Statistiques Multivariées

Hicham Janati

hjanati@insea.ac.ma

INSEA

Table des matières

INSEA

Statistiques Multivariées

Introduction

Hicham Janati

hjanati@insea.ac.ma

INSEA

Pourquoi ce cours ?

Données multi-variées = Plusieurs variables observées en même temps

- Données médicales physiologiques (Température, pression artérielle, cholesterol …)
- Données météorologiques (vent, précipitation, température, couverture nuageuse …)
- Données financières (prix d'actifs financiers, volumes de transactions …)
- Données images

Comment est-ce que l'analyse univariée des données peut-elle conduire à des biais / conclusions erronées ?

L'effet des interactions entre les variables doit être modélisé.

INSEA

Exemple 1: Inférence

Essai clinique: Efficacité d'un médicament pour traiter l'hypertension

Blood Pressure	Cholesterol Level	Group
118.29	198.12	Control
122.35	210.54	Control
116.58	180.73	Treatment
119.02	193.47	Control
121.94	202.21	Treatment
124.68	195.18	Treatment
⋮	⋮	⋮

Table 1: Sample of the Multivariate Data

Peut-on effectuer une analyse univariée i.e variable par variable ?

Exemple 1: Inférence / Tests

Essai clinique: Efficacité d'un médicament pour traiter l'hypertension

Pour chaque variable, on teste l'hypothèse d'égalité des moyennes

$$H_0 : \mu_{\text{control}} = \mu_{\text{treatment}}$$

Variable	T-statistic	P-value
Blood Pressure	1.17	0.24
Cholesterol Level	1.92	0.056

Table 1: Results of the Univariate T-Tests

Basé sur la distribution Gaussienne multivariée

Problème 1: Conclusions différentes !

On effectue **le test multivarié de Hotelling** (*qu'on découvrira plus tard*) :

Test	T-squared Statistic	P-value
Hotelling T-squared Test	4.01	0.023

Table 1: Results of the Hotelling T-squared Test

Exemple 1: Inférence / Tests

Essai clinique: Efficacité d'un médicament pour traiter l'hypertension

Pour chaque variable, on teste l'hypothèse d'égalité des moyennes

Problème 2: Supposons que les données contiennent 100 variables. Le seuil de tolérance est de 5%. On effectue 100 tests de Student. On obtient donc 100 p-valeurs triées:

Variable	P-value
V4	0.001
V30	0.004
V2	0.01
V8	0.02
V74	0.025
V18	0.054
V19	0.07
⋮	⋮

Que peut-on dire ?

Tests multiples univariés: il faut tenir compte du nombre de tests effectués ! La p-value est la probabilité de rejeter à tort l'hypothèse nulle. Avec 100 tests, et un seuil de 5%, on s'attend à avoir 5 faux positifs.

Exemple 2: Tests multivariés

Tests d'égalités multiples

Nous avons les données d'observations de plusieurs lancers d'un dé dont les probabilités sont notés $p_1, p_2, p_3, p_4, p_5, p_6$

On souhaite écarter l'hypothèse que le dé soit équitable.

Quelle analyse proposez-vous ?

Idéalement on veut tester les égalités:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6$$

Réécrivez cette hypothèse en multivarié i.e avec une seule égalité faisant intervenir le vecteur

$$p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{pmatrix}$$

$$\mathrm{C} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$H_0 : Cp = 0$$

INSEA

Exemple 2: Tests multivariés

Tests de covariances

Nous avons les données de retour sur investissement de plusieurs actifs financiers. Vous voulez évaluer si la corrélation entre ces variables a changé après un événement majeur (covid).

Avant la crise

Date	AAPL	MSFT	TSLA	BTC-USD	ETH-USD
01/01/2020	0.01	0.03	-0.02	0.05	0.04
02/01/2020	-0.01	0.04	0.01	0.03	0.06
03/01/2020	0.00	-0.02	0.02	0.04	-0.01
04/01/2020	0.02	0.01	-0.01	-0.02	0.03
05/01/2020	-0.03	0.03	0.00	0.06	0.02

$\longrightarrow \Sigma_{\text{avant}}$

Après la crise

Date	AAPL	MSFT	TSLA	BTC-USD	ETH-USD
01/01/2024	0.01	0.03	-0.02	0.05	0.04
02/01/2024	-0.01	0.04	0.01	0.03	0.06
03/01/2024	0.00	-0.02	0.02	0.04	-0.01
04/01/2024	0.02	0.01	-0.01	-0.02	0.03
05/01/2024	-0.03	0.03	0.00	0.06	0.02

$\longrightarrow \Sigma_{\text{après}}$

On teste l'égalité des matrices de covariances:

$$H_0 : \Sigma_{\text{avant}} = \Sigma_{\text{après}}$$

INSEA

Pourquoi ce cours ?

Récap des tests:

L'analyse univariée séquentielle (variable l'une après l'autre) peut se faire mais:

1. Elle ignore les corrélations et dépendances entre les variables
2. Elle nécessite la réalisation de plusieurs tests statistiques qui augmente la probabilité de rejeter à tort une ou plusieurs hypothèses*.
3. Certaines hypothèses complexes nécessitent une formulation multivariée (ex: tests multiples, tests de covariances)

* On peut quand même y remédier en utilisant une correction pour tests multiples (Bonferroni, Benjamini-Hochberg ..)

INSEA

Pourquoi ce cours ?

La Gausienne (loi normale) multivariée joue un rôle centrale dans les tests d'hypothèse mais pas que …

Exemple 3: Détection de spam

Les données sont des SMS avec un label: spam / non-spam

Texte SMS	**Spam (1) / Non-Spam (0)**
Salut, tu viens ce soir ?	0
Votre rendez-vous est confirmé pour demain.	0
Vous avez un colis en attente, cliquez ici pour récupérer.	**1**
Bonjour, comment vas-tu aujourd'hui ?	0
...	**...**
Inscrivez-vous maintenant pour gagner un iPhone !	**1**
J'ai bien reçu ton email, merci !	0
On se retrouve à 14h devant la gare.	0
N'oublie pas notre réunion demain matin.	0

Comment transformer les lignes texte en données numériques ?

Exemple 3: Détection de spam

Les données sont des SMS avec un label: spam / non-spam

La transformation TF-IDF

ID SMS	Type	Salut	tu	viens	...	colis	gagner	email	réunion
1	Ok	0.2	0.2	0.2	...	0	0	0	0
2	Ok	0	0	0	...	0	0	0	0
3	Spam	0	0	0	...	0.17	0	0	0
4	Ok	0	0.167	0	...	0	0	0	0
5	Spam	0	0	0	...	0	0.2	0	0
6	Ok	0	0	0	...	0	0	0.25	0
7	Ok	0	0	0	...	0	0	0	0.2
8	Ok	0	0	0	...	0	0	0	0.167
...	...	...	...	...	...	...	...	...	...

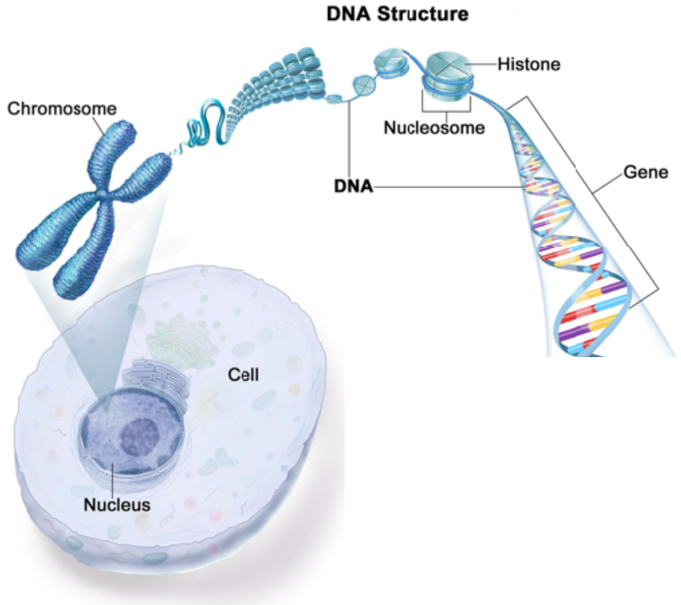
$$\text{Spam} \sim \mathcal{N}(\mu_s, \text{Id})$$
$$\text{Ok} \sim \mathcal{N}(\mu_o, \text{Id})$$

1. Quelle est la dimension de ces données ?

2. Pensez vous que la classification (détection de spam) nécessite un modèle simple ou sophistiqué ?

Un modèle Gaussien multivarié simple est suffisant (Naive Bayes)

Exemple 4: Clustering



Données génomiques

L'ADN humain compte environ 20000 gènes codant des protéines

Cellule	Gène 1	Gène 2	...	Gène 20000
1	2.3	1.5	...	0.6
2	5.2	4.1	...	3.1
3	2.5	1.6	...	0.7
4	5.1	4.0	...	3.2
5	2.4	1.4	...	0.5
6	5.3	4.2	...	3.0
7	2.6	1.7	...	0.8
8	5.0	3.9	...	3.3
...	...	...	...	...

Table 1: Données d'expression génique

$$X \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

Comment identifier les groupes de cellules ayant des profils géniques similaires ?

Modèles de classification non-supervisée (Mélange de Gaussiennes multivariées)

INSEA

Exemple 5: Régression linéaire

Objectif: estimer l'intensité d'activation en chaque point du cerveau donnée par le vecteur θ

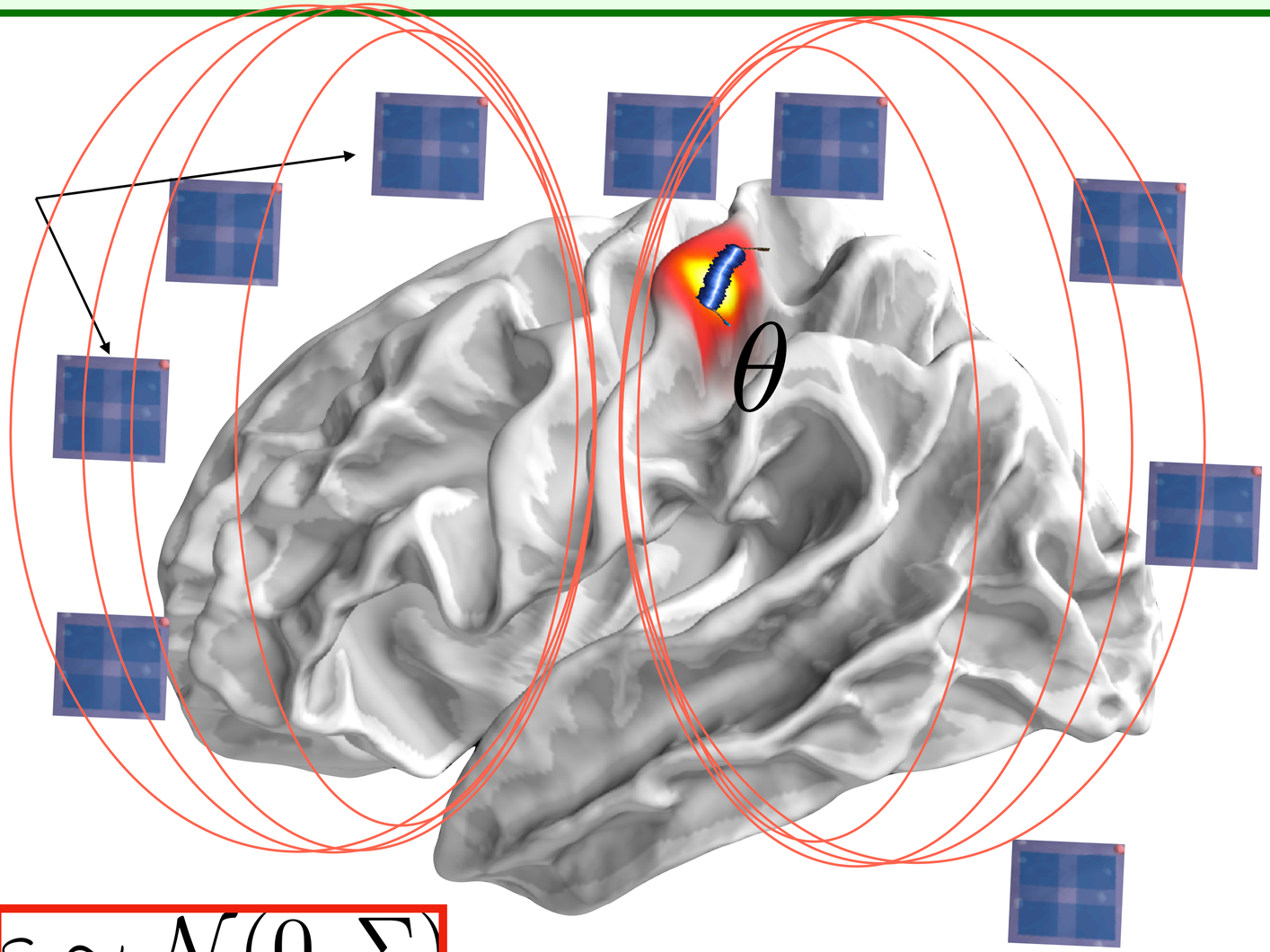
n capteurs magnétiques

Il faut tenir compte de l'erreur

$$= y = X\theta$$

Équations de Maxwell

$$\in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p}$$



$$\varepsilon \sim \mathcal{N}(0, \Sigma)$$

Programme

Rappels d’algèbre et de probabilités

I - Loi normale (Gaussienne) multivariée

II - Inférence et test d’hypothèses

III - Modèles de prédiction / classification

INSEA

Statistiques Multivariées

Rappels

Hicham Janati

hjanati@insea.ac.ma

INSEA

Soit $a, b \in \mathbb{R}^d$

Le produit scalaire entre a et b s'écrit : $\langle a, b \rangle = \sum_{i=1}^{d} a_i b_i$

et peut aussi s'écrire: $(a_1, \ldots, a_d) \begin{pmatrix} b_1 \\ \vdots \\ b_d \end{pmatrix} = a^\top b$:

On lit: "a transposée b"

Par convention, $\|a\|$ correspond à la norme 2 (Euclidienne) : $\|a\| = \sqrt{\sum_{i=1}^{d} a_i^2}$

Une matrice $\mathbf{X} \in \mathbb{R}^{n \times d}$ a n lignes et d colonnes.

On note sa $i-$ème ligne (resp. j-ème colonne) par $\mathbf{X}_{i.}$ (resp. $\mathbf{X}_{.j}$)

Soit $a \in \mathbb{R}^d$

$\text{diag}(a)$ correspond à la matrice diagonale $\begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \vdots & 0 \\ 0 & \dots & 0 & a_d \end{pmatrix}$

$\mathbb{1}_d$ correspond au vecteur de dimension d avec des 1 partout $\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}^\top$

La matric Identité est donc $\mathrm{I}_d = \text{diag}(\mathbb{1}_d)$

Le déterminant d'une matrice carrée $\mathbf{A} \in \mathbb{R}^{d \times d}$ est noté par $\det(\mathbf{A})$ ou $|\mathbf{A}|$

Soit $\mathbf{A} \in \mathbb{S}_d$ i.e une matrice symmétrique

Le théorème spectral garantit l'existence de d valeurs propres λ_i associées à d vecteurs propres orthogonaux e_i qui vérifient pour tout $i = 1..d$:

$$\mathbf{A}e_i = \lambda_i e_i$$

Matriciellement, il existe une matrice diagonale Λ et une matrice orthogonale $\mathbf{P}$ telles ques:

$$\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}^\top$$

Comment a-t-on obtenu l'écriture matricielle ?

Interprétation du théorème spectral

Toute matrice semi-définie positive peut s'écrire comme somme pondérée de matrices de rang 1

En plus, $\mathbf{A}$ est semi-définie positive (resp. définie positive) ssi pour tout $i = 1..d$ $\lambda_i \geq 0$ (resp. $\lambda_i > 0$)

Le rang de $\mathbf{A}$ noté par $\text{rank}(\mathbf{A})$ est égal au nombre de valeurs propres non nulles.

La trace de $\mathbf{A}$ est notée par $\text{tr}(\mathbf{A}) = \sum_{i=1}^{d} \lambda_i$

Son déterminant est $\det(\mathbf{A}) = \prod_{i=1}^{d} \lambda_i$

Soit $\mathbf{A} \in \mathbb{S}_d$ i.e une matrice symmétrique

Matriciellement, il existe une matrice diagonale Λ et une matrice orthogonale $\mathbf{P}$ telles ques:

$$\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}^\top$$

Et donc, comme $\mathbf{P}\mathbf{P}^\top = \mathrm{I}_d$, pour tout $q \in \mathbb{N}_*$, on a:

$$\mathbf{A}^q = \underbrace{\mathbf{P}\Lambda\mathbf{P}^\top\mathbf{P}\Lambda\mathbf{P}^\top \dots \mathbf{P}\Lambda\mathbf{P}^\top}_{q \text{ fois}} = \mathbf{P}\Lambda^q\mathbf{P}^\top$$

Si $\mathbf{A}$ est semi-définie positive, on définit la racine carrée matricielle:

$$\mathbf{A}^{\frac{1}{2}} = \mathbf{P}\Lambda^{\frac{1}{2}}\mathbf{P}^\top$$

Où $\Lambda^{\frac{1}{2}} = \mathrm{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_d})$

Soit $\mathbf{A} \in \mathbb{S}_d$ i.e une matrice symmétrique

Matriciellement, il existe une matrice diagonale Λ et une matrice orthogonale $\mathbf{P}$ telles ques:

$$\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}^\top$$

Même si A n'est pas inversible, on définit la pseudo-inverse:

$$\mathbf{A}^+ = \mathbf{P}\Lambda^+\mathbf{P}^\top$$

$$\text{Où } \Lambda^+ = \text{diag}(\lambda_1^+, \ldots, \lambda_d)$$

$$\text{Avec } \lambda_i^+ = \frac{1}{\lambda_i} \text{ si } \lambda_i \neq 0 \text{ et } \lambda_i^+ = 0 \text{ sinon.}$$

Soit $\mathbf{A} \in \mathbb{S}_d$ i.e une matrice symmétrique

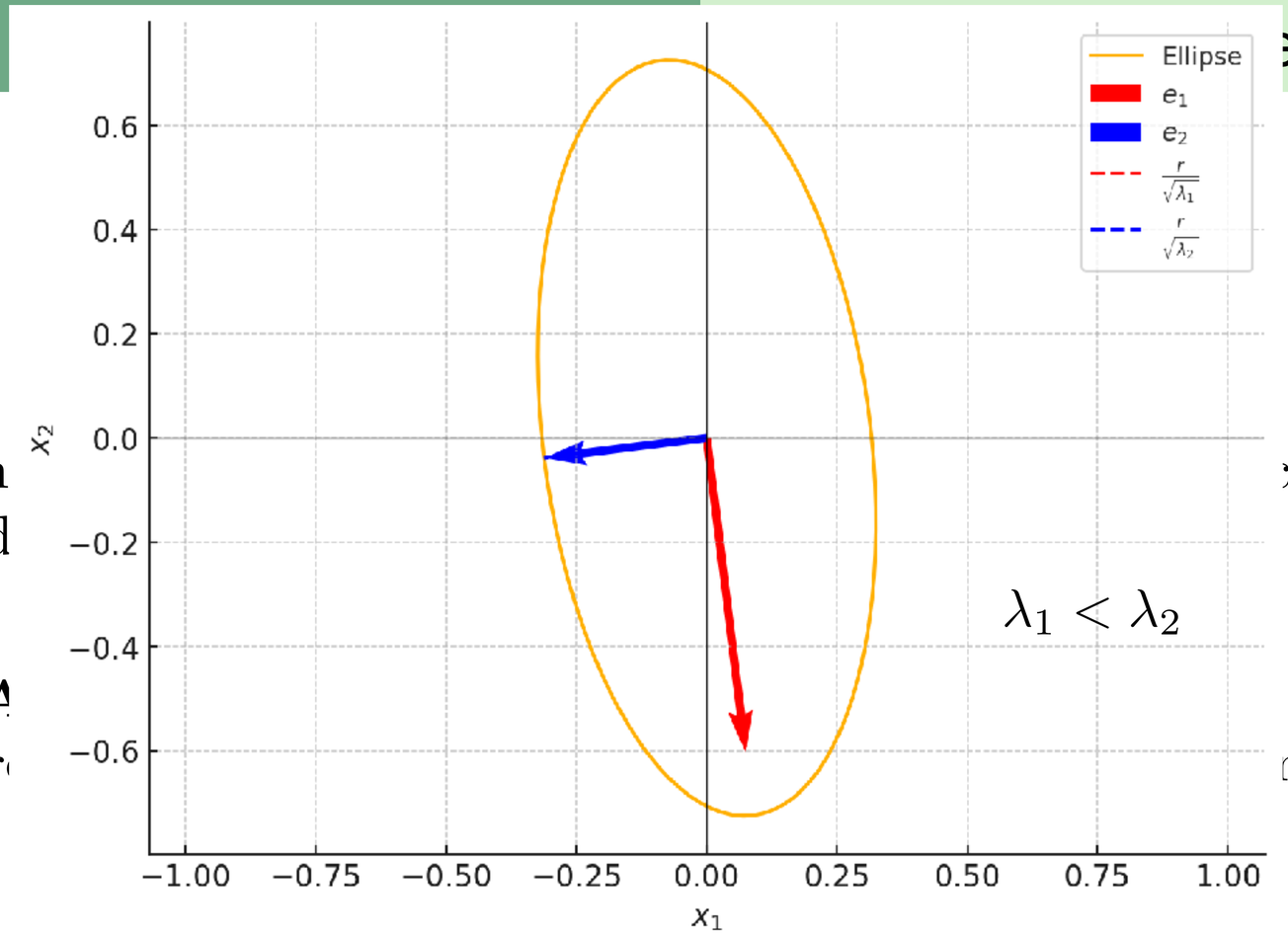
La fonction de la forme:

$$Q : x \mapsto x^\top \mathbf{A} x$$

est appelée forme quadratique, elle est définie (semi-définie) ssi $\mathbf{A}$ est définie (semi-définie) positive ssi pour tout $x \neq 0 \quad Q(x) > 0$ $(Q(x) \geq 0)$

On se positionne dans $\mathbb{R}^2$. Quelle est la matrice des formes quadratiques suivantes:

1. $x \mapsto x_1^2 + x_2^2$
2. $x \mapsto x_1^2 - 4x_1x_2 + x_2^2$
3. $x \mapsto x_1^2 + x_1x_2 + 4x_2^2$



La fon ce si et seulement si $\mathbf{A}$ est d

Avec A

Et la r ntrée en x_0 de rayon r.

En général la région $(x - x_0)^\top \mathbf{A}(x - x_0) = r$ correspond à une ellipse dont les axes principaux sont donnés par e_i les vecteurs propres de $\mathbf{A}$.

La longueur de ses demi-axes est donnée par $\frac{r}{\sqrt{\lambda_i}}$.

Prenons désormais un vecteur aléatoire en dimension d noté par $\mathbf{X} = (X_1, \ldots, X_d)^\top$.

Sa fonction de répartition (*cumulative distribution function* / cdf) est définie par:

$$F : x \in \mathbb{R}^d \mapsto \mathbb{P}(\mathbf{X} \leq x) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) \in [0, 1]$$

$\mathbf{X}$ est continue ssi il existe une densité de probabilité (pdf) (*probability density function*) f telle que:

$$F(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(u_1, \ldots, u_d) \mathrm{d}u_1 \ldots \mathrm{d}u_d$$

Notons les k premières composantes de $\mathbf{X}$ par $\mathbf{A}$: $\mathbf{X} = \left(\underbrace{X_1, \ldots, X_k}_{\mathbf{A}}, X_{k+1}, \ldots, X_d \right)$

Comment peut on obtenir la densité de A ?

La fonction $F_{\mathbf{A}}(a) = \mathbb{P}(\mathbf{A} \leq a) = F(a_1, \ldots, a_k, +\infty, \ldots, +\infty)$ est dite fonction de répartition *marginale*

Sa densité (si $\mathbf{X}$ est continue) est obtenue en intégrant le reste des variables:

$$f_{\mathbf{A}}(a) = \int_{\mathbb{R}^{d-k}} f(a, u_{k+1}, \ldots, u_d) \mathrm{d}u_{k+1}, \ldots, \mathrm{d}u_d$$

Soit $\mathbf{X}$ un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)^\top$ dont la densité est notée f.

Son espérance est donnée par:

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_d))^\top = \int x f(x) \mathrm{d}x = \begin{pmatrix} \int x_1 f(x) \mathrm{d}x \\ \vdots \\ \int x_d f(x) \mathrm{d}x \end{pmatrix}$$

Avec la linéarité de l'espérance, pour toute matrice $A \in \mathbb{R}^{n \times d}$:

$$\mathbb{E}(A\mathbf{X}) = A\mathbb{E}(\mathbf{X})$$

Soit $\mathbf{X}, \mathbf{Y}$ deux vecteurs aléatoires de dimension d dont les espérances sont $\mu_{\mathbf{X}}$ et $\mu_{\mathbf{Y}}$

$$\text{La matrice } \mathbb{V}(\mathbf{X}) = \Sigma = \mathbb{E}\left((\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^\top\right)$$

est appelée la matrice de covariance de $\mathbf{X}$ qui correspond également à:

$$\Sigma = \begin{pmatrix} \mathbb{V}(X_1) & \cdots & \sigma_{X_1 X_p} \\ \vdots & \ddots & \vdots \\ \sigma_{X_p X_1} & \cdots & \mathbb{V}(X_p) \end{pmatrix}.$$

On définit la covariance entre $\mathbf{X}$ et $\mathbf{Y}$ par:

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left((\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})^\top\right) = \mathbb{E}(\mathbf{X}\mathbf{Y}^\top) - \mu_{\mathbf{X}}\mu_{\mathbf{Y}}^\top$$

Si $\mathbf{X}$ et $\mathbf{Y}$ sont indépendants alors $\mathbb{E}(\mathbf{X}\mathbf{Y}^\top) = \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{Y}^\top)$

INSEA

Propriétés de la variance et covariance

$$\Sigma = \mathbb{E}(\mathbf{X}\mathbf{X}^\top) - \mu_{\mathbf{X}}\mu_{\mathbf{X}}^\top$$

$$\Sigma \geq 0$$

$$\mathbb{V}(a^\top \mathbf{X}) = a^\top \mathbb{V}(\mathbf{X})a$$

$$\mathbb{V}(A\mathbf{X} + b) = A\mathbb{V}(\mathbf{X})A^\top$$

$$\mathbb{V}(\mathbf{X} + \mathbf{Y}) = \mathbb{V}(\mathbf{X}) + \mathbb{V}(\mathbf{Y}) + \mathrm{Cov}(\mathbf{X}, \mathbf{Y}) + \mathrm{Cov}(\mathbf{Y}, \mathbf{X})$$

Théorème de Cramer-Wold

La distribution de $\mathbf{X} \in \mathbb{R}^d$ *est entièrement déterminée par l'ensemble de toutes les distributions (univariées) de* $t^\top \mathbf{X}$ *où* $t \in \mathbb{R}^d$.

Ce théorème signifie que nous pouvons déterminer la distribution de $\mathbf{X}$ dans $\mathbb{R}^d$ en spécifiant toutes les distributions unidimensionnelles des combinaisons linéaires

$$\sum_{j=1}^{d} t_j X_j = t^\top \mathbf{X}, \quad t = (t_1, t_2, \ldots, t_d)^\top.$$

Fonction caractéristique

$$\phi_{\mathbf{X}} : t \in \mathbb{R}^d \mapsto \mathbb{E}(e^{it^\top \mathbf{X}}) = \int e^{it^\top x} f(x)\, dx \in \mathbb{C}$$

Fonction caract. marginale

Si $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top$, alors pour $t = (t_1, t_2, \ldots, t_d)^\top$

$$\phi_{X_1}(t_1) = \phi_{\mathbf{X}}(t_1, 0, \ldots, 0), \quad \ldots, \quad \phi_{X_d}(t_d) = \phi_{\mathbf{X}}(0, \ldots, 0, t_d).$$

Fonction caract. et indépendance

Si $X_1, \ldots, X_d$ sont des variables aléatoires indépendantes, alors pour $t = (t_1, t_2, \ldots, t_d)^\top$

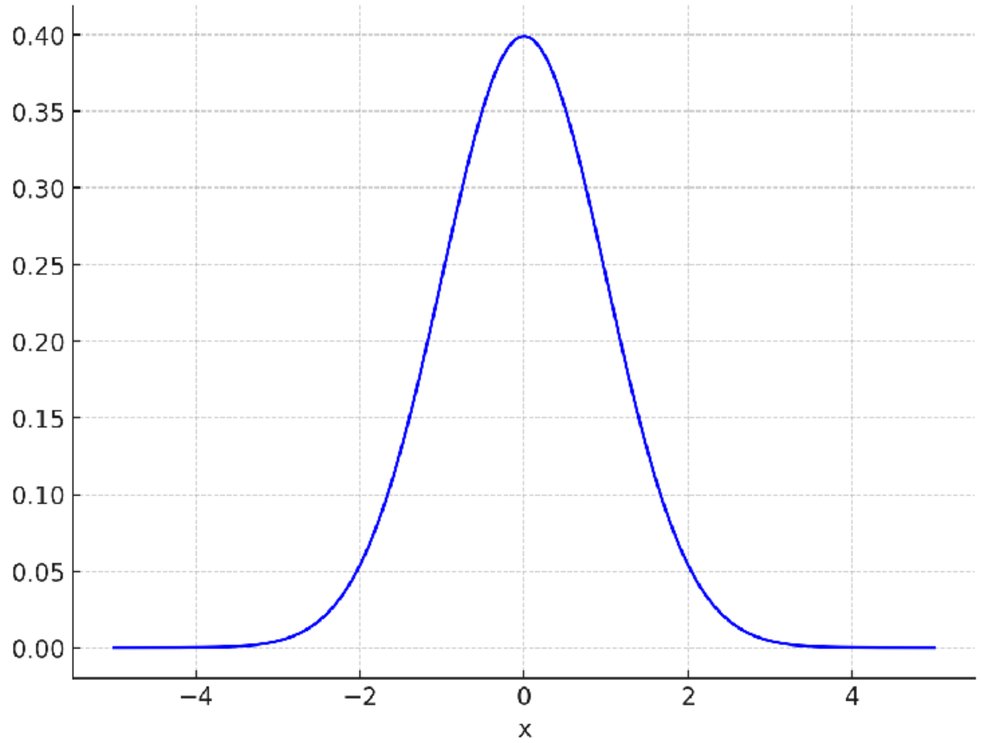
$$\phi_{\mathbf{X}}(t) = \phi_{X_1}(t_1) \cdots \phi_{X_d}(t_d).$$

INSEA

I - Loi normale multivariée

Soit X une gaussienne univariée de moyenne μ et de variance σ^2. On note $X \sim \mathcal{N}(\mu, \sigma^2)$. Sa densité est donnée par:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$



Soit $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_d)$ un vecteur aléatoire de moyenne $\mathbb{E}(\mathbf{X}) = \mu$ et de variance $\mathbb{E}(\mathbf{X}\mathbf{X}^\top) - \mu\mu^\top = \Sigma$

Définition

$\mathbf{X}$ est un vecteur Gaussien ssi toute combinaison linéaire des $\mathbf{X}_i$ est une Gaussienne univariée:

$$(\forall \alpha \in \mathbb{R}^d)(\exists \mu_*, \sigma_*) \quad \alpha^\top \mathbf{X} = \sum_{i=1}^{d} \alpha_i \mathbf{X}_i \sim \mathcal{N}(\mu_*, \sigma_*^2)$$

Exercice

Soit $X \sim \mathcal{N}(0,1)$ et une variable binaire $\varepsilon \sim \mathcal{U}(\{-1,1\})$ deux variables aléatoires indépendantes. On définit $Y = \varepsilon X$ et $Z = (X, Y)$.

