Data Science Course (ATI) Notes on 1st Part

S. MARCHAND-MAILLET 2019

Chapter ① Linear Algebra

Main Concepts : - What is a vector / vector space ?
- what does the scalar product represent ?
- Why is projection an important concept ?
- What is critical about eigen-things ?

1. Vector spaces and vectors

Given a field K ($\mathbb{R}$ or $\mathbb{C}$) and a set V, a vector space is using the addition on V and the external product (s-multiplication) to transform vectors (slide 4)
The axioms to be satisfied by the addition and the s-multiplication are generally enumerated as CANI-ADDU.
The important point is that one can compute linear combinations of vectors Eg :

$$\vec{w} = \sum_{i=1}^{n} \alpha_i \vec{v}_i \qquad \alpha_i \in K \quad \vec{v}_i \in V$$

In particular

$$\vec{g} = \sum_{i=1}^{n} \frac{1}{n} \vec{v}_i = \frac{1}{n} \sum_{i=1}^{n} \vec{v}_i$$

In turn this defines the idea of a set of linearly independant vectors (slide 6)

If $\exists \alpha_i$'s such that $\sum_{i=1}^{n} \alpha_i \vec{v}_i = 0$ and $\exists j \,/\, \alpha_j \neq 0$

$$\Rightarrow \sum_{\substack{i=1 \\ i \neq j}} \vec{v}_i + \alpha_j \vec{v}_j = 0 \Rightarrow \vec{v}_j = \frac{1}{\alpha_j} \sum_{\substack{i=1 \\ i \neq j}} \alpha_i \vec{v}_i$$

dependent

→ linear independent vectors cannot be made up from others

Bases: this notion of linear independence induces that of a (Hamel) basis:

A set of linearly independent vectors $B = \{e_1 \dots e_n\}$ is a basis of V iff any vector $v \in V$ can be combined from the vectors in B. ie:

$$\exists\ \alpha_1 \dots \alpha_n \in K \qquad v = \sum_{i=1}^{n} \alpha_i e_i$$

The number $n \in \mathbb{N}$ of basis vectors e_i is the dimension of the vector space. n is both the minimum number of vectors necessary to represent any $v \in V$ and it is also a maximal number of linearly independent vectors in V.

In turn this defines the notion of components as the values $\alpha_i \in K$, which completely characterize v in B this is why any vector in V can be identified to its components in B, the "vector" of K^n $(\alpha_1 \dots \alpha_n)^T$ as a representation of v.

Hence, clearly the same vector v may have different representations, depending on the choice of the basis.

Given 2 bases $B = \{e_1 \dots e_n\}$ and $B' = \{e_1' \dots e_n'\}$, every e_j can be written in B' as

$$e_j = \sum_{i=1}^{n} P_{ij} e_i'$$

Hence if we have a vector $v \in V$,

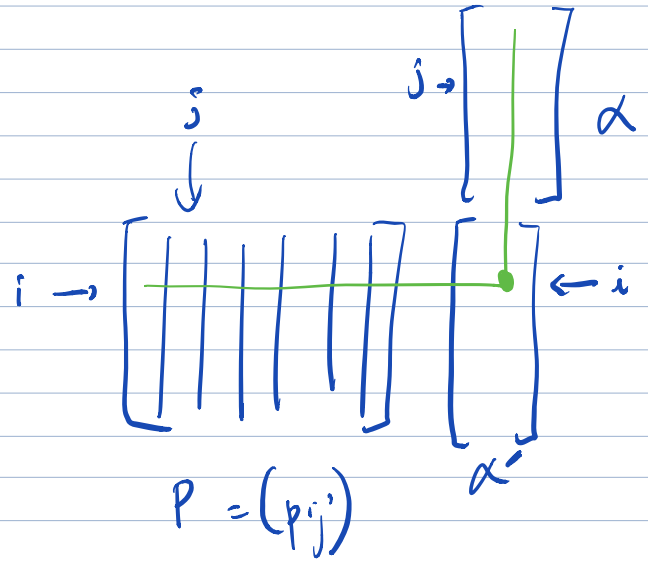
in B :

$$v = \sum_{j=1}^{n} \alpha_j e_j = \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{n} p_{ij} e'_i = \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \alpha_j p_{ij} \right] e'_i$$

in B' :

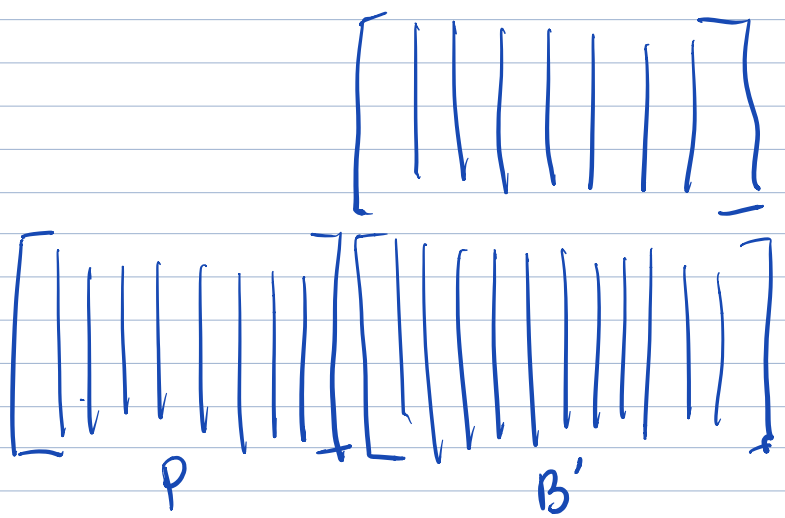
$$v = \sum_{i=1}^{n} \alpha'_i e'_i$$

$$\Rightarrow \alpha'_i = \sum_{j=1}^{n} p_{ij} \alpha_j$$



⇒ la matrice de passage P qui transforme les coordonnées de v dans B (α) en coordonnées de v dans B' (α') a comme colonnes $P_{\cdot j}$ les coordonnées de e_j dans B'

$$\Rightarrow \alpha' = P \alpha$$



More generally $B' = PB$ (matrix form)

2 - Scalar product

This is a structure added to the vector space.

