|\cos(\theta)}_{\text{length}} \cdot \underbrace{\frac{\underline{w}}{\|\underline{w}\|}}_{\text{direction}} = \|\underline{v}\| \cdot \frac{\underline{v}^T\underline{w}}{\|\underline{v}\| \cdot \|\underline{w}\|} \cdot \frac{\underline{w}}{\|\underline{w}\|}$$

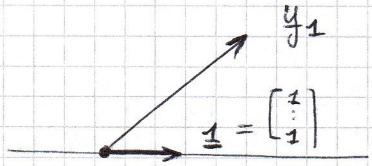
$$= \frac{\underline{v}^T\underline{w}}{\|\underline{w}\|^2} \cdot \underline{w} = \frac{\underline{w}^T\underline{v}}{\underline{w}^T\underline{w}} \cdot \underline{w} = \frac{\underline{w}\,\underline{w}^T}{\underline{w}^T\underline{w}} \cdot \underline{v}$$

ORTHOGONAL PROJECTOR: matrix that project any $\underline{v}$ on $\mathcal{L}(\underline{w})$

$\Longrightarrow$ going back to $X = [y_1 \dots y_p]$, $y_j \in \mathbb{R}^n$ sample from X_j

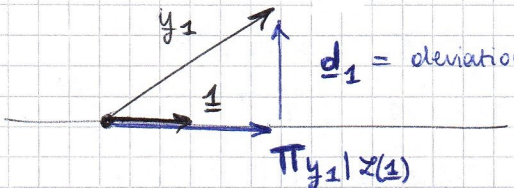
CONSIDER 1 VARIABLE:

(For instance) $y_1 \in \mathbb{R}^n$: y_1 = sample of heights



$\mathcal{L}(\underline{1})$ = space with "no statistic" (because there is no variability, all the components are the same)
If $\underline{v} \in \mathcal{L}(\underline{1}) \Longrightarrow \underline{v} = c\,\underline{1}$, $c \in \mathbb{R} \Longrightarrow \underline{v} = \begin{bmatrix} c \\ \vdots \\ c \end{bmatrix}$

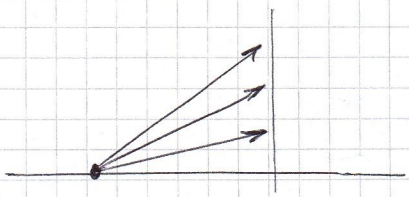
What is the closest vector in $\mathcal{L}(\underline{1})$ to y_1? (The closer approximation where there is no statistics) $\longrightarrow$ the orthogonal projection:



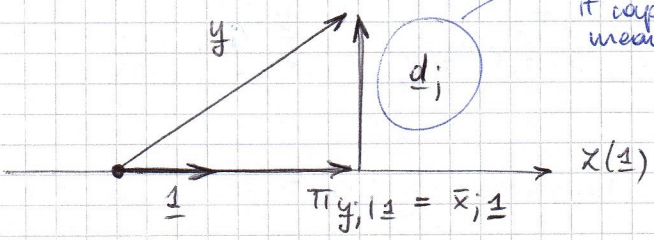
$$\Pi_{y_1|\underline{1}} = \frac{\underline{1}\cdot\underline{1}^T}{\underline{1}^T\cdot\underline{1}}\cdot y_1 = \frac{\sum_{i=1}^n (y_1)_i}{n}\,\underline{1} = \frac{\sum_{i=1}^n x_{i1}}{n}\,\underline{1} = \bar{x}_n \cdot \underline{1} = \begin{bmatrix} \bar{x}_n \\ \vdots \\ \bar{x}_n \end{bmatrix}$$

$\Longrightarrow$ the best approximation we can get for the vector of observations when we require that there's no variability $= \bar{x}_n \cdot \underline{1}$

But we're not capturing the deviations:



Many vectors can have the same mean (the same projection on $\mathcal{L}(\underline{1})$) but all these vectors have different deviation $\longrightarrow$ the mean is not a sufficient summary



DEVIATION VECTOR
it captures informations that mean doesn't give

$$\underline{d}_j = y_j - \bar{x}_j\underline{1} = \begin{bmatrix} x_{1j} - \bar{x}_j \\ x_{2j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix}$$

The longer the vector $\underline{d}_j$ is the worse the approximation is:

$$\|\underline{d}_j\| = \sqrt{\underline{d}_j^T\underline{d}_j} = \sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} = \sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} = \sqrt{n}\sqrt{s_{jj}}$$

$\Longrightarrow$ $\boxed{\|\underline{d}_j\| = \sqrt{n}\sqrt{s_{jj}}}$

this is why s_{jj} talks about variability

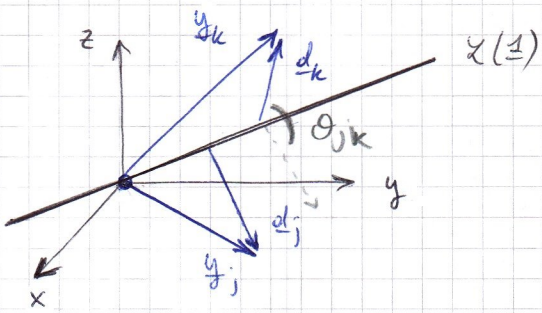
CONSIDER 2 VARIABLES :

y_j , y_k

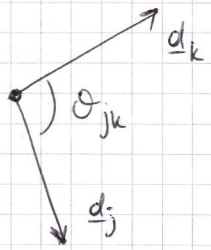
$y_j = \bar{x}_j \underline{1} + \underline{d}_j$

$y_k = \bar{x}_k \underline{1} + \underline{d}_k$

$\in \mathcal{X}(\underline{1})$ $\quad$ $\in \mathcal{X}(\underline{1})^\perp$



Consider the angle θ_{jk} :



(1) $\theta_{jk} = 0$: $\underline{d}_j \in \mathcal{X}(\underline{d}_k) \Longrightarrow \underline{d}_j = \beta \underline{d}_k$, $\beta \in \mathbb{R}$

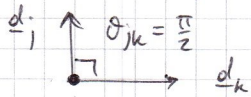
$\Longrightarrow y_j - \bar{x}_j \underline{1} = \beta (y_k - \bar{x}_k \underline{1})$

$\Longrightarrow y_j = \bar{x}_j \underline{1} + \beta y_k - \beta_k \bar{x}_k \underline{1}$

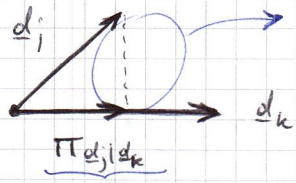
$\Longrightarrow$ there is a linear relation between y_j and y_k ($\nexists$ new informations in y_j that is not contained in y_k :

: if we know one variable then we know them both

(2) $\theta_{jk} = \frac{\pi}{2}$: $\nexists$ informations that are contained in $\underline{d}_k$ related to $\underline{d}_j$:



(3) $0 < \theta_{jk} < \frac{\pi}{2}$: there is some information of $\underline{d}_j$ in $\underline{d}_k$ (more exactly $\pi_{\underline{d}_j | \underline{d}_k}$) :



$\underline{d}_j - \pi_{\underline{d}_j | \underline{d}_k}$ = informations of $\underline{d}_j$ that are not in $\underline{d}_k$

informations contained in $\underline{d}_k$ about $\underline{d}_j$

We need to compute θ_{jk} to know how much one variable talks about the other :

$$\cos\theta_{jk} = \frac{\underline{d}_j^T \underline{d}_k}{\|\underline{d}_j\| \cdot \|\underline{d}_k\|} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\left(\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2\right)\left(\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2\right)}} \cdot \frac{\frac{1}{n}}{\frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}}}$$

$$= \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}$$

$$= \mathrm{Corr}(X_j, X_k) = r_{jk}$$

that's why $r_{jk} \in [-1, 1]$, it's a cosine !