1. Montrez que les composantes de Z sont des gaussiennes univariées.
2. Z peut-il être un vecteur Gaussien ?

Densité

Soit $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_d)^\top$ un vecteur aléatoire Gaussien avec une moyenne $\boldsymbol{\mu} \in \mathbb{R}^d$ et une matrice de covariance $\Sigma \in \mathbb{R}^{d \times d}$. On note $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Si Σ est inversible, $\mathbf{X}$ admet une densité donnée par:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^d,$$

Exemple dans $\mathbb{R}^2$ avec $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

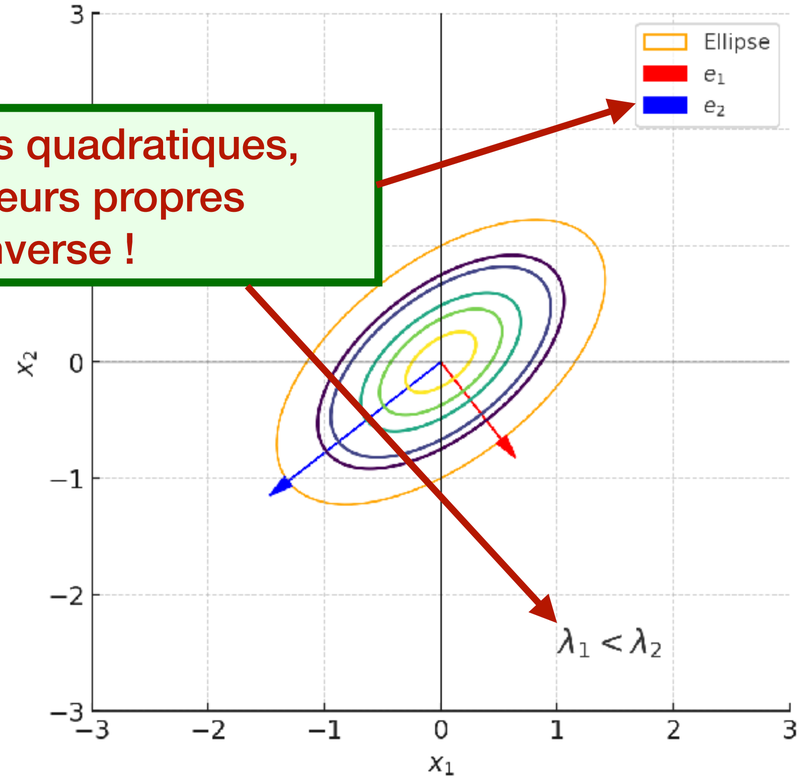
$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^2,$$

On note $\lambda_1 \leq \lambda_2$ et e_1, e_2 les valeurs et vecteurs propres de Σ.

Q9. Détermine
densité, c-à-d
constante:

Attention ! contrairement à la slide des formes quadratiques,
On s'intéresse désormais aux vecteurs/valeurs propres
de la covariance et non pas de son inverse !

$$(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = r^2$$



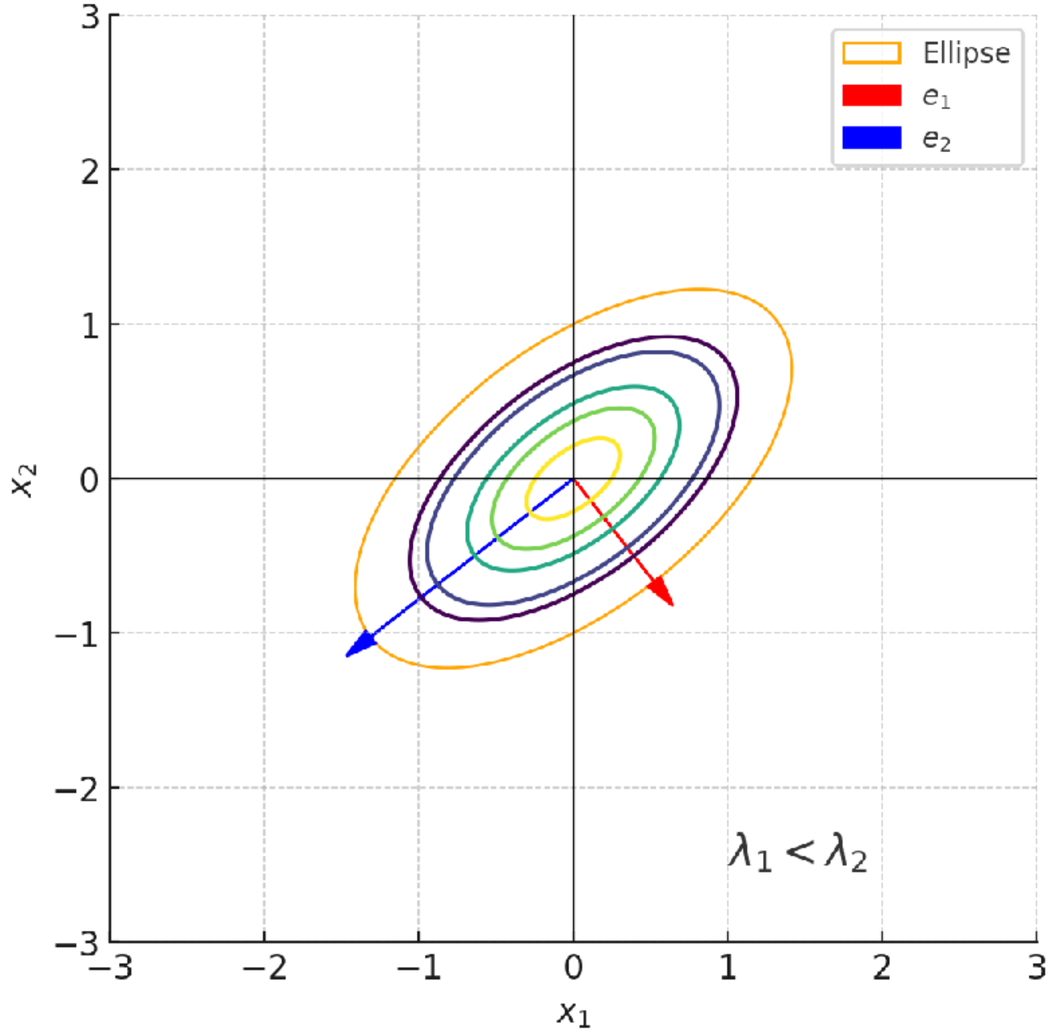
Densité

La région de la forme

$$(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = r^2$$

correspond donc à une ellipse dans les demi-axes sont donnés par les vecteurs propres de la covariance $\sum$

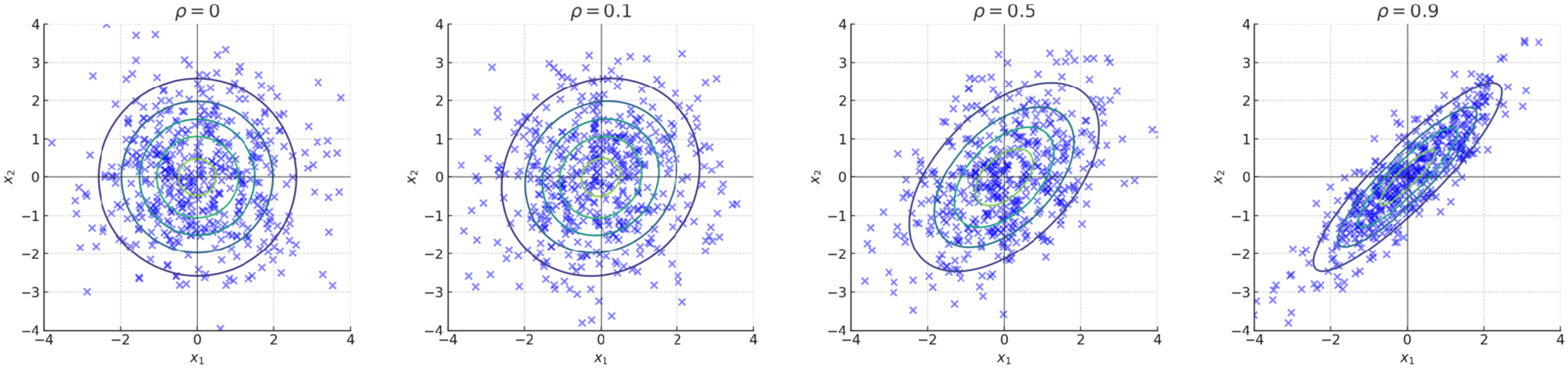
Leur longueur est égale à $\frac{r}{\sqrt{1/\lambda_i}} = r\sqrt{\lambda_i}$



Densité

On génère des échantillons i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}(0, \Sigma)$ où $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

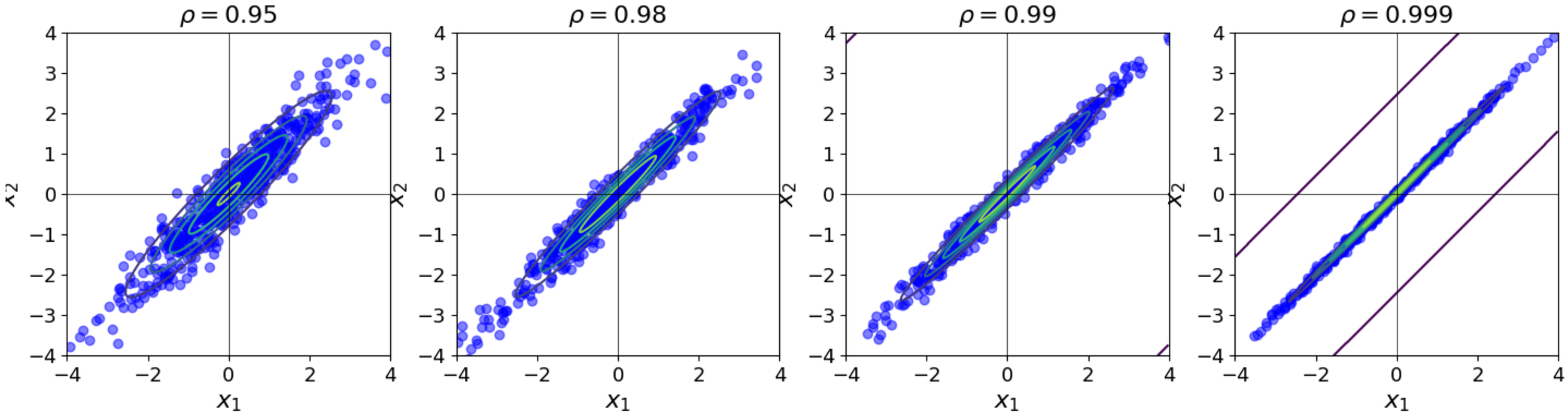
On visualise les données pour plusieurs valeurs de la covariance ρ:



Que se passe-t-il lorsque la corrélation = 1 ?

Densité

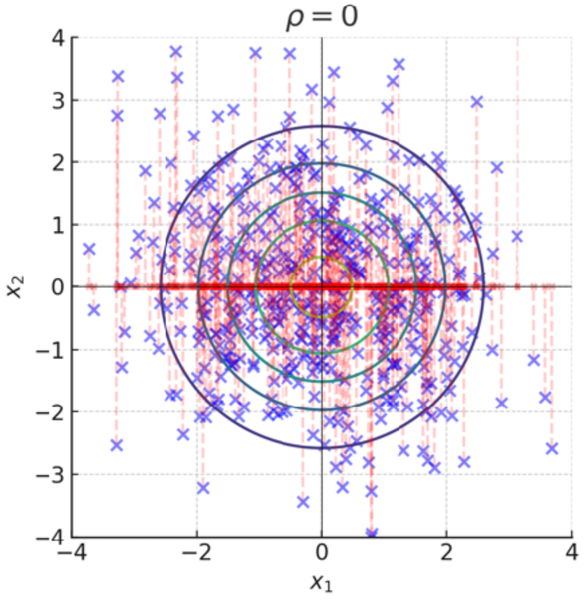
On génère des échantillons i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}(0, \Sigma)$ où $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

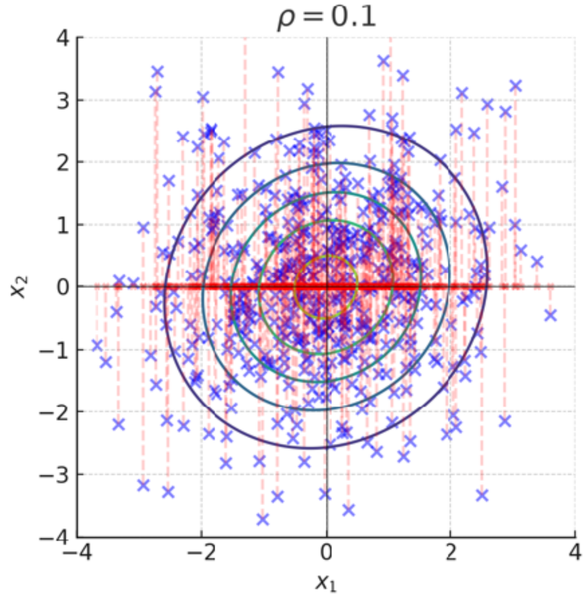


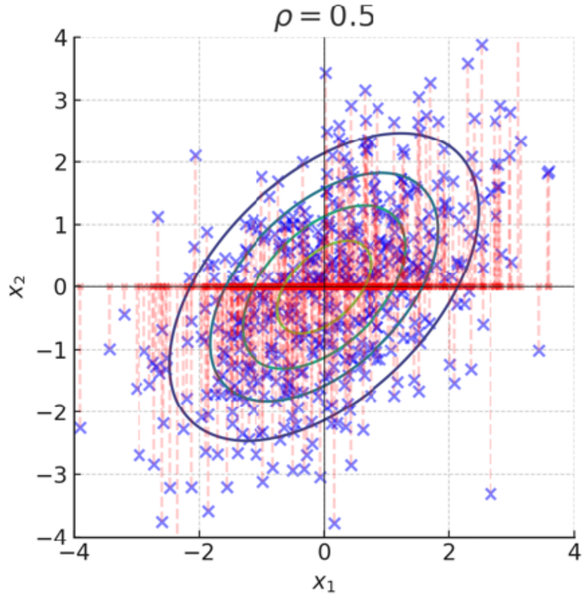
La covariance n'est plus inversible: la distribution devient dégénérée

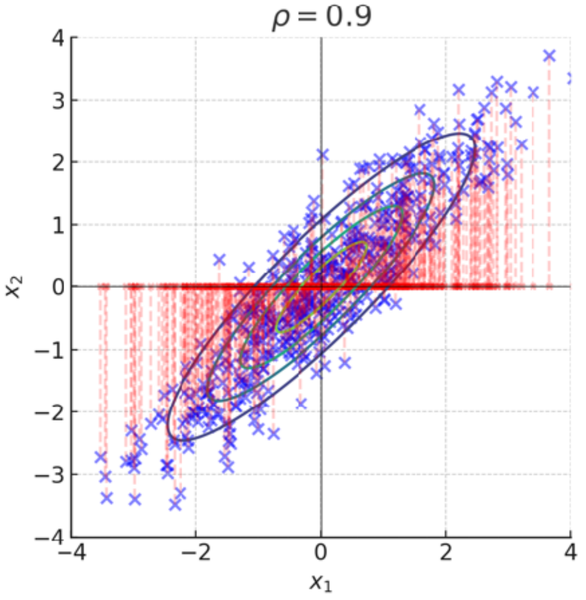
On génère des échantillons i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}(0, \Sigma)$ où $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

On projette les points pour obtenir la première coordonnée



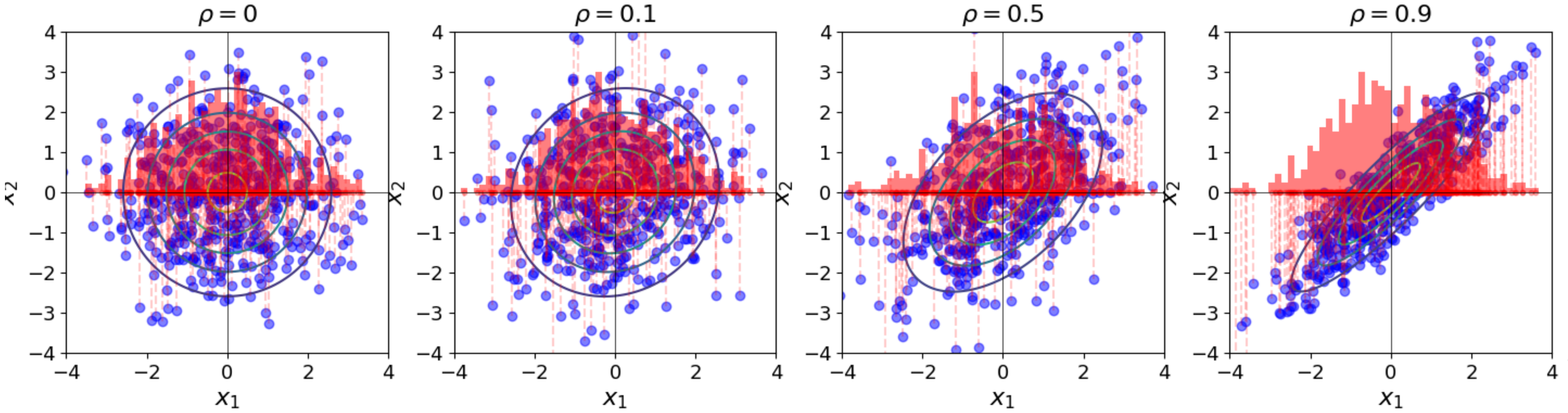






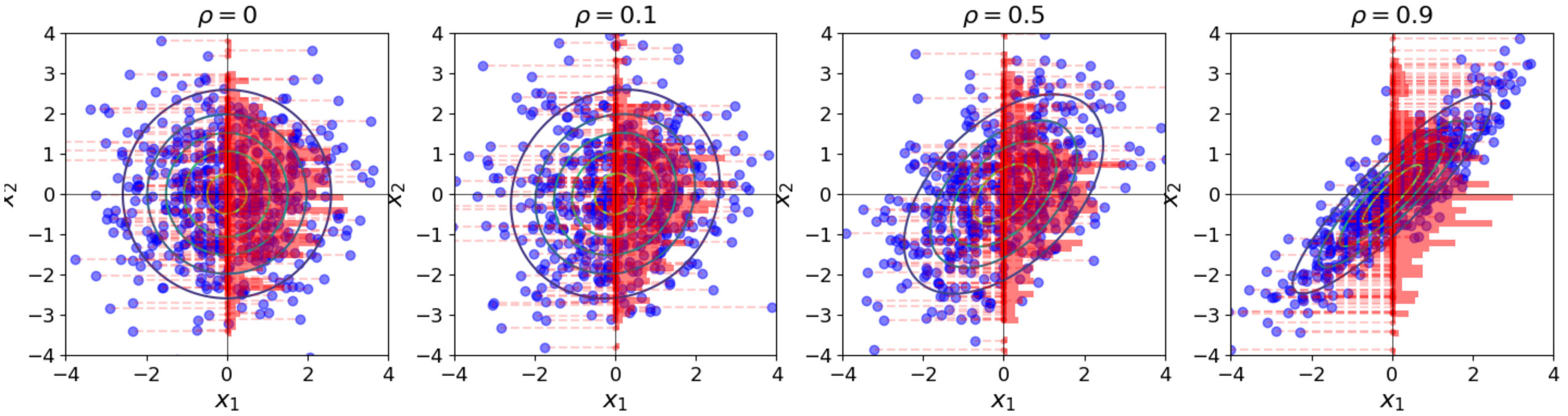
Densité

On génère des échantillons i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}(0, \Sigma)$ où $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.



Densité

On génère des échantillons i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}(0, \Sigma)$ où $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.



Fonction caractéristique

Soit $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$. Pour $t \in \mathbb{R}^d$ la fonction caractéristique est donnée par:

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[e^{i\mathbf{t}^\top \mathbf{X}}\right] = \exp\left(i\mathbf{t}^\top \boldsymbol{\mu} - \tfrac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}\right).$$

On considère maintenant $\mathbf{X} = (\underbrace{X_1, \ldots, X_r}_{X_a}, \underbrace{X_{r+1}, \ldots, X_d}_{X_b}) \sim \mathcal{N}(\mu, \Sigma)$

Où $\mu = (\mu_a, \mu_b)^\top$ et $\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$

Exercice

Montrez que X_a et X_b sont indépendants si et seulement si $\Sigma_{ab} = 0$.

La Gaussienne est la seule loi pour laquelle la dépendance équivaut à la corrélation

Whitening

Soit $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$. Pour $t \in \mathbb{R}^d$ la fonction caractéristique est donnée par:

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[e^{i\mathbf{t}^\top \mathbf{X}}\right] = \exp\left(i\mathbf{t}^\top \boldsymbol{\mu} - \tfrac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}\right).$$

Exercice

1. $\mathcal{A} \in \mathbb{R}^{n \times d}$ et $b \in \mathbb{R}^n$. Trouver la loi de $\mathcal{A}\mathbf{X} + b$.
2. Quelle transformation doit-on appliquer à $\mathbf{X}$ pour obtenir $\mathcal{N}(0, \mathrm{I}_d)$.

Utilité: Certains modèles (ex: ICA) et algorithmes d'optimisation (ex: SGD) nécessitent des variables décorrélées en input.

Whitening

Propriété

Soit $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma), \mathcal{A} \in \mathbb{R}^{n \times d}$ et $b \in \mathbb{R}^n$. Alors:

$$\mathcal{A}\mathbf{X} + b \sim \mathcal{N}(\mathcal{A}\mu + b, \mathcal{A}\Sigma\mathcal{A}^\top)$$

Ainsi, pour $\mathcal{A} = \Sigma^{-\frac{1}{2}}$ et $b = -\mathcal{A}\mu$ on obtient:

$$\Sigma^{-\frac{1}{2}}(\mathbf{X} - \mu) \sim \mathcal{N}(0, \mathrm{I}_d)$$

Soit $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$.

Rappel: Soit $Z_1, \ldots, Z_p \sim \mathcal{N}(0,1)$ des variables i.i.d. Alors $\sum_{i=1}^{p} Z_i^2 \sim \chi^2(p)$

Montrez que $(\mathbf{X} - \mu)^\top \Sigma^{-1} (\mathbf{X} - \mu) \sim \chi^2(d)$

Théorème central limite

Soient $\mathbf{X}_1, \ldots, \mathbf{X}_n$ des observations i.i.d suivant une distribution quelconque de moyenne μ et de variance Σ. On pose $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Alors:

$$\sqrt{n}(\bar{X} - \mu) \underset{n\to\infty}{\sim} \mathcal{N}(0, \Sigma)$$

Corollaire

Si $\hat{\Sigma}_u$ est un estimateur consistent (convergent et de biais nul) vers Σ, alors:

$$\sqrt{n}\hat{\Sigma}_u^{-\frac{1}{2}}(\bar{X} - \mu) \underset{n\to\infty}{\sim} \mathcal{N}(0, \mathrm{I}_d)$$

II - Inférence et tests statistiques

Partie 1 - Cas univarié

Soient $X_1, \ldots, X_n$ des variables univariées i.i.d suivant $\mathcal{N}(\mu, \sigma^2)$

Par indépendance, la vraisemblance du modèle est donnée par:

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}|\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\boldsymbol{\sigma}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x}_i - \boldsymbol{\mu})^2\right)$$

En dérivant, le maximum est atteint en:

$$\hat{\boldsymbol{\mu}} = \bar{X_n} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad \hat{\boldsymbol{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^2$$

Mais l'estimateur de la variance est biaisé: $\mathbb{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$

Soient $X_1, \ldots, X_n$ des variables univariées i.i.d suivant $\mathcal{N}(\mu, \sigma^2)$

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

$$\hat{\boldsymbol{\sigma}_u}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{X}}_n)^2$$

Propriété (Lemme de Fisher)

Biais nul: $\mathbb{E}(\bar{\mathbf{X}}_n) = \mu$ $\quad$ $\mathbb{E}(\hat{\sigma_u}^2) = \sigma^2$

Distribution: $\bar{\mathbf{X}}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\quad$ $(n-1)\frac{\hat{\sigma}_u^2}{\sigma^2} \sim \chi^2_{n-1}$

Indépendance: $\bar{\mathbf{X}}_n$ et $\hat{\sigma}_u^2$ sont indépendants

INSEA

Problématique

Vous êtes un nouveau gamer de Call of Duty. Votre dernier score K/D (kill/death) est 2. Vous pensez avoir le niveau pour commencer votre carrière pro sur Twitch. Comment pouvez-vous vérifier ou rejeter cette hypothèse ?

1. Débutant: On calcule des moyennes approximatives:
Le KD moyen de vos 5 streamers préférés: 2.9

Votre K/D **2.** est inférieur à **2.9** Vous renoncez à la carrière pro.

2. Amateur: Mais vous "savez" que vous êtes mieux qu'un gamer amateur, car la moyenne du KD de tous les gamers est 1.2

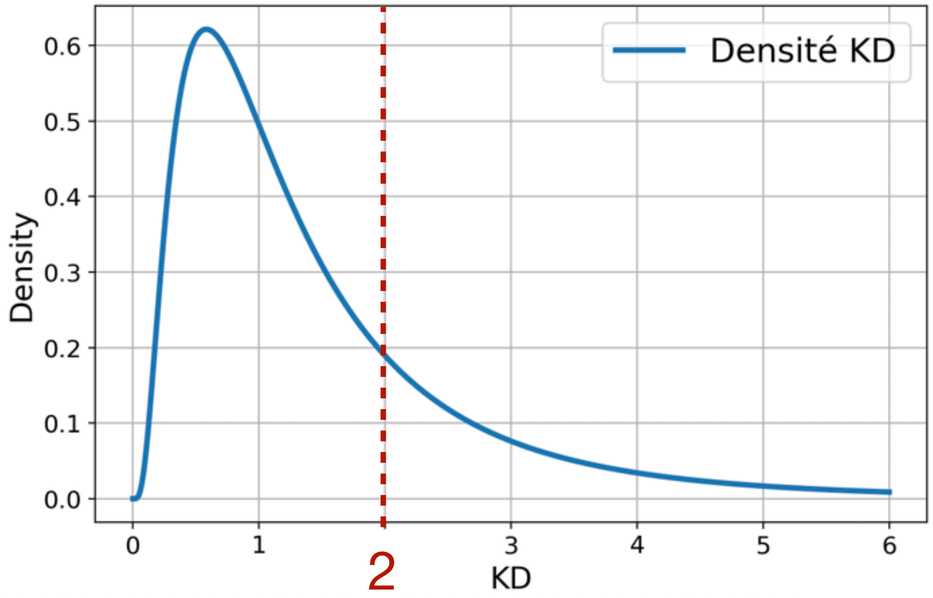
Votre K/D **2** est supérieur à **1.2**. Vous faites carrière pro ?

Il se peut qu'on soit le pire "des pros" mais un pro quand même. Ou inversement, le meilleur amateur, mais pas assez pour en faire une carrière. Comment quantifier à quel point votre K/D est "exceptionnel" ?

Comment quantifier à quel point votre K/D est "exceptionnel" -> "exceptionnellement **rare**" ?

3. Avancé:

On a une **distribution** du K/D des amateurs avec une moyenne **1.2**. **Si** on est un amateur, **alors** notre observation **2.0** "est un échantillon de **cette loi**". Vérifions visuellement si cette hypothèse peut être vraie.



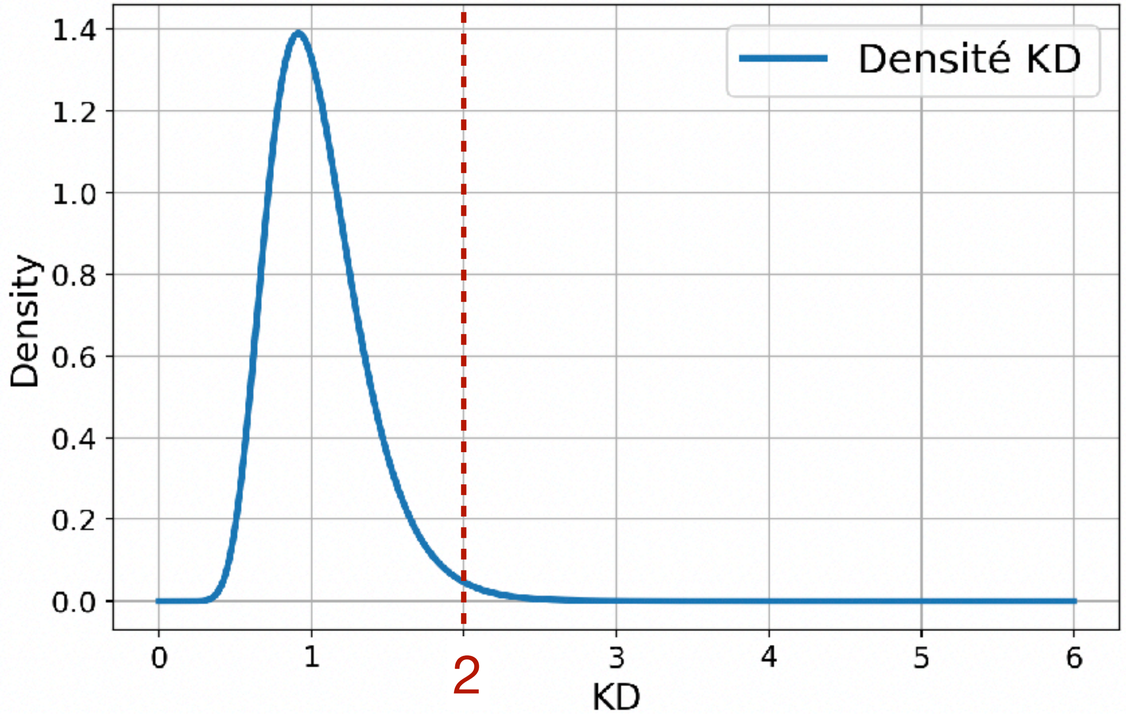
Notre valeur n'est pas "très grande + rare" pour cette distribution, 2. semble plausible. Notre K/D moyen peut très bien suivre cette distribution.