It defines:

(Slides 8-9)
- orthogonality: $F^{\perp} := \{x \in E / \langle x, y \rangle = 0 \ \forall y \in F\}$
- a norm: $\|x\|^2 = \langle x, x \rangle$

This is related by the Pythagoras theorem:

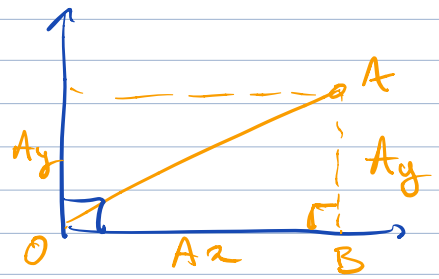
If x and y are orthogonal then

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2$$

Since

$$\|x+y\|^2 = \langle x+y, x+y \rangle = \underbrace{\langle x, x \rangle}_{\|x\|^2} + \underbrace{\langle y, y \rangle}_{\|y\|^2} + 2\underbrace{\langle x, y \rangle}_{0}$$

In a Euclidean space with Cartesian coordinates this helps deriving



$$\|\vec{OA}\|^2 = \langle \vec{OA}, \vec{OA} \rangle$$

Here $\|\vec{OA}\|^2 = A_x^2 + A_y^2$ by Pythagoras theorem for triangle (OAB)

Thus $\|\vec{OA}\|^2 = \langle \vec{OA}, \vec{OA} \rangle = A_x A_x + A_y A_y$

which generalizes in more dimensions and for different vectors as $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$

Note: You may also related this to (in fact start from) duality via $\langle x, y \rangle = x^T y$ with transpose a duality

A way to think of the scalar product is via the formula:

$$\langle x, y \rangle = \|x\| \cdot \|y\| \cdot \cos(\widehat{x, y})$$

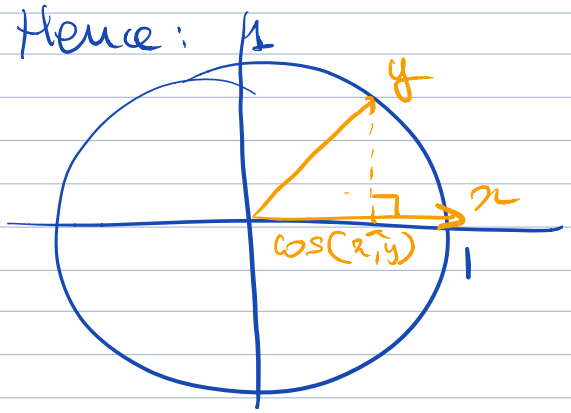
Hence:

$$\langle x, x \rangle = \|x\| \cdot \|x\| \cdot \cos(0) = \|x\|^2$$

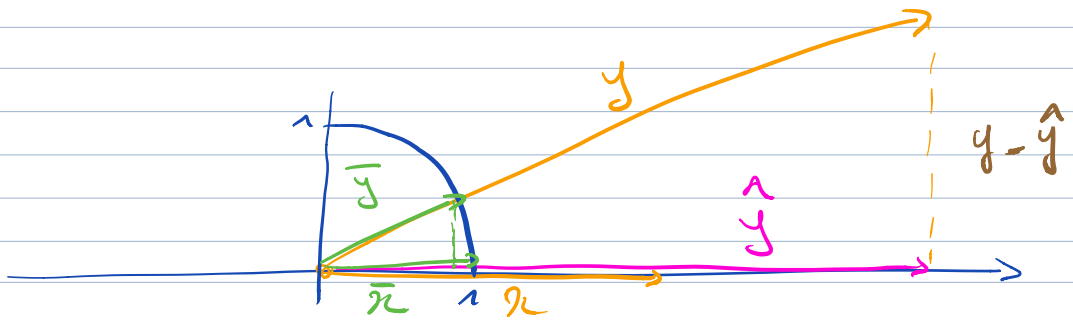
x, y orthogonal: $\langle x, y \rangle = \|x\| \cdot \|y\| \cdot \cos\left(\frac{\pi}{2}\right) = 0$

x, y of norm 1: $\langle x, y \rangle = \cos(\widehat{x, y})$

Hence:



leading to projection of a vector y of any norm onto a vector x of any norm:



$$\langle x, y \rangle = \|x\| \cdot \|y\| \cos(\widehat{x, y})$$

Thales: $\|\hat{y}\| = \|y\| \cos(\widehat{x, y}) = \langle \frac{x}{\|x\|}, y \rangle$ and $\hat{y} = \alpha x$

$$\hat{y} = \langle \frac{x}{\|x\|}, y \rangle x = \frac{x}{\|x\|} \langle x, y \rangle$$

Note that the norm of the projection vector, ie the distance between the point and its projection is: $\|y - \hat{y}\|$ with

$$y - \hat{y} = y - \frac{x}{\|x\|} \langle x, y \rangle$$

3. Orthogonal basis – Canonical basis

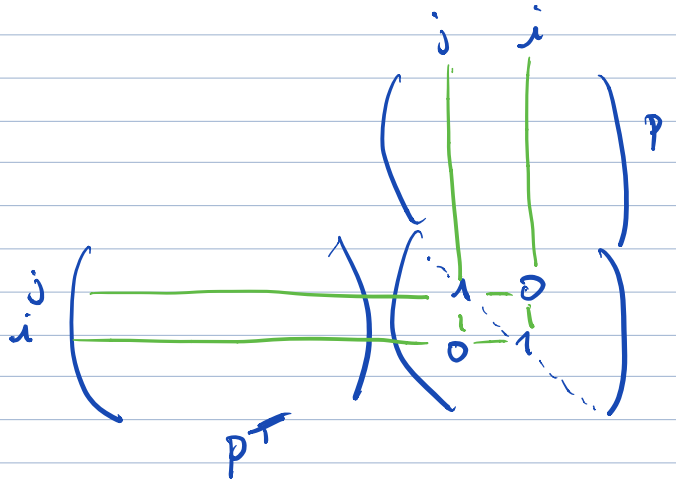
Combining the scalar product and linear independence, we can construct orthogonal basis by choosing