$\Longrightarrow$ $\boxed{\cos(\theta_{jk}) = r_{jk}}$:

- $\theta_{jk} = 0 \Longrightarrow \cos(\theta_{jk}) = 1 \Longrightarrow r_{jk} = 1 \Longrightarrow \underline{d}_j \in \mathcal{X}(\underline{d}_k)$
- $\theta_{jk} = \frac{\pi}{2} \Longrightarrow \cos(\theta_{jk}) = 0 \Longrightarrow r_{jk} = 0 \Longrightarrow \underline{d}_j \perp \underline{d}_k$

Conclusion : mean and variance (standard deviation)

2. BY ROWS

Now let's look at $\mathbb{X}$ by rows : $\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix}$ $\underline{x}_i \in \mathbb{R}^p$ $i = 1, \dots, n$

We need some models for these data:
we'll assume that $\underline{x}_i$ are the realizations of a random vector $\underline{X}_i$ and the basic assumption is that :

$$\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X} \in \mathbb{R}^p$$

We need tools to manage with random vectors: $\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ → random variable

When we work with random variables we need to introduce the probability law (≡distribution) of $\underline{X}$ (it gives all the informations about the random variable) :

$$\mathcal{L}_{\underline{X}} : \mathcal{B}(\mathbb{R}^p) \longrightarrow [0,1] \quad ; \quad \mathcal{L}_{\underline{X}} = \mathbb{P}(\underline{X} \in B) \quad \forall B \in \mathcal{B}(\mathbb{R}^p)$$

$\mathcal{L}_{\underline{X}}$: law of $\underline{X}$ — $\mathcal{B}(\mathbb{R}^p)$: Borel sets of $\mathbb{R}^p$

We'll always be in the special case : $\mathcal{L}_{\underline{X}}(B) = \int_B f_{\underline{X}}(\underline{t})\, d\underline{t} \quad \forall B \in \mathcal{B}(\mathbb{R}^p)$

$f_{\underline{X}}$: DENSITY OF $\underline{X}$

Besides the density we want something that summarize the entire distribution:

$$\mathbb{E}[\underline{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_p] \end{bmatrix} = \underline{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

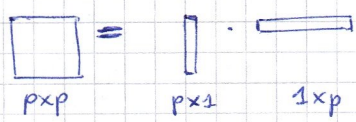
MEAN of $\underline{X}$
Expected value of $\underline{X}$

$$\sigma_{jk} = \mathbb{E}[(X_j - \mu_j)(X_k - \mu_k)] = [\Sigma]_{jk} \qquad j,k = 1, \dots, p$$

COVARIANCE between X_j and X_k

$\Sigma = [\sigma_{jk}] \in \mathbb{R}^{p \times p}$ matrix := COVARIANCE MATRIX of $\underline{X}$

$$\Sigma = \mathbb{E}[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T]$$



Notice that :

$$\sigma_{jj} = \mathbb{E}[(X_j - \mu_j)^2] = \mathrm{Var}(X_j)$$

Define now :

$$V = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{pp} \end{bmatrix} = \mathrm{diag}(\sigma_{11}, \dots, \sigma_{pp}) \qquad \sigma_{ii} \geq 0 \quad \forall i$$

$$V^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & & \\ & \ddots & \\ & & \sqrt{\sigma_{pp}} \end{bmatrix} \xrightarrow{\sigma_{ii} > 0 \ \forall i} V^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$\rho = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}}$$

CORRELATION MATRIX
$\in \mathbb{R}^{p \times p}$, symmetric

We want to work with LINEAR COMBINATIONS OF the components of $\underline{X}$:

$\underline{c} \in \mathbb{R}^p$

$$\underline{c}^T \underline{X} = c_1 X_1 + c_2 X_2 + \dots + c_p X_p \in \mathbb{R}$$

- $\mathbb{E}[\underline{c}^T \underline{X}] = c_1 \mathbb{E}[X_1] + \dots + c_p \mathbb{E}[X_p]$
  $= c_1 \mu_1 + \dots + c_p \mu_p$
  $= \underline{c}^T \underline{\mu}$
- $\text{Var}(\underline{c}^T \underline{X}) = \underline{c}^T \Sigma \underline{c}$

(**Note**: $\text{Var}(c_1 X_1 + c_2 X_2) = c_1^2 \text{Var}(X_1) + c_2^2 \text{Var}(X_2) + 2 c_1 c_2 \text{Cov}(X_1, X_2)$)

Not only linear combinations, we want to consider K LINEAR COMBINATIONS of $\underline{X}$:

$C \in \mathbb{R}^{k \times p}$ matrix

$$C = \begin{bmatrix} \underline{c}_1^T \\ \vdots \\ \underline{c}_p^T \end{bmatrix} \quad \underline{c}_i \in \mathbb{R}^p \implies C\underline{X} = \begin{bmatrix} \underline{c}_1^T \underline{X} \\ \vdots \\ \underline{c}_p^T \underline{X} \end{bmatrix}$$

k - linear combinations of the components of $\underline{X}$

- $\mathbb{E}[C\underline{X}] = C\underline{\mu}$
- $\text{Cov}(C\underline{X}) = C \Sigma C^T$ → Particular case: $C = \underline{c}^T = [c_1 \dots c_p]$ (1xp matrix)
  $\implies \text{Cov}(C\underline{X}) = \text{Var}(\underline{c}^T \underline{X})$
  $= C \Sigma C^T$
  $= \underline{c}^T \Sigma \underline{c}$
  $\implies$ coherent with the case before

ESTIMATORS

We have data $\mathbb{X}$ ($\in \mathbb{R}^{n \times p}$) and the model $\underline{X}$ random vector.
Can we do inference with this? Can we use the data to estimate $\underline{\mu}$, Σ?
Can we say something from the sample that is true for the entire population?

Estimator of $\underline{\mu}$:

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \xleftarrow{\text{estimated}} \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \Longleftarrow \text{Data}$$

Why does it work?
We again consider $\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X}$ that generate $\underline{x}_1, \dots, \underline{x}_n$ (Data):

$$\bar{\underline{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \xleftarrow{\text{realization of}} \bar{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$$

Note: there is a difference between the ESTIMATOR (algorithm) and the ESTIMATE (product (output) of the algorithm). We want to know how good is an estimator, not an estimate (an estimate is good if it's produced by a good estimator).

Prop. Assuming $\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X}$, $\mathbb{E}[\underline{X}] \underline{\mu}$, $\text{Cov}(\underline{X}) = \Sigma$:

1. $\mathbb{E}[\bar{\underline{X}}] = \underline{\mu}$ (unbiased estimator: on average this estimator is good)
2. $\text{Cov}(\bar{\underline{X}}) = \frac{1}{n} \Sigma$

proof.

(1) $$\mathbb{E}[\bar{\underline{X}}] = \mathbb{E}\begin{bmatrix} \frac{1}{n} \sum_i X_{i1} \\ \vdots \\ \frac{1}{n} \sum_i X_{ip} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \sum_i \mu_1 \\ \vdots \\ \frac{1}{n} \sum_i \mu_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}.$$

(2) $\mathrm{Cov}(\bar{\underline{X}}) = \mathbb{E}\left[(\bar{\underline{X}}-\underline{\mu})(\bar{\underline{X}}-\underline{\mu})^T\right]$

$= \mathbb{E}\left[\left(\frac{1}{n}\sum_i(\underline{X}_i)-\underline{\mu}\right)\left(\frac{1}{n}\sum_i(\underline{X}_i)-\underline{\mu}\right)^T\right]$ $\quad \bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^n \underline{X}_i$

$= \mathbb{E}\left[\frac{1}{n}\left(\sum_i(\underline{X}_i-\underline{\mu})\right)\frac{1}{n}\left(\sum_i(\underline{X}_i-\underline{\mu})\right)^T\right]$

$= \frac{1}{n^2}\,\mathbb{E}\left[\sum_{\substack{i=1\\k=1}}^n(\underline{X}_i-\underline{\mu})(\underline{X}_k-\underline{\mu})^T\right]$

$= \frac{1}{n^2}\sum_{\substack{i=1\\k=1}}^n \underbrace{\mathbb{E}\left[(\underline{X}_i-\underline{\mu})(\underline{X}_k-\underline{\mu})^T\right]}$

$$\mathbb{E}\left[(\underline{X}_i-\underline{\mu})(\underline{X}_k-\underline{\mu})^T\right] = \begin{cases} 0 = \mathbb{E}[(\cdot)(\cdot)] = \mathbb{E}[(\cdot)]\,\mathbb{E}[(\cdot)] & i \neq k \\ \Sigma = \mathbb{E}\left[(\underline{X}_i-\underline{\mu})(\underline{X}_i-\underline{\mu})^T\right] & i = k \end{cases}$$

$= \frac{1}{n^2}\sum_{i=1}^n(\Sigma) = \frac{n\Sigma}{n^2} = \frac{1}{n}\Sigma$ ■

Estimator of Σ:

$$S = \frac{1}{n}\sum_{i=1}^n(\underline{X}_i-\bar{\underline{X}})(\underline{X}_i-\bar{\underline{X}})^T \qquad \in \mathbb{R}^{p\times p} \quad \text{(random matrix)}$$

$$S = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & & \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} \longleftarrow \text{realization of } S$$

sample covariance matrix

Prop. $\mathbb{E}[S] = \frac{n-1}{n}\Sigma$

Corollary. $\mathbb{E}\left[\frac{n}{n-1}S\right] = \Sigma \Longrightarrow \frac{n}{n-1}S$ is an unbiased estimator of Σ

$$\frac{n-1}{n}S = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n(\underline{X}_i-\bar{\underline{X}})(\underline{X}_i-\bar{\underline{X}})^T = \frac{1}{n-1}\sum_{i=1}^n(\underline{X}_i-\bar{\underline{X}})(\underline{X}_i-\bar{\underline{X}})^T$$

$\Longrightarrow$ from now on: $S = \frac{1}{n-1}\sum_{i=1}^n(\underline{X}_i-\bar{\underline{X}})(\underline{X}_i-\bar{\underline{X}})^T$

(we want the covariance of the population, not of the sample, when we'll need S (the old) we'll write S_n. Note that BIG DATA uses an n so large that there is no difference)