L'hypothèse que notre KD suit la loi en bleu n'est pas absurde. On ne peut donc pas rejeter cette hypothèse.

Comment quantifier à quel point votre K/D est "exceptionnel" -> "exceptionnellement **rare**" ?

3. Avancé:

On a une **distribution** du K/D des amateurs avec une moyenne **1.2**. **Si** on est un amateur, **alors** notre observation **2.0** "est un échantillon de **cette loi**". Vérifions visuellement si cette hypothèse peut être vraie.



Et dans ce cas ?

Notre valeur observée 2 est tellement grande qu'elle devient rare pour cette distribution.

Tellement rare qu'on remet en question l'hypothèse effectuée: l'hypothèse est t**rès probablement** fausse.

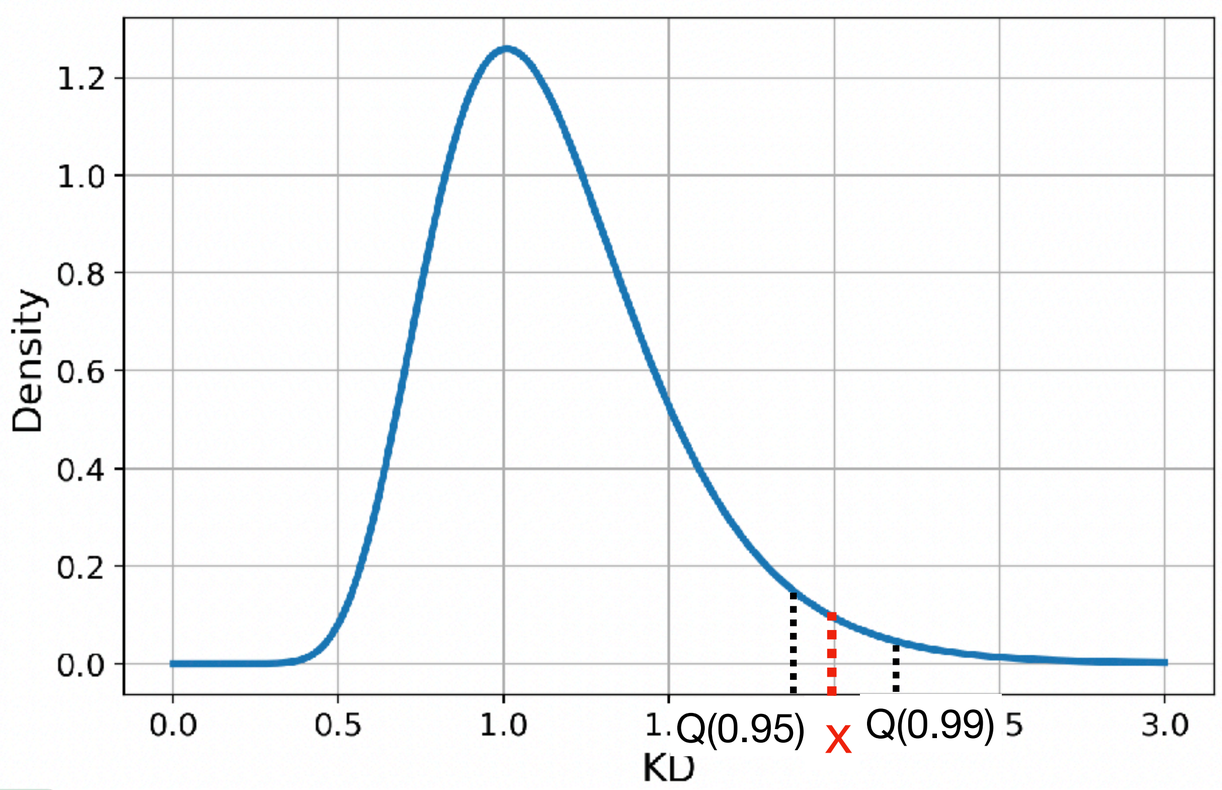
Comment développer une règle générale ?

4. Expert: supposons X = KD ~ une loi connue: **on a sa CDF + quantiles (CDF^{-1})**. On observe x = 2.0

"x est trop grand s'il a dépassé un c (x > c) tel que la probabilité de le dépasser est très petite, disons 0.05"

$$\mathbb{P}(X \geq c) = 0.05 \Leftrightarrow 1 - F_X(c) = 0.05 \Leftrightarrow F_X(c) = 0.95 \Leftrightarrow c = F_X^{-1}(0.95) = Q_X(0.95) = 1.9$$

"x est trop grand s'il a dépassé 1.9"



On a supposé que notre x soit issu de cette loi

Selon cette loi, P(observer qqch > 1.9) = 0.05

Observer x = 2 est encore plus rare

On **rejette** donc notre hypothèse.

Et si on change le seuil à 0.01 ?

x = 2 < Q(0.99) = 2.3

x n'est plus assez grand,
On ne peut **pas rejetter** l'hypothèse.

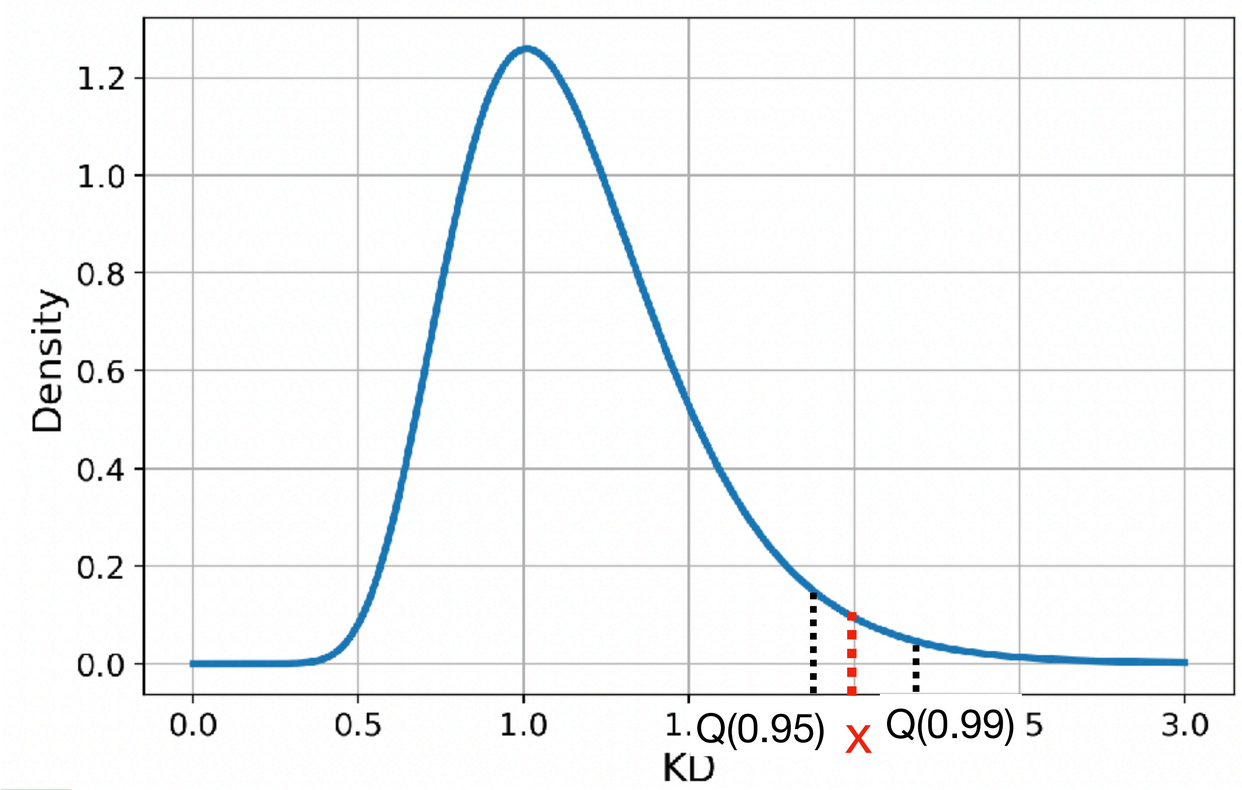
INSEA

4. Expert: supposons X = KD ~ une loi connue: **on a sa CDF + quantiles (CDF^{-1})**. On observe x = 2.0

"x est trop grand s'il a dépassé un c (x > c) tel que la probabilité de le dépasser est très petite, disons 0.05"

$$\mathbb{P}(X \geq c) = 0.05 \Leftrightarrow 1 - F_X(c) = 0.05 \Leftrightarrow F_X(c) = 0.95 \Leftrightarrow c = F_X^{-1}(0.95) = Q_X(0.95) = 1.9$$

"x est trop grand s'il a dépassé 1.9"



On fixe un seuil $\alpha \Rightarrow$ on compare $x > Q_X(1-\alpha)$?

Peut-on faire l'inverse ?

On calcule $\alpha^\star$ tel que $x = Q_X(1-\alpha^\star)$

$\Downarrow$

on compare $\alpha^\star$ à un seuil choisi

Comment l'interpréter ?

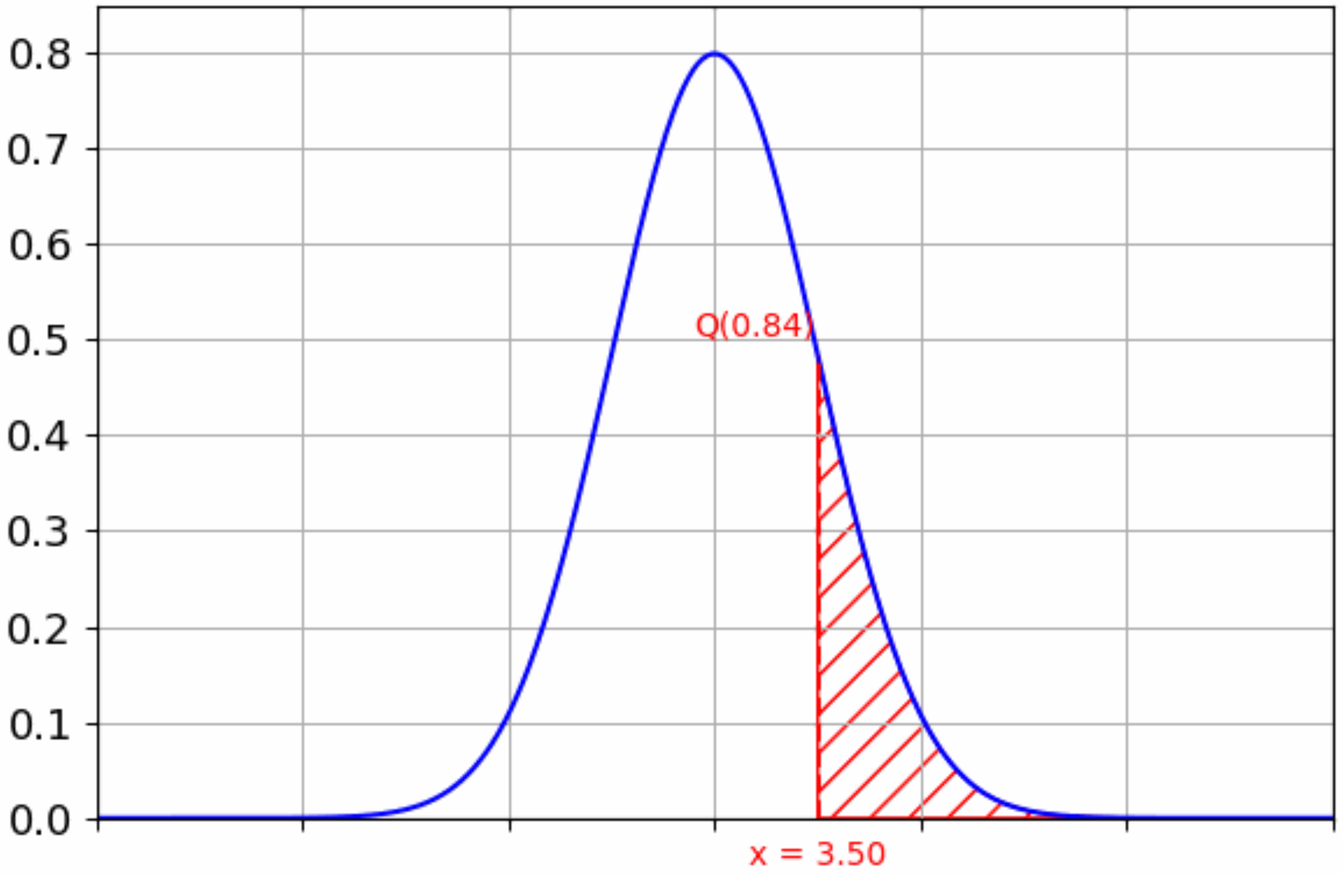
$$x = Q_X(1-\alpha^\star) \Leftrightarrow \mathbb{P}(X \geq x) = \alpha^\star$$

Probabilité d'observer qqch **plus rare** que l'observation x.

C'est la p-valeur

INSEA

Qu'est-ce-qu'une p-valeur ?



Pour un x observé, elle correspond graphiquement à l'aire sous la courbe dans la région extrême = c'est l'aire de la région de rejet pour le seuil x.

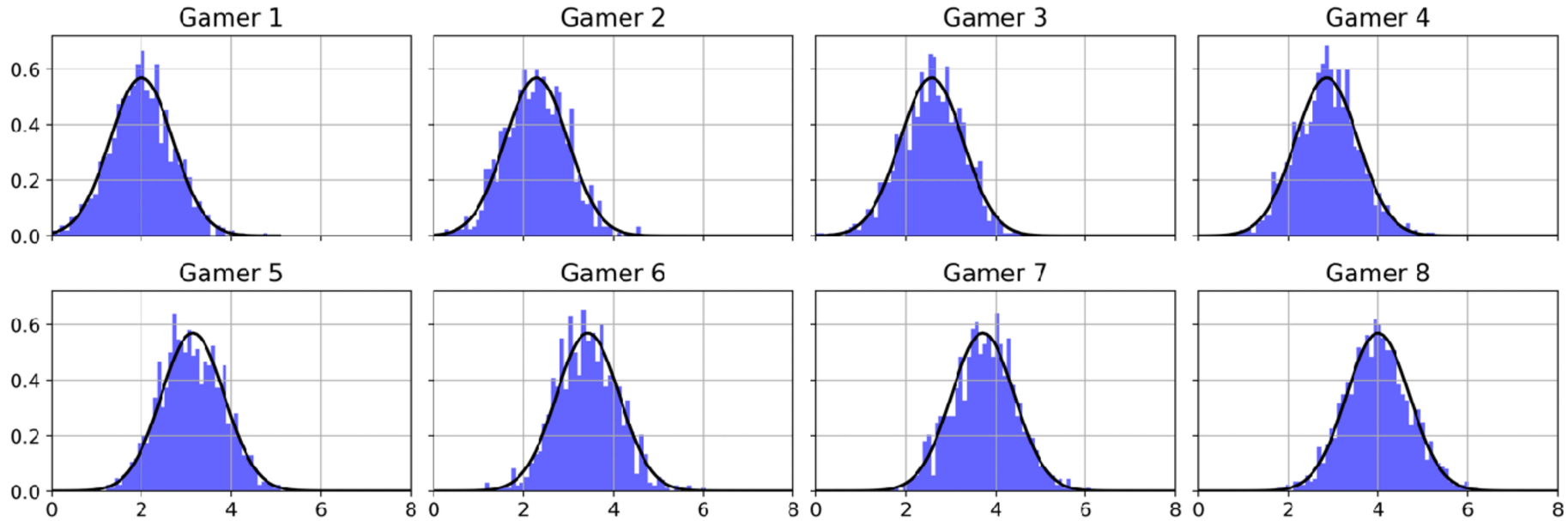
Ici, c'est la probabilité d'avoir un KD plus "rare/extrême" que le x observé.

Intuitivement c'est donc la probabilité de se tromper en rejetant l'hypothèse.

Ici, elle est donnée par c'est la 1- fonction de répartition(x), ou 1-*cumulative distribution function* (cdf).

5. Statisticien

Vous n'avez aucune idée quelle est la bonne distribution théorique du KD. Vous scrapez les données de tous les matchs de tous les gamers. Vous visualisez les distributions:



La loi Normale semble être un bon modèle pour ces données.

La moyenne diffère d'un gamer à l'autre, mais la variance est la même.

On suppose alors que $\mathrm{KD} \sim \mathcal{N}(\mu, \sigma^2)$ avec $\sigma^2 = 0.7$

Vous modélisez donc votre KD par X $\sim \mathcal{N}(\mu, \sigma^2)$ avec $\sigma^2 = 0.7$

La moyenne de tous les gamers amateurs est $\mu_0 = 1.2$

Pour avoir l'espoir d'être un pro, il faut écarter l'hypothèse $\mu = \mu_0$

en faveur de l'hypothèse $\mu > \mu_0$

Vous observez $n = 10$ valeurs de KD qu'on suppose i.i.d. $X_1, \dots, X_n$.

1. Quelle est la distribution de $\bar{X}$?
2. En **supposant** que l'hypothèse $H_0 : \mu = \mu_0$ est vraie, définir une variable Z en fonction $\bar{X}$ telle que $Z \sim \mathcal{N}(0, 1)$.
3. Définir la région de rejet de H_0 en faveur de $H_1 : \mu > \mu_0$ pour le seuil 0.01.
4. Empiriquement, on trouve $\bar{X} = 2.7$. Peut-on rejeter H_0 avec le seuil de 1% ?
5. Trouver la p-valeur de ce test.
6. On suppose désormais que l'on ne connait pas la valeur de σ^2. On remplace σ par son estimateur empirique $\hat{\sigma}$ estimé à partir des X_i observés. Cela pose-t-il problème ?

Python

INSEA

Avec le théorème central limite:

$$Z_n = \sqrt{n}\frac{(\bar{X} - \mu_0)}{\hat{\sigma}_u} \overset{n\to+\infty}{\sim} \mathcal{N}(0,1)$$

1. On considère que n est assez grand. Quel est le résultat du test asymptotique ?

2. Vous voulez vérifier qu'on peut appliquer le TCL avec n petit. Pour différentes valeurs de n, générez 1000 observations de Z_n pour visualiser son histograme. Comparez avec la densité de la distribution asymptotique.

3. Si n est petit et que le régime asymptotique n'est pas atteint, quel test peut-on effectuer ?

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

On peut par exemple, essayer de trouver la loi exacte de:

$$\sqrt{n}\frac{(\bar{X}_n - \mu)}{\hat{\sigma}_u} \quad \text{Où } \hat{\sigma}_u \overset{\text{def}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

On démontre que cette variable a pour densité:

$$f(z) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1 + \frac{z^2}{n-1}\right)^{-\frac{n}{2}}$$

On peut l'intégrer et trouver ses quantiles.

Bingo ! Vous avez inventé un test statistique.

Qui est “Student” ?

C’est exactement ce qu’a fait le chimiste William Sealy Gosset en 1908. Son employeur l’autorise à publier ses résultats sous le pseudonyme de “Student”.

Plus tard, Ronald Fisher démontra rigoureusement les résultats de W. Gosset en utilisant la lettre “t”: *“Student’s t-distribution”*

Propriété: loi de student

Soit $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2_p$ deux variables indépendantes. Alors: $\dfrac{X}{\sqrt{\frac{Y}{p}}} \sim t_p$

Corollaire 1: One sample Student's t-test

Soit $X_1, \ldots, X_n$ i.i.d suivant $\mathcal{N}(\mu, \sigma^2)$. Alors: $\sqrt{n}\dfrac{(\bar{X}-\mu)}{\hat{\sigma}_u} \sim t_{n-1}$

Indication: On rappelle que le lemme de Fisher donne: $(n-1)\dfrac{\hat{\sigma}^2_u}{\sigma^2} \sim \chi^2_{n-1}$

INSEA

Exercice

1. Dans votre bilan sanguin vous lisez: niveau de Potassium: 2.0mg /dL ce qui est trop faible par rapport à la valeur saine de référence 3.5 mg / dL. Avant de prendre des suppléments, vous voulez vous assurer que cette valeur est “trop faible”, vous effectuez quatre mesures supplémentaires. Les cinq mesures donnent: 2.0, 2.4, 2.8, 1.6. 3.2. Faites l’étude statistique adéquate.

2. Que devrait-on modifier pour démontrer que l’on a un surplus de Potassium ?

3. Que devrait-on modifier pour démontrer que l’on a un taux anormal de Potassium ?

4. Peut-on utiliser une telle étude pour confirmer statistiquement qu’on est en bonne santé ?

Problématique

Vous remarquez que lorsque vous jouez à Call of duty chez votre ami votre performance est médiocre. Vous pensez que son débit internet en est la raison. La moyenne de votre ping est 54ms. Celle de votre ami est 75ms.

En général, on peut supposer que le ping suit toujours une loi normale et que seule la moyenne diffère d'une installation à une autre. Comment peut-on vérifier si cette différence est statistiquement significative ?

Vous prenez alors des échantillons des pings de chaque côté:

$$X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma^2) \qquad Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$$

On suppose que les X_i sont indépendants des Y_i.

Exercice

On suppose l'hypothèse $H_0 : \mu_1 = \mu_2$.

1. Quel est la loi de $\bar{X} - \bar{Y}$? En déduire, en fonction de σ, une variable $\sim \mathcal{N}(0, 1)$.

2. Montrez que $\hat{\sigma}^2_{u,a} = \frac{(n_1-1)\hat{\sigma}^2_{u,X} + (n_2-1)\hat{\sigma}^2_{u,Y}}{n_1+n_2-2}$ estime σ^2 sans biais.

3. En utilisant le lemme de Fisher, en déduire que: $\frac{\bar{X}-\bar{Y}}{\hat{\sigma}_{u,a}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

INSEA

Corollaire 2: Two samples Student's t-test

Soit $X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ et $Y_i \sim \mathcal{N}(\mu_2, \sigma^2)$ des observations i.i.d. On suppose que les X_i sont indépendants des Y_i. On définit la variance empirique agrégée: $\hat{\sigma}^2_{u,a} = \frac{(n_1-1)\hat{\sigma}^2_{u,X} + (n_2-1)\hat{\sigma}^2_{u,Y}}{n_1+n_2-2}$. Alors:

$$\frac{\bar{X} - \bar{Y}}{\hat{\sigma}_{u,a}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Exercice

Voici les pings observés:

X	62	49	74	90	55	70
Y	89	78	71	53	80	101

Les deux connexions internet ont-elle des pings moyens différents ?

INSEA

Problématique

Vous remarquez que lorsque vous utilisez une souris pro, votre score KD (Call of Duty, toujours) est mieux. Vous voulez évaluer si cette différence est significative. Vous avez les scores de plusieurs matchs avant et après l'achat de la souris: $A_1, \dots, A_n$ (after) et $B_1, \dots, B_n$ (before).

On peut supposer que les matchs sont indépendants les uns des autres. Mais on ne peut pas supposer l'indépendance de A et B. Quelle procédure statistique peut-on appliquer ?

On considère les différences $Z_i = A_i - B_i$ qu'on modélise suivant $\mathcal{N}(\mu, \sigma^2)$.

1. Quelles Hypothèses H_0 et H_1 sont adéquates ici ?
2. Définir la statistique du test sous H_0.
3. Quelle hypothèse peut-on relâcher si on suppose que n est très grand ?

Test de Student: Résumé

Test statistique	Conditions	Hypothèse nulle	Statistique du test
1. Test à un échantillon: Comparaison avec une moyenne connue μ_0	$X_1, \dots, X_n$ i.i.d $\sim \mathcal{N}(\mu, \sigma^2)$	$\mu = \mu_0$	$\sqrt{n}\frac{(\bar{X}-\mu_0)}{\hat{\sigma}_u} \sim t_{n-1}$
2. Test à deux échantillons indépendants: Comparaison de deux moyennes empiriques	$X_1, \dots, X_n$ i.i.d $\sim \mathcal{N}(\mu_1, \sigma^2)$ $Y_1, \dots, Y_n$ i.i.d $\sim \mathcal{N}(\mu_2, \sigma^2)$ $X_i \perp Y_i$	$\mu_1 = \mu_2$	$\frac{\bar{X}_1-\bar{X}_2}{\hat{\sigma}_{u,agr}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$ Où $\hat{\sigma}^2_{u,agr} = \frac{(n_1-1)\hat{\sigma}^2_{u,X}+(n_2-1)\hat{\sigma}^2_{u,Y}}{n_1+n_2-2}$
3. Test de student à deux échantillons dépendants (paires)	$Z_i = X_i - Y_i \sim \mathcal{N}(\mu, \sigma^2)$ $Z_1, \dots, Z_n$ i.i.d	$\mu = 0$	$\sqrt{n}\frac{\bar{Z}}{\hat{\sigma}_{u,Z}} \sim t_{n-1}$

Remarque: Si n est assez grand, la condition de normalité n'est plus nécessaire. La loi du test est $\mathcal{N}(0,1)$.

INSEA

Quand est-ce que n est “assez grand” ?

Python

Simuler des échantillons $X_1, \ldots, X_n$ i.i.d de distribution quelconque de moyenne μ_0

Et comparer la distribution de $\sqrt{n}\frac{(\bar{X}-\mu_0)}{\hat{\sigma}_u}$ à la distribution asymptotique $\mathcal{N}(0,1)$

pour des valeurs différentes de n. À partir de quel n, la distribution semble-t-elle avoir convergé ?

INSEA

Quand est-ce que n est “assez grand” ?

On va simuler des échantillons $X_1, \ldots, X_n$ i.i.d de distribution quelconque de moyenne μ_0

Et comparer la distribution de $\sqrt{n}\frac{(\bar{X}-\mu_0)}{\hat{\sigma}_u}$ à la distribution asymptotique $\mathcal{N}(0,1)$

pour des valeurs différentes de n.

Comment quantifier la proximité i.e la convergence vers la loi normale standard?

Il nous faut une mesure de distance entre des distributions de probabilité.

On peut utiliser la distance de Wasserstein (*Earth mover distance - EMD*)

(En python la librairie POT — *python optimal transport*)

INSEA

Conclusion

1. Le but d'un test statistique est de quantifier si la différence observée D est signficative ou pas.
2. L'hypothèse H_0, souvent, suppose: "aucune différence".
3. L'alternative H_1 est donnée par le contexte: on veut rejeter H_0 si H_1 est vrai.
4. Ainsi, les valeurs D pour lesquelles on rejette H_0 (zone de rejet) définissent H_1.
5. Sous l'hypothèse H_0 : "aucune différence", D suit une loi $\mathcal{L}$ donnée (Normale, Student, ..).
6. Si n est très grand et D converge en loi vers $\mathcal{L}$, on parle alors de test asymptotique.
7. Si la valeur observée $D = d$ est en faveur de H_1 et a une très faible probabilité selon la loi $\mathcal{L}$, on rejette H_0.
8. Sinon, H_0 reste plausible.
9. La p-valeur correspond à la probabilité d'observer une valeur plus extrême que d avec H_0 vrai.
10. C'est la probabilité de se tromper en rejetant H_0: On rejette donc H_0 pour des p-valeurs très petites.

II - Inférence et tests statistiques

Partie 2 - Cas multivarié

Maximum de vraisemblance

$$\text{Soient } \mathbf{X}_1, \ldots, \mathbf{X}_n \text{ des variables i.i.d suivant } \mathcal{N}(\mu, \Sigma)$$

Par indépendance, la vraisemblance du modèle est donnée par:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right)$$

On peut démontrer que son maximum est atteint en:

$$\hat{\boldsymbol{\mu}} = \bar{X_n} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

Preuve en TD

Soit n observations i.i.d $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$$

$$\hat{\mathbf{\Sigma}}_u = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)^\top$$

Propriété (Lemme de Fisher)

Biais nul: $\mathbb{E}(\bar{\mathbf{X}}_n) = \mu$ $\quad$ $\mathbb{E}(\hat{\Sigma}_u) = \Sigma$

Distribution: $\bar{\mathbf{X}}_n \sim \mathcal{N}\left(\mu, \frac{1}{n}\Sigma\right)$ $\quad$ $(n-1)\,\hat{\Sigma}_u \sim W_d(n-1, \Sigma)$

Loi de Wishart

Indépendance: $\bar{\mathbf{X}}_n$ et $\hat{\Sigma}_u$ sont indépendants

INSEA

Loi de Wishart

Définition

La loi de Wishart est une distribution dans l'espace des matrices définies poitives. Soit $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(0, \Sigma)$ des vecteurs Gaussiens i.i.d. Alors:

$$\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \sim W_d(n, \Sigma)$$

À quoi correspond cette loi en dimension 1 ?