$B = \{e_1 \dots e_n\}$ with e_i such that

$$\langle e_i, e_j \rangle = 0 \quad \forall\, i \neq j$$

and $\langle e_i, e_i \rangle = \|e_i\|^2 = 1$ (for orthonormal basis)

in that case, the transition matrix P containing as columns the coordinates of e'_j in B is orthogonal since



$$P^T P = Id$$

$$\Rightarrow (P^T P) = P P^T = (Id)^T = Id$$

$$\rightarrow P^T = P^{-1}$$

As particular (canonical) basis we can choose:

$B = \{e_1 \dots e_n\}$ such that

$$e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i$$

which is orthogonal

4- Representation of linear maps-

Given $f: V \longrightarrow W$ a map from K-vector space V to K-vector space W, f is said to be linear iff:

$f(\alpha u + \beta v) = \alpha f(u) + \beta f(v) \in W \quad \forall u, v \in V \quad \alpha, \beta \in K$

Given $B = \{e_1 \dots e_n\}$ a basis on V

and $B' = \{e'_1 \dots e'_m\}$ a basis on W

We have

$\forall v \in V \quad \exists \alpha_1 \dots \alpha_n \text{ s.t } \quad v = \sum_{i=1}^{n} \alpha_j e_j$

by linearity: $f(v) = f\left(\sum_{i=1}^{n} \alpha_j e_j\right) = \sum_{i=1}^{n} \alpha_j f(e_j)$

similarly since $f(e_j) \in W \quad \exists a_{1j} \dots a_{mj} \text{ s.t } f(e_j) = \sum_{i=1}^{m} a_{ij} e'_i$

Together: $f(v) = \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{m} a_{ij} e'_i = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \alpha_j e'_i$

Denoting $w = f(v) = \sum_{i=1}^{m} \alpha'_i e'_i$

Hence in matrix form:

$$\underset{m \times 1}{\begin{bmatrix} \alpha'_1 \\ \vdots \\ \alpha'_m \end{bmatrix}} = \underset{m \times n}{\begin{bmatrix} a_{ij} \end{bmatrix}} \underset{n \times 1}{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}}$$

$$m\left(\begin{array}{c} n \\ A \end{array}\right)\begin{matrix} a(v) \\ (w) \end{matrix}$$

Hence the $m \times n$ coefficients of matrix $A = (a_{ij})_{\substack{i=1 \dots m \\ j=1 \dots n}}$

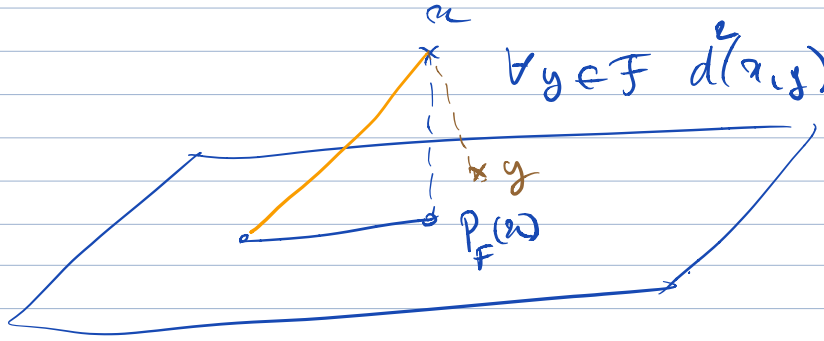
Completely determine the linear map f

5 - Projection:

Projection is a basic concept in data analysis and geometry in general.

The projection defines a notion of distance between a point x and a subspace F

$$d(x, F) = d(x, P_F(x)) \quad \text{where} \quad P_F(x) = \underset{y \in F}{\operatorname{argmin}}\, d^2(x, y)$$



$$\forall y \in F \quad d^2(x, y) \geq d^2(x, P_F(x))$$

if we define $d^2(x, y) = \|x - y\|^2 = \langle x - y, x - y \rangle$ we can use scalar product to compute the orthogonal projection.

Basically: $P_F(x)$ is the best representative (approximation) of x on F. This relates to operations that use the distance between x and $P_F(x)$

(Slide 39) - MSE Criterion (linear systems, regression)

(Lecture probas) - Computation of Variance (F "mean subspace")

This definition of projection is equivalent to always considering an orthogonal projection as shown next.

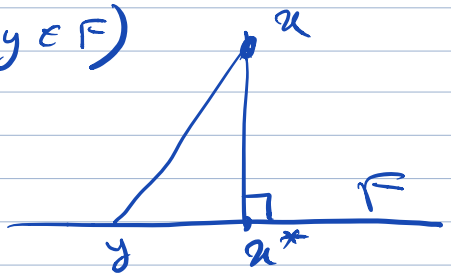
Note: $d(x,F) = d(x, P_F(x))$ avec $P_F(x) = \underset{y \in F}{\operatorname{argmin}}\, d^2(x,y)$

$\Rightarrow P_F(x)$ is such that

$$d^2(P_F(x), x) = \min_{y \in F} d^2(x,y)$$

If F is flat then by Pythagoras: $\forall y \in F$

$d^2(x,y) = d^2(x,x^*) + d^2(y,x^*)$ where x^* is such that $(x - x^*) \perp F$ $(\Rightarrow \langle x - x^*, y \rangle = 0 \ \forall y \in F)$



Hence since $d^2(y, x^*) \geq 0$

$\min_{y \in F} d^2(x,y) = d^2(x, x^*)$

$\Rightarrow P_F(x) = x^*$

More generally consider the curve $\gamma : [0,1] \longrightarrow F$, $t \longmapsto \gamma(t) \in F$

such that $\exists\, t^* \in [0,1] \ / \ \gamma(t^*) = P_F(x)$

$$\Rightarrow P_F(x) = \underset{y \in \gamma}{\operatorname{argmin}}\, d^2(x,y) \Rightarrow t^* = \underset{t \in [0,1]}{\operatorname{argmin}}\, d^2(x, \gamma(t))$$