Focus on S:

$$\underline{d}_j = \underline{y}_j - \Pi_{\underline{y}_j\mathbf{1}}\underline{y}_j = \underline{y}_j - \frac{\underline{1}\cdot\underline{1}^T}{\underline{1}^T\cdot\underline{1}}\underline{y}_j = \underbrace{\left(I - \frac{\underline{1}\cdot\underline{1}^T}{\underline{1}^T\underline{1}}\right)}\underline{y}_j$$

orthogonal projector on $\mathcal{L}(\underline{1})$

$$\underline{d} = [\underline{d}_1 \cdots \underline{d}_p] \longrightarrow \boxed{\underline{d} = \left(I - \frac{\underline{1}\cdot\underline{1}^T}{\underline{1}^T\underline{1}}\right)X}$$

$\underline{d}_1$: deviation of variable 1 from its mean

DEVIATION MATRIX ($n\times p$ matrix)

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ \vdots & & & \\ s_{n1} & \cdots & & s_{np} \end{bmatrix} = \frac{1}{n-1} \begin{bmatrix} \sum_i (x_{i1}-\bar{x} & \sum_i (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) & \cdots \\ \vdots & & \end{bmatrix}$$

$$= \frac{1}{n} \begin{bmatrix} \underline{d}_1^T \underline{d}_1 & \underline{d}_2^T \underline{d}_2 & \cdots & \underline{d}_1^T \underline{d}_p \\ \vdots & & & \vdots \\ & & & \underline{d}_p^T \underline{d}_p \end{bmatrix}$$

$$\Longrightarrow S = \frac{1}{n} d^T d = \frac{1}{n} \mathbb{X}^T \underbrace{\left(I - \frac{\underline{1}\underline{1}^T}{\underline{1}^T\underline{1}}\right)^T}_{\text{since it's an orthogonal projector then it's symmetric } (T^T = T) \text{ and idempotent } (T^2 = T)} \left(I - \frac{\underline{1}\underline{1}^T}{\underline{1}^T\underline{1}}\right) \mathbb{X}$$

$$\Longrightarrow \boxed{S = \frac{1}{n-1} \mathbb{X}^T \left(I \frac{\underline{1}\underline{1}^T}{\underline{1}^T\underline{1}}\right) \mathbb{X}}$$

VARIABILITY IN A MULTIVARIATE SENSE

There are two possibilities:
- Generalized variance : $\text{Det}(S)$
- Total variance : $\text{Tr}(S)$

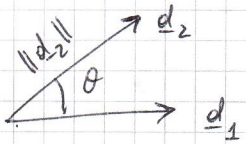
In $\mathbb{R}$ the variability is simply captured by variance.

- **$\text{Det}(S)$**

Consider the case $p = 2$:

$$S = \frac{1}{n-1} d^T d = \frac{1}{n-1} \begin{bmatrix} \underline{d}_1^T \underline{d}_1 & \underline{d}_1^T \underline{d}_2 \\ \underline{d}_2^T \underline{d}_1 & \underline{d}_2^T \underline{d}_2 \end{bmatrix}$$

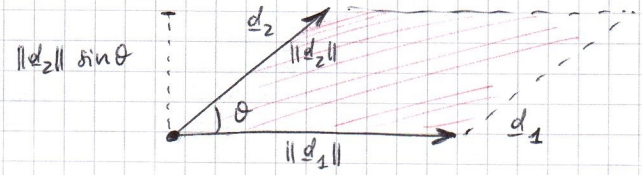
$$= \frac{1}{n-1} \begin{bmatrix} \|\underline{d}_1\|^2 & \|\underline{d}_1\| \cdot \|\underline{d}_2\| \cos\theta \\ \|\underline{d}_1\| \cdot \|\underline{d}_2\| \cos\theta & \|\underline{d}_2\|^2 \end{bmatrix}$$



$$\text{Det}(S) = \left[\|\underline{d}_1\|^2 \|\underline{d}_2\|^2 - \|\underline{d}_1\|^2 \|\underline{d}_2\|^2 \cos^2(\theta)\right] \left(\frac{1}{n-1}\right)^2$$

$$= \|\underline{d}_1\|^2 \|\underline{d}_2\|^2 \left(1 - \cos^2\theta\right) \left(\frac{1}{n-1}\right)^2$$

$$= \|\underline{d}_1\|^2 \|\underline{d}_2\|^2 \sin^2\theta \left(\frac{1}{n-1}\right)^2$$



$\Longrightarrow \text{Det}(S) \propto (\text{Area of the parallelogram } (\underline{d}_1, \underline{d}_2))^2$

$\Longrightarrow$ [$\text{Det}(S)$ increase $\Longrightarrow$ Area of the parallelogram increase]

can increase for two reasons : ① because of the angle (max area with $\theta = \frac{\pi}{2}$) or ② because of the length of $\underline{d}_1, \underline{d}_2$

$\Longrightarrow$ If $\text{Det}(S) = 0 \Longrightarrow$ $(\|\underline{d}_1\| \vee \|\underline{d}_2\|) = 0$ or $\theta = 0$ ($\Longrightarrow$ correlation = 1)

$\Longrightarrow$ If $\text{Det}(S) = 0$ it doesn't mean that there is no variability

- **Tr(S)**

$$Tr(S) = \left(\|\underline{d}_1\|^2 + \|\underline{d}_2\|^2\right)\frac{1}{n-1}$$

The total variance is capturing the sum of the marginal variabilities of the variables.

In general case (general p) :
- $Det(S) \propto Vol^2(\text{parallelotope}(\underline{d}_1, .., \underline{d}_p))$
- $Tr(S) = \|\underline{d}_1\|^2 + \|\underline{d}_2\|^2 + \dots + \|\underline{d}_p\|^2$

Note : both (generalized variance and total variance) are interesting. One is talking about volume, the other about perimeter. We can't, for instance, decide to use only generalized variance (since it depends also on correlation).

Prop. $Det(S) = 0 \iff \underline{d}_1, .., \underline{d}_p$ are linearly dependent
(i.e. $\exists \underline{c} \neq \underline{0}$ s.t. $d\underline{c} = \underline{0} = c_1\underline{d}_1 + c_2\underline{d}_2 + .. + c_p\underline{d}_p$)

proof.

($\Leftarrow$) suppose $\underline{d}_1, .., \underline{d}_p$ are linearly dependent ($\exists \underline{c} \in \mathbb{R}^p$ s.t. $d\underline{c} = \underline{0}$)

$S = \frac{1}{n-1} d^T d \implies S\underline{c} = \frac{1}{n-1} d^T d\underline{c} = \underline{0}$

$\implies$ the columns of S are linearly dependent $\implies det(S) = 0$

($\Rightarrow$) suppose $det(S) = 0 \implies \exists \underline{c} \neq \underline{0}$ s.t. $S\underline{c} = \underline{0}$
(column must be linearly dependent)

$\implies \underline{c}^T S \underline{c} = 0 \implies \frac{1}{n-1} \underline{c}^T d^T d \underline{c} = 0 \implies \frac{1}{n-1} \|d\underline{c}\|^2 = 0$

$\implies d\underline{c} = \underline{0} \implies \underline{d}_1, .., \underline{d}_p$ are linearly dependent ■

So, $Det(S) = 0 \implies \underline{d}_1, .., \underline{d}_p$ are linearly dependent ($c_1\underline{d}_1 + .. + c_p\underline{d}_p = \underline{0} \quad \underline{c} \neq \underline{0}$),
without loss of generality let's assume that $c_1 \neq 0$:

$$\implies \underline{d}_1 = -\sum_{i=2}^{p} \frac{c_i}{c_1} \underline{d}_i \qquad (\underline{d}_k = \underline{y}_k - \bar{x}_k \underline{1})$$

$$\implies \boxed{\underline{y}_1 = \bar{x}_1 \underline{1} - \sum_{i=2}^{p} \frac{c_i}{c_1} (\underline{y}_i - \bar{x}_i \underline{1})} \implies$$

there is a perfect linear relationship between the var 1 and all the other variables, so the informations coming from $\underline{y}_1$ are useless if we have the ones from $\underline{y}_j$, $j = 2, .., p$

Prop. $\mathbb{X} \in \mathbb{R}^{n \times p}$, if $p \geq n \implies det(S) = 0$

(If we have too many variables wrt the sample size then it's sure that we'll find a perfect linear relationship between them)

proof.