On retrouve la loi du Chi-2 en prenant également des projections en 1D:

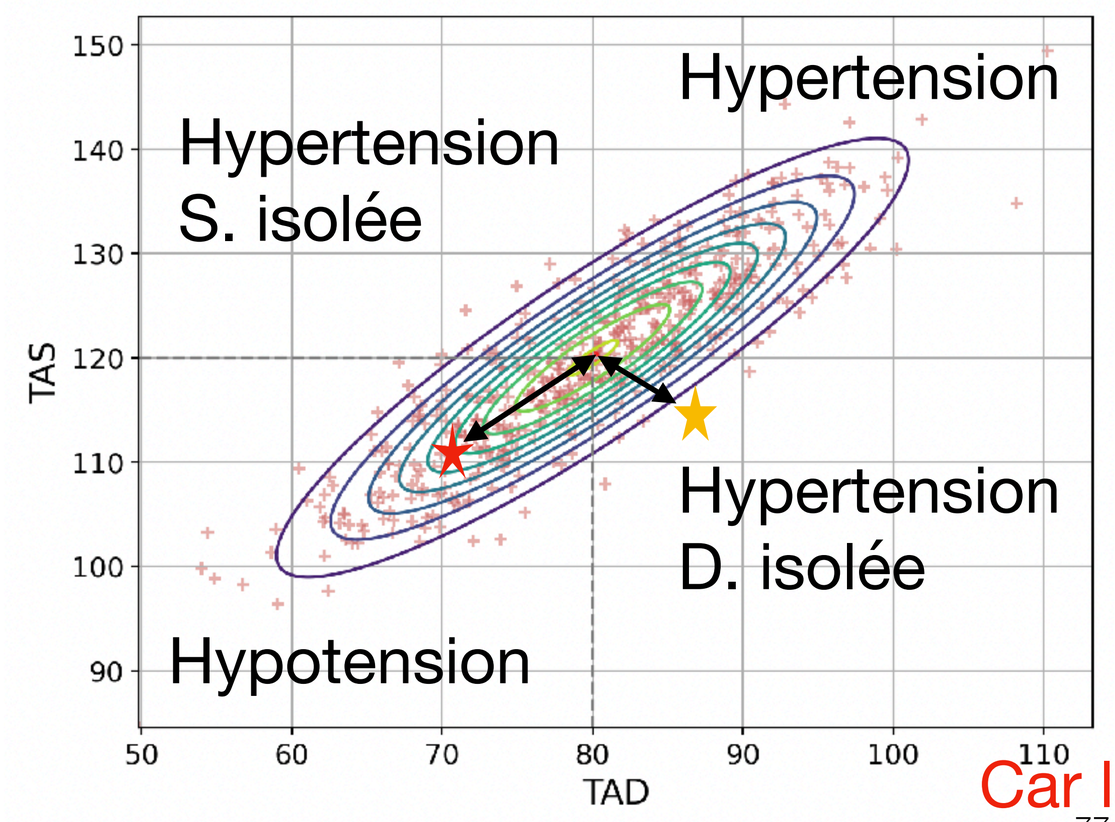
Propriété

Soit $a \in \mathbb{R}^d$ tel que $a^\top \Sigma a \neq 0$ et $V \sim W_d(n, \Sigma)$ alors: $\dfrac{a^\top V a}{a^\top \Sigma a} \sim \chi^2(n)$

Problématique 2

La tension artérielle se mesure deux fois:
- Tension artérielle systolique (TAS): Pendant les battements cardiaques (environ 120mmHg)
- Tension artérielle diastolique (TAD): Entre deux battements cardiaques (environ 80mmHg)

Dans une population saine, on suppose qu'on observe la distribution suivante:



On a les mesures de deux personnes:

Personne A: 70 et 110

Personne B: 87 et 113

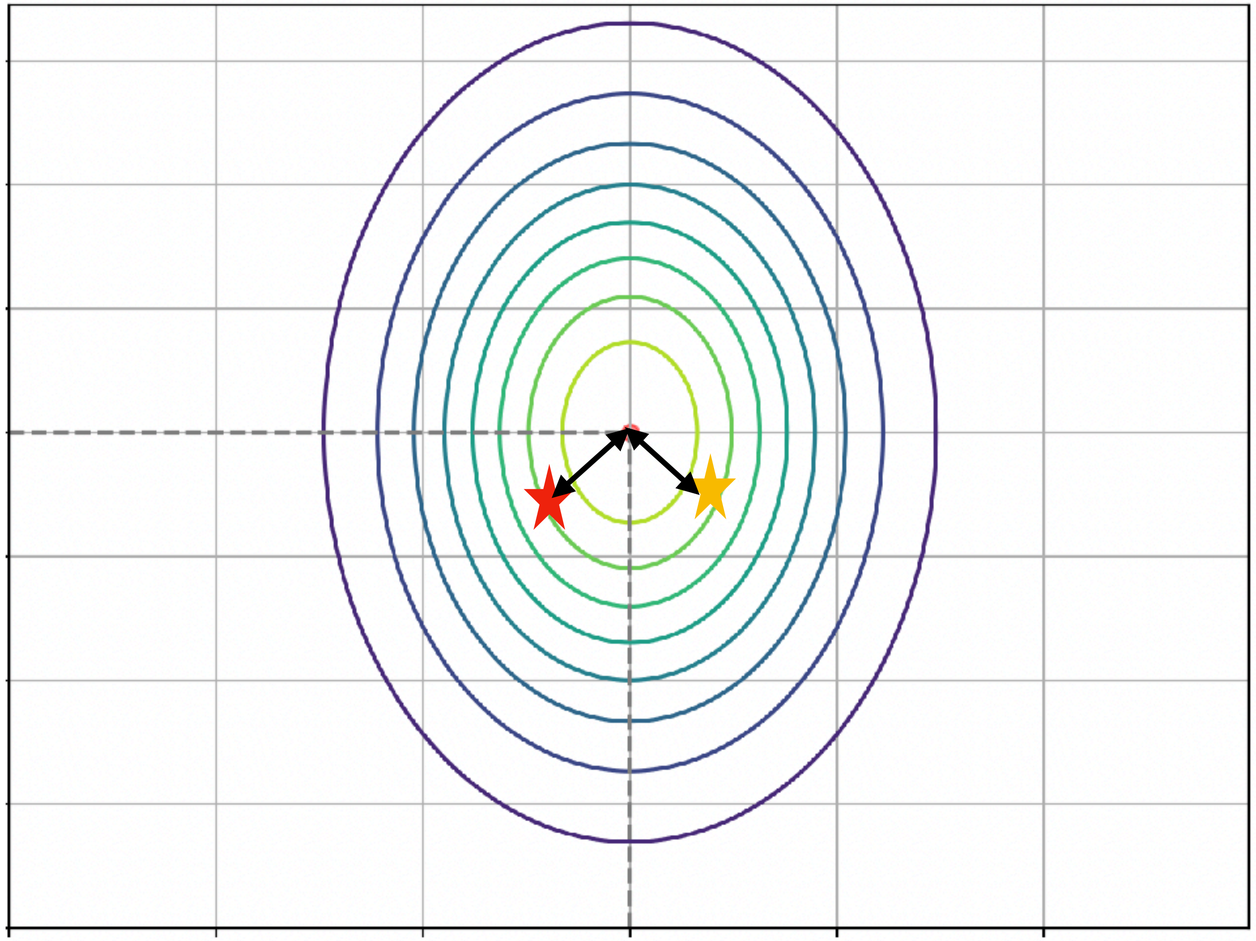
Quel personne a la plus grande probabilité d'être en bonne santé ?

Pourtant, les mesures de **A** sont plus lointaines de la moyenne que celles de **B** !

Car la distance Euclidienne ignore la corrélation

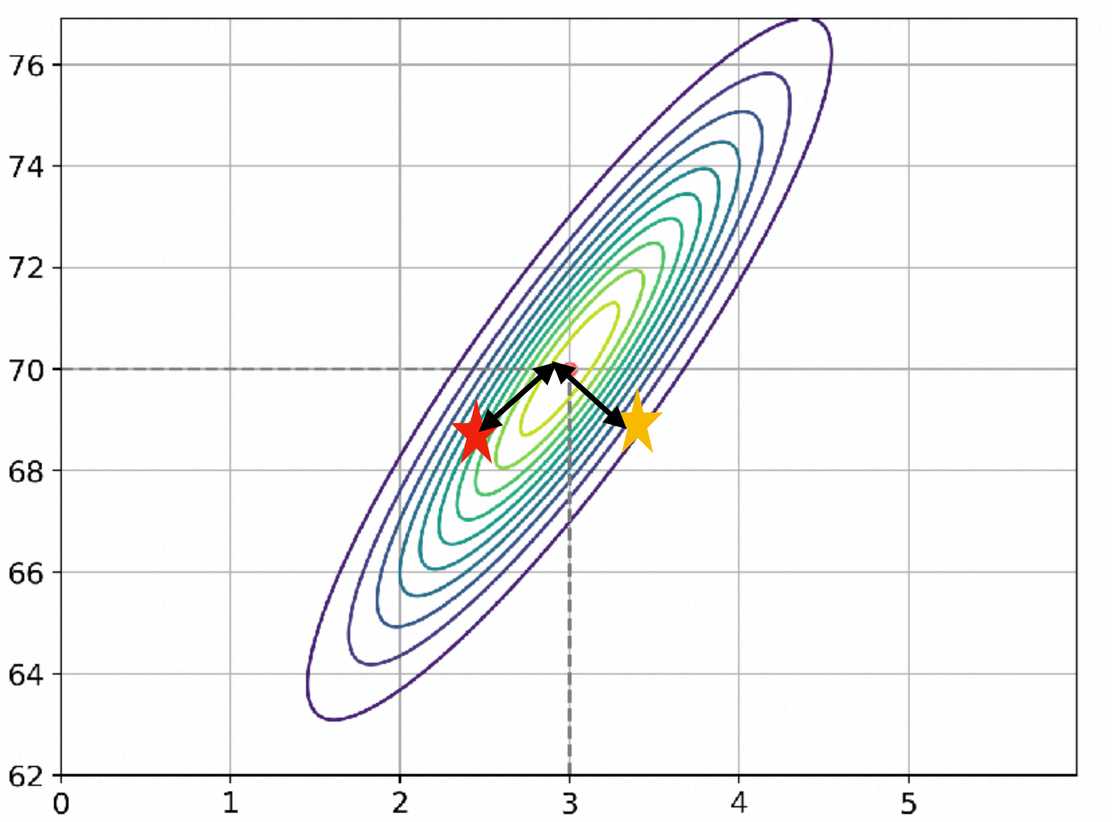
Distance de Mahalanobis

Si on n’avait pas de corrélation, les deux scénarios seraient équivalents:



Distance de Mahalanobis

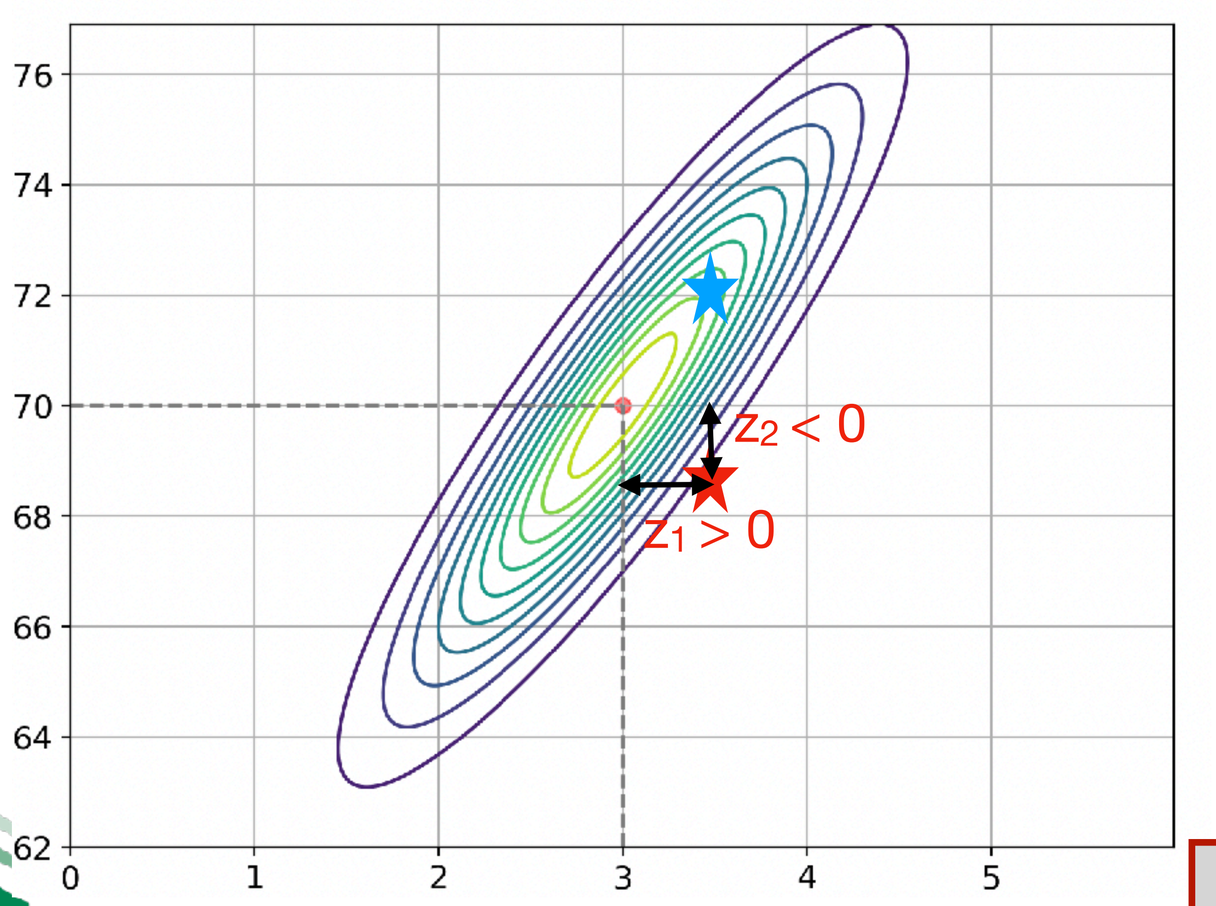
Avec la corrélation, le scénario **B** a une densité plus faible: il est plus rare



Distance de Mahalanobis

Pour cela, on utilise la distance de Mahalanobis qui prend en compte la covariance:

$$d_\Sigma(\mathbf{x}, \mu) = \sqrt{(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)}$$



Exemple avec c > 0:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix} \quad \mathbf{\Sigma}^{-1} = \frac{1}{1-c^2} \begin{pmatrix} 1 & -c \\ -c & 1 \end{pmatrix}$$

$$z = \mathbf{x} - \mu = (z_1, z_2)^\top$$

$$d_\Sigma(\mathbf{x}, \mu)^2 = z^T \mathbf{\Sigma}^{-1} z = \frac{1}{1-c^2}(z_1^2 + z_2^2 - 2c z_1 z_2)$$

$-2cz_1z_2 > 0 \Leftrightarrow z_1$ et z_2 ont des signes opposés.

$-2cz_1z_2 < 0 \Leftrightarrow z_1$ et z_2 ont même signe.

Remarque: cette distance est inversement liée à la densité

Soit n observations i.i.d $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(\mu, \Sigma)$. On suppose Σ connue, on a donc $\bar{\mathbf{X}} \sim \mathcal{N}(\mu, \frac{\Sigma}{n})$. Pour évaluer l'hypothèse $H_0 : \mu = \mu_0$, on calcule donc la distance:

$$d_{\frac{\Sigma}{n}}(\bar{\mathbf{X}}, \mu_0)^2 = (\bar{\mathbf{X}} - \mu_0)^\top \left(\frac{\Sigma}{n}\right)^{-1} (\bar{\mathbf{X}} - \mu_0)$$

Qu'est ce qu'il nous faut pour en faire un test statistique ?

Il faut connaître sa distribution

$$d_{\frac{\Sigma}{n}}(\bar{\mathbf{X}}, \mu_0)^2 = \underbrace{\left(\left(\frac{\Sigma}{n}\right)^{-1/2} (\bar{\mathbf{X}} - \mu_0)\right)}_{\stackrel{\text{def}}{=} Z \sim \mathcal{N}(0, I)}^\top \left(\left(\frac{\Sigma}{n}\right)^{-1/2} (\bar{\mathbf{X}} - \mu_0)\right) = \sum_{j=1}^{d} Z_j^2 \sim \chi^2_d$$

Et si on ne connait pas Σ ?

INSEA

Théorème central limite

Soit n observations i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}(\mu, \Sigma)$. On suppose Σ connue, on a donc $\bar{\mathbf{X}} \sim \mathcal{N}(\mu, \frac{\Sigma}{n})$. Pour évaluer l'hypothèse $H_0 : \mu = \mu_0$, on calcule donc la distance:

$$d_{\frac{\Sigma}{n}}(\bar{\mathbf{X}}, \mu_0)^2 = (\bar{\mathbf{X}} - \mu_0)^\top \left(\frac{\Sigma}{n}\right)^{-1} (\bar{\mathbf{X}} - \mu_0)$$

Et si on ne connait pas Σ ? Peut-on la remplacer par $\hat{\Sigma}_u$?

Si n est assez grand, alors par le thèorème CTL:

$$d_{\frac{\hat{\Sigma}_u}{n}}(\bar{\mathbf{X}}, \mu_0) \overset{n \to \infty}{\sim} \chi^2_d$$

Sinon, il nous faut la loi exacte de $d_{\frac{\hat{\Sigma}_u}{n}}(\bar{\mathbf{X}}, \mu_0)$

INSEA

Hotelling généralise la loi t de Student au cadre multivarié:

Définition

Soit $\mathbf{Z} \sim \mathcal{N}(0, \Sigma)$ indépendant de $M \sim \mathcal{W}_d(m, \Sigma)$ alors:

$$m\mathbf{Z}^\top M^{-1}\mathbf{Z} \sim T^2(d, m)$$

Loi T^2 de Hotelling

Soit $\bar{\mathbf{x}}$ la moyenne empirique d'un vecteur $\mathcal{N}(\mu, \Sigma)$ et S_u la covariance empirique.

1. Montrez que $n(\bar{\mathbf{x}} - \mu)^\top S_u^{-1}(\bar{\mathbf{x}} - \mu) \sim T^2(d, n-1)$
2. Que devient ce résultat en dimension 1 ?

Test de Hotelling de la moyenne à un échantillon

INSEA

La loi de Hotelling est directement liée à la F-distribution de Fisher:

Théorème

$$T^2(d,n) = \frac{dn}{n-d+1} F_{d,n-d+1}$$

Où la statistique F est la loi du ratio de deux Chi-2 indépendantes et normalisées:

$$F_{p,q} = \frac{\chi^2(p)/p}{\chi^2(q)/q}$$

Corollaire 1: One sample Hotelling's T^2 test

Soit $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d suivant $\mathcal{N}(\mu, \Sigma)$. Alors:

$$n(\bar{\mathbf{X}} - \mu)^\top S_u^{-1}(\bar{\mathbf{X}} - \mu) \sim T^2(d, n-1) \ = \frac{d(n-1)}{n-d} F_{d,n-d}$$

Exercice

On applique le test de Hotelling sur des échantillons simulés en prenant une distribution $\mathcal{N}(\mu, \Sigma)$ avec:

$$\mu = (0.2, -0.2)^\top \qquad \Sigma = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}$$

1. Avec $c = 0$, simulez $n = 10$ observations suivant cette loi.
2. Effectuez le test de Hotelling $H_0 : \mu = (0, 0)^\top$ contre $H_1 : \mu \neq 0$.
3. Effectuez le test de Student équivalent sur chaque variable. La conclusion est-elle la même ?
4. Que se passe-t-il lorsqu'on prend $c = 0.95$?
5. Pour différents $c \in [-1, 1]$ Lancer l'expérience 100 fois. Tracez les moyennes des pvaleurs en fonction de c.

Corollaire 2: Two samples Hotelling's T^2 test

Soit $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1} \sim \mathcal{N}(\mu, \Sigma)$ i.i.d et $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2} \sim \mathcal{N}(\mu, \Sigma)$ i.i.d et $X_i \perp Y_i$.
Alors:

$$\frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})^\top \hat{\Sigma}_{u,agr}^{-1}(\bar{\mathbf{X}}-\bar{\mathbf{Y}}) \sim T^2(d, n_1+n_2-2) = \frac{d(n_1+n_2-2)}{n_1+n_2-d-1} F_{d,n_1+n_2-1-d}$$

$$\text{Où } \hat{\Sigma}_{u,agr} = \frac{(n_1-1)\hat{\Sigma}_{u,X}+(n_2-1)\hat{\Sigma}_{u,Y}}{n_1+n_2-2}$$

Problématique

Vous voulez tester si l'effet d'un médicament sur 3 patients est significatif ou non. Vous prenez des mesures de la tension artérielle et du taux de cholestérol avant et après traitement. Quelle procédure statistique est adéquate ?

1. Définir l'hypothèse nulle et l'hypothèse alternative.
2. Définir la zone de rejet à 5%.
3. Effectuez le test pour les observations suivantes:

Patient	TA Avant (mmHg)	TA Après (mmHg)	Cholestérol Avant (mg/dL)	Cholestérol Après (mg/dL)
1	130	128	200	198
2	135	134	210	208
3	125	124	190	189

Paired samples Hotelling's T^2 test

Soit $\mathbf{A}_1, \ldots, \mathbf{A}_n \sim \mathcal{N}(\mu_1, \Sigma)$ i.i.d et $\mathbf{B}_1, \ldots, \mathbf{B}_n \sim \mathcal{N}(\mu_2, \Sigma)$ i.i.d. On suppose que les observations sont pairées càd, par exemple, on teste l'effet d'un nouveau traitement sur les **mêmes patients**. On considère alors les différences $X_i = A_i - B_i$ et on teste l'hypothèse $\mu = \mu_1 - \mu_2 = 0$ en appliquant le test de Hotelling à un échantillon.

Hotelling vs Student

Test statistique	Conditions	Hypothèse nulle	Statistique du test
1. Test à un échantillon: Comparaison avec une moyenne connue μ_0	$X_1, \dots, X_n$ i.i.d $\mathcal{N}(\mu, \Sigma)$	$\mu = \mu_0$	$n(\bar{\mathbf{X}} - \mu)^\top S_u^{-1}(\bar{\mathbf{X}} - \mu) \sim T^2(d, n-1)$
2. Test à deux échantillons indépendants: Comparaison de deux moyennes empiriques	$X_1, \dots, X_n$ i.i.d $\mathcal{N}(\mu_1, \Sigma)$ $Y_1, \dots, Y_n$ i.i.d $\mathcal{N}(\mu_2, \Sigma)$ $X_i \perp Y_i$	$\mu_1 = \mu_2$	$\frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})^\top \hat{\Sigma}_{u,agr}^{-1}(\bar{\mathbf{X}}-\bar{\mathbf{Y}}) \sim T^2(d, n_1+n_2-2)$ Où $\hat{\Sigma}_{u,agr} = \frac{(n_1-1)\hat{\Sigma}_{u,X}+(n_2-1)\hat{\Sigma}_{u,Y}}{n_1+n_2-2}$
3. Test de student à deux échantillons dépendants (paires)	$Z_i = X_i - Y_i \sim \mathcal{N}(\mu, \Sigma)$ $Z_1, \dots, Z_n$ i.i.d	$\mu = 0$	$n\bar{\mathbf{Z}}^\top \hat{\Sigma}_{u,Z}^{-1}\bar{\mathbf{Z}} \sim T^2(d, n-1)$

Remarque: Si n est assez grand, la condition de normalité n'est plus nécessaire. La loi du test est $\chi^2(d)$.

INSEA

Intuition

Soit n observations i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n$ suivant une loi paramétrée par $\omega \in \mathbb{R}^d$.

On veut quantifier à quel point l'hypothèse $H_0 : \omega \in \Omega_0$ vs $H_1 : \omega \in \Omega_1$ est plausible.

On définit l'ensemble des paramètres non contraints (*full model*): $\Omega_f = \Omega_0 \cup \Omega_1$

On évalue le rapport des maximums de vraisemblance sur les deux ensembles:

$$\text{LRT} \quad \stackrel{\text{def}}{=} \quad \frac{\max_{\omega \in \Omega_0} \mathcal{L}(\omega | \mathbf{X}_i)}{\max_{\omega \in \Omega_f} \mathcal{L}(\omega | \mathbf{X}_i)} = \frac{\mathcal{L}_0}{\mathcal{L}_f}$$

1. LRT est-il borné ?
2. Discuter les valeurs possibles de LRT selon la plausibilité de l'hypothèse nulle.

Si H_0 est très probable, alors les deux maximums sont très proches, donc LRT $\to 1$.

Sinon, $\mathcal{L}_0$ serait très petit par rapport à $\mathcal{L}_f$ donc LRT $\to 0$

Intuition

Soit n observations i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n$ suivant une loi paramétrée par $\omega \in \mathbb{R}^d$.

On veut quantifier à quel point l'hypothèse $H_0 : \omega \in \Omega_0$ vs $H_1 : \omega \in \Omega_1$ est plausible.

On définit l'ensemble des paramètres non contraints (*full model*): $\Omega_f = \Omega_0 \cup \Omega_1$

$$\text{LRT} \quad \stackrel{\text{def}}{=} \quad \frac{\max_{\omega \in \Omega_0} \mathcal{L}(\omega | \mathbf{X}_i)}{\max_{\omega \in \Omega_f} \mathcal{L}(\omega | \mathbf{X}_i)} = \frac{\mathcal{L}_0}{\mathcal{L}_f}$$

On rejette donc H_0 si LRT est très *petit*.. mais à quel point "petit" est assez pour rejeter H_0?

On a besoin de connaître la loi de LRT

Parfois, en calculant le LRT on reconnaît des lois usuelles (Hotelling, Chi-2..)

Sinon, si n est assez grand, on peut appliquer le thèorème de Wilks:

Théorème de Wilks

Soit n observations i.i.d $\mathbf{X}_1, \ldots, \mathbf{X}_n$ suivant une loi paramétrée par $\omega \in \mathbb{R}^D$.

On considère le test $H_0 : \omega \in \Omega_0$ vs $H_1 : \omega \in \Omega_1$ avec $\Omega_f \overset{\text{def}}{=} \Omega_0 \cup \Omega_1$ tel que:

$d_0 \overset{\text{def}}{=} \dim(\Omega_0) < \dim(\Omega_f) \overset{\text{def}}{=} d_f \leq D$ et on pose $r = d_f - d_0$.

Alors:

$$-2\log(\text{LRT}) \overset{n \to +\infty}{\sim} \chi^2(r)$$

$$\text{Où LRT} \quad \overset{\text{def}}{=} \quad \frac{\max_{\omega \in \Omega_0} \mathcal{L}(\omega|\mathbf{X}_i)}{\max_{\omega \in \Omega_f} \mathcal{L}(\omega|\mathbf{X}_i)} = \frac{\mathcal{L}_0}{\mathcal{L}_f}$$

1. Test de moyenne avec variance connue

Exercice

Soit n observations i.i.d $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(\mu, \Sigma)$

On s'intèresse au test $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

1. On suppose Σ connue. Quels sont les ensembles Ω_0 et Ω_f et leurs dimensions ?
2. Calculez $-2\log(\text{LRT})$. Reconnaissez-vous sa distribution ?
3. En déduire les zones de rejet de niveau $(1-\alpha) \times 100\% = 95\%$.

2. Test de moyenne exact avec variance inconnue

Propriété

Soit n observations i.i.d $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(\mu, \Sigma)$ avec $\mu_0 \in \mathbb{R}^d$

On note $\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \mu_0)(\mathbf{X}_i - \mu_0)^\top$ et $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$

$$\hat{T}^2 = n(\bar{\mathbf{X}} - \mu_0)\hat{\Sigma}_u^{-1}(\bar{\mathbf{X}} - \mu_0)^\top$$

On s'intèresse au test $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

Alors le test de rapport de vraisemblance est donné par:

$$\text{LRT} \quad \stackrel{\text{def}}{=} \quad \frac{\max_\Sigma \mathcal{L}(\mu_0, \Sigma | \mathbf{X}_i)}{\max_{\mu,\Sigma} \mathcal{L}(\mu, \Sigma | \mathbf{X}_i)} = \left(\frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma}_0)} \right)^{\frac{n}{2}} = \left(1 + \frac{\hat{T}^2}{n-1} \right)^{-\frac{n}{2}}$$

$$\text{Ainsi:} \quad \hat{T}^2 = (n-1)\left(\frac{\det(\hat{\Sigma}_0)}{\det(\hat{\Sigma})} - 1 \right)$$

3. Test de covariance asymptotique

Exercice

Soit n observations i.i.d $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(\mu, \Sigma)$ et $\Sigma_0 \in \mathbb{S}_+^d$

On s'intèresse au test $H_0 : \Sigma = \Sigma_0$ contre $H_1 : \Sigma \neq \Sigma_0$.

1. Quels sont les ensembles Ω_0 et Ω_f et leurs dimensions ?
2. Calculer $-2\log(\text{LRT})$ en fonction de Σ_0 et $\hat{\Sigma}_u$.
3. Donner la formule du test asymptotique.

Conclusion

1. La loi de Wishart généralise la loi χ^2 en multivarié. Elle est définie pour les matrices définies positives.
2. Les tests de Hotelling généralisent les tests de student en testant plusieurs variables jointement.
3. La statistique T^2 de Hotelling en dimension 1 est équivalente au t de Student au carré.
4. Avec Hotelling, il n'est pas évident de tester des alternatives unilatérales de type $\mu > 0$.
5. Les tests multivariés tiennent compte de la corrélation pour déterminer si la statistique est extrême ou pas.
6. Lorsque $n \to \infty$, le théorème central limite donne que $T^2(d, n)$ converge vers une $\chi^2(d)$
7. Le test de rapport de vraisemblance donne un test asymptotique dans un cadre général
8. Dans le cas de la comparaison de la moyenne d'une normale, il est exact et équivaut au test de Hotelling.
9. Le test LRT permet de tester si un modèle simple avec moins de paramètres (H_0) est assez bon ou non.

III - Modèles probabilistes

Partie 1 - Introduction à l'apprentissage

Learning paradigms

Les données disponibles à l'entraînement

Unsupervised learning

Large unlabeled data

$(X_1, \ldots, X_n)$

$\sim p(X)$

Semi-supervised learning

Small labeled data

$(X_1, y_1) \ldots (X_l, y_l)$

$\sim p(X, Y)$

Large unlabeled data

$(X_{l+1}, \ldots, X_{l+u})$

$\sim p(X)$

Supervised learning

Large labeled data

$(X_1, y_1) \ldots (X_l, y_l)$

$\sim p(X, Y)$

Self-supervised learning

Prédire des parties masquées de X_i

Labeliser des données coûte cher

But: apprendre des patterns / groupes de la distribution marginale

But: apprendre une fonction $f : \mathcal{X} \mapsto \mathcal{Y}$ qui généralise à des données non vues $\sim p(X, Y)$

Supervised learning

Exemple: *Churn* prediction

Un opérateur téléphonique a les données historiques sur ses clients.

Dependents	TechSupport	Contract	InternetService	Months	MonthlyCharges	Churn
0	1	0	1	12	75.65	0
1	0	0	0	24	89.50	0
0	0	0	1	6	65.25	1
0	1	1	0	48	35.30	?
1	0	0	1	48	85.81	?

Churn = 1: client a annulé son abonnement

L'entreprise souhaite anticiper le "churn" avec un algorithme de prédiction pour cibler les clients concernés

$$\mathbf{X} = (\mathbf{X}^1, \ldots, \mathbf{X}^6) \rightarrow y \in \{0, 1\}$$

On cherche une fonction f telle que: $f(\mathbf{X}) \approx y$

$$\min_f \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$$ Erreur de prédiction

f doit donner 1 ou 0, on considère alors des fonctions de type: $f(\mathbf{x}) = \mathbb{1}_{g(\mathbf{x}) \geq 0}$

On ne peut pas chercher g dans la totalité de l'espace des fonctions (dimension infinie), il faut paramétriser g

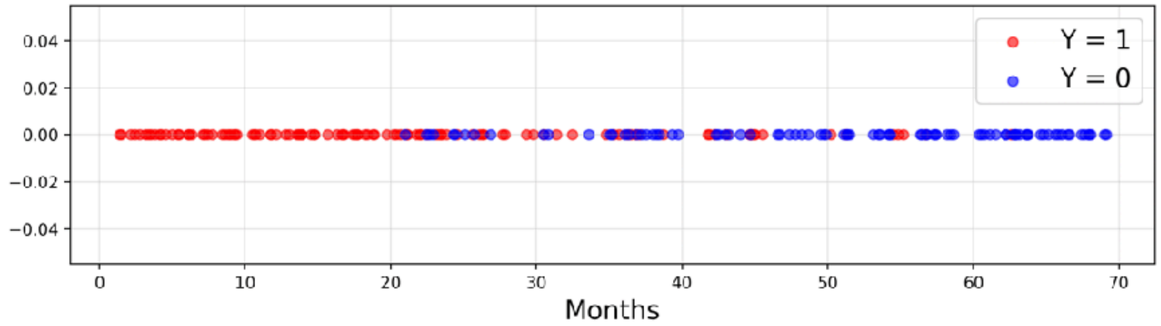
séparateur linéaire en dimension 1

f doit donner 1 ou 0, on considère alors des fonctions de type: $f(\mathbf{x}) = \mathbb{1}_{g(\mathbf{x})\geq 0}$

On ne peut pas chercher g dans la totalité de l'espace des fonctions (dimension infinie), il faut paramétriser g

On considère une seule variable "Months" qui donne la durée du contrat:

$\mathbf{x} = \text{Months} \in \mathbb{R}$



Quelle serait la fonction paramétrée g la plus simple ici ?

$$g(\mathbf{x}) = \beta_1 \mathbf{x} + \beta_0, \quad \beta_0, \beta_1 \in \mathbb{R}$$

Chercher la meilleure f = chercher le meilleur β: $\min_{\beta\in\mathbb{R}^2} \sum_{i=1}^{n} (\mathbb{1}_{\{\beta_1 \mathbf{x}_i + \beta_0 \geq 0\}} - y_i)^2$

Pouvez-vous donner des estimations vagues de ces paramètres ?