Let $v(t) = \gamma(t) - x \Rightarrow d^2(x, \gamma(t)) = \|v(t)\|^2 = \langle v(t), v(t) \rangle$

$\Rightarrow x$ being fixed $\left.\frac{\partial v(t)}{\partial t}\right|_{t=t_1}$ is the tangent vector to γ at $\gamma(t_1)$

Now $P_F(x)$ is such that $\left.\frac{\partial \langle v(t), v(t) \rangle}{\partial t}\right|_{t=t^*} = 0$

By linearity: $\frac{\partial \langle v(t), v(t) \rangle}{\partial t} = \left\langle \frac{\partial v(t)}{\partial t}, v(t) \right\rangle + \left\langle v(t), \frac{\partial v(t)}{\partial t} \right\rangle$

$$= 2 \left\langle \frac{\partial v(t)}{\partial t}, v(t) \right\rangle$$

$\rightarrow$ at t^* : $\left\langle \frac{\partial v(t)}{\partial t}, v(t) \right\rangle \Big|_{t=t^*} = 0 \Rightarrow (x - x^*) \perp F$

The projection happens perpendicularly to F

5. Eigen-things

The eigenvector of a linear map is defined as a vector that is only rescaled by the map f

u eigenvector $\Rightarrow \exists \lambda > 0 \;/\; f(u) = \lambda u$

λ is the eigenvalue associated to eigenvector u

In matrix terms this translates to

$$Au = \lambda u \qquad \text{(Slide 19)}$$

The set of all eigen-pairs can be gathered into one system $AU = U\Lambda$ where

a) Columns of U are eigenvectors u_i

b) $\Lambda = \text{diag}(\lambda_1 \dots \lambda_n)$

a) $\Rightarrow$ $U^T U = Id$ since $(U^T U)_{ij} = u_i^T u_j = \delta_i^j$

$\Rightarrow U^{-1} = U^T$

$\Rightarrow A = U\Lambda U^{-1} = U\Lambda U^T$

One immediate consequence is that

$$Tr(A) = Tr(U\Lambda U^{-1}) = Tr(\Lambda(U^{-1}U)) = Tr(\Lambda) = \sum_{i=1}^{\hat{n}} \lambda_i$$

(Slide 17)

ie the Trace is invariant to a change of Basis

Also $A \cdot A = (U\Lambda U^{-1})(U\Lambda U^{-1}) = U\Lambda\Lambda U^{-1} = U\Lambda^2 U^{-1}$

Hence $A^k = U\Lambda^k U^{-1}$ with $\Lambda^k = \text{diag}(\lambda_1^k \dots \lambda_n^k)$

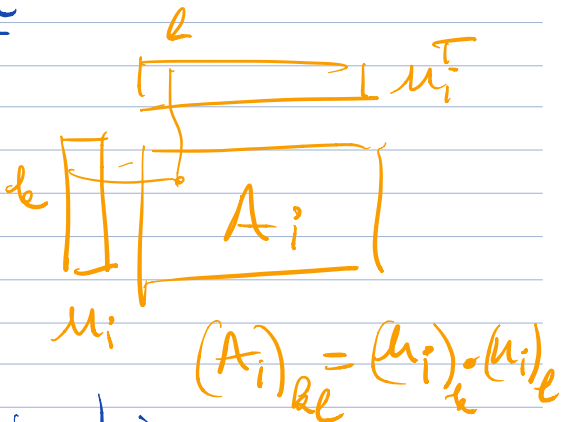
Note that $A^k v$ is attracted by eigenvectors

($A^k v \underset{k\to\infty}{\to} u_i$ for some i)

Reconstruction and Approximation

From $A = U \Lambda U^T$

we can write $A = \sum_{i=1}^{n} \lambda_i \underbrace{u_i u_i^T}_{A_i}$

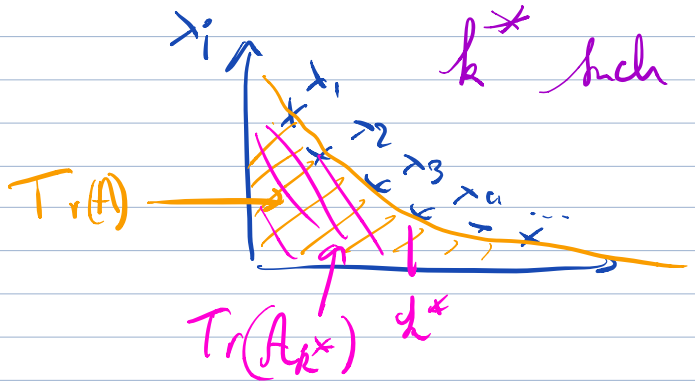


$(A_i)_{kl} = (u_i)_k \cdot (u_i)_l$

This is the basis for approximation:

$A_k = \sum_{i=1}^{k} \lambda_i u_i u_i^T = \sum_{i=1}^{k} \lambda_i A_i$ (& any other combination)

the choice of approximation may be motivated by conservation of the trace:



k^* such that $Tr(A_{k^*}) \geq \alpha \, Tr(A)$

$\alpha \in [0,1]$

Interpretation:

A corresponds to the representation of a linear map $f: \mathbb{R}^n \to \mathbb{R}^n$ in the original coordinate system: Λ is the view of that map along coordinate vectors u_i's.

A^k is therefore a "simplification" of this map in some optimal sense (eg such that it operates only within the $\mathbb{R}^{k^*}$ subspace of basis $(u_1 \dots u_{k^*})$ (→ cf applications for PCA, LSI)

More generally, given a linear endomorphism $f: \mathbb{R}^n \to \mathbb{R}^n$, given a basis $B = \{e_1 \dots e_n\}$ one can write

$$f(e_i) = \sum_{j=1}^{n} a_{ij} e_j$$

As we know, the n^2 numbers a_{ij} completely describe the action of a linear endomorphism

If we happen to choose $B = \{u_1 \dots u_n\}$ we have

$$f(u_i) = \lambda_i u_i \quad \text{(by definition of a eigenvector)}$$

Hence

$$f(v) = f\left(\sum_{i=1}^{n} v_i u_i\right) = \sum_{i=1} v_i f(u_i) = \sum_{i=1}^{n} v_i \lambda_i u_i$$

Therefore, in the eigenspace (the space whose basis is eigenvectors) only n numbers (λ_i) are needed to represent the action of the linear map.