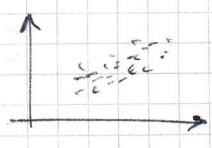
$d = [\underline{d}_1 \cdots \underline{d}_p] \quad \underline{d}_i \in \mathbb{R}^n$ and $\underline{d}_i \in \mathcal{Y}^{\perp}(\underline{1})$

$dim(\mathcal{Y}^{\perp}(\underline{1})) = n-1 \implies$ we can have at most $n-1$ linearly independent variables

If $p \geq n \implies \underline{d}_1, .., \underline{d}_p$ are linearly dependent $\implies det(S) = 0$ ■

Example : Raw data : (received data ≠ data matrix)

x	y
x_1	y_1
⋮	⋮
x_n	y_n



↑ we can create the data matrix that we want (we are free to take what variables we want : $x, x^2, ..$)

Data matrix $\mathbb{X} = \begin{bmatrix} x_1 & x_1^2 & \cdots & x_1^p & y_1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_n & x_n^2 & & x_n^p & y_n \end{bmatrix}$ $\underbrace{\qquad\qquad}_{(*)}$

Consider for instance $p = n-1$ $\rightarrow$ # columns of $\mathbb{X} = p+1 = n$
$\Longrightarrow$ $\det(S) = 0$.
Consider for instance $(*)$ to be linearly independent
$\rightarrow$ $y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p$
$\rightarrow$ perfect fit! Are we happy with that? **No**, this is an overfitting
(we always have to think at the bias-variance trade-off)
$\Longrightarrow$ this model is terrible for prediction

Necessary condition for $\text{Det}(S) > 0$: $n \geq p+1$

SPECTRAL DECOMPOSITION OF S

Since S is a $p \times p$ symmetric and real ($s_{ij} \in \mathbb{R}$) matrix

$\Longrightarrow$ $\begin{cases} \exists\, \lambda_1, \dots, \lambda_p \in \mathbb{R} \\ \exists\, \underline{e}_1, \dots, \underline{e}_p \in \mathbb{R}^p \end{cases}$ s.t. $\underline{e}_j^T \underline{e}_i = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$ (orthonormal system)

s.t. $S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T$ (spectral decomposition of a symmetric real matrix : in fact $(\lambda_i, \underline{e}_i)$ are an eigenvalue and eigenvector couple for S: $S\underline{e}_i = \lambda_i \underline{e}_i$ $i = 1, \dots, p$)

$P := [\underline{e}_1, \dots, \underline{e}_p]$, $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$ $\Longrightarrow$ $\boxed{S = P \Lambda P^T}$

$\Longrightarrow$ $\begin{cases} \det(S) = \prod_{i=1}^{p} \lambda_i & \text{(generalized variance)} \\ \text{Tr}(S) = \sum_{i=1}^{p} \lambda_i & \text{(total variance)} \end{cases}$

Prop. S is positive semi-definite. $(\lambda_i \geq 0 \ \forall i)$
If $\text{Det}(S) \neq 0 \Longrightarrow S$ is positive definite. $(\lambda_i > 0 \ \forall i)$

proof.

We need to prove that $\underline{c}^T S \underline{c} \geq 0 \quad \forall \underline{c} \in \mathbb{R}^p$.

$\underline{c}^T S \underline{c} = \frac{1}{n-1} \underline{c}^T D D^T \underline{c} = \frac{1}{n-1} \| D\underline{c} \|^2 \geq 0$

Suppose $\exists\, \underline{c} \neq \underline{0}$ s.t. $\underline{c}^T S \underline{c} = 0$ (semi-def.)

$\Longrightarrow$ $\| D\underline{c} \| = 0 \Longrightarrow D\underline{c} = \underline{0} \Longrightarrow c_1 \underline{d}_1 + \cdots + c_p \underline{d}_p = \underline{0}$

$\Longrightarrow$ $\underline{d}_1, \dots, \underline{d}_p$ are linearly dependent $\Longrightarrow$ $\text{Det}(S) = 0$

So if $\text{Det}(S) \neq 0 \Longrightarrow \nexists\, \underline{c} \Longrightarrow S$ positive definite $\Longrightarrow$ $\text{Det}(S) > 0$ (not only $\neq 0$) ■

Notations : from now on $S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

INDUCED DISTANCE (induced from S)

Assume that $\text{Det}(S) > 0$ $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0)$

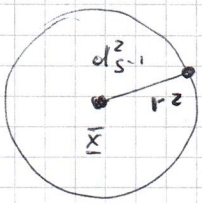
$S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T \Longrightarrow S^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i^T$ $\leftarrow$ induces a metric on $\mathbb{R}^p$

$\underline{x}, \underline{y} \in \mathbb{R}^p$: $\boxed{d^2_{S^{-1}}(\underline{x}, \underline{y}) = (\underline{x} - \underline{y})^T S^{-1} (\underline{x} - \underline{y})}$

MAHALANOBIS DISTANCE
(it's a (distance)^2 in $\mathbb{R}^p$, we can prove it with spectral decomposition)

Let's consider the neighborhood with this distance:

$$\mathcal{E}_{r^2, S^{-1}}(\bar{x}) := \{ x \in \mathbb{R}^p : d^2_{S^{-1}}(x, \bar{x}) \leq r^2 \}$$



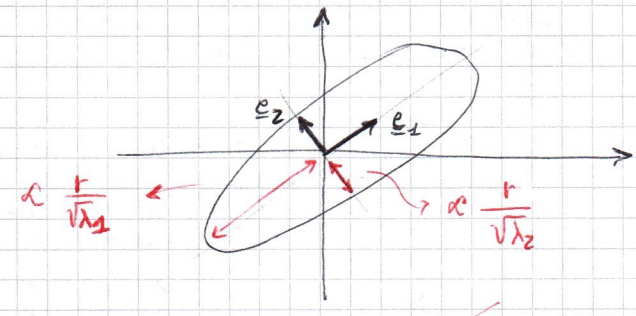
$\mathcal{E}_{r^2, S^{-1}}$ is a circle when we're using Mahalanobis' glasses, but what do we see if we use the euclidean glasses?

$$\mathcal{E}_{r^2, S^{-1}}(\bar{x}) = \{ x \in \mathbb{R}^p : \underbrace{(x - \bar{x})^T S^{-1} (x - \bar{x})}_{\text{in the euclidean sense this is an ellipse centered in } \bar{x}} \leq r^2 \}$$

Generic case: if we consider a generic positive definite matrix B:

$B = \sum_{i=1}^{p} \lambda_i e_i e_i^T$

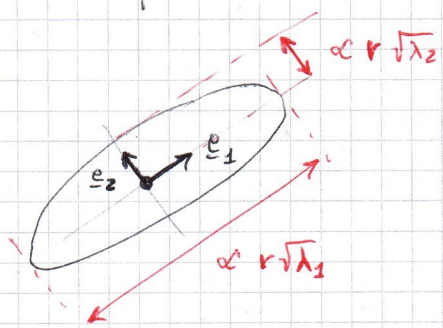
$x^T B x = r^2 \longrightarrow$



Our case:

$S^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} e_i e_i^T$

$r^2, S^{-1}(x) \longrightarrow$



$$\text{Volume}\left(\mathcal{E}_{r^2, S^{-1}}(\bar{x})\right) = K_p\, r^p \sqrt{\prod_{i=1}^{p} \lambda_i} = K_p\, r^p \sqrt{\text{Det}(S)}$$

(in 2D it is the area)

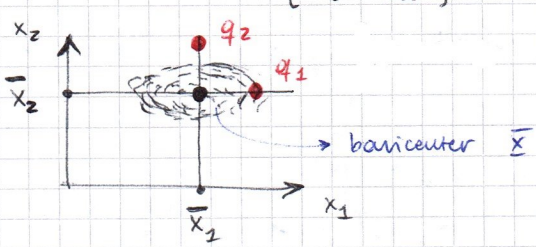
K_p: constant depending on the dimension p

$\sqrt{\text{Det}(S)}$: the larger the generalized variance, the larger is the neighborhood (given the same r) around the mean

$\Longrightarrow$ the neighborhood depends on the generalized variance.

But, why the Mahalanobis distance is the right one for capturing the distance between points considering the variability of the data?

Example: $p = 2$, $S = \begin{bmatrix} s_{11} & 0 \\ 0 & s_{22} \end{bmatrix}$ $\quad s_{11} > s_{22}$



q_1 and q_2 are s.t. $d(\bar{x}, q_1) = d(\bar{x}, q_2)$, so they're equivalently far from $\bar{x}$ considering the euclidean distance. But are they the same from a statistical point of view? The individual q_1 has the same characteristics wrt (x_1, x_2) as individual q_2? **No**, q_1 is in the cloud, q_2 not $\Longrightarrow$ the euclidean distance is not a good distance

$\Longrightarrow$ Standardize data first (Why? Because we notice that the variance of x_1 and x_2 is not the same. We want to measure distances in terms of statistical units : std deviations, not in other units (meter, kg, ...))

$$q_1 = \begin{bmatrix} q_1 \\ \bar{x}_2 \end{bmatrix} \longrightarrow \underset{\text{euclidean}}{d}\left(std(q_1), std(\underline{\bar{x}})\right) = \frac{|q_1 - \bar{x}_1|}{\sqrt{s_{11}}} \quad (1)$$

$$q_2 = \begin{bmatrix} \bar{x}_1 \\ q_2 \end{bmatrix} \longrightarrow d\left(std(q_2), std(\underline{\bar{x}})\right) = \frac{|q_2 - \bar{x}_2|}{\sqrt{s_{22}}} \quad (2)$$

$\Longrightarrow$ since $s_{11} > s_{22} \longrightarrow (1) < (2)$

In the general case (general point) :

$$q = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \longrightarrow d\left(std(q), std(\underline{\bar{x}})\right) = \sqrt{\frac{(q_1 - \bar{x}_1)^2}{s_{11}} + \frac{(q_2 - \bar{x}_2)^2}{s_{22}}} = \sqrt{(\underline{q} - \underline{\bar{x}})^T S^{-1} (\underline{q} - \underline{\bar{x}})}$$

(MAHALANOBIS DISTANCE)