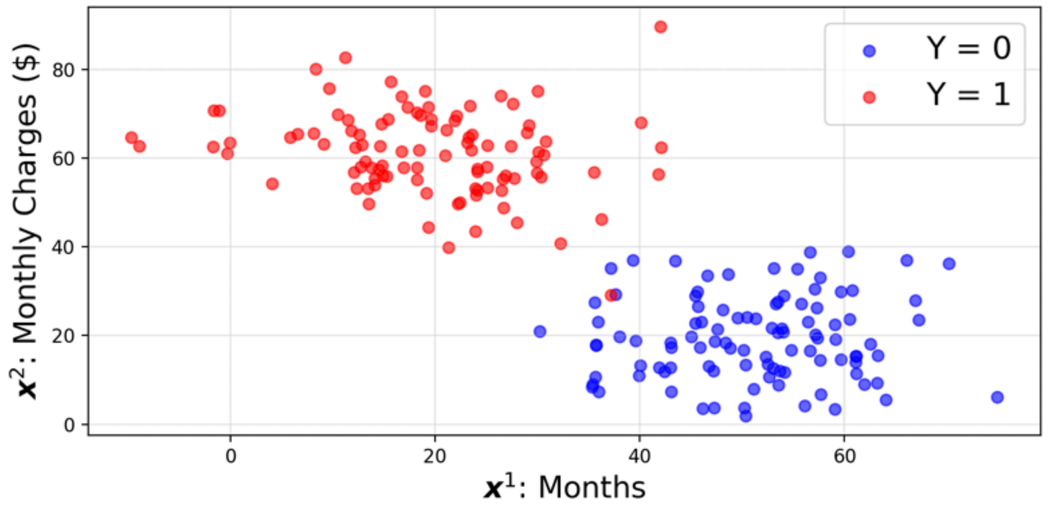
INSEA

Machine learning classique: zero-to-hero

séparateur linéaire en dimension 2

On considère une deux variables: “Months” et “MonthlyCharges”:

$\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2) \qquad f(\mathbf{x}) = \mathbb{1}_{g(\mathbf{x})\geq 0}$



Quelle serait la fonction paramétrée g la plus simple ici ?

$$g(\mathbf{x}) = \alpha + \beta_1 \mathbf{x}^1 + \beta_2 \mathbf{x}^2, \quad \alpha, \beta_1, \beta_2 \in \mathbb{R}$$

$$g(\mathbf{x}) = \alpha + \langle \beta, \mathbf{x} \rangle, \quad \alpha \in \mathbb{R}, \beta \in \mathbb{R}^2$$

$$g(\mathbf{x}) = \alpha + \beta^\top \mathbf{x}, \quad \alpha \in \mathbb{R}, \beta \in \mathbb{R}^2$$

$$\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^2} \sum_{i=1}^{n} (\mathbb{1}_{\{\alpha + \beta^\top \mathbf{x}_i \geq 0\}} - y_i)^2$$

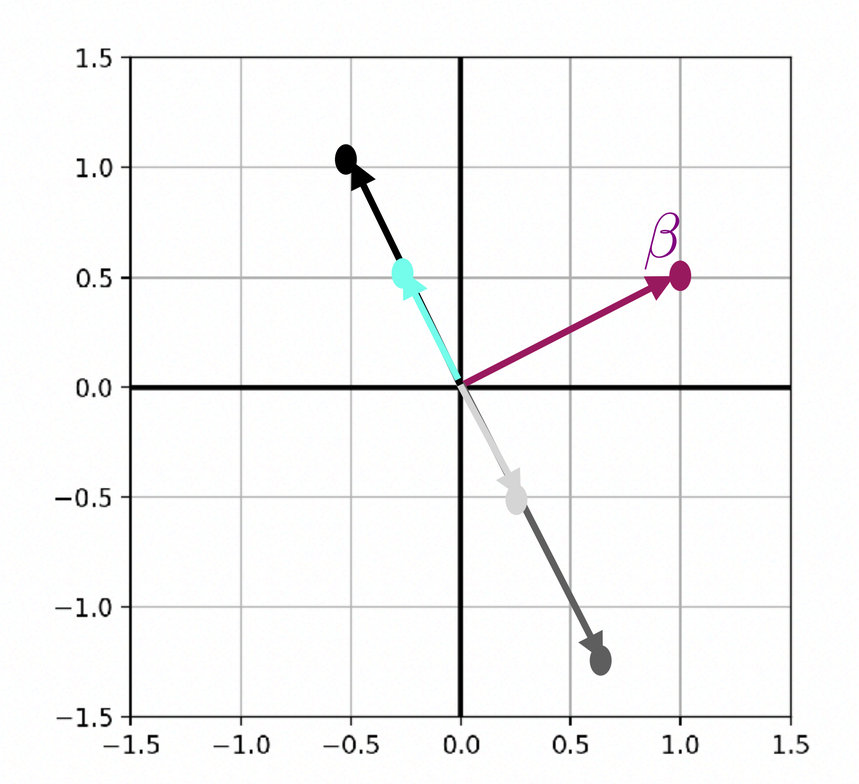
À quoi ressemble l’ensemble des fonctions g ?

On considère $g : \mathbf{x} \mapsto \beta^\top \mathbf{x}$. Étudions ses courbes de niveaux, c-à-d pour $c \in \mathbb{R}$ les ensembles: $\{\mathbf{x} | g(\mathbf{x}) = c\}$.

INSEA

séparateur linéaire en dimension 2

On considère $g : \mathbf{x} \mapsto \beta^\top \mathbf{x}$. Étudions ses courbes de niveaux, c-à-d pour $c \in \mathbb{R}$ les ensembles: $\{\mathbf{x} | g(\mathbf{x}) = c\}$.



Exemple avec $\beta = (1, 0.5)^\top$ et $c = 0$.

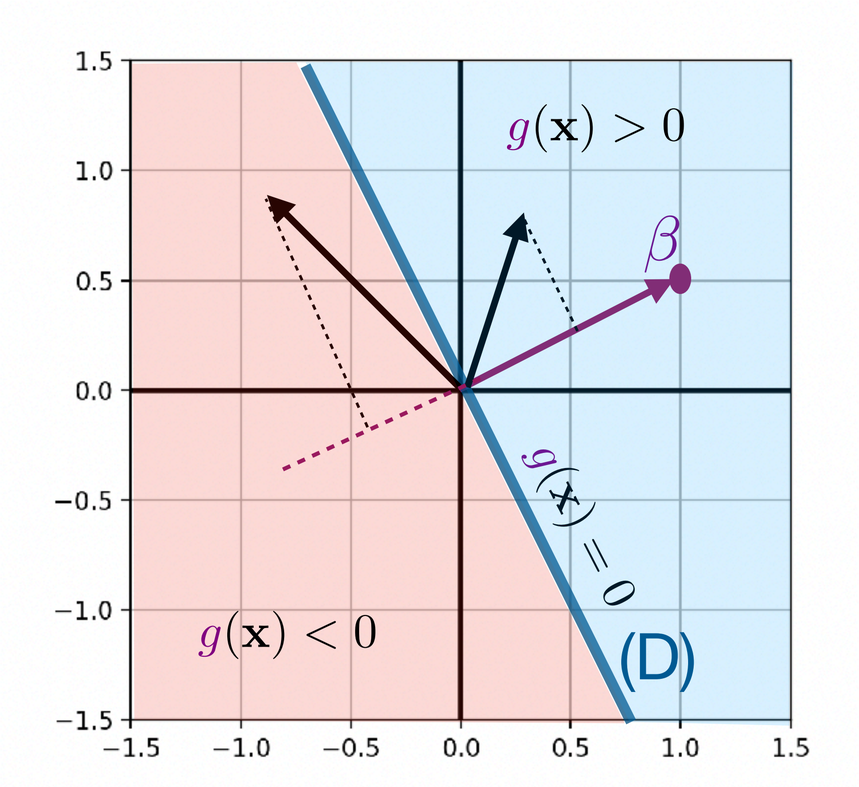
Quels sont les $\mathbf{x}$ tels que $\beta^\top \mathbf{x} = 0$?

Tous les vecteurs orthogonaux à β.

$\{\mathbf{x} \in \mathbb{R}^2 | \beta^\top \mathbf{x} = 0\}$ est la droite perpendiculaire à β.

séparateur linéaire en dimension 2

On considère $g : \mathbf{x} \mapsto \beta^\top \mathbf{x}$. Étudions ses courbes de niveaux, c-à-d pour $c \in \mathbb{R}$ les ensembles: $\{\mathbf{x} | g(\mathbf{x}) = c\}$.



Exemple avec $\beta = (1, 0.5)^\top$ et $c = 0$.

Quels sont les $\mathbf{x}$ tels que $\beta^\top \mathbf{x} = 0$?

Tous les vecteurs orthogonaux à β.

$\{\mathbf{x} \in \mathbb{R}^2 | \beta^\top \mathbf{x} = 0\}$ est la droite perpendiculaire à β.

à droite de (D), $\beta^\top \mathbf{x} > 0$

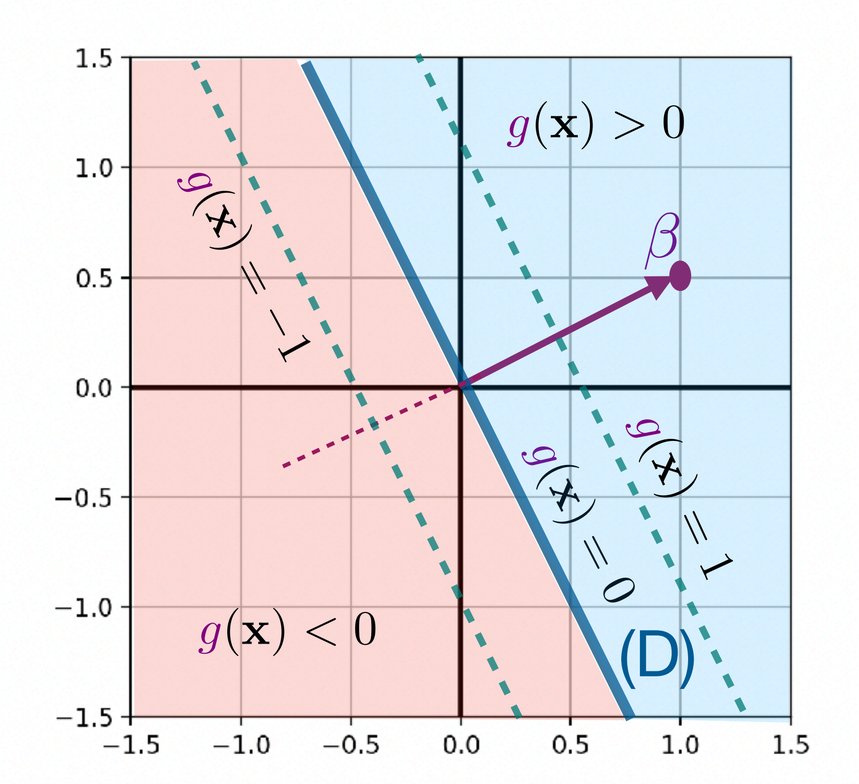
à gauche de (D), $\beta^\top \mathbf{x} < 0$

et si $c = 1$? ou $c = -1$?

INSEA

séparateur linéaire en dimension 2

On considère $g : \mathbf{x} \mapsto \beta^\top \mathbf{x}$. Étudions ses courbes de niveaux, c-à-d pour $c \in \mathbb{R}$ les ensembles: $\{\mathbf{x} | g(\mathbf{x}) = c\}$.



Exemple avec $\beta = (1, 0.5)^\top$ et $c = 0$.

Quels sont les $\mathbf{x}$ tels que $\beta^\top \mathbf{x} = 0$?

Tous les vecteurs orthogonaux à β.

$\{\mathbf{x} \in \mathbb{R}^2 | \beta^\top \mathbf{x} = 0\}$ est la droite perpendiculaire à β.

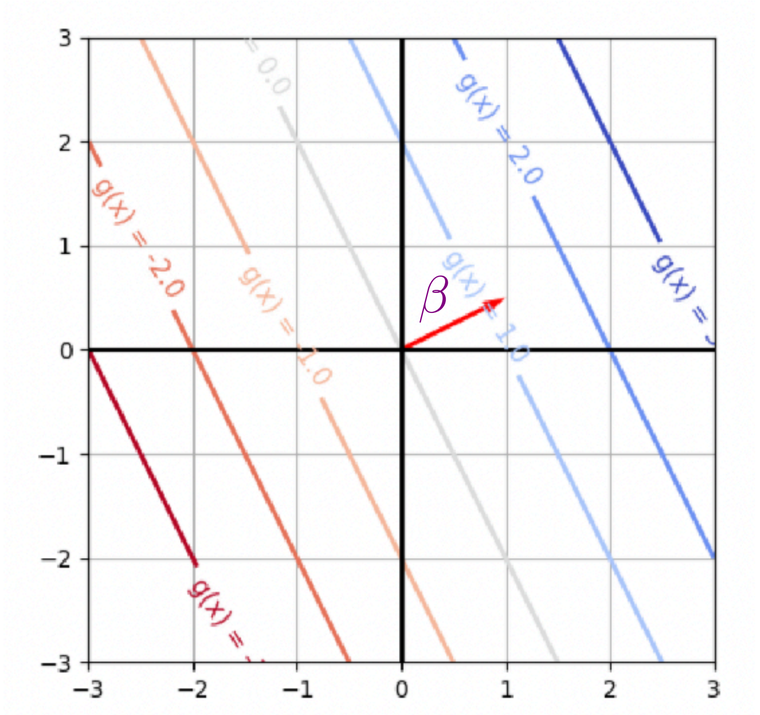
à droite de (D), $\beta^\top \mathbf{x} > 0$

à gauche de (D), $\beta^\top \mathbf{x} < 0$

et si $c = 1$? ou $c = -1$?

INSEA

séparateur linéaire en dimension 2

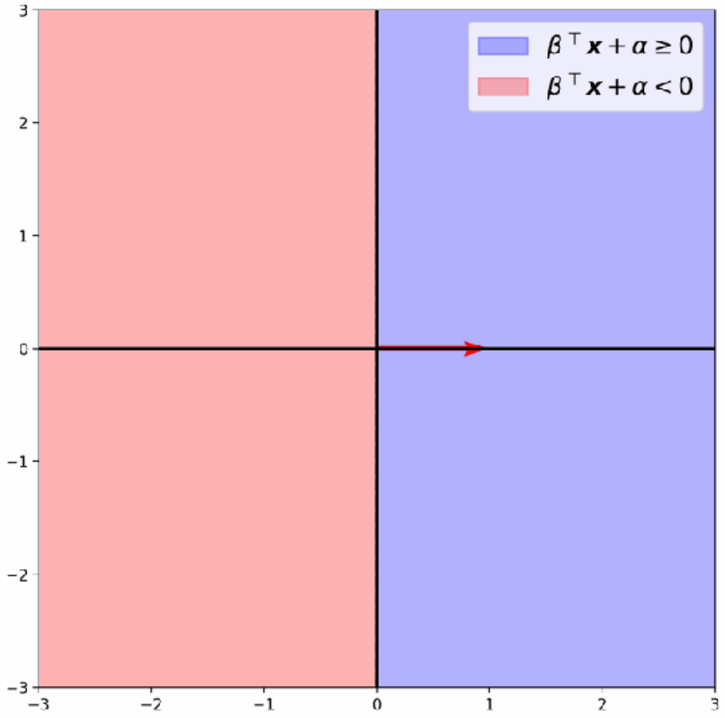


Comment change la fonction de prédiction $f : \mathbb{1}_{\{\alpha+\beta^\top \mathbf{x} \geq 0\}}$ en fonction de α et β ?

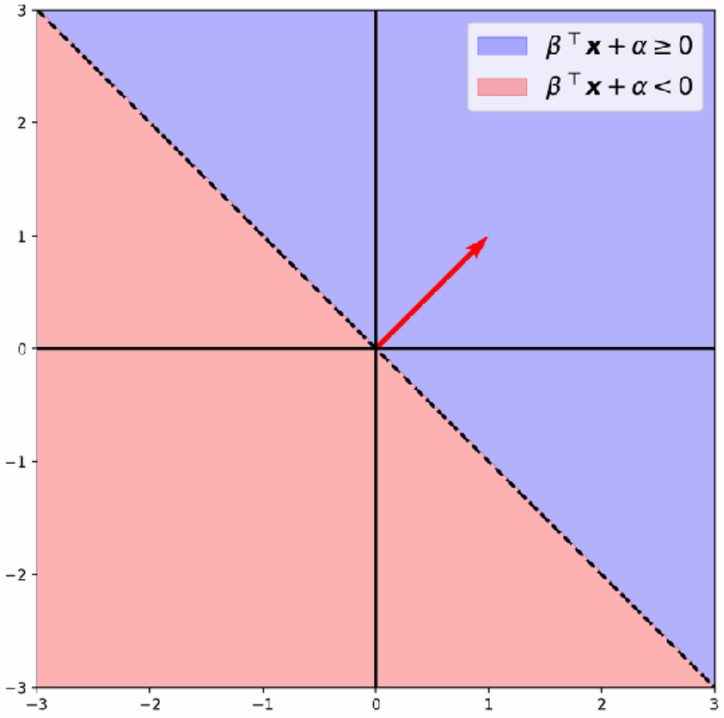
INSEA

séparateur linéaire en dimension 2

$\alpha = 0$, β varie:



α varie, $\beta = [1, 1]$:

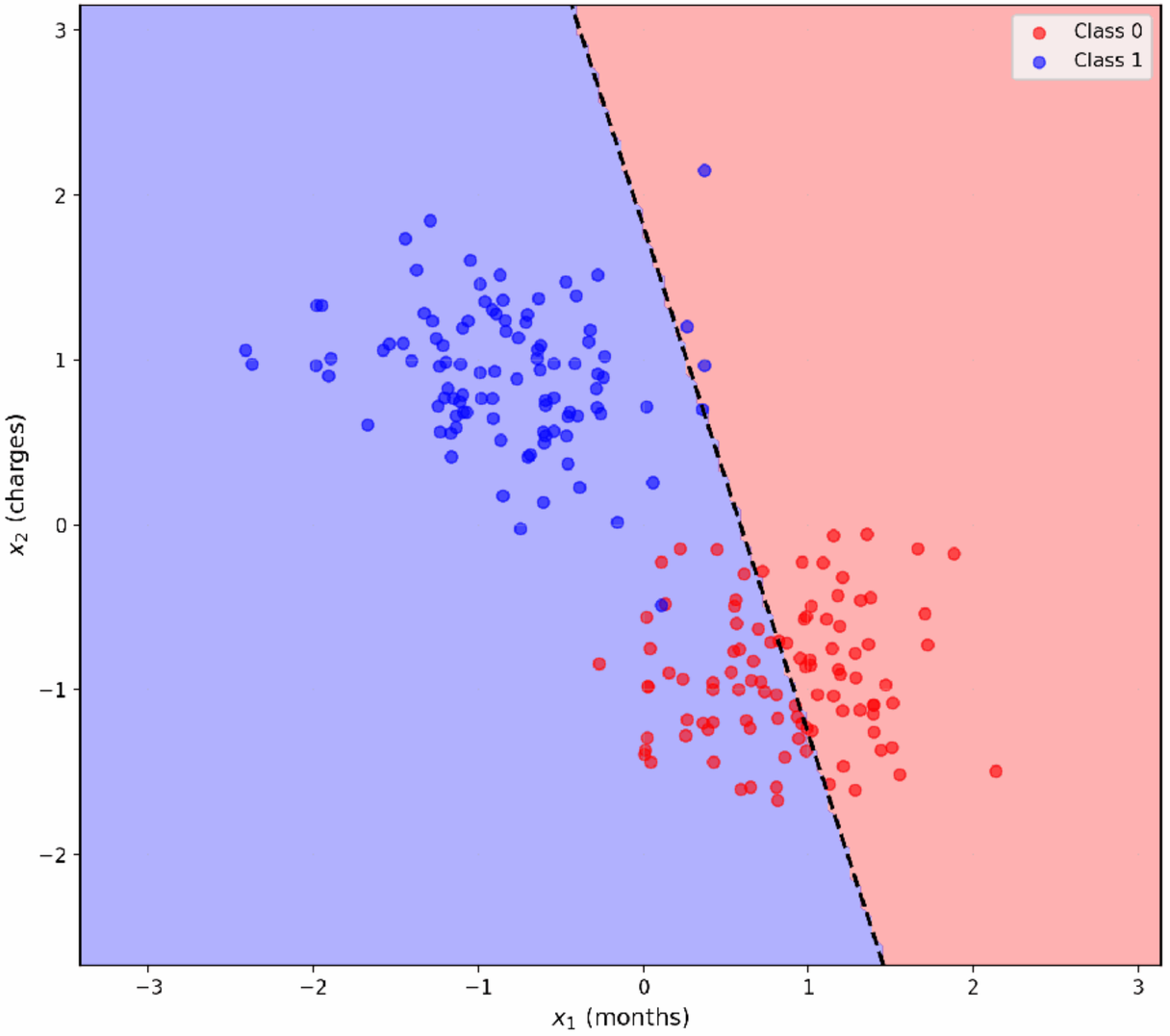


Comment change la fonction de prédiction $f : \mathbb{1}_{\{\alpha + \beta^\top \mathbf{x} \geq 0\}}$ en fonction de α et β ?

INSEA

séparateur linéaire en dimension 2

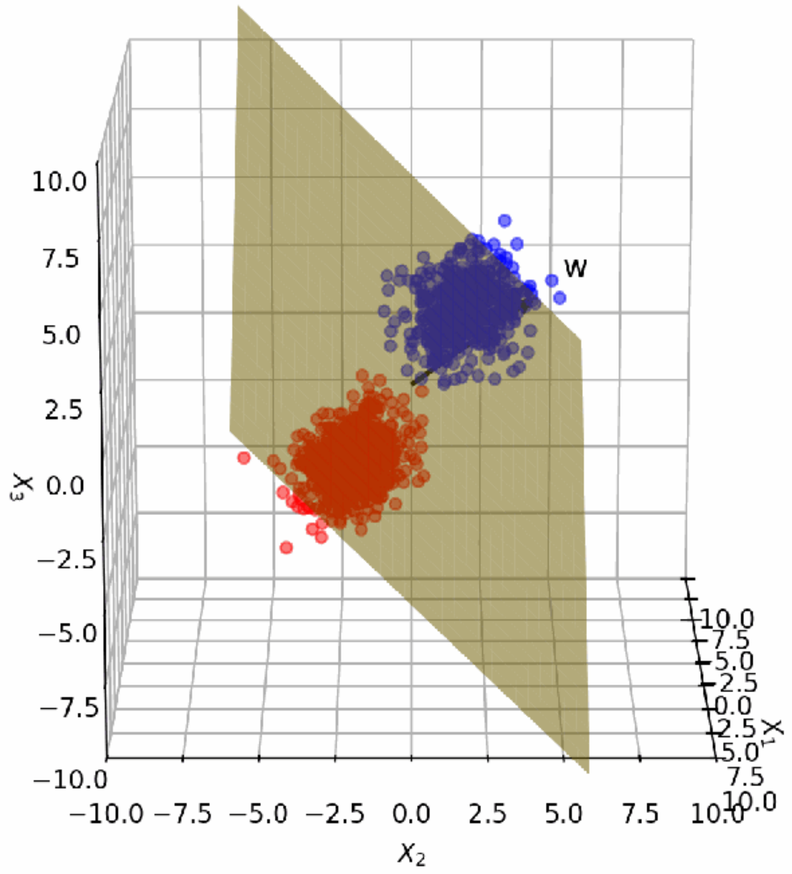
$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^2} \sum_{i=1}^{n} (\mathbb{1}_{\{\alpha+\beta^\top \mathbf{x}_i \geq 0\}} - y_i)^2$$



INSEA

séparateur linéaire en dimension d

Et si on utilise trois variables:



$$g(\mathbf{x}) = \alpha + \beta_1 \mathbf{x}^1 + \beta_2 \mathbf{x}^2 + \beta_3 \mathbf{x}^3$$

$$g(\mathbf{x}) = \alpha + \beta^\top \mathbf{x}$$

Que forment les $\mathbf{x}$ tels que $\{g(\mathbf{x}) = 0\}$?

En dimension d: $g(\mathbf{x}) = \alpha + \beta^\top \mathbf{x}, \quad \beta \in \mathbb{R}^d$

Que forment les $\mathbf{x}$ tels que $\{g(\mathbf{x}) = 0\}$?

Un espace de dimension d-1: un hyperplan

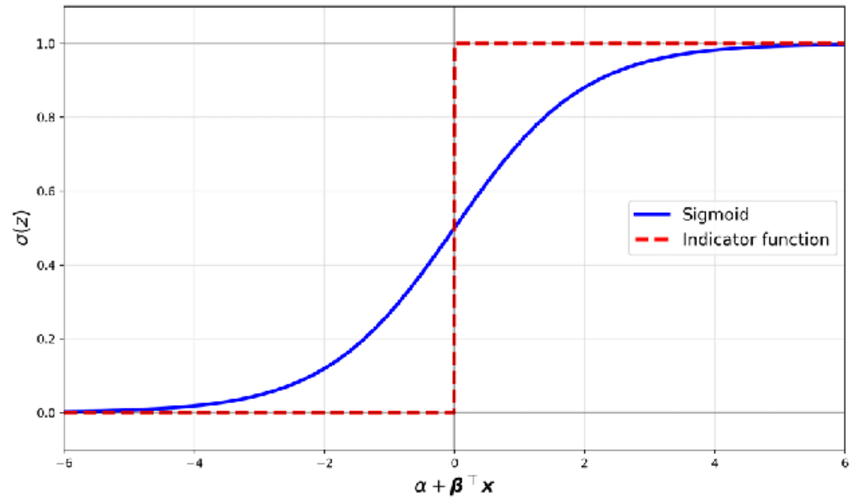
$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^d} \sum_{i=1}^{n} (\mathbb{1}_{\{\alpha+\beta^\top \mathbf{x}_i \geq 0\}} - y_i)^2$$

Fonction non différentiable (discontinue même) difficile à optimiser

Au lieu de prendre le signe, transformer les scores $\alpha + \beta^\top \mathbf{x}_i$ vers $[0, 1]$ et modéliser des probabilités

$$\text{sigmoid}: \ t \mapsto \frac{1}{1+e^{-t}} \quad \text{(logistique)}$$

$$p_i \overset{\text{def}}{=} \mathbb{P}(y_i = 1|\mathbf{x}_i) = \text{sigmoid}(\alpha + \beta^\top \mathbf{x}_i)$$



On peut comparer les p_i avec les y_i avec la *cross-entropy*:

$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^d} -\sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i)$$

On a donc une fonction de prédiction: $f^\star(\mathbf{x}_i) = 1 \Leftrightarrow \text{sigmoid}(\alpha^\star + \beta^{\star\top}\mathbf{x}_i) \geq \frac{1}{2}$

Modèle de régression logistique

Train and test error

$$p_i \overset{\text{def}}{=} \mathbb{P}(y_i = 1|\mathbf{x}_i) = \text{sigmoid}(\alpha + \beta^\top \mathbf{x}_i)$$

$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^d} -\sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i)$$

Optimisation faite sur $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

“Training” data

$\mathbf{x}_1$	y_1
$\vdots$	$\vdots$
$\mathbf{x}_n$	y_n

→ “Training” → “Learned” f* →

predictions	true labels
$f^\star(\mathbf{x}_1)$	y_1
$\vdots$	$\vdots$
$f^\star(\mathbf{x}_n)$	y_n

→ “Train” error

Est-ce une bonne manière d’évaluation la performance du modèle ?

L’erreur de prédiction sur ces données est optimisée: elle est forcément petite.

Il faut évaluer la performance du modèle sur des données nouvelles non vues à l’entraînement: “Test data”

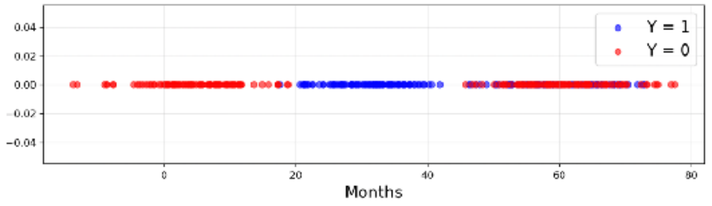
predictions	true labels
$f^\star(\mathbf{x}'_1)$	y'_1
$\vdots$	$\vdots$
$f^\star(\mathbf{x}'_m)$	y'_m

→ “Test” error

Train error goes to zero

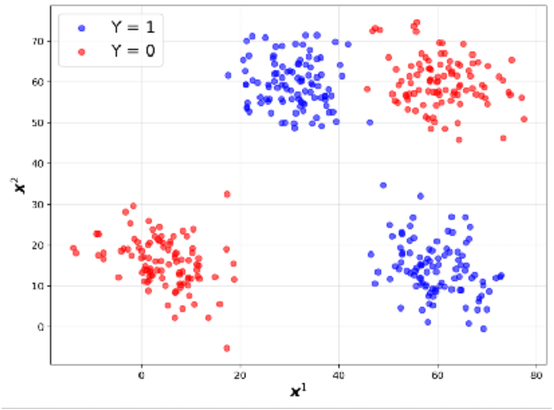
Peut-on séparer les classes avec une séparation linéaire dans ces cas ?

d = 1



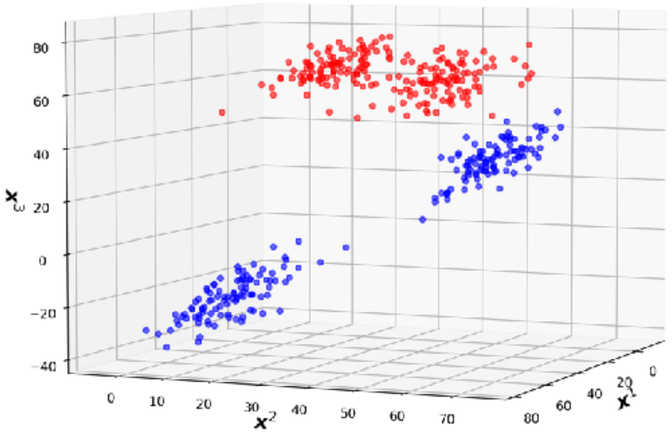
Non !

d = 2



Non !

d = 3



Oui !