Clearly the interpretation of $A = P^{-1} \Lambda P$ is $f = g \circ \lambda \circ g^{-1}$ where g is a rotation (represented by orthogonal matrix P) taking a vector in basis $\{e_1 \dots e_n\}$ and placing it into basis $\{u_1 \dots u_n\}$

Then λ is a scaling transform of matrix $\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$ applying the map and g^{-1} then returns this result into basis $\{e_1 \dots e_n\}$.

Chapter ② Proba & Stats

Data Analysis often starts with a set of documents (images, texts, objects, ...) to which we must associate a representation to immerse them into a mathematical space within which the analysis is performed. In general it is assumed that the representation is so that the space is a topological space (defining a notion of neighborhood) and a vector space (so we can sum and scale elements).
The topology is in general constructed as the standard topology where open sets are open balls for some given distance function.
(Slide 8)

The distance function can be used as a notion of similarity. This view also leads to the idea of a probabilistic space where the occurence of a "document" or a group of "similar document" may occur with more or less "chance".

A probability space can be thought of as a space of neighbourhoods (with certain properties - of a σ-algebra) with each associated with a positive measure P summing to 1
(Slide 14)

1. Probability function:

P(A) can be thought of as the area of A if the total area is 1

Eg

the intersection counted twice

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2. Conditional probability

the information an event (A) brings about another (B) $P(B|A)$ (slide 15)

$$P(A,B) = P(B|A)\,P(A) = P(A|B)\,P(B)$$

- $P(A,B)$: Prob of A and B
- $P(B|A)$: Prob of B knowing A
- $P(A)$: realise A first
- $P(A|B)$: Prob of A knowing B first
- $P(B)$: realise B first

Exple: Two tokens, one red, one blue in envelopes and given to X and Y

A = {X gets red}

B = {Y gets blue}

$P(A) = \frac{1}{2}$ uncertain $\quad P(B) = \frac{1}{2}$ uncertain $\quad P(B|A) = 1$ the knowledge of A removes all uncertainty

$P(A,B) = \frac{1}{2}$ (other possibility is X → blue, Y → red)

$$P(B|A) = \frac{P(A,B)}{P(B)} = \frac{1/2}{1/2} = 1$$

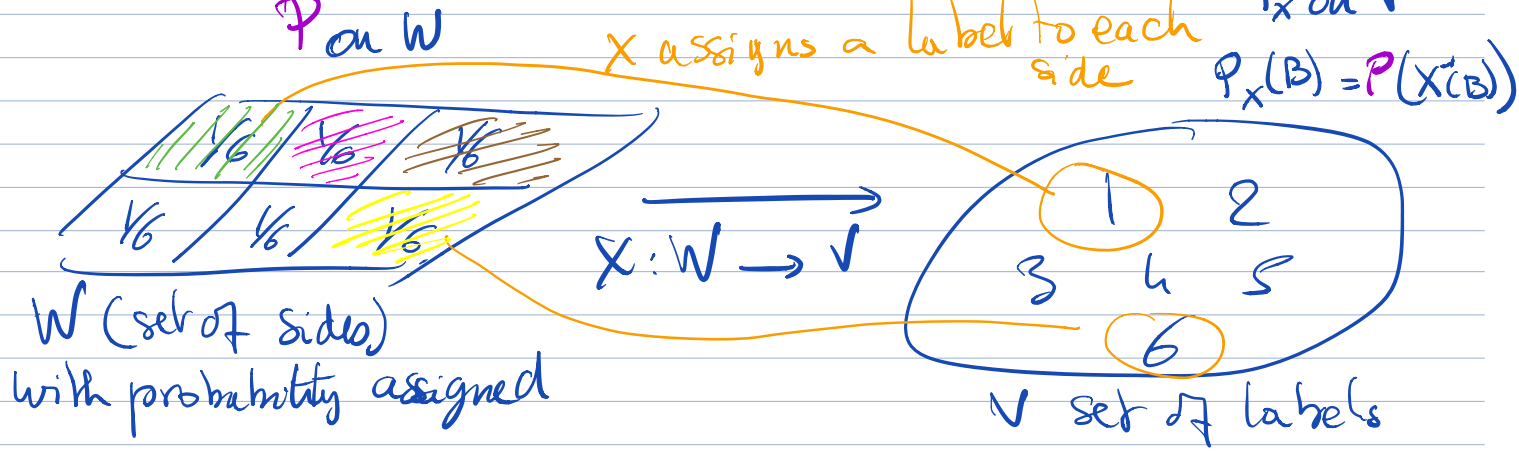
3- Random variable.

Creating a r.v is to give each event a label (on which comparisons can be made) and transfering the probability function on the space of labels.

Exple: Dice. (Exple of a discrete case)

Formally a dice is a cube. The raw event is $w =$ "the cube will fall on that side", $w \in W$ where a probability function is defined giving $1/6$ to each side.

A random variable X is a function from Ω to E a set of labels assigned to each face



Hence I can ask $X(w) \leq 3$ (or $X \leq 3$) this answer "sides green ($X=1$), Pink ($X=2$), brown ($X=3$)" and use this to assign a probability. (slide 17)

$$P_X(X \leq 3) := P(X^{-1}(1,2,3)) = P(\text{green, Pink, brown}) = \frac{3}{6}$$

Note: writing "$X \leq 3$" is not possible before assigning labels to "sides"

This is why V is generally chosen as a space where an order or comparison can be made ($\mathbb{N}, \mathbb{Z}$ for discrete, $\mathbb{R}$ for continuous)

Continuous example: Temperature

X assigns a scale (degrees) to the height of mercury so I can mention "$X < 0$" and compute

$P_X(X<0) = P(\underbrace{\text{"all height below that labeled 0"}}_{X^{-1}(0)})$

Note: for integration consistency

$P_X(x) = 0$ (lebesgue notion "a.c.")