$\Longrightarrow$ the Mahalanobis' distance is just the euclidean distance after we standardize the data. So we'll always consider :

$$\mathcal{E}_{r^2, S^{-1}}(\underline{\bar{x}}) = \left\{ \underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\bar{x}})^T S^{-1} (\underline{x} - \underline{\bar{x}}) \leq r^2 \right\}$$

$$\Longrightarrow \text{Area}\left(\mathcal{E}_{r^2, S^{-1}}(\underline{\bar{x}})\right) \propto r^2 \sqrt{Det(S)} = r^2 \sqrt{s_{11} s_{22}}$$

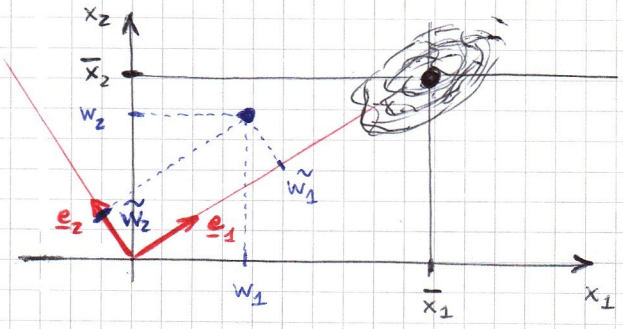
So, If $Det(S)$ or $Tr(S)$ increase $\Longrightarrow$ $\text{Area}\left(\mathcal{E}_r(\underline{\bar{x}})\right)$ increase

(larger the variability $\Longrightarrow$ larger the neighborhood (same r))

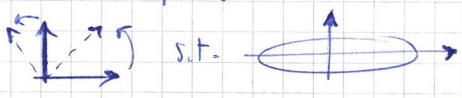
Is this example a special case?

- $p = 2 \longrightarrow$ easy to extend to general $p \geq 1$
- S diagonal is special, but ...

Let $p \geq 1$ and let S be s.t. $Det(S) > 0$



We can always find a reference system s.t. S is diagonal w.r.t. that system



We want to introduce the system identified by the eigenvalues of S $\left(= \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T\right)$

OLD SYSTEM	NEW SYSTEM
$\underline{w} = [x_1 \; x_2 \; \dots \; x_p]^T$	$\tilde{\underline{w}} = [\underline{e}_1^T \underline{w} \;\; \underline{e}_2^T \underline{w} \dots \underline{e}_p^T \underline{w}]^T = P^T \underline{w}$
$\mathcal{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix}$	$\tilde{\mathcal{X}} = \begin{bmatrix} (P^T \underline{x}_1)^T \\ \vdots \\ (P^T \underline{x}_n)^T \end{bmatrix} = \begin{bmatrix} \underline{x}_1^T P \\ \vdots \\ \underline{x}_n^T P \end{bmatrix} = \mathcal{X} P$
$S = \frac{1}{n-1} \mathcal{X} \left(I - \frac{\underline{1}\underline{1}^T}{\underline{1}^T \underline{1}}\right) \mathcal{X}$	$\tilde{S} = \frac{1}{n-1} \tilde{\mathcal{X}}^T \left(I - \frac{\underline{1}\underline{1}^T}{\underline{1}^T \underline{1}}\right) \tilde{\mathcal{X}}$ $= \frac{1}{n-1} P^T \mathcal{X}^T \left(I - \frac{\underline{1}\underline{1}^T}{\underline{1}^T \underline{1}}\right) \mathcal{X} P$ $= P^T S P = P^T P \Lambda P^T P = \Lambda$ (*) $= [\lambda_1 \dots \lambda_p]$

(*) $P^T P = I$, why? I. P is an orthogonal matrix

II. $P^T P = \begin{bmatrix} \underline{e}_1^T \\ \vdots \\ \underline{e}_p^T \end{bmatrix} [\underline{e}_1, \dots, \underline{e}_p] = \begin{bmatrix} \underline{e}_1^T \underline{e}_1 = 1 & \underline{e}_1^T \underline{e}_2 = 1 & \dots & \underline{e}_1^T \underline{e}_p = 1 \\ \underline{e}_2^T \underline{e}_1 = 1 & \underline{e}_2^T \underline{e}_2 = 1 & \dots & \\ & & \dots & \end{bmatrix} = I$

$\Longrightarrow$ In this new system (reference system of eigenvectors) the covariances are simply 0, the only thing left are only the variances.

Take home message: there is always a reference system for which the coordinates are uncorrelated $\Longrightarrow$ correlation is not a property of the data.

Do we lose some informations using this system? No.

- Generalized variance: $\text{Det}(S) = \prod_{i=1}^{p} \lambda_i = \text{Det}(\hat{S}) = \text{Det}(\Lambda) = \prod_{i=1}^{p} \lambda_i$
- Total variance: $\text{Tr}(S) = \sum_{i=1}^{p} \lambda_i = \text{Tr}(\hat{S}) = \text{Tr}(\Lambda) = \sum_{i=1}^{p} \lambda_i$

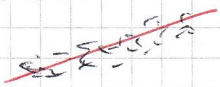
Observation: $\|\underline{w}\|^2 = \underline{w}^T \cdot \underline{w} = \tilde{\underline{w}}^T \tilde{\underline{w}} = \|\tilde{\underline{w}}\|^2 = \underline{w}^T P P^T \underline{w}$

↑ since we're just changing reference, the length of the vectors remain the same

$\longrightarrow$ $\underline{w}^T \cdot \underline{w} = \underline{w}^T P P^T \underline{w}$ (!)

PRINCIPAL COMPONENT ANALYSIS (PCA)

Consider the milky way:



the linear structure that we see is not the correlation of the stars, it's just the first principal component.

Let's refer to the model (not the data):

$\underline{X}$ = random vector in $\mathbb{R}^p$: $\mathbb{E}[\underline{X}] = \underline{\mu}$, $\text{Cov}(\underline{X}) = \Sigma$

If we take $\underline{a} \in \mathbb{R}^p$ we can do a linear comb.:

$$\underline{a}^T \underline{X} = a_1 X_1 + \dots + a_p X_p$$

Problem: find $\underline{a}$ s.t. $\text{Var}(\underline{a}^T \underline{X})$ is maximum.

(We want to find the linear combination of the components of the vector s.t. the variability is maximized)

Is the problem well posed?

Suppose we find the solution: $\underline{a} = [10, 0, \dots, 0]^T \Longrightarrow$ max $\text{Var}(\underline{a}^T \underline{X})$ is reached by: $\underline{a}^T \underline{X} = 10 X_1$

But: $10\,\underline{a} = [100, 0, \dots, 0]^T \Longrightarrow \text{Var}(10\,\underline{a}^T \underline{X}) = 100\,\text{Var}(\underline{a}^T \underline{X})$

$\Longrightarrow$ $\underline{a}$ cannot be the max $\Longrightarrow$ the problem is bad posed

$\Longrightarrow$ **Problem**: find $\underline{a}$ s.t. $\|\underline{a}\| = 1$ and $\text{Var}(\underline{a}^T \underline{X})$ is maximized

Note that:

$$\max_{\underline{a} \in \mathbb{R}^p : \|\underline{a}\| = 1} \text{Var}(\underline{a}^T \underline{X}) = \max_{\underline{a} \in \mathbb{R}^p : \|\underline{a}\| = 1} \underline{a}^T \Sigma\, \underline{a} = \max_{\underline{a} \in \mathbb{R}^p} \frac{\underline{a}^T \Sigma\, \underline{a}}{\underline{a}^T \underline{a}}$$

$\text{Var}(\underline{a}^T \underline{X}) = \underline{a}^T \Sigma\, \underline{a}$

equivalent since $\left\| \frac{\underline{a}}{\|\underline{a}\|} \right\| = 1$ (we transfer $\|\underline{a}\| = 1$ in the objective function)

Lemma: Let $B \in \mathbb{R}^{p \times p}$ positive semidefinite and $B = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T$ its spectral decomp. Then:

1. $\max_{\underline{x} \in \mathbb{R}^p} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_1$, $\operatorname{argmax}(\cdot\cdot) = \underline{e}_1$

2. $\max_{\substack{\underline{x} \in \mathbb{R}^p \\ \underline{x} \perp \underline{e}_1}} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_2$, $\operatorname{argmax}(\cdot\cdot) = \underline{e}_2$

⋮

p. $\max_{\substack{\underline{x} \in \mathbb{R}^p \\ \underline{x} \perp \underline{e}_1 \perp \dots \perp \underline{e}_{p-1}}} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_p = \min_{\underline{x} \in \mathbb{R}^p} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}}$

proof.