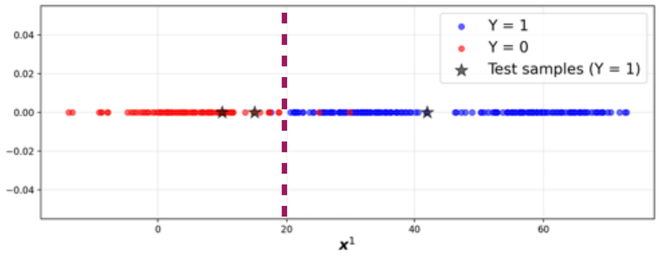
d + 1 représente le nombre de paramètres à estimer: plus d est grand, plus le modèle est riche, complexe.

Comment évolue l'erreur sur le train au fur-et-à mesure que la dimension d augmente ?

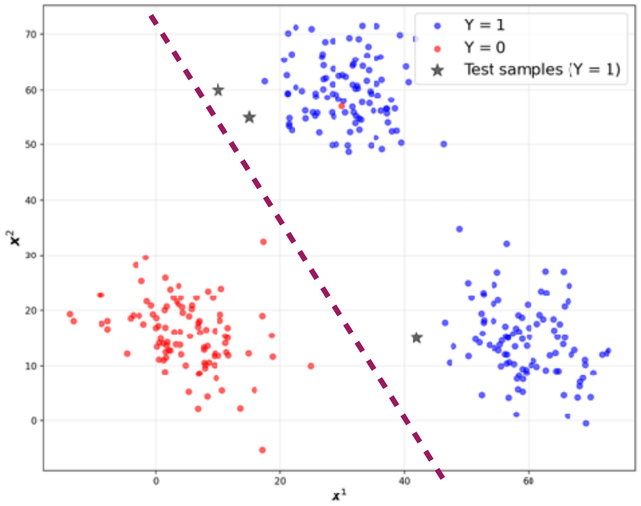
Test error makes a U shape

Quelle est la meilleure séparation linéaire sur ces données ?

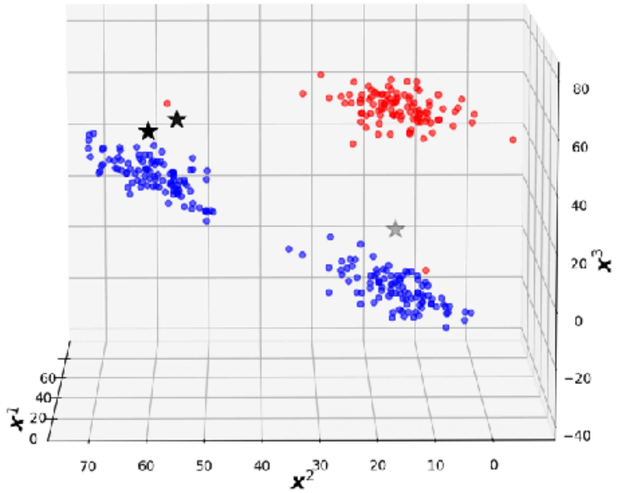
d = 1

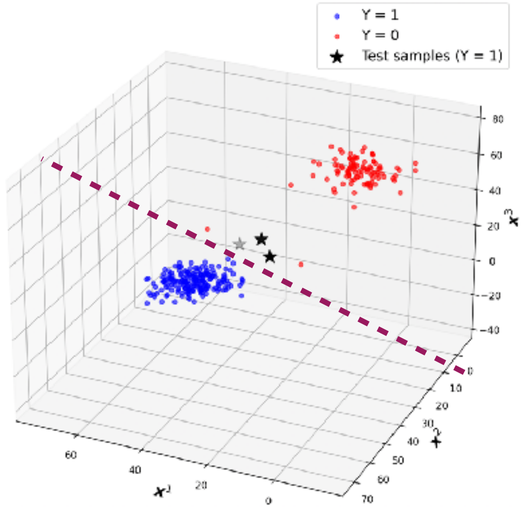


d = 2



d = 3





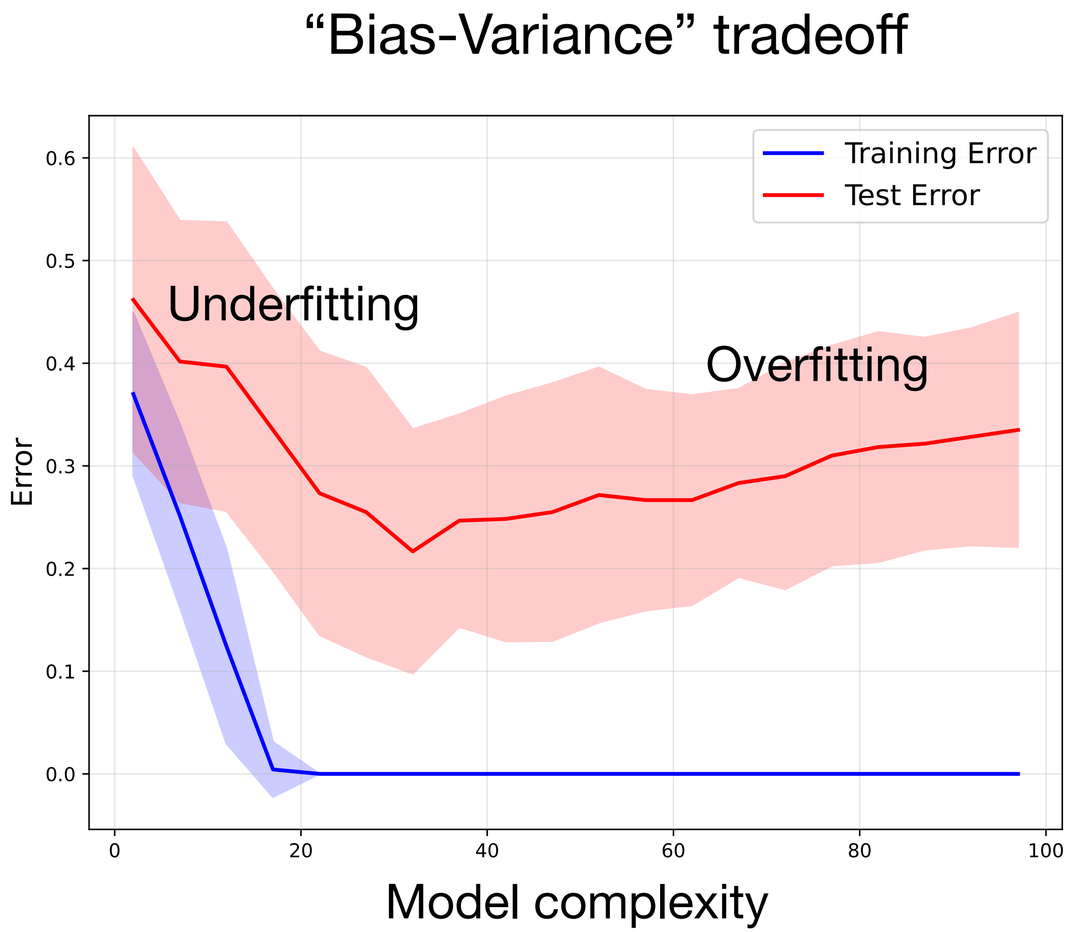
Calculer l'erreur de train et de test.

d = 3 donne la meilleure erreur de train = 0

d = 2 donne la meilleure erreur de test = 0

"La meilleure" séparation linéaire sur le train n'est pas la meilleure sur le test: elle est biaisée par les outliers

Une grande dimension peut causer **l'overfitting**



Underfitting correspond à:

Grand biais ou grande variance ?

Variance nulle = prédiction constante
= underfitting

Pour réduire l'overfitting, on peut réduire l'espace d'optimisation en privilégiant des coefficients simples:

$$\min_{\substack{\beta\in\mathbb{R}^{d+1}\\ \|\beta\|_2^2\leq C}} -\sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i)$$

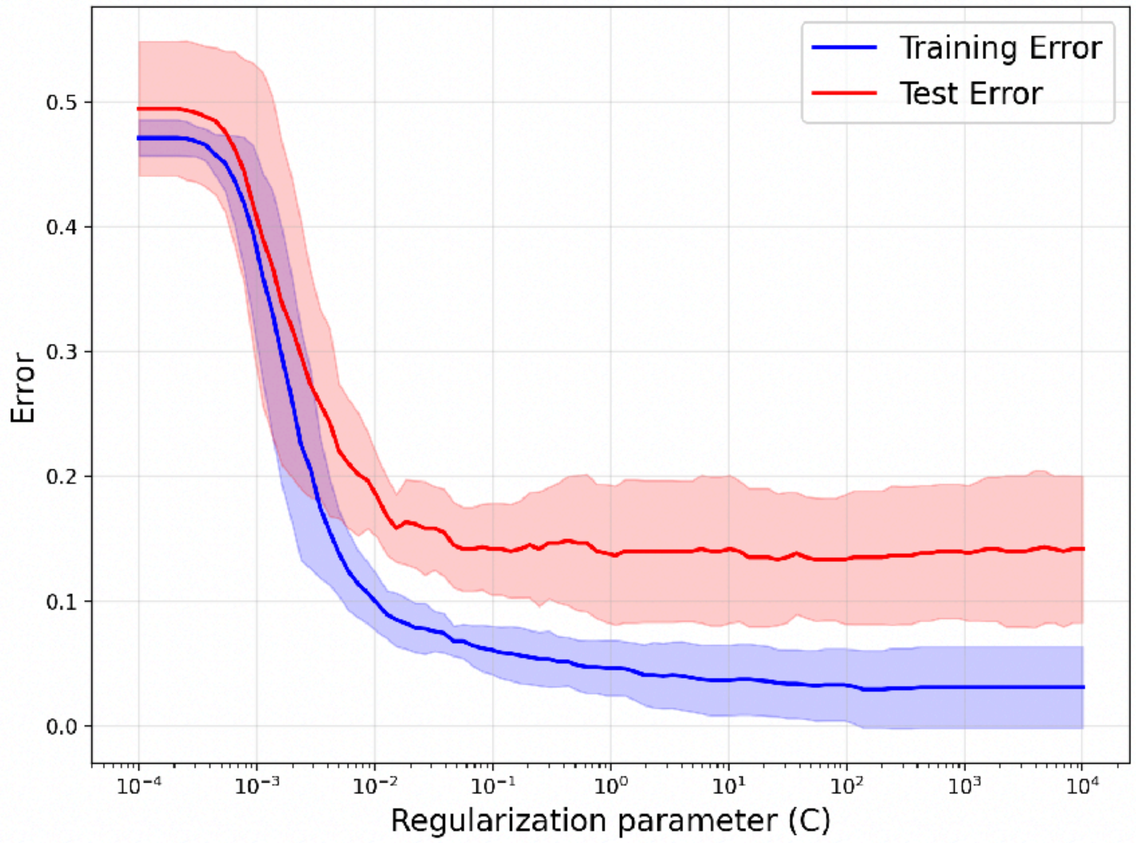
Ce problème n'est pas facile à résoudre (contrainte quadratique), on peut montrer que ce problème est équivalent:

$$\min_{\beta\in\mathbb{R}^{d+1}} -\sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i) + \frac{1}{C}\|\beta\|_2^2$$

Dans les deux cas plus C est petit plus on minimise $\|\beta\|$: "Plus on régularise".

$C \to 0$? le β optimal est le vecteur nul: la fonction de prédiction est constante: underfitting

$C \to +\infty$? l'optimisation est sur $\mathbb{R}^{d+1}$ en entier: risque d'overfitting.


Training Error
Test Error
Error
0.0
0.1
0.2
0.3
0.4
0.5
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
Regularization parameter (C)

Tout modèle de machine learning (supervisé) cherche une fonction de prédiction f.

Supposons qu'elle est paramétrée par $\theta \in \mathbb{R}^p$.

Tout modèle de machine learning cherche un compromis entre:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f_\theta(\mathbf{x}_i), y_i) \quad + \quad \frac{1}{C} \text{pénalité}(\theta)$$

Minimiser l'erreur de prédiction sur les données "train"

des paramètres "simples" pour généraliser à des données nouvelles test (éviter l'overfitting)

C contrôle la complexité du modèle

La fonction "pénalité" est aussi appelée "régularisation": elle vient simplifier (régulariser) la fonction de prédiction

Ce type de régularisation (+ pénalité) est dit: régularisation de Tikhonov

Comment choisir la pénalité ?

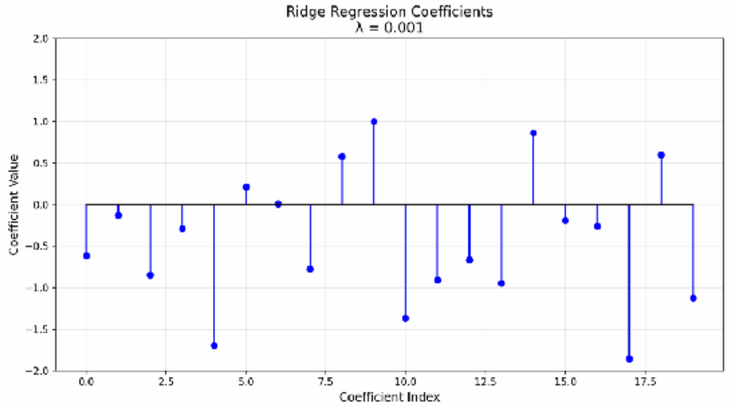
Comment choisir C ?

Comment choisir la pénalité ?

Les pénalités les plus utilisés sont:

1. pénalité Ridge $/\ell_2$: $\|\theta\|_2^2$

1. Facile à optimiser (différentiable)
2. Toutes les variables contribuent au modèle
3. Peut être ajoutée à n'importe quel modèle

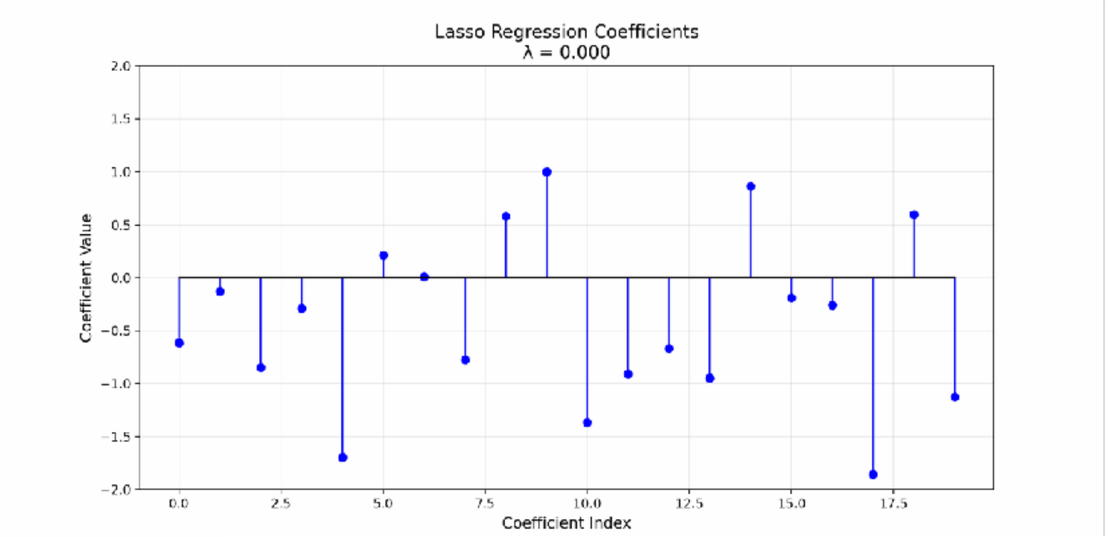


$$\lambda = \frac{1}{C}$$

Ridge: "*shrink*" toutes les coordonnées lentement vers zéro (sans l'atteindre)

2. pénalité Lasso $/\ell_1$: $\|\theta\|_1$

1. Moins facile à optimiser (non-différentiable)
2. Permet d'avoir des coefficients "sparses" (beaucoup de 0): utile pour la sélection de variables pertinentes



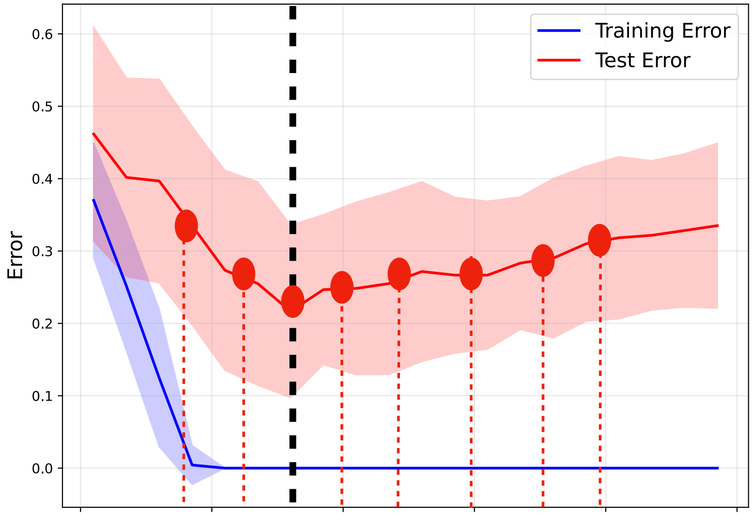
Lasso: annule les coefficients un par un

3. pénalité Elastic net : $\delta\|\theta\|_2^2 + (1-\delta)\|\theta\|_1$

Souvent, prendre un mélange des deux permet d'obtenir de meilleures performances.

Comment choisir C ?

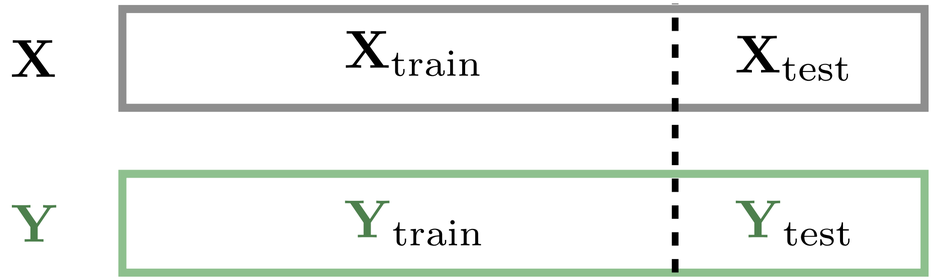
On veut le C qui donne la meilleure performance sur le test



C optimal

Pour cela on peut:

1. Couper le dataset en deux train et test:



2. Choisir une liste de valeurs de C, par ex: [0.01, 0.05, 0.1, 1., 10]

Pour chaque C:

1. Optimiser sur $\mathbf{X}_{\text{train}}$ $\mathbf{Y}_{\text{train}}$ $\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f_\theta(\mathbf{x}_i), y_i) + \frac{1}{C} \text{pénalité}(\theta)$

2. Évaluer l'erreur de prédiction sur $\mathbf{X}_{\text{test}}$ $\mathbf{Y}_{\text{test}}$

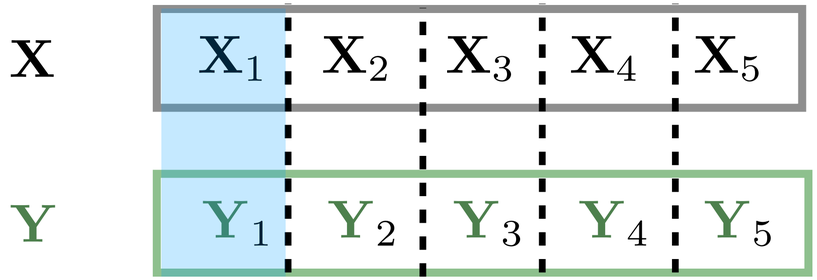
3. Choisir la valeur de C avec la plus petite erreur de prédiction sur le test

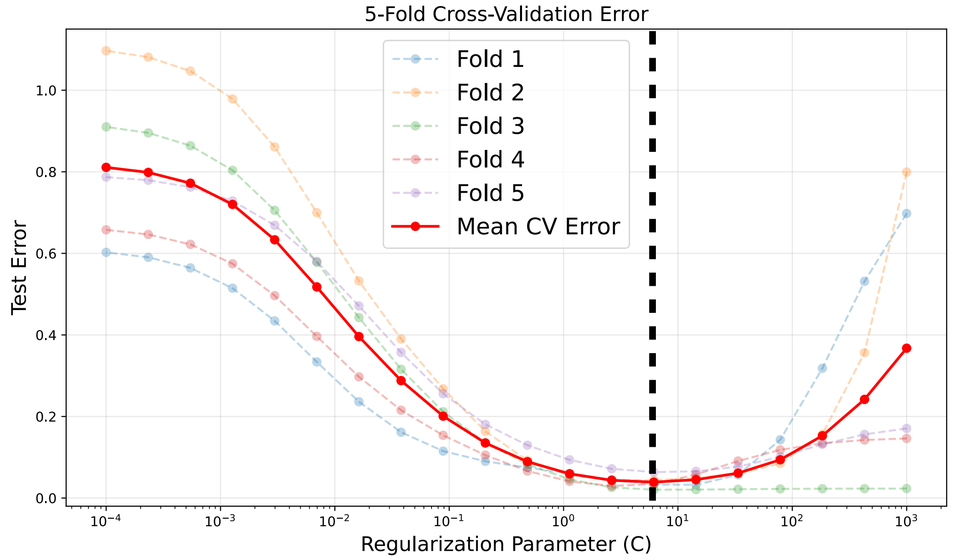
Quel est l'inconvénient principal de cette méthode ? Le C choisi dépend du découpage aléatoire train / test

Cross-validation

Idée: Effectuer plusieurs découpages et moyenner l'erreur de test

1. Couper le dataset en 5 parties (folds)





C optimal

2. Choisir une liste de valeurs de C, par ex: [0.01, 0.05, 0.1, 1., 10]

3. Pour chaque **k** in [1, 2, 3, 4, 5], créer un découpage train/test

$$\mathbf{X}_{\text{test}}, \mathbf{Y}_{\text{test}} = \mathbf{X}_k, \mathbf{Y}_k \qquad \mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}} = [\mathbf{X} \text{ sans } \mathbf{X}_k], \ldots, [\mathbf{Y} \text{ sans } \mathbf{Y}_k]$$

Pour chaque C:

1. Optimiser sur $\mathbf{X}_{\text{train}}$ $\mathbf{Y}_{\text{train}}$ $\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f_\theta(\mathbf{x}_i), y_i) + \frac{1}{C} \text{pénalité}(\theta)$

2. Évaluer l'erreur de prédiction sur $\mathbf{X}_{\text{test}}$ $\mathbf{Y}_{\text{test}}$

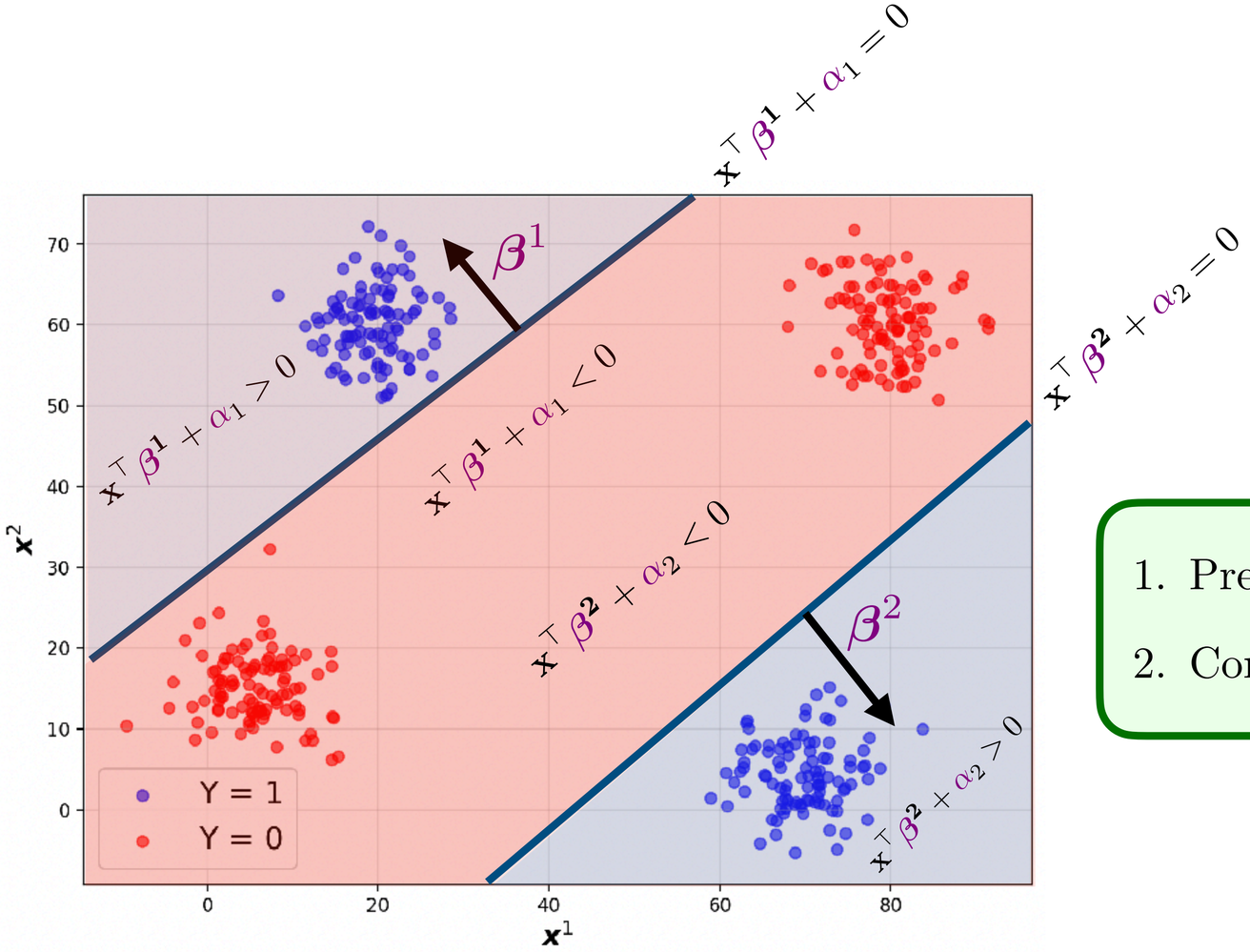
4. Pour chaque C, calculer l'erreur de prédiction moyenne

5. Choisir le C avec l'erreur de prédiction moyenne la plus petite

C'est l'erreur de validation croisée

5-Fold cross validation

Et si les données ressemblent à ceci ?



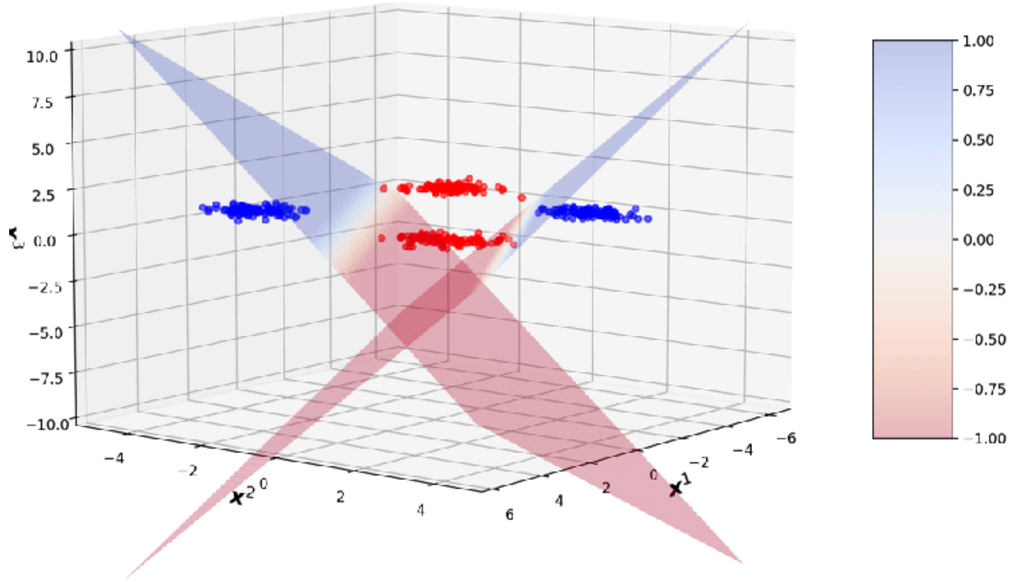
Aucune fonction linéaire ne peut séparer les classes

Idée: “combiner” plusieurs fonctions linéaires

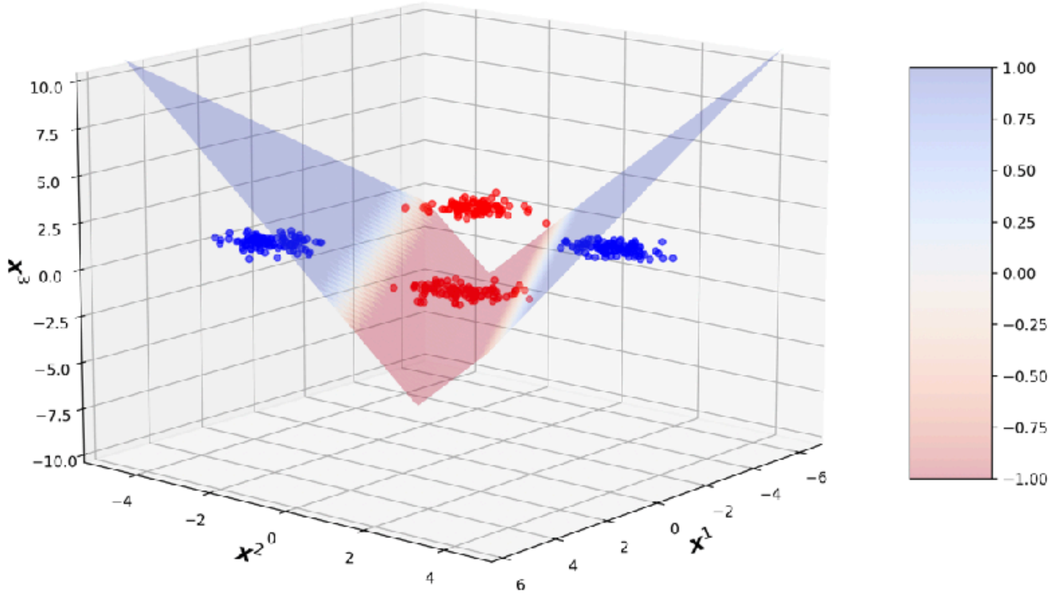
$$z_1 \overset{\text{def}}{=} \mathbf{x}^\top\boldsymbol{\beta^1}+\alpha_1 \qquad z_2 \overset{\text{def}}{=} \mathbf{x}^\top\boldsymbol{\beta^2}+\alpha_2$$

1. Prendre des $\mathbf{x}$ dans $\mathbb{R}^2$ et étudier les signes possibles de z_1, z_2.
2. Comment peut-on prédire $Y = 1$ à partir des z_i ?