Hence we extend the notion by defining

$f(x) = P_X([x; x+dx[)$ (dx: infinitesimal)

Summary: Defining a random variable allows to install a probability function P_X on a space (V) where an order may be defined. and then define entities like:

$$\sum_{x \leq \tau} x \cdot P_X(x) \qquad \text{discrete}$$

$$\int_{x \leq \tau} x \cdot f(x)\, dx \qquad \text{Continuous}$$

4 - Expectation, Variance.

The Expectation $E(X)$ of a random variable is the "center of mass" of the probability P_X

(slide 60) $$E(X) = \sum_{x \in V} x P_X(x) \qquad E(X) = \int_V x f(x)\, dx$$

Note: $E(X)$ is not the "most likely event".

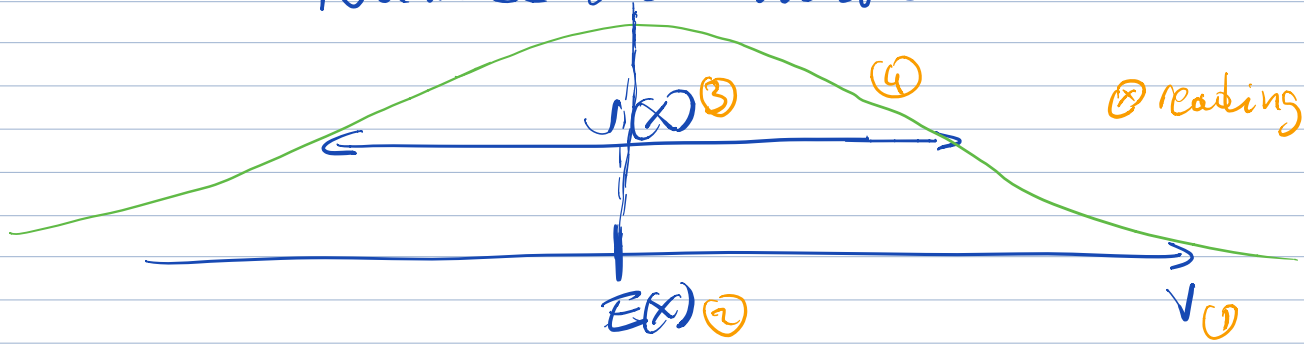
Eg: Fair dice: $E(X) = \sum_{x=1}^{6} x \cdot \frac{1}{6} = \frac{21}{6} = 3{,}5$ ← not an event value

The variance is related to the expectation of deviation from this Center of Mass:

$$V(X) = E\left((X - E(X))^2\right)$$

E: Expectation; $X - E(X)$: deviation from $E(X)$

Note: Considering a r.v X with only non-zero Expectation and Variance (all other moments taken to 0) leads naturally to the Normal distribution



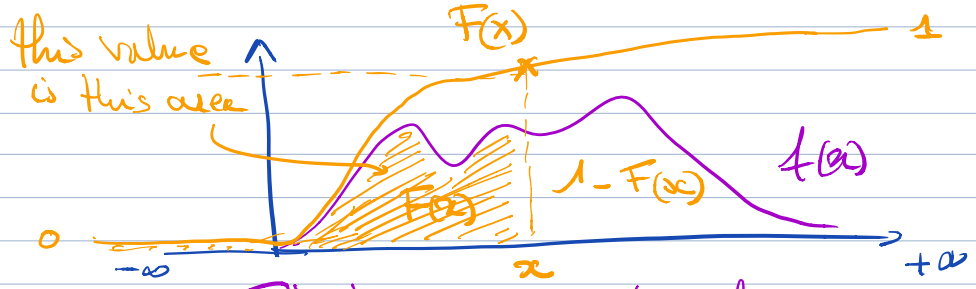
② reading order

This can model a phenomenon whose "normal" (regular) realisation is $E(X)$ and samples deviate in average of $V(X)$

5 - Simulating a Probability Density Function.

$X : W \to V$ thanks to the order on V we can define the Cumulative function F that tells us the Cumulative chance of realization of an event w whose label $y = X(w)$ is less than $x \in V$.

$$F(x) = \sum_{\substack{y \in V \\ y \leq x}} P_X(y) \quad \text{or} \quad F(x) = \int_{-\infty}^{x} f(a)\,da$$

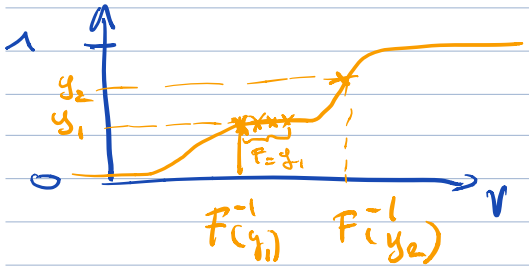


Total area under f sums to 1

$F : V \longrightarrow [0,1]$

(Slides 30 - 31)

We define the inverse of F as



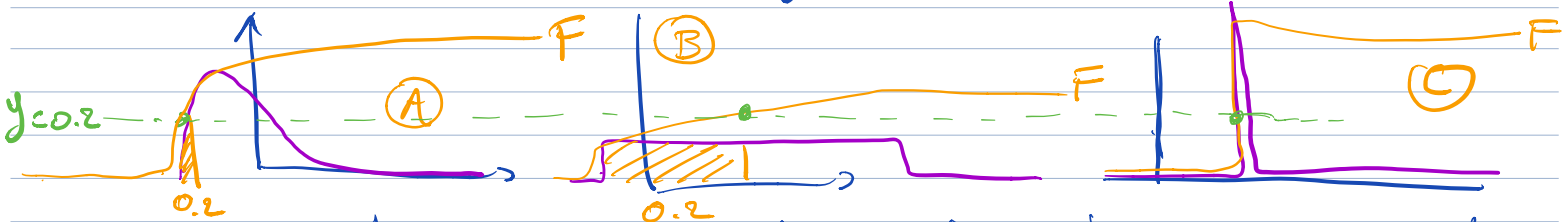
$$F^{-1} : [0,1] \longrightarrow V$$

$$y \longmapsto F^{-1}(y) = \inf\{x \in V / F(x) = y\}$$

if F is not strictly monotone

Since $F(x) = y$ is the 'number' of realizations of "labels" less than x, F (and F^{-1}) "encode" f (or P_X)

Take $y = 0.2$ depending on P_X $F^{-1}(y)$ will be different



F^{-1} "remaps" uniform values to the original distribution. In extreme case, all goes to one value (C)

Chapter (3) High dimensional data

This is to motivate the idea that one should be careful when choosing the representation space for the documents since the structure of the resulting (vector) space, including the dimension will influence the statistics over the data.