1. $B = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T = P \Lambda P^T$

$$\longrightarrow \quad \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \frac{\underline{x}^T P \Lambda P^T \underline{x}}{\underline{x}^T P P^T \underline{x}} \qquad \forall \underline{x} \in \mathbb{R}^p \qquad (*)$$

we proved: $\underline{w}^T \underline{w} = \underline{w}^T P P^T \underline{w}$

$$\underline{y} := P^T \underline{x} \implies (*) = \frac{\underline{y}^T \Lambda \underline{y}}{\underline{y}^T \underline{y}} = \frac{\sum_{i=1}^{p} (\lambda_i y_i)^2}{\sum_{i=1}^{p} (y_i)^2} \leq \lambda_1 \frac{\sum_{i=1}^{p} (y_i)^2}{\sum_{i=1}^{p} (y_i)^2}$$

since they're ordered

If $\underline{x} = \underline{e}_1$: $\frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \frac{\underline{e}_1^T B \underline{e}_1}{\underline{e}_1^T \underline{e}_1} = \lambda_1 \underline{e}_1^T \underline{e}_1 = \lambda_1$

$$\implies \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} \leq \lambda_1 \quad \text{but for } \underline{x} = \underline{e}_1 : \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_1$$

$$\implies \max_{\underline{x} \in \mathbb{R}^p} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_1 \quad \text{and} \quad \operatorname{argmax}(\cdot\cdot) = \underline{e}_1$$

2. $\frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \frac{\underline{x}^T P \Lambda P^T \underline{x}}{\underline{x}^T P P^T \underline{x}} = \frac{\underline{y}^T \Lambda \underline{y}}{\underline{y}^T \underline{y}} = \frac{\sum_{i=2}^{p} \lambda_i (y_i)^2}{\sum_{i=2}^{p} (y_i)^2} \leq \lambda_2 \qquad \forall \underline{x} \perp \underline{e}_2$

$$\underline{y} := P^T \underline{x} = \begin{bmatrix} \underline{e}_1^T \\ \vdots \\ \underline{e}_p^T \end{bmatrix} \underline{x} = \begin{bmatrix} 0 \\ \underline{e}_2^T \underline{x} \\ \underline{e}_p^T \underline{x} \end{bmatrix}$$

because we're looking for $\underline{x}$ s.t. $\underline{x} \perp \underline{e}_1$

$$\underline{x} = \underline{e}_2 \implies \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \frac{\underline{e}_2^T B \underline{e}_2}{\underline{e}_2^T \underline{e}_2} = \lambda_2 \underline{e}_2^T \underline{e}_2 = \lambda^2$$

$$\implies \max_{\underline{x} \in \mathbb{R}^p,\ \underline{x} \perp \underline{e}_1} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_2 \ , \quad \operatorname{argmax}(\cdot\cdot) = \underline{e}_2$$

[..]

p. (proof of the min)

$$\frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \frac{\underline{y}^T \Lambda \underline{y}}{\underline{y}^T \underline{y}} = \frac{\sum_{i=1}^{p} \lambda_i (y_i)^2}{\sum_{i=1}^{p} (y_i)^2} \geq \lambda_p$$

$\underline{y} := P^T \underline{x}$

$$\underline{x} = \underline{e}_p \implies \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_p \implies \min_{\underline{x} \in \mathbb{R}^p} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_p$$

■

Back to PCA:

$$\max_{a \in \mathbb{R}^p : \|a\|=1} Var(a^T X) = \max_{a \in \mathbb{R}^p} \frac{a^T \Sigma a}{a^T a} = \lambda_1 \qquad \left(\Sigma = \sum_{i=1}^{p} \lambda_i e_i e_i^T\right)$$

$$\arg\max_{a \in \mathbb{R}^p : \|a\|=1} Var(a^T X) = e_1$$

Def. PC1: $Y_1 = e_1^T X$ is the first principal component $\left(/\ Y_1 = e_1^T (X - \mu)\right)$
(projection of X on the first eigenvector: along this direction we have max variability)

We want to find the **second principal component**:

Problem:

$$\max_{\substack{a \in \mathbb{R}^p : \|a\|=1 \\ Cov(a^T X, e_1^T X) = 0}} Var(a^T X) = \max_{\substack{a \in \mathbb{R}^p \\ Cov(a^T X, e_1^T X) = 0}} \frac{a^T \Sigma a}{a^T a} \qquad (*)$$

we want to find an other direction of max variability but we want the projection on this direction to be uncorrelated with those we have already found

$Cov(a^T X, b^T X) = ?$

$C \in \mathbb{R}^{k \times p}$: $Cov(CX) = C \Sigma C^T$

$C = \begin{bmatrix} a^T \\ b^T \end{bmatrix} \in \mathbb{R}^{2 \times p}$; $Cov(CX) = Cov\left(\begin{pmatrix} a^T X \\ b^T X \end{pmatrix}\right) = \begin{bmatrix} a^T \\ b^T \end{bmatrix} \Sigma \begin{bmatrix} a & b \end{bmatrix} = \begin{bmatrix} a^T \Sigma a & a^T \Sigma b \\ b^T \Sigma a & b^T \Sigma b \end{bmatrix}$

$\Longrightarrow Cov(a^T X, b^T X) = a^T \Sigma b$

Hence:

$$0 = Cov(a^T X, e_1^T X) = a^T \Sigma e_1 = \lambda_1 a^T e_1 \Longrightarrow a \perp e_1$$

$$\Longrightarrow (*) = \max_{a \in \mathbb{R}^p : a \perp e_1} \frac{a^T \Sigma a}{a^T a} = \lambda_2, \quad \arg\max(\cdot\cdot) = e_2$$

Def. PC2: $Y_2 = e_2^T X$ is the second principal component $\left(/\ Y_2 = e_2^T (X - \mu)\right)$

So, for the **j-th** principal component:

Problem:

$$\max_{\substack{a \in \mathbb{R}^p : \|a\|=1 \\ Cov(a^T X, e_i^T X) = 0 \quad \forall i = 1, \cdots, j-1}} Var(a^T X) = \lambda_j, \quad \arg\max(\cdot\cdot) = e_j$$

Def. PCj: $Y_j = e_j^T X$ $\left(/\ Y_j = e_j^T (X - \mu)\right)$ j-th principal component

$\Longrightarrow Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix}$ vector of PC's (p principal components) $= P^T X$

we simply project X on the eigenvectors of Σ to have PCs

Characteristics of Y:

Prop.

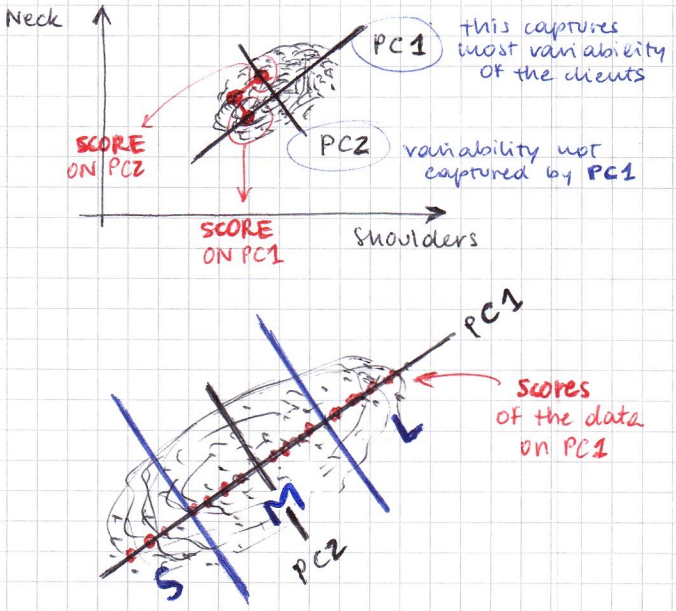
- $\mathbb{E}[Y] = \mathbb{E}[P^T X] = P^T \mu$ (If $Y = P^T (X - \mu) \Rightarrow \mathbb{E}[Y] = 0$)
- $Cov(Y) = Cov(P^T X) = P^T \Sigma P = P^T P \Lambda P^T P = \Lambda$

$$\Longrightarrow \begin{cases} Cov(Y_i, Y_j) = 0 & \forall i \neq j \\ Var(Y_i) = \lambda_i & \forall i \end{cases} \qquad i, j = 1, \cdots, p$$

Observations :

- there is no correlation between the coordinates of the r.v. $\underline{Y}$
- we have ordered the components : the first component is the one with the larger variability, the second is the second larger and so on. We can capture most of the variability with the first components (wrt this system), we can forget about the last components because they express small variability.

Example : We want to industrialize the process of shirts making, so we want to categorize them in groups (S, M, L). How do we do this?



in order to divide properly the shirts we split the clients according to their value on PC1 (score on PC1)

There will be some variability left out, but most of the variability will be captured by PC1

So, what is Y_1? It's a linear combination of Neck and Shoulders :

$$Y_1 = X_1 e_{11} + X_2 e_{21}$$

(if we want to know in which category we are, we compute Y_1)

Observations :

- Are we losing some variability looking at data in this system?

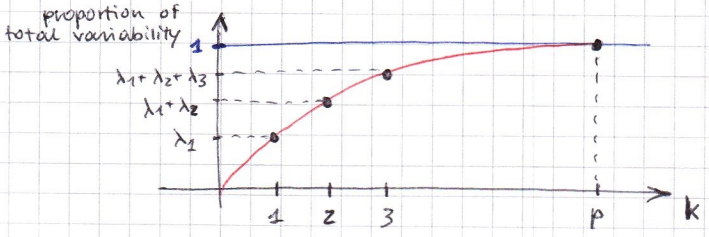
  Generalized variance $(\underline{Y}) = \text{Det}(\Lambda) = \prod_{i=1}^{p} \lambda_i$ $\left(= \text{Det}(\Sigma) = \text{generalized var}(\underline{X})\right)$

  Total variance $(\underline{Y}) = \text{Tr}(\Lambda) = \sum_{i=1}^{p} \lambda_i$ $\left(= \text{Tr}(\Sigma) = \text{total variance}(\underline{X})\right)$

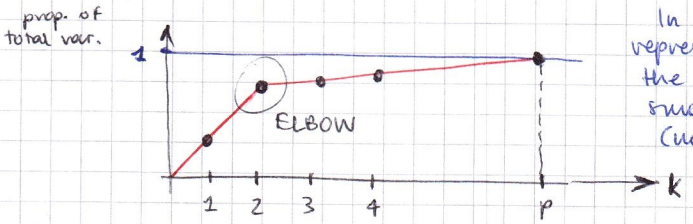
  $\Longrightarrow$ we've not losing informations.