Surfaces des hyperplans z_1, z_2



Surface de $\max(z_1, z_2)$



Prédire $Y = 1$ si l'un des z_i est positif $\Leftrightarrow \max(z_1, z_2) > 0$

$$f_{\alpha,\beta}(\mathbf{x}) = \mathbb{1}_{\{\max(\mathbf{x}^\top \boldsymbol{\beta^1} + \alpha_1, \mathbf{x}^\top \boldsymbol{\beta^2} + \alpha_2) > 0\}}$$

Linear functions

Comment entraîner ce modèle, c-à-d optimiser α, β ?

+

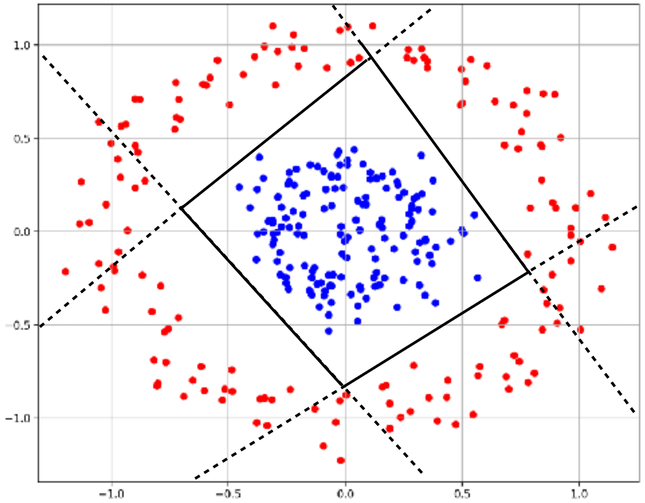
$$p_i \overset{\text{def}}{=} \mathbb{P}_{\alpha,\beta}(Y = 1|\mathbf{x}_i) = \text{sigmoid}(\max(\mathbf{x}_i^\top \boldsymbol{\beta^1} + \alpha_1, \mathbf{x}_i^\top \boldsymbol{\beta^2} + \alpha_2))$$

Non-linearity

Comme la régression logistique:

$$\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d} -\sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

Comment adapter ce modèle à des données plus complexes ?



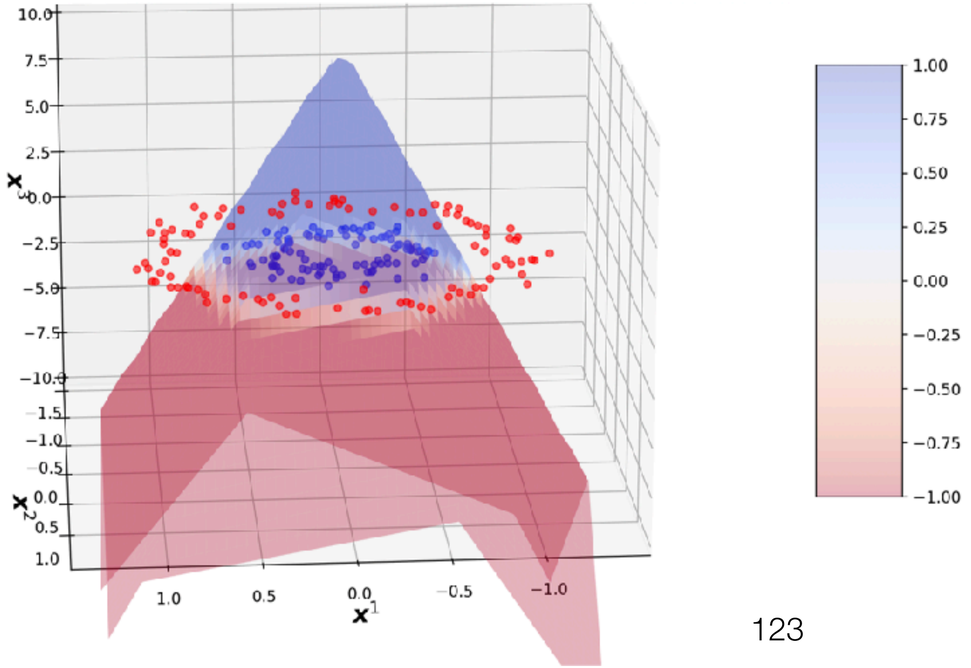
Linéarités

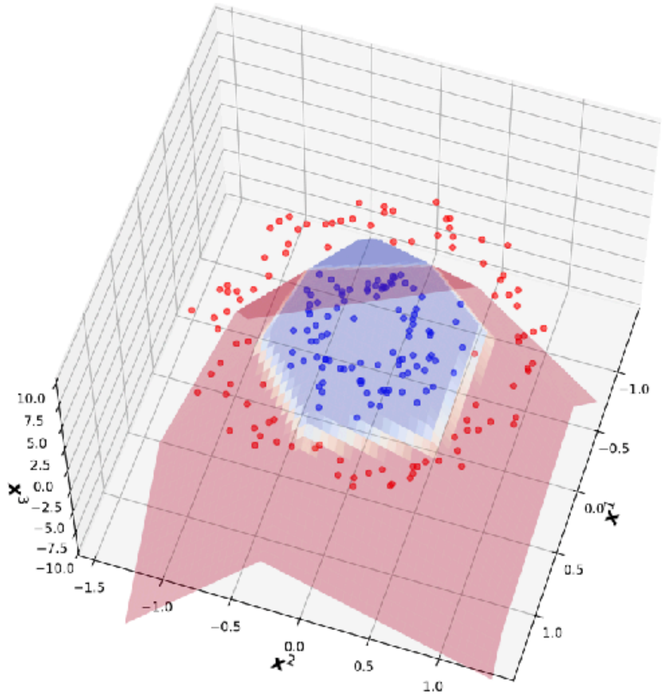
$$z_1 \overset{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta^1} + \alpha_1$$
$$z_2 \overset{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta^2} + \alpha_2$$
$$\vdots$$
$$z_p \overset{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta^p} + \alpha_p$$

non-linéarité

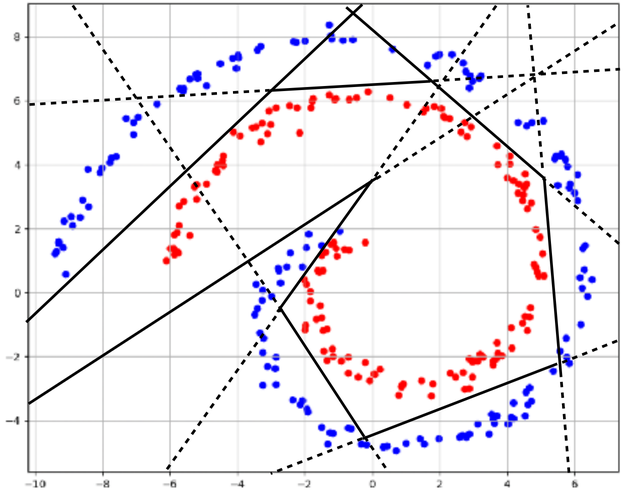
$$\max(z_1, \ldots, z_p)$$

sigmoid





Comment adapter ce modèle à des données plus complexes ?



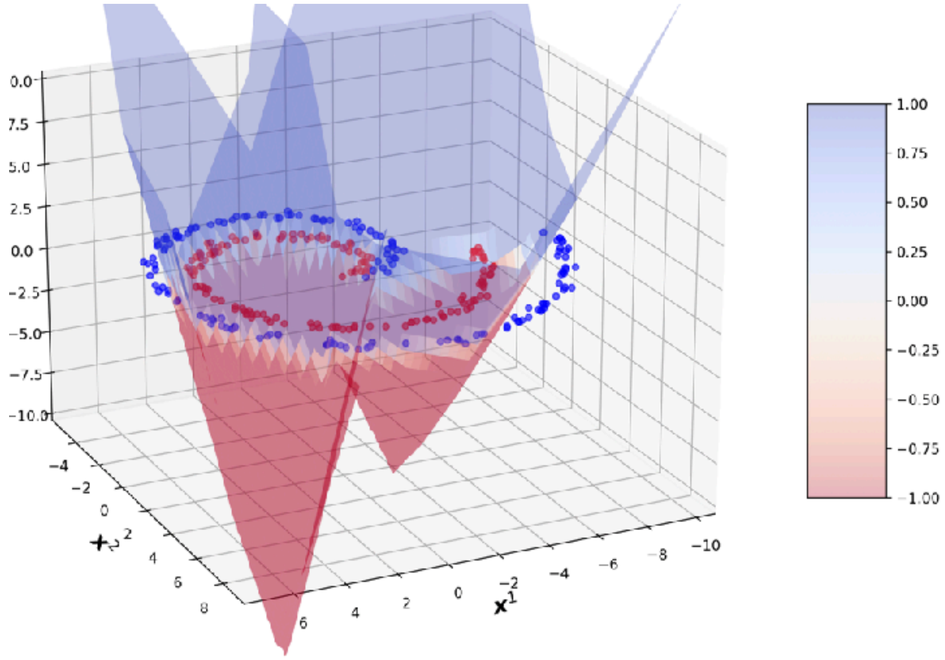
Linéarités

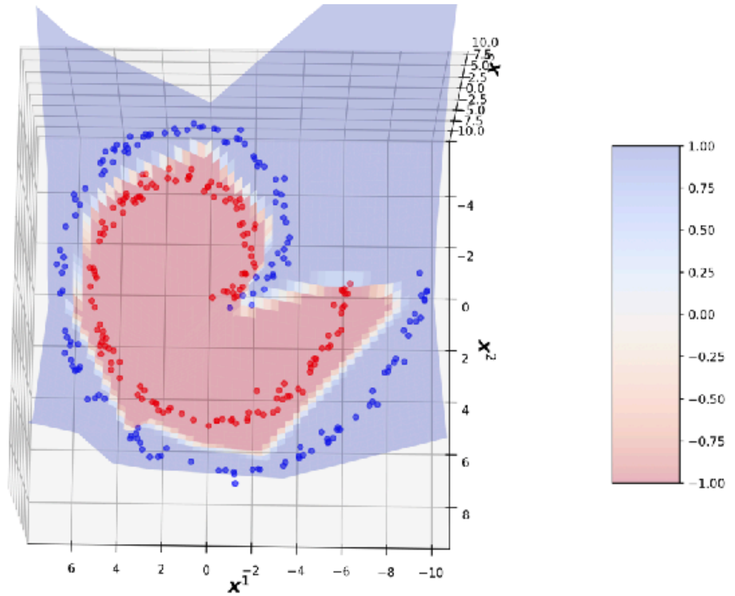
$$z_1 \overset{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta^1} + \alpha_1$$
$$z_2 \overset{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta^2} + \alpha_2$$
$$\vdots$$
$$z_p \overset{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta^p} + \alpha_p$$

non-linéarité

$$\max(z_1, \ldots, z_p)$$

sigmoid





Linéarités

$$z_1 \overset{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta^1} + \alpha_1$$
$$\vdots$$
$$z_p \overset{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta^p} + \alpha_p$$

non-linéarité

$$\max(z_1, \ldots, z_p) \longrightarrow \text{sigmoid}$$

En pratique, ce modèle ne fonctionne pas pour ces données complexes. Pourquoi à votre avis ?

1. On n'utilise qu'une seule non-linéarité
2. Elle est fixée par la fonction max: on ne l'apprend pas

Il faudrait donc: utiliser plusieurs non-linéarités simples + les combiner pour apprendre des fonctions non-linéaires complexes

Idée:

1. Appliquer plusieurs non-linéarités h plus tôt
2. Combiner les z_j linéairement avec w_j à optimiser

$$z_1 \overset{\text{def}}{=} h(\mathbf{x}^\top \boldsymbol{\beta^1} + \alpha_1)$$
$$\vdots$$
$$z_p \overset{\text{def}}{=} h(\mathbf{x}^\top \boldsymbol{\beta^p} + \alpha_p)$$

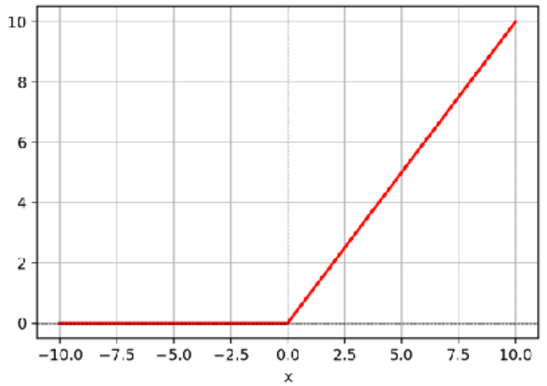
$$\sum_{j=1}^{p} \omega_j z_j + \omega_0 \longrightarrow \text{sigmoid}$$

$$z_1 \overset{\text{def}}{=} h(\mathbf{x}^\top \boldsymbol{\beta^1} + \alpha_1)$$
$$\vdots$$
$$z_p \overset{\text{def}}{=} h(\mathbf{x}^\top \boldsymbol{\beta^p} + \alpha_p)$$

$$\text{sigmoid}\left(\sum_{j=1}^{p} \omega_j z_j + \omega_0\right) = \text{sigmoid}(\boldsymbol{z}^\top \omega + \omega_0)$$

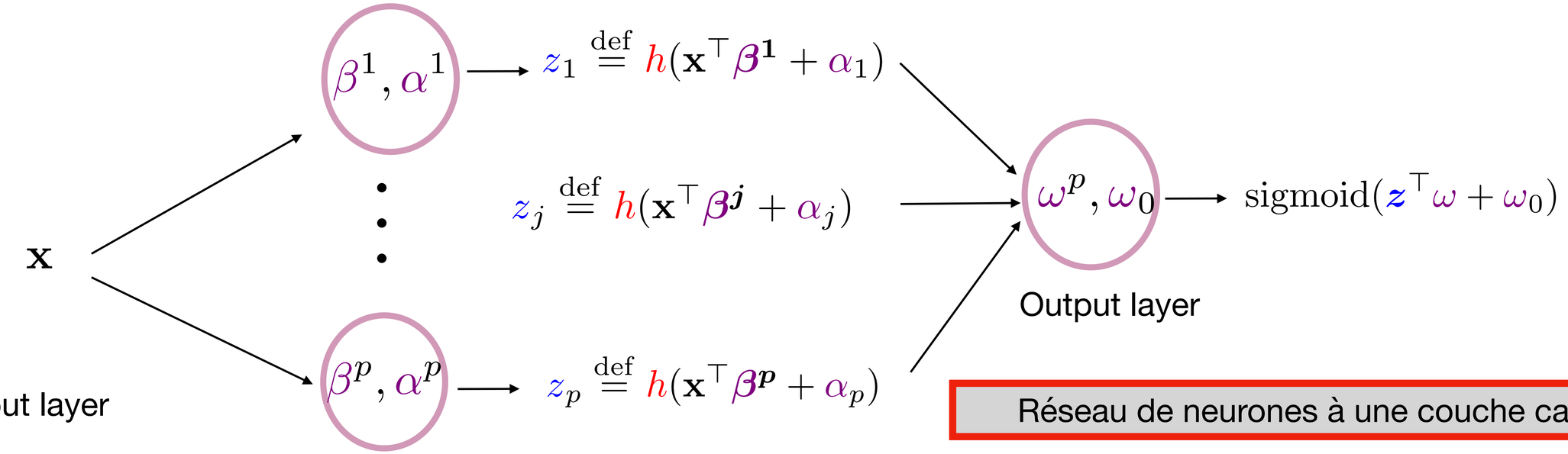
Quelle est la fonction non-linéaire h la plus simple possible ?

$$h(x) = \max(x, 0)$$



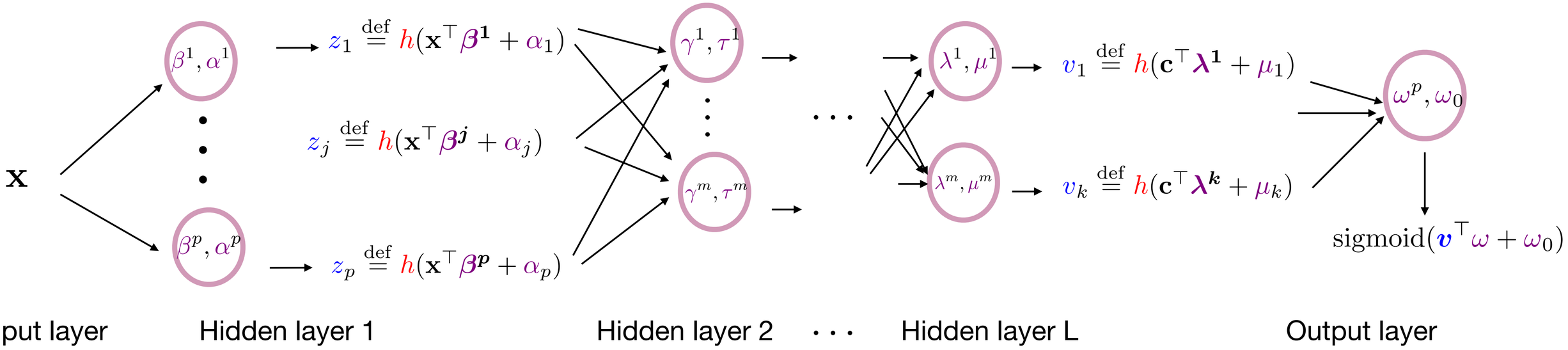
ReLU: Rectified Linear Unit

On représente ce type de modèle sous forme de graphe avec des “unités” de calcul simples: fonction linéaire + non-linéarité. Unité = un neurone:



Réseau de neurones à une couche cachée

On peut augmenter la complexité du modèle à l'infini…



Input layer | Hidden layer 1 | Hidden layer 2 | · · · | Hidden layer L | Output layer

Deep neural networks = many layers / many neurons

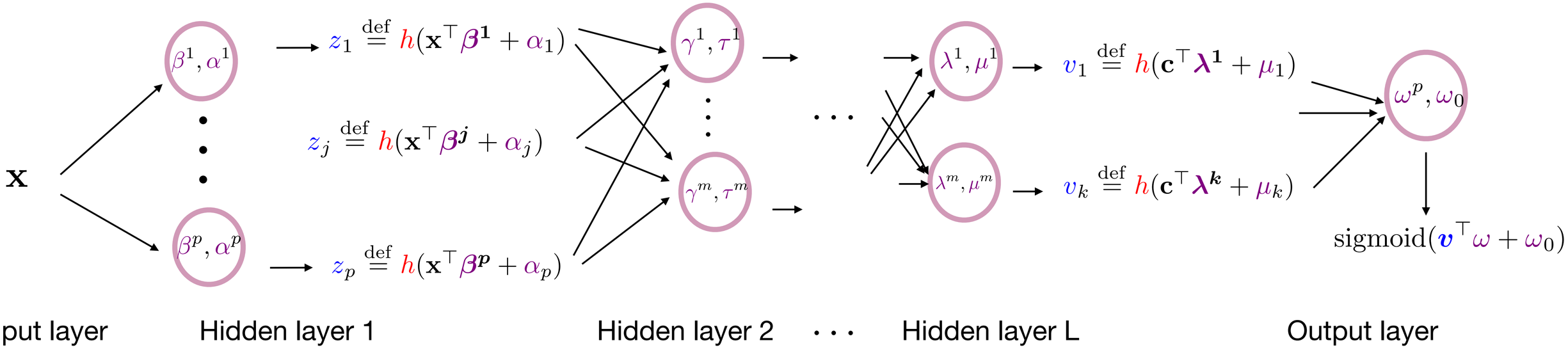
This is a "general purpose" neural network (NN) known as "fully connected multilayer perceptron" (MLP)

On considère un problème de classification binaire avec le réseau ci-dessus optimisé avec une très bonne performance.

On définit la transformation des données en s'arrêtant à l'avant dernier layer: $g : \mathbf{x} \in \mathbb{R}^d \mapsto \mathbf{v} \in \mathbb{R}^k$

Apprendre à classifier les $g(\mathbf{x}_i)$ est-il plus facile ou plus difficile que classifier les $\mathbf{x}_i$?

On peut augmenter la complexité du modèle à l'infini…



Input layer | Hidden layer 1 | Hidden layer 2 | · · · | Hidden layer L | Output layer

On considère un problème de classification binaire avec le réseau ci-dessus optimisé avec une très bonne performance.

On définit la transformation des données en s'arrêtant à l'avant dernier layer: $g : \mathbf{x} \in \mathbb{R}^d \mapsto \mathbf{v} \in \mathbb{R}^k$

Apprendre à classifier les $g(\mathbf{x}_i)$ est-il plus facile ou plus difficile que classifier les $\mathbf{x}_i$?

Plus facile: car un seul neurone (output) a suffit pour les classifier: ils sont **forcément** linéairement séparables

$g(\mathbf{x}_i)$ est un *embedding* ou une *représentation vectorielle* de $\mathbf{x}_i$

III - Modèles probabilistes

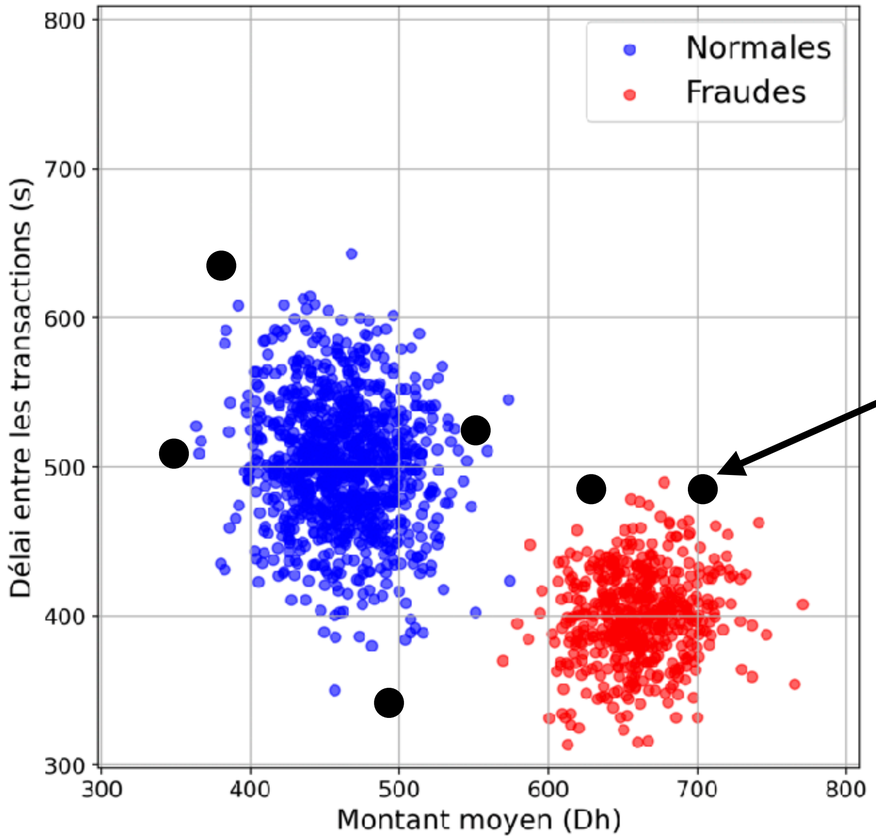
Partie 2 - Apprentissage supervisé

Intuition

Nous avons une base de données de l'utilisation quotidienne des cartes bancaires de plusieurs clients.

On observe deux variables:
(1) le montant moyen des transactions (2) le délai moyen entre les transactions

Après vérification des réclamations des clients, les transactions frauduleuses ont été marquées.
On observe les nuages de points:



On souhaite anticiper les fraudes et les détecter rapidement.

Une nouvelle observation dans la base de donnée est faite, quelle devrait être sa classe: fraude ou normale ?

Idée: on compare les probabilités

$$\mathbb{P}(fraude|\text{montant, délai}) \text{ et } \mathbb{P}(normale|\text{montant, délai})$$

Exemple Gaussien

Pour cela on besoin d'un modèle. On note $\mathbf{X} = (\text{Montant}, \text{Délai})^\top$ et $\mathbf{Y} \in \{F, N\}$.

Nous allons modéliser les probabilités conditionnelles $\mathbb{P}(\mathbf{X}|Y = F)$ et $\mathbb{P}(\mathbf{X}|Y = N)$ en utilisant des lois connues.

Dans cet exemple on peut considérer les modèles:

$$\mathbb{P}(\mathbf{X}|Y = F) = \mathcal{N}(\mu_F, \Sigma_F) \text{ et } \mathbb{P}(\mathbf{X}|Y = N) = \mathcal{N}(\mu_N, \Sigma_N).$$

Où $\mu_F, \Sigma_F, \mu_N, \Sigma_N$ sont appelés "paramètres" du modèle et doivent être estimés à partir des données.

On observe un nouveau $\mathbf{x}$ pour lequel on ne connait pas le y correspondant. Pour le prédire, on doit comparer:

$$\mathbb{P}(Y = F|\mathbf{X} = \mathbf{x}) \text{ et } \mathbb{P}(Y = N|X = x).$$

Comment peut-on les calculer ?

En utilisant le théorème de Bayes deux fois:

$$\mathbb{P}(Y = N|\mathbf{X} = \mathbf{x}) = \frac{f_{(\mathbf{X},Y)}(\mathbf{x},N)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}|Y=N}(\mathbf{x})\mathbb{P}(Y=N)}{f_{\mathbf{X}}(\mathbf{x})}$$

Et pareil:

$$\mathbb{P}(Y = F|\mathbf{X} = \mathbf{x}) = \frac{f_{(\mathbf{X},Y)}(\mathbf{x},F)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}|Y=F}(\mathbf{x})\mathbb{P}(Y=F)}{f_{\mathbf{X}}(\mathbf{x})}$$

Or, avec notre modèle:

$$f_{\mathbf{X}|Y=N}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\det(\Sigma_N)} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mu_N)^\top {\Sigma_N}^{-1}(\mathbf{x}-\mu_N)\right)$$

$$f_{\mathbf{X}|Y=F}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\det(\Sigma_F)} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mu_F)^\top {\Sigma_F}^{-1}(\mathbf{x}-\mu_F)\right)$$

Exercice

On suppose le modèle simplifié: $\Sigma_F = \Sigma_N = \sigma^2 I_2$. On note $p_N = \mathbb{P}(Y = N)$ et $p_F = \mathbb{P}(Y = F)$.

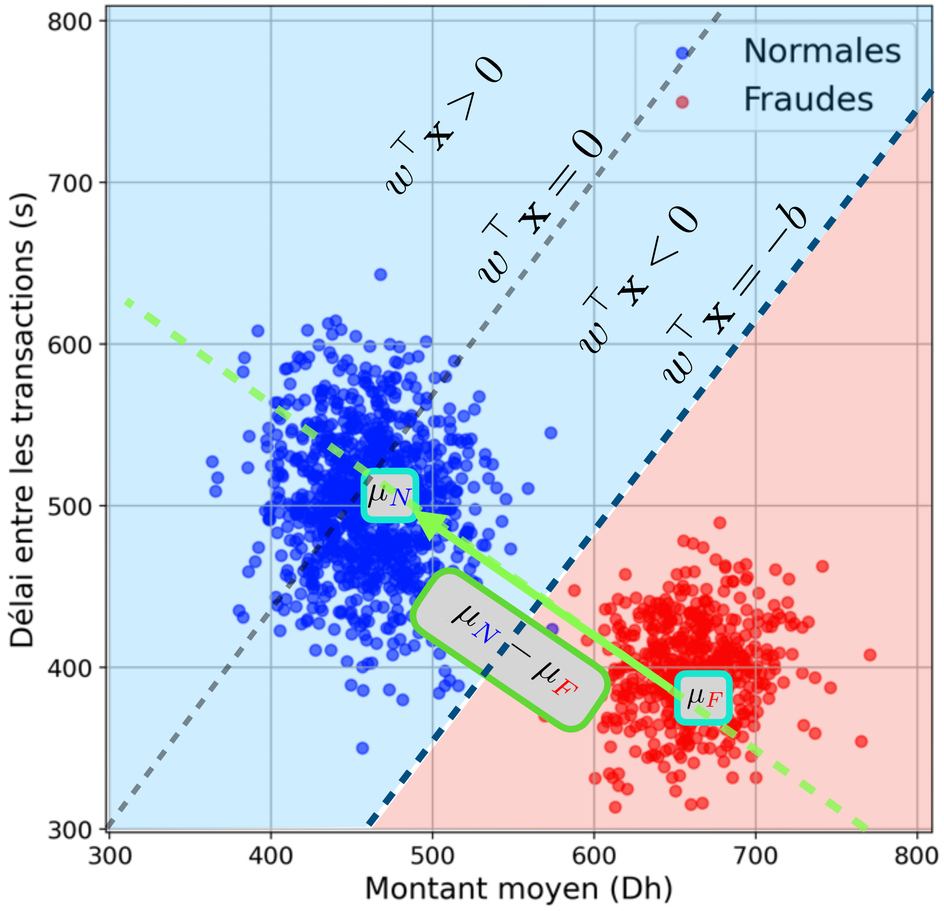
On observe les données $(X_1, y_1), \ldots, (X_n, y_n)$ i.i.d. On cherche à apprendre un classifieur $g : \mathbb{R}^2 \to \{N, F\}$.

On suppose que parmi les observations y_i, il y a autant de N que de F.

1. Proposez des estimateurs des paramètres du modèle μ_F, μ_N, p_N, p_F et σ^2 en fonction des observations (X_i, y_i).
2. Trouver une fonction $\varphi : \mathbb{R}^2 \to \mathbb{R}$ telle que $\phi(\mathbf{x}) \geq 0 \Leftrightarrow \mathbb{P}(Y = N|\mathbf{X} = \mathbf{x}) \geq \mathbb{P}(Y = F|\mathbf{X} = \mathbf{x})$
3. En déduire une fonction de prédiction $g : \mathbb{R}^2 \to \{N, F\}$.