This is an artifact of:

- the way we compute distances:
  Minkowsky norms sum over dimensions hence add up this effect
- the fact we only access the space (and the probability density function) via a sample and the dimension has an effect on how large this sample should be.

(Slides 40–56)

Chapter ④ Principal Component Analysis

We look at a data sample $\mathcal{X}$ whose representation is $X = \{x_i\}_{i=1\dots M}$ in vector space $\mathbb{R}^N$.

$\mathbb{R}^N$ is equipped with scalar product

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{N} x_i y_i$$

1- Data statistics

By way of representation, each sample is associated to a vector $x_i \in \mathbb{R}^N$. The given of many samples allows to compare statistics over each dimension. The more sample, the better the statistics.

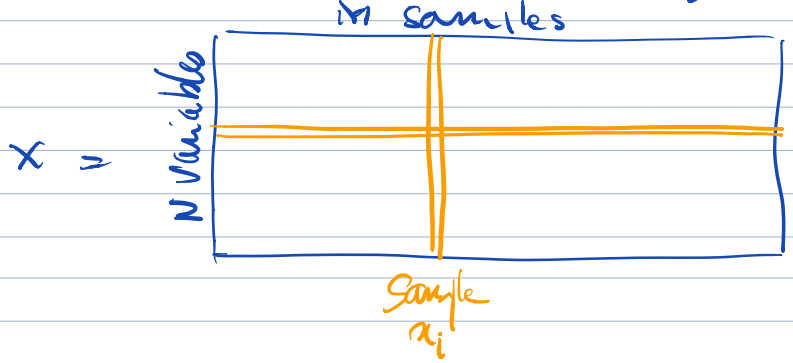
If we consider the additive noise model

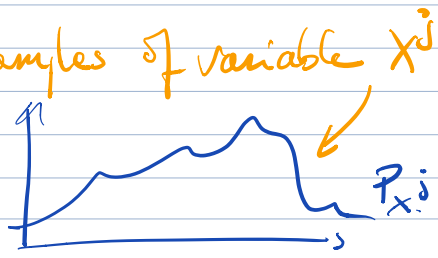
$$\hat{x} = x + \varepsilon \leftarrow \text{noise} \quad \begin{cases} - \text{measure} \\ - \text{formation} \\ - \dots \end{cases}$$

$\hat{x}$: sample observed; x: True, never observed

and associate a random variable X^j to each dimension, and consider these N random variables independant, then, we can compute the statistics of the data sample as the statistic of the N-dimensional random variable whose sample is X (the random variable will then also be called X).

(Slide 30)





2 - Notion of inertia (physics)

Identify samples with points of unit mass in a N-dimensional space

Center of mass : $g = \frac{1}{\sum_{i=1}^{M} m_i} \sum_{i=1}^{M} m_i x_i = \frac{1}{M} \sum_{i=1}^{M} x_i$

Define the inertia of the sample wrt a point $a \in \mathbb{R}^N$ as $I_a(X) = \sum_{i=1}^{M} d^2(x_i, a)$ (slide 42)

we can extend the definition to the inertia wrt a subspace $F \subset \mathbb{R}^N$

$I_F(X) = \sum_{i=1}^{M} d^2(x_i, Proj_F(x_i))$

(cf properties of projection above)

Note : given X, point a is the point minimizing the inertia of X if

$$a^* = \operatorname{argmin}_a I_a(X) \iff \left.\frac{\partial I_a(X)}{\partial a}\right|_{a=a^*} = 0$$

$$\frac{\partial I_a(X)}{\partial a} = \frac{\partial}{\partial a}\left(\sum_i (x_i - a)^T (x_i - a)\right) = \sum_i \frac{\partial}{\partial a}(x_i - a)^T (x_i - a)$$

$$= \sum_i \frac{\partial}{\partial a}\left(x_i^T x_i - 2 x_i^T a + a^T a\right) = \sum_i -2x_i + 2a$$

$$\Rightarrow \sum_i -2x_i + 2a^* = 0 \Rightarrow \sum_i x_i = \sum_i a^* \text{ and } a^* = \frac{1}{M}\sum_i x_i$$

$\Rightarrow$ the center of mass g is the point of minimum inertia

$\rightarrow$ it makes sense to center our space at g ($x_i \leftarrow x_i - g$)

$\rightarrow$ similarly we will be interested in subspaces F_g containing g

$\Rightarrow$ relation to Huygens theorem - slide 43

As we know, the space can be decomposed into two orthogonal components: a subspace F_g and its orthogonal $F_g^\perp$: $V = F_g \oplus F_g^\perp$

we can always choose (by rotation) a basis such that F is the subspace spanned by the first N_F components and $F^\perp$ by the rest.

$d^2(x,y) = (x-y)^T(x-y) = \sum_{i=1}^{N} (x_i - y_i)(x_i - y_i)$

decomposes into $\sum_{i=1}^{N_F} (x_i - y_i)^2 + \sum_{i=N_F+1}^{N}$ so that

$I_g(X)$ decomposes into:

$$I_g(X) = I_F(X) + I_{F^\perp}(X)$$

I_F is the explained inertia (the inertia within the considered subspace F)

$I_{F^\perp}$ is the remain, called the residual inertia

3- relation to statistics

Clearly there is a analogy between a physical system and statistics. Considering the m of a particle as related to its probability, one can associate:

g Center of mass $\hookrightarrow$ $E(X) = \mu$ Expectation

$X - g$ Centered data $\hookrightarrow$ $X - \mu$ Centered data

$I_g(X) = \sum_{i=1}^{M} d^2(x_i, g)$ Inertia $\hookrightarrow$ $M\left[E((X-\mu)^2)\right]$ Variance

In matrix form, with centered data

$$I_g = Tr(XX^T) = \sum_{i=1}^{N} X^i X^{iT}$$

4- PCA

PCA seeks a new coordinate system into which the data will be better legible. Here "legible" means "incrementally decomposed". The central notion is Variance / Inertia. PCA seeks to incrementally construct subspaces which will retain the maximum of the original variance.