- $\text{Var}(Y_1) = \lambda_1 \Longrightarrow Y_1$ is capturing $\frac{\lambda_1}{\sum_i \lambda_i}$ of total variance

  $\text{Var}(Y_2) = \lambda_2 \Longrightarrow Y_1$ and Y_2 are capturing $\frac{\lambda_1 + \lambda_2}{\sum \lambda_i}$ of total var.



Usually it looks like :



In this case we chose $k=2$ to represent the data. Why? Because the marginal gain get smaller and smaller (not worthy)

When we should stop? 80% / after an elbow (thumb rules)

PCA (continued)

23/03

$\underline{X}$ random vector : $\mathbb{E}[\underline{X}] = \underline{\mu}$, $Cov(\underline{X}) = \Sigma$

$\Sigma = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T$ spectral decomposition (s.t. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$)

$\boxed{Y_i = \underline{e}_i^T \underline{X}}$ $\quad (Y_i = \underline{e}_i^T (\underline{X} - \underline{\mu}))$

i-th principal component

$\boxed{Y_i = e_{1i} X_1 + e_{2i} X_2 + \dots + e_{pi} X_p}$

$:=$ LOADINGS (WEIGHTS) : How to interpret them?

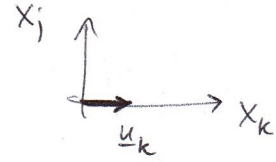
we're still working with the model, not the data

projection on the system generated by the eigenvalues of Σ

Prop $\boxed{Corr(Y_i, X_k) = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}}$ $\quad i,k = 1, \dots, p$

proof-

$$Corr(Y_i, X_k) = \frac{Cov(Y_i, X_k)}{\sqrt{\lambda_i \sigma_{kk}}}$$



$$Cov(Y_i, X_k) = Cov(\underline{e}_i^T \underline{X}, \underline{u}_k^T \underline{X})$$

$\underline{u}_k = [0, 0, \dots, 0, \underset{k}{1}, 0, \dots, 0]^T$ $\quad (\underline{u}_k^T \underline{X} = X_k)$

$$Cov(Y_i, X_k) = Cov(\underline{e}_i^T \underline{X}, \underline{u}_k^T \underline{X}) = \underline{e}_i^T \Sigma \underline{u}_k = \underline{u}_k^T \Sigma \underline{e}_i = \lambda_i \underline{u}_k^T \underline{e}_i$$

$\perp \lambda_i e_{ki}$

$$Corr(Y_i, X_k) = \frac{\lambda_i e_{ki}}{\sqrt{\lambda_i \sigma_{kk}}} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} e_{ki}$$

k-th component of the i-th eigenvector

■

Maybe we don't want to compute the PCA wrt the original variables, but first we standardize the variables.

$\underline{X}$

$$\underline{Z} = V^{-\frac{1}{2}}(\underline{X} - \underline{\mu}) = \left[\frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \dots, \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}\right]^T$$

$\left(V = \begin{bmatrix} \sigma_{11} & & \emptyset \\ & \ddots & \\ \emptyset & & \sigma_{pp} \end{bmatrix}, \quad \Sigma = [\sigma_{ij}] \right)$

$\mathbb{E}[\underline{Z}] = \underline{0}$

$Cov(\underline{Z}) = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}} = \rho$ **CORRELATION MATRIX** (covariance matrix of the standardized variables)

why? to have a way to compare variables

$\Longrightarrow \quad \rho = \sum_{i=1}^{p} \lambda_i \underline{e}_i^T \underline{e}_i$ $\quad \lambda_i$ eigenvalue of ρ, $\underline{e}_i$ eigenvectors of ρ

$\boxed{Y_i = \underline{e}_i^T \underline{Z} = \underline{e}_i^T V^{-\frac{1}{2}} (\underline{X} - \underline{\mu})}$

Observations :

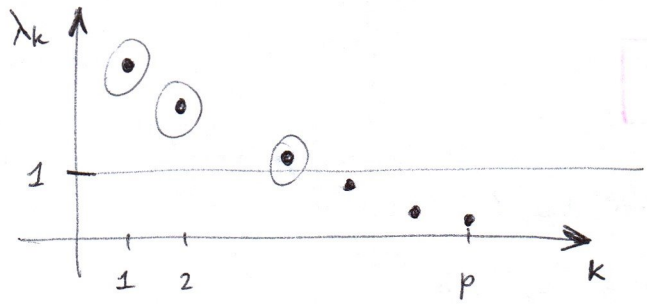
1. $\sum_i^p Var(Y_i) = Tr(\rho) = \sum_{i=1}^p Var(Z_i) = p$
2. $Corr(Y_i, Z_k) = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{1}} = e_{ki}\sqrt{\lambda_i}$

3. Proportion of total variability explained by the first k-components:

$$\frac{\sum_{i=1}^{k} \lambda_i}{p}$$

4. $\bar{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p} = \frac{Tr(\rho)}{p} = \frac{p}{p} = 1$ (average value for the eigenvalues for ρ)

RULE OF THUMB ($\Rightarrow$ Chose Y_i if $\lambda_i > 1$)



If we know that the average variability is 1 (and we've looking for the max variability λ_i) we select all the λ_i above the average.

Example:

Why do PCA on Σ is different from PCA on ρ?

$$\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix} \qquad \rho = \begin{bmatrix} 1 & 0,4 \\ 0,4 & 1 \end{bmatrix}$$

Why doing PCA on Σ is not the right thing? ($100 >> 1$)

$$\Sigma = Cov\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right)$$

Suppose that X_1 is in cm, X_2 is in mm; even if the range of variability maybe is the same, the $Var(X_2)$ is 100 higher than $Var(X_1)$ simply for the units of measure!

PCA of Σ:

eigenvalues & vectors:

$\lambda_1 = 100.16$ $\quad \underline{e}_1 = [0.04, 0.999]^T$

$\lambda_2 = 0.84$ $\quad \underline{e}_2 = [0.999, -0.04]^T$

$Y_1 = 0.04 X_1 + 0.999 X_2$ basically X_2

$Y_2 = 0.999 X_1 - 0.04 X_2$ basically X_1

and in this case the $Var(X_2)$ completely masks $Var(X_1)$!

PCA of ρ:

$\lambda_1 = 1.4$ $\quad \underline{e}_1 = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$

$\lambda_2 = 0.6$ $\quad \underline{e}_2 = \left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right]^T$

$Y_1 = \frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2$

$\downarrow$

$\frac{1}{\sqrt{2}} \left(\frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} + \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)$ $\neq Y_1$ of Σ's PCA

$Y_2 = \frac{1}{\sqrt{2}} Z_1 - \frac{1}{\sqrt{2}} Z_2$

$\downarrow$

$\frac{1}{\sqrt{2}} \left(\frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} - \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)$ $\neq Y_2$ of Σ's PCA

$Var(X_2)$ is so big that if we are looking for the direction where there is maximum variability the algorithm tells us that it is the direction of X_2, same for the direction of minimum variability (X_1)

- PCA(Σ) $\neq$ PCA(ρ)
- if we think that there are variables with very different variabilities (maybe because of the units of measure or maybe because of the phenomena) then it will be better PCA(ρ)

Now we move on **Data**. How do we use data to perform PCA?

Usually μ and Σ are unknown but we have data:

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix}$$

$\underline{x}_i$ realizations of $\underline{X}$;

$\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X}$

We want to use data to estimate $\underline{\mu}$ and Σ

Σ estimated by S

$\underline{\mu}$ estimated by $\bar{\underline{X}}$.

Perform PCA on S. (and not on Σ, since we don't know it)

$$S = \sum_{i=1}^{p} \lambda_i \underline{e}_i^T \underline{e}_i \Longrightarrow \text{PCA}: \quad Y_i \text{ projection on } \underline{e}_i$$

$\underline{x}_i$ (what we observed for the i-th stat. unit) $\xrightarrow{\text{PCA}}$ $y_i = \begin{bmatrix} \underline{e}_1^T \underline{x}_i \\ \vdots \\ \underline{e}_p^T \underline{x}_i \end{bmatrix}$:= **SCORES**

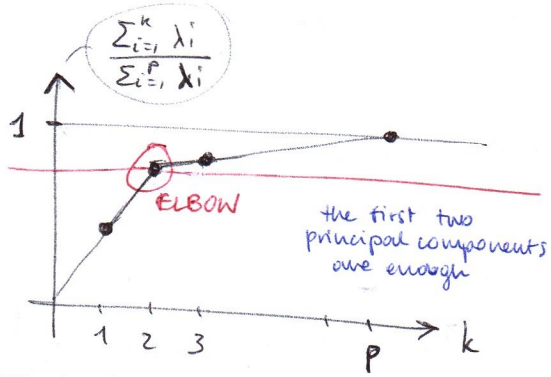
$\mathbb{X}$ (data) $= \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} \xrightarrow{\text{PCA}} \begin{bmatrix} y_1^T \\ \vdots \\ y_n^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & & y_{2p} \\ \vdots & & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$

Y_1 Y_2 ... Y_p

k | $p-k$

Scores on the first principal component

$\dfrac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq$ threshold (0.8) $\Longrightarrow$ We can get rid of part of the matrix (and we know that by doing this we're just missing 20% of the variability of the data)



SCREE PLOTS

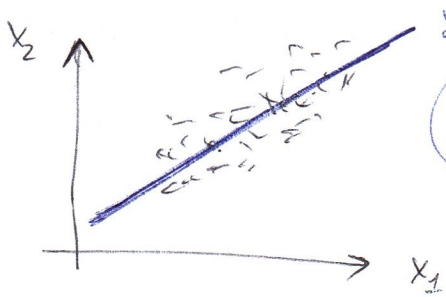
Observation **1**: PCA for categorical variables is called CORRESPONDENCE ANALYSIS and it's performed on the table of joint frequencies among variables. (CONTINGENCY TABLES)

Observation **2**: look also at the smaller λ_i: if $\lambda_p \sim 0$ it means that $\exists$ linear relationship between $X_1, \dots, X_k$

OPTIMAL ORTHONORMAL BASIS

$\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}^p$ (data) $\quad \mathcal{L}$

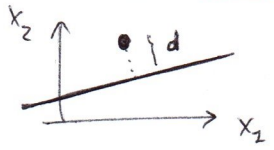
Problem: find a linear space V of dimension $k \leq p$ which best approximate the data, i.e. s.t. $\mathcal{L}$ is "closest" to the data



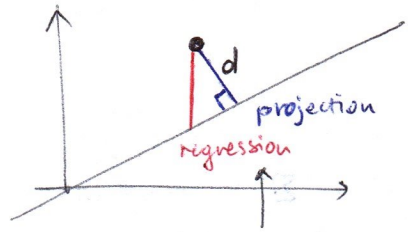
Regression line? **NO**

We want to find a linear space of dimension 1 (since we're in $\mathbb{R}^2$) which is the closest to the data. What is this linear space? The regression line? **No**

What is "close"?



the regression line minimize the distance between what we're trying to predict and the observation (vertical distance). the projection is minimizing the distance between points and line



that's why they're different ($\neq$ regression line)

uses the axes (the regression), the projection doesn't depend on the system (on the way we're looking on the data, it's a property of the data)

We want to identify $\mathcal{L}$:

consider $\eta_1, \dots, \eta_k$:= orthonormal basis spanning $\mathcal{L}$

$\mathcal{L} = \text{span}(\eta_1, \dots, \eta_k)$

Problem: find $\eta_1, \dots, \eta_k$ orthonormal basis s.t.

$$\left(d^2 :=\right) \left\| (\underline{x}_i - \bar{\underline{x}}) - \sum_{j=1}^{k} \eta_j \, \eta_j^T (\underline{x}_i - \bar{\underline{x}}) \right\|^2 = \left(\begin{array}{c} \text{distance between the signal datum (after centering) and the projection on the linear space generated by } \eta_1, \dots, \eta_k \end{array} \right)^2$$

projecting on η_j

$$\Rightarrow \sum_{i=1}^{n} \left\| (\underline{x}_i - \bar{\underline{x}}) - \sum_{j=1}^{k} \eta_j \, \eta_j^T (\underline{x}_i - \bar{\underline{x}}) \right\|^2 \text{ is minimum} \qquad (*)$$

Let's define: $\underline{v}_i = \underline{x}_i - \bar{\underline{x}}$

$$(*) = \left\| \underline{v}_i - \sum_{j=1}^{k} \eta_j \, \eta_j^T \underline{v}_i \right\|^2 = \left(\underline{v}_i - \sum_{j=1}^{k} \eta_j \, \eta_j^T \underline{v}_i \right)^T \left(\underline{v}_i - \sum_{j=1}^{k} \eta_j \, \eta_j^T \underline{v}_i \right)$$

$$= \underline{v}_i^T \underline{v}_i - \sum_{j=1}^{k} \underline{v}_i^T \eta_j \, \eta_j^T \underline{v}_i - \sum_{j=1}^{k} \underline{v}_i^T \eta_j \, \eta_j^T \underline{v}_i + \left(\sum_{j=1}^{k} \eta_j \, \eta_j^T \underline{v}_i \right)^T \cdot \left(\sum_{t=1}^{k} \eta_t \, \eta_t^T \underline{v}_i \right)$$

$$= \underline{v}_i^T \underline{v}_i - \sum_{j=1}^{k} (\eta_j^T \underline{v}_i)^2 - \sum_{j=1}^{k} (\eta_j^T \underline{v}_i)^2 + \sum_{j=1}^{k} (\eta_j^T \underline{v}_i)^2$$

$\eta_j \cdot \eta_t = \begin{cases} 0 & j \neq t \\ 1 & j = t \end{cases}$

$$= \underline{v}_i^T \underline{v}_i - \sum_{j=1}^{k} (\eta_j^T \underline{v}_i)^2$$

Problem:
Find $\eta_1, \dots, \eta_k$ o.b. s.t. $\sum_{i=1}^{n} \left(\underline{v}_i^T \underline{v}_i - \sum_{j=1}^{k} (\eta_j^T \underline{v}_i)^2 \right)$ is **min**

$$\Leftrightarrow \boxed{\sum_{i=1}^{n} \sum_{j=1}^{k} (\eta_j^T \underline{v}_i)^2} \text{ is } \textbf{max} \qquad (**)$$

$$(\star) = \sum_{j=1}^{k}\left(\sum_{i=1}^{n} \eta_j^T \underline{v}_i \underline{v}_i^T \eta_j\right) = \sum_{j=1}^{k} \eta_j^T \underbrace{\left(\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T\right)}_{=(n-1)S} \eta_j$$

$$= (n-1) \sum_{j=1}^{k} \eta_j^T S \eta_j$$

CONCLUSION

So if we're looking for a linear space that is close to the data (in terms of minimizing the dist.) this linear space of dimension k is given by the first k principal components (direction)

$k=1$: $\max_{\eta:\ \|\eta\|=1} \eta^T S \eta = \lambda_1$, $\operatorname{argmax} = \underline{e}_1$

by induction on k : $\Longrightarrow$ $\eta_1 = \underline{e}_1$, ... , $\eta_k = \underline{e}_k$

s.t. $S = \sum \lambda_i \underline{e}_i \underline{e}_i^T$

Observation: What is the error of approximation?

$$\sum_{i=1}^{n} d^2_{\text{euclidean}}\left(\underline{x}_i - \bar{\underline{x}}, \sum_{j=1}^{k} \underline{e}_j \underline{e}_j^T (\underline{x}_i - \bar{\underline{x}})\right) =$$

$$= \underbrace{\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})^T (\underline{x}_i - \bar{\underline{x}})}_{\in \mathbb{R}\ (\star)} - \underbrace{(n-1) \sum_{j=1}^{k} \underline{e}_j^T S \underline{e}_j}_{(\star\star)}$$

using what we did before with $\underline{v}_i$ & Co.

$Tr(a) = a$ if $a \in \mathbb{R}$

$$(\star) = Tr\left(\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})^T (\underline{x}_i - \bar{\underline{x}})\right) = \sum_{i=1}^{n} Tr\left((\underline{x}_i - \bar{\underline{x}})^T (\underline{x}_i - \bar{\underline{x}})\right)$$

$Tr(ABC) = Tr(CAB) = Tr(BCA)$

$$= \sum_{i=1}^{n} Tr\left((\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T\right)$$

$$= Tr\left(\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T\right)$$

$$= Tr\left((n-1) S\right)$$

$$= (n-1) \sum_{i=1}^{p} \lambda_i$$

$$\Longrightarrow (\star\star) = (n-1) \sum_{j=1}^{k} \underline{e}_j^T S \underline{e}_j = (n-1) \sum_{j=1}^{k} \lambda_j \underline{e}_j^T \underline{e}_j = (n-1) \sum_{j=1}^{k} \lambda_j$$

$$\Rightarrow \sum_{i=1}^{n} d^2_{\text{eucl.}}\left(\underline{x}_i - \bar{\underline{x}}, \sum_{j=1}^{k} \eta_j \eta_j^T (\underline{x}_i - \bar{\underline{x}})\right) =$$

$$= (n-1) \sum_{j=1}^{p} \lambda_j - (n-1) \sum_{j=1}^{k} \lambda_j = \underbrace{(n-1) \sum_{j=k+1}^{p} \lambda_j}_{\text{error of approximation}}$$

sum of the λ_j that we're leaving out

Extensions:

- **ICA** : INDIPENDENT COMPONENT ANALYSIS
- **NON-LINEAR** dimensional reduction

the directions are not $\perp$, they're stochastically independent