⇒ PCA seeks axis maximising the variance of the data projected onto that axis

From the above analogy, because:

- Variance of projected data ~ Explained Inertia
- (Explained + residual = total) inertia
- Inertia ~ Trace (invariant) (Slide 67)

⇒ line $\vec{u}$ maximising variance of $Proj_{\vec{u}}(X)$ = line $\vec{u}$ minimizing the residual inertia

$$\Rightarrow \vec{u}^* = \underset{\|\vec{u}\|=1}{\operatorname{argmin}} \sum_{i=1}^{M} \|x_i - \underbrace{\langle x_i, \vec{u} \rangle \vec{u}}_{\text{projection}}\|^2$$

$$\|x_i - \langle x_i, u \rangle u\|^2 = \langle x_i - \langle x_i, u \rangle u, x_i - \langle x_i, u \rangle u \rangle$$

Constant wrt u $\rightarrow$ $= \langle x_i, x_i \rangle - 2\langle \langle x_i, u \rangle u, x_i \rangle + \langle u, u \rangle$ ($\langle u, u \rangle \rightarrow 1$)

$$= -2\langle x_i, u \rangle \langle u, x_i \rangle = -2 \langle u, x_i \rangle \langle x_i, u \rangle$$ (since scalars)

$$= -2 u^T x_i x_i^T u$$

$$\Rightarrow \underset{u^T u = 1}{\operatorname{argmin}} \sum_i Tr(u^T X X^T u) \Rightarrow J = u^T X X^T u - \lambda(1 - u^T u)$$

$$\frac{\partial J}{\partial u} = 0 \Rightarrow X X^T u = \lambda u \Rightarrow u \text{ eigvect of } \underbrace{X X^T}_{\text{cov}}$$

In summary:

on subspace Δ_1 (line of direction $\vec{u}$ going through) the projected dataset retains the maximum of the variance of the original X.

the rest of the variance is on the orthogonal subspace $\Delta_1^\perp$ of dimension $N-1$

The procedure is iterated over $Proj_{\Delta_1^\perp}(X)$ to find Δ_2 ... etc

In effect, this corresponds to finding all eigenvectors u_i of the covariance matrix of the centered data $\Sigma = XX^T$ (slide 69)

u_i = Principal Component associated to $\lambda_i = I_{\Delta_i}$

5 - Properties of PCA

One can summarize PCA as:

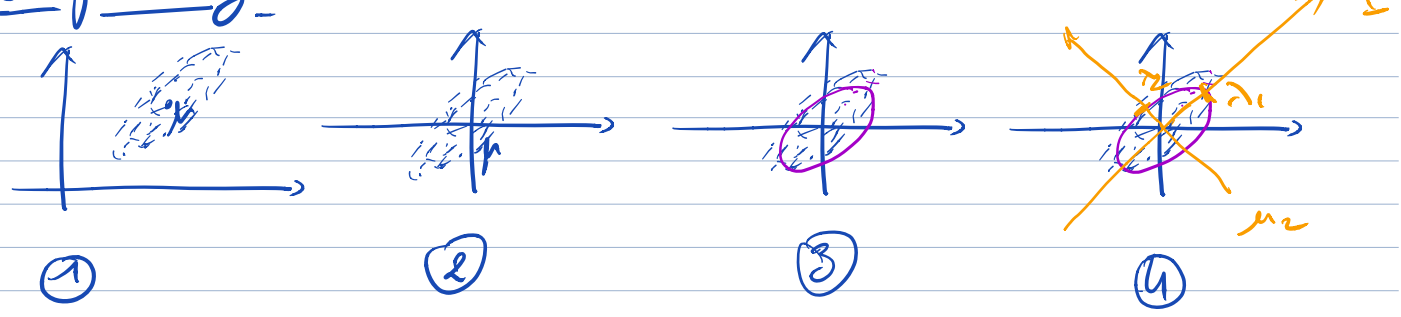
① Get data $X \in \mathbb{R}^{N \times M}$

② $X \leftarrow X - \mu$ Centering

③ $\Sigma \leftarrow XX^T$ Covariance

④ $\Sigma = U \Lambda U^T$ gives vectors.

Graphically:



we have $XX^T = U \Lambda U^T \Rightarrow \Lambda = U^T X X^T U$

$\Rightarrow \Lambda = (U^T X)(U^T X)^T$

Define new representation $Y = U^T X$ $(y_i = U^T x_i)$

the covariance matrix in that new representation

$\Rightarrow YY^T = \Lambda$ (diagonal)

$\Rightarrow$ Data components are decorrelated

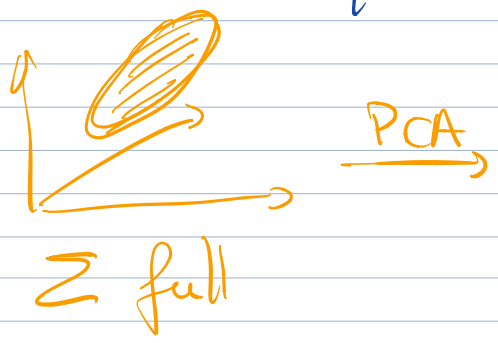
given a new point $x_{new} \Rightarrow$ its representation is

$y_{new} = U^T(x_{new} - \mu) \Leftrightarrow x_{new} = U y_{new} + \mu$.

Note: By considering mean (μ) and variance (Σ), PCA considers the data distributed with an underlying Gaussian (normal) distribution.

Principal directions u_i are in the directions of maximally decreasing variance. It makes it simple and optimal to decimate the dimensions of the space

Eg: Take data distributed on a 2D ellipse add one component and rotate and translate



Σ full

$\Lambda = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}$