Séparation linéaire

Pour un $\mathbf{x}$ observé, on prédit $y = N$ si et seulement si $\varphi(\mathbf{x}) = (\mu_N - \mu_F)^\top \mathbf{x} + \frac{\|\mu_F\|^2 - \|\mu_N\|^2}{2} \geq 0$



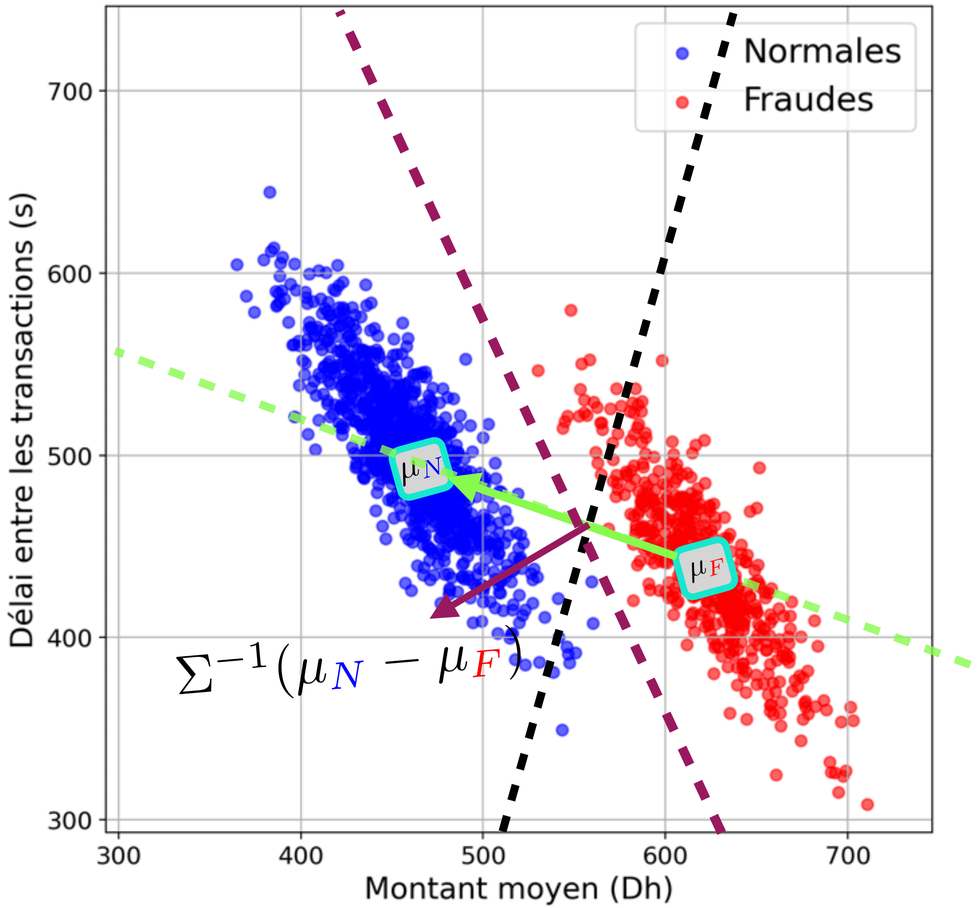
$$\Leftrightarrow w^\top \mathbf{x} + b \geq 0$$

avec $w = \mu_N - \mu_F$ et $b = \frac{\|\mu_F\|^2 - \|\mu_N\|^2}{2} > 0.$

Une fonction de prédiction peut donc être donnée par: $g : \mathbf{x} \to \text{sign}(w^\top \mathbf{x} + b)$

Séparation linéaire

On observe désormais les données suivantes:



Est-ce la meilleure séparation linéaire ?

Quelle hypothèse n'est plus vérifiée ?

Les variables ont une corrélation négative:

Σ n'est plus diagonale $\neq \sigma^2 \mathrm{I}$!

Que devient la fonction de décision φ si

$$\Sigma_N = \Sigma_F = \Sigma \ ?$$

$$\varphi(\mathbf{x}) = (\Sigma^{-1}(\mu_N - \mu_F))^\top \mathbf{x} + \frac{\mu_F^\top \Sigma^{-1} \mu_F - \mu_N^\top \Sigma^{-1} \mu_N}{2}$$

$$= w^\top \mathbf{x} + b$$

Soit $\mathbf{X}$ un vecteur aléatoire dans $\mathbb{R}^d$ et Y une variable aléatoire dans $\{-1, 1\}$.

On observe n paires i.i.d $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.

On suppose que: $\mathbb{P}(\mathbf{X}|Y = -1) = \mathcal{N}(\mu_{-1}, \Sigma_{-1})$ et $\mathbb{P}(\mathbf{X}|Y = 1) = \mathcal{N}(\mu_1, \Sigma_1)$.

On note $p_1 = \mathbb{P}(Y = 1)$ et $p_{-1} = \mathbb{P}(Y = -1)$.

$$\text{On suppose } \Sigma_1 = \Sigma_{-1} = \Sigma$$

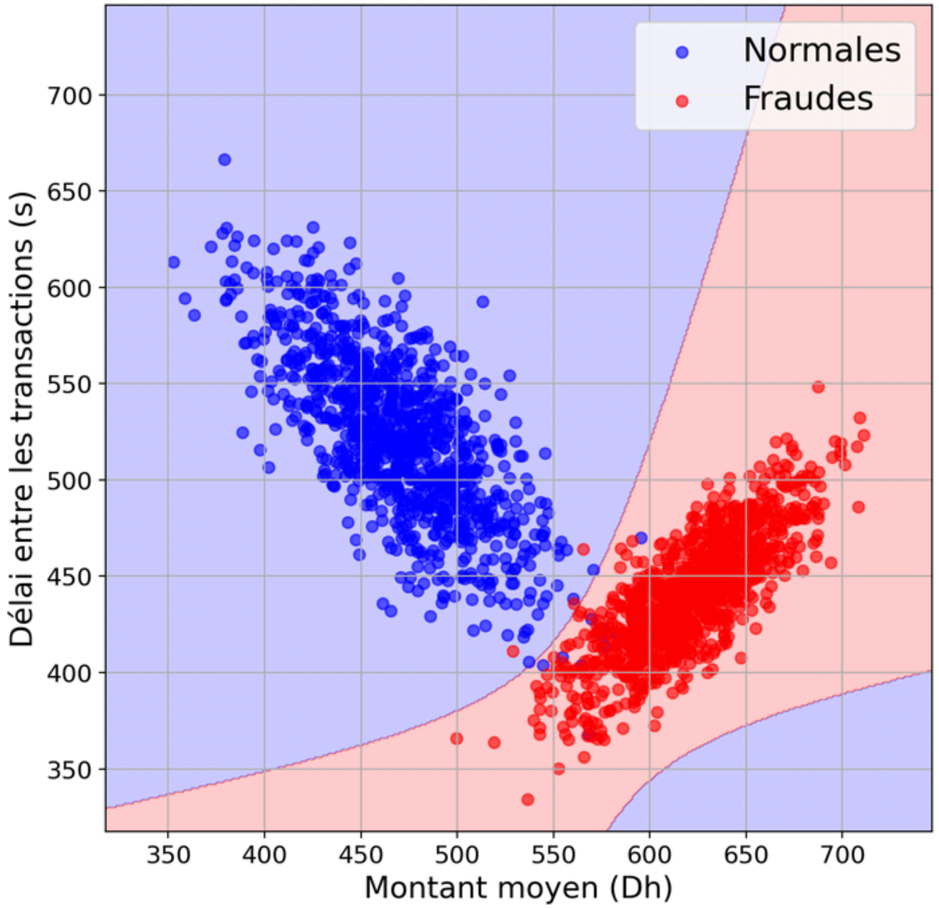
LDA

Alors la fonction de décision définie par: $\varphi : \mathbf{x} \mapsto \log \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) - \log \mathbb{P}(Y = -1|\mathbf{X} = \mathbf{x})$

est linéaire et peut s'écrire: $\varphi(\mathbf{x}) = w^\top \mathbf{x} + b$

Où $w \in \mathbb{R}^d$ et $b \in \mathbb{R}$ dépendent des paramètres μ_1, μ_{-1}, Σ et p_1, p_{-1}.

On observe désormais les données suivantes:



Quelle hypothèse n'est plus vérifiée ici ?

$$\Sigma_N \neq \Sigma_F !$$

Que devient la fonction de décision φ dans le cas général ?

Une fonction quadratique de la forme: $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{A}\mathbf{x} + w^\top \mathbf{x} + b$

Quadratic discriminant analysis (QDA)

Soit $\mathbf{X}$ un vecteur aléatoire dans $\mathbb{R}^d$ et Y une variable aléatoire dans $\{-1, 1\}$.

On observe n paires i.i.d $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.

On suppose que: $\mathbb{P}(\mathbf{X}|Y = -1) = \mathcal{N}(\mu_{-1}, \Sigma_{-1})$ et $\mathbb{P}(\mathbf{X}|Y = 1) = \mathcal{N}(\mu_1, \Sigma_1)$.

On note $p_1 = \mathbb{P}(Y = 1)$ et $p_{-1} = \mathbb{P}(Y = -1)$.

QDA

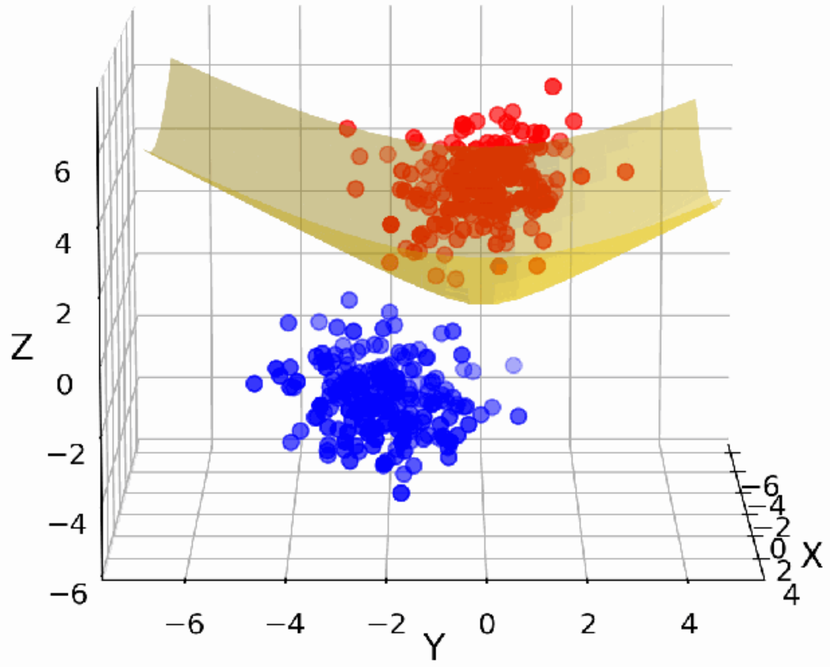
Alors la fonction de décision définie par: $\varphi : \mathbf{x} \mapsto \log \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) - \log \mathbb{P}(Y = -1|\mathbf{X} = \mathbf{x})$

est quadratique et peut s'écrire: $\varphi(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + w^\top \mathbf{x} + b$

Où $\mathbf{A} \in \mathbb{S}_d$, $w \in \mathbb{R}^d$ et $b \in \mathbb{R}$ dépendent des paramètres $\mu_1, \mu_{-1}, \Sigma_1, \Sigma_{-1}$ et p_1, p_{-1}.

INSEA

Quadratic discriminant analysis (QDA)



INSEA

Soit $\mathbf{X}$ un vecteur aléatoire dans $\mathbb{R}^d$ et Y une variable aléatoire dans $\{0, 1, \ldots, K-1\}$.

On observe n paires i.i.d $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.

On suppose que: $\mathbb{P}(\mathbf{X}|Y = k) = \mathcal{N}(\mu_k, \Sigma_k)$

On note $p_k = \mathbb{P}(Y = k)$.

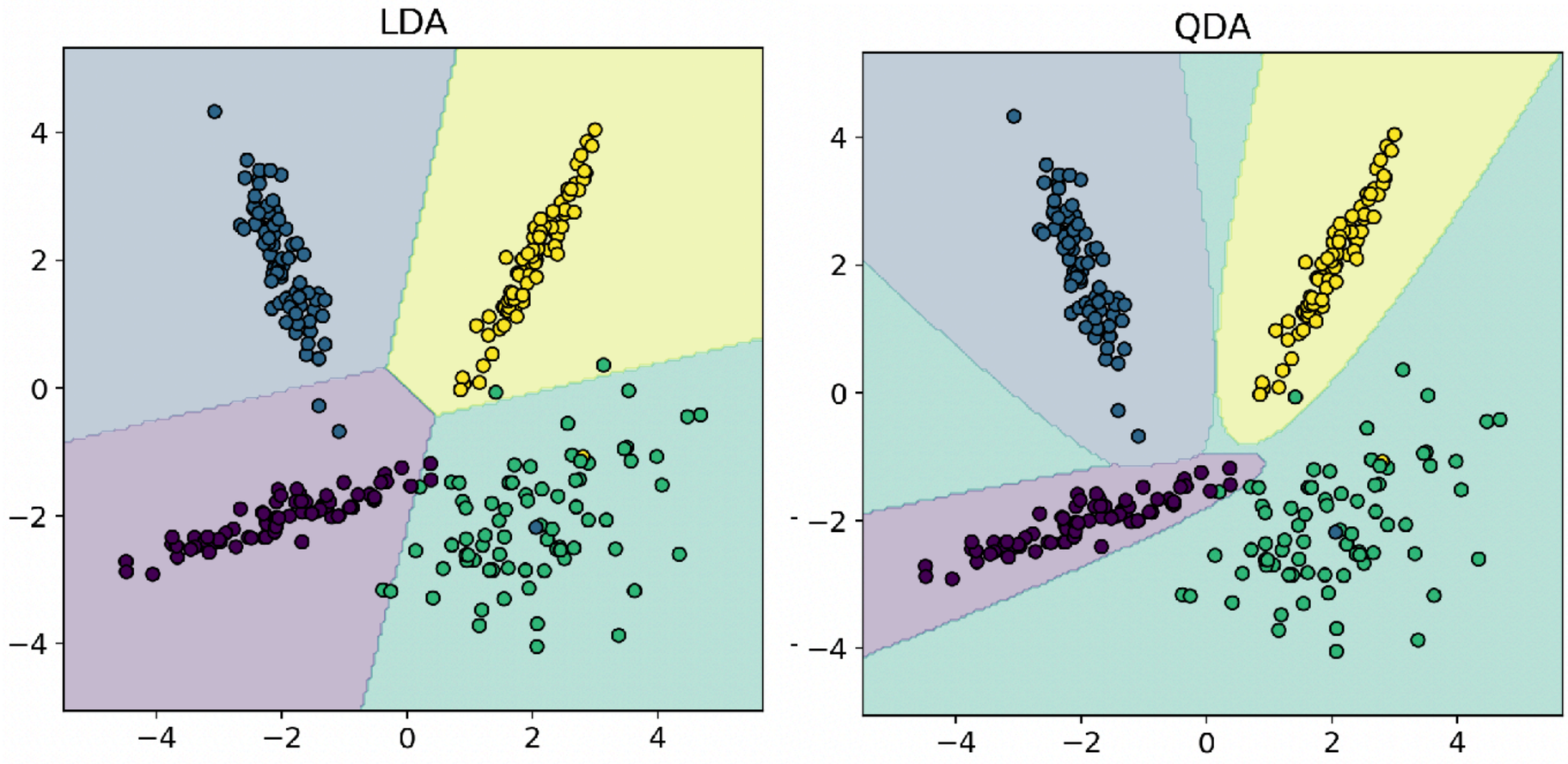
$$\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x}) = \frac{f_{\mathbf{X}|Y=k}(\mathbf{x})\mathbb{P}(Y=k)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}|Y=k}(\mathbf{x})p_k}{f_{\mathbf{X}}(\mathbf{x})}$$

On prédit la classe k si et seulement si:

$$\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x}) = \max_{j \in \{0,\ldots,K-1\}} \mathbb{P}(Y = j|\mathbf{X} = \mathbf{x})$$

$$\Leftrightarrow \log\left(f_{\mathbf{X}|Y=k}(\mathbf{x})p_k\right) = \max_{j \in \{0,\ldots,K-1\}} \log\left(f_{\mathbf{X}|Y=j}(\mathbf{x})p_j\right)$$

Exemple avec 4 classes



Résumé

Hypothèses	Nom du modèle	Séparation
Covariances identiques	LDA	**Linéaire**
Covariances différentes	QDA	**Quadratique**
Covariances diagonales et identiques	Naive Bayes Gaussien	**Linéaire**
Covariances diagonales et différentes	Naive Bayes Gaussien	**Quadratique**

Classification (apprentissage supervisé)

Soit $\mathbf{X}$ un vecteur aléatoire dans $\mathbb{R}^d$ et Y une variable aléatoire dans $\{0, 1, \ldots, K-1\}$.

On observe n paires i.i.d $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. On note $p_k = \mathbb{P}(Y = k)$.

On suppose que l'on peut modéliser les lois conditionelles: $\mathbb{P}(\mathbf{X} = \mathbf{x}|Y = k) = f_{a_k}(\mathbf{x})$

$$\text{Bayes:} \quad \mathbb{P}(Y = k|\mathbf{X} = \mathbf{x}) = \frac{f_{\mathbf{X}|Y=k}(\mathbf{x})\mathbb{P}(Y=k)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}|Y=k}(\mathbf{x})p_k}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{a_k}(\mathbf{x})p_k}{f_{\mathbf{X}}(\mathbf{x})}$$

La fonction de prédiction est donnée par:

$$g : \mathbf{x} \mapsto \underset{j \in \{0,\ldots,K-1\}}{\operatorname{argmax}} \log\left(f_{a_j}(\mathbf{x})p_j\right)$$

Ces modèles sont aussi appelés: *Generative models*

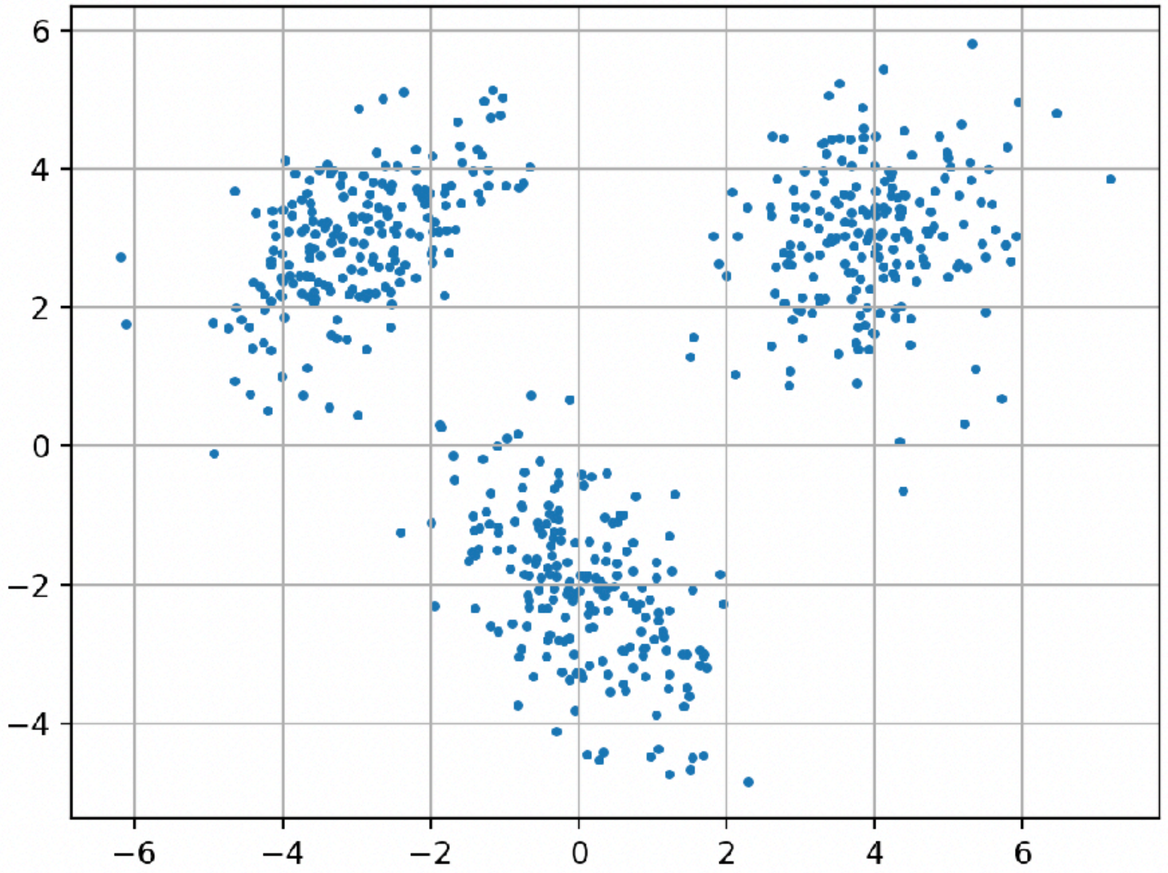
INSEA

III - Modèles probabilistes

Partie 3 - Apprentissage non-supervisé

On souhaite identifier des groupes différents dans des données **sans labels:**

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$$



Comment peut-on modéliser cette distribution ?

Idée: si on connaît le label Z, alors on peut modéliser les lois conditionnelles pour chaque cluster par une Gaussienne:

$$\mathbf{X}|Z = 0 \sim \mathcal{N}(\mu_0, \Sigma_0)$$

$$\mathbf{X}|Z = 1 \sim \mathcal{N}(\mu_1, \Sigma_1)$$

$$\mathbf{X}|Z = 2 \sim \mathcal{N}(\mu_2, \Sigma_2)$$

Soit Z une variable discrète dans $\{1, \ldots, K\}$ telle que $\mathbb{P}(Y = k) = \pi_k$

Avec $\sum_{k=1}^{K} \pi_k = 1$

Et soit $\mathbf{X}$ un vecteur aléatoire tel que $\mathbb{P}(\mathbf{X}|Z = k) = \mathcal{N}(\mu_k, \Sigma_k)$

Alors, on dit que $\mathbf{X}$ est un mélange de Gaussiennes avec K composantes et on écrit:

$$\mathbf{X} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

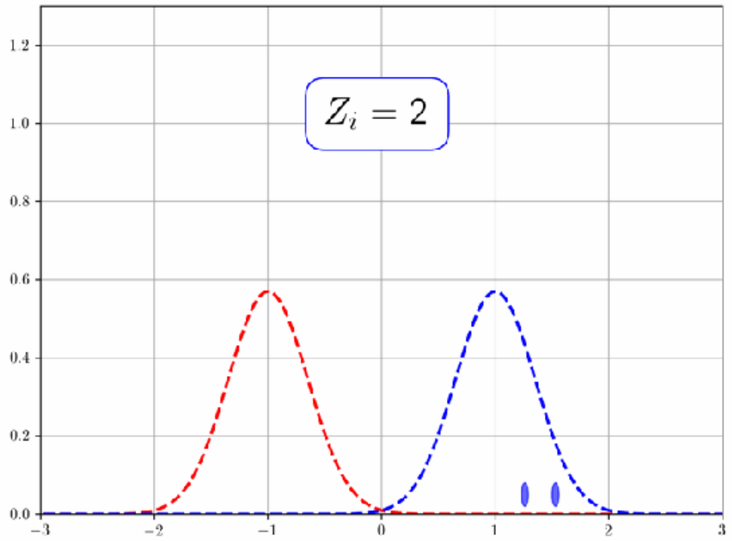
La variable Z, en général non observée est dite: variable **latente**

Soit Z une variable discrète dans $\{1, 2\}$ avec $\pi_1 = \pi_2 = \frac{1}{2}$

On considère deux Gaussiennes univariées $\mathcal{N}(-1, 0.5^2)$ et $\mathcal{N}(1, 0.5^2)$

On génère X_i suivant la valeur de Z_i:

$$\mathbf{X}|Z = 1 \sim \mathcal{N}(-1, 0.5^2) \qquad \mathbf{X}|Z = 2 \sim \mathcal{N}(1, 0.5^2)$$



INSEA

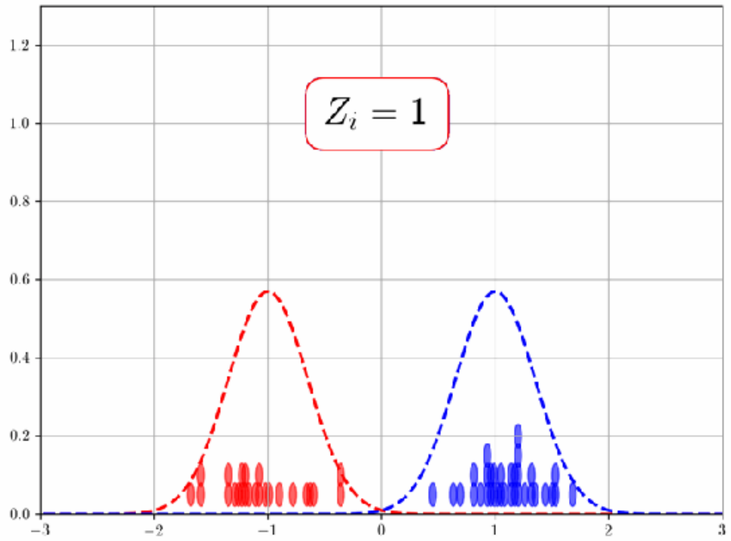
Exemple en 1D

Soit Z une variable discrète dans $\{1, 2\}$ avec $\pi_1 = \pi_2 = \frac{1}{2}$

On considère deux Gaussiennes univariées $\mathcal{N}(-1, 0.5^2)$ et $\mathcal{N}(1, 0.5^2)$

On génère X_i suivant la valeur de Z_i:

$$\mathbf{X}|Z = 1 \sim \mathcal{N}(-1, 0.5^2) \qquad \mathbf{X}|Z = 2 \sim \mathcal{N}(1, 0.5^2)$$



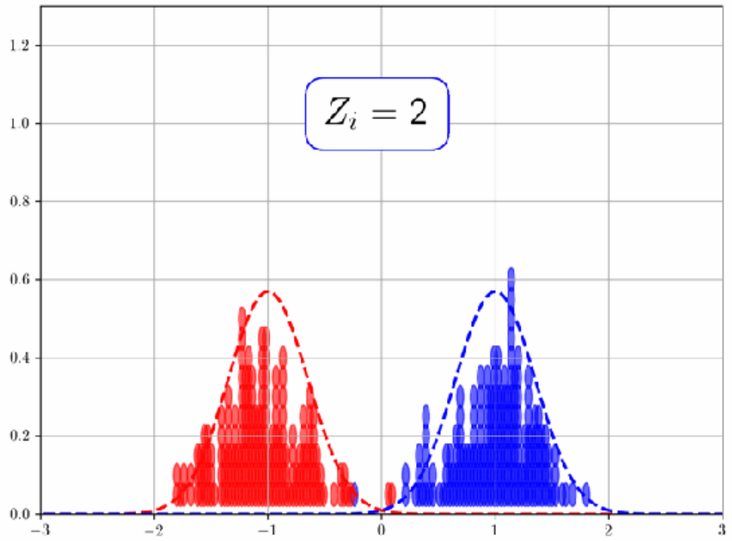
INSEA

Soit Z une variable discrète dans $\{1, 2\}$ avec $\pi_1 = \pi_2 = \frac{1}{2}$

On considère deux Gaussiennes univariées $\mathcal{N}(-1, 0.5^2)$ et $\mathcal{N}(1, 0.5^2)$

On génère X_i suivant la valeur de Z_i:

$$\mathbf{X}|Z = 1 \sim \mathcal{N}(-1, 0.5^2) \qquad \mathbf{X}|Z = 2 \sim \mathcal{N}(1, 0.5^2)$$

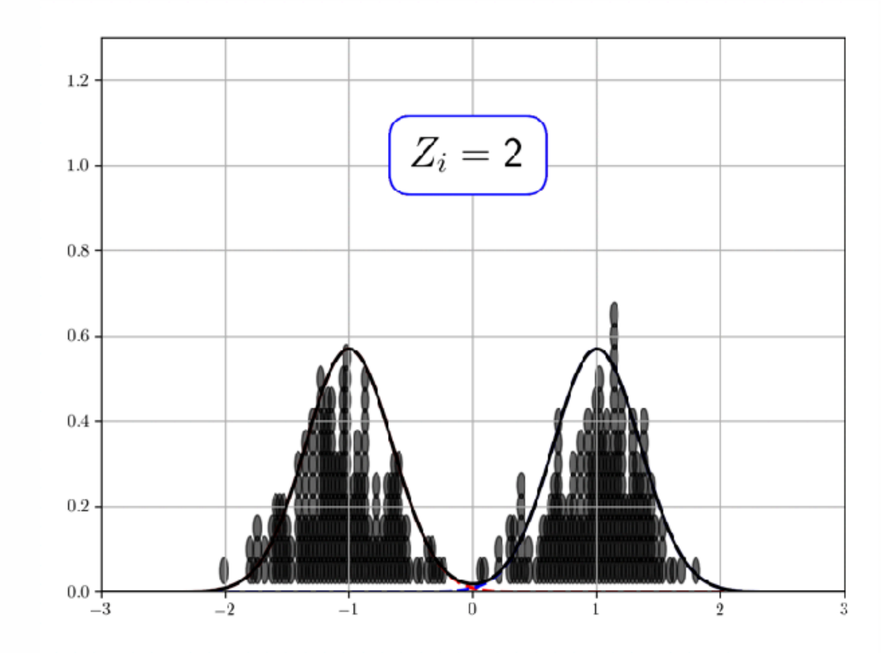


Soit Z une variable discrète dans $\{1, 2\}$ avec $\pi_1 = \pi_2 = \frac{1}{2}$

On considère deux Gaussiennes univariées $\mathcal{N}(-1, 0.5^2)$ et $\mathcal{N}(1, 0.5^2)$

On génère X_i suivant la valeur de Z_i:

$$\mathbf{X}|Z = 1 \sim \mathcal{N}(-1, 0.5^2) \qquad \mathbf{X}|Z = 2 \sim \mathcal{N}(1, 0.5^2)$$



Quelle est la loi de X ?

INSEA

Soit Z une variable discrète dans $\{1, 2\}$ avec $\pi_1 = \pi_2 = \frac{1}{2}$

$$\mathbf{X}|Z = 1 \sim \mathcal{N}(-1, 0.5^2) \qquad \mathbf{X}|Z = 2 \sim \mathcal{N}(1, 0.5^2)$$

Quelle est la loi de X ?

Avec la formule des probabilités totales:

$$\mathbb{P}(\mathbf{X} \leq x) = \mathbb{P}(\mathbf{X} \leq x|Z = 1)\mathbb{P}(Z = 1) + \mathbb{P}(\mathbf{X} \leq x|Z = 2)\mathbb{P}(Z = 2)$$

$$= \mathbb{P}(\mathbf{X} \leq x|Z = 1)\pi_1 + \mathbb{P}(\mathbf{X} \leq x|Z = 2)\pi_2$$

$$= \pi_1 \int_{-\infty}^{x} f_{\mathcal{N}(-1,0.5^2)}(u)\mathrm{d}u + \pi_2 \int_{-\infty}^{x} f_{\mathcal{N}(1,0.5^2)}(u)\mathrm{d}u$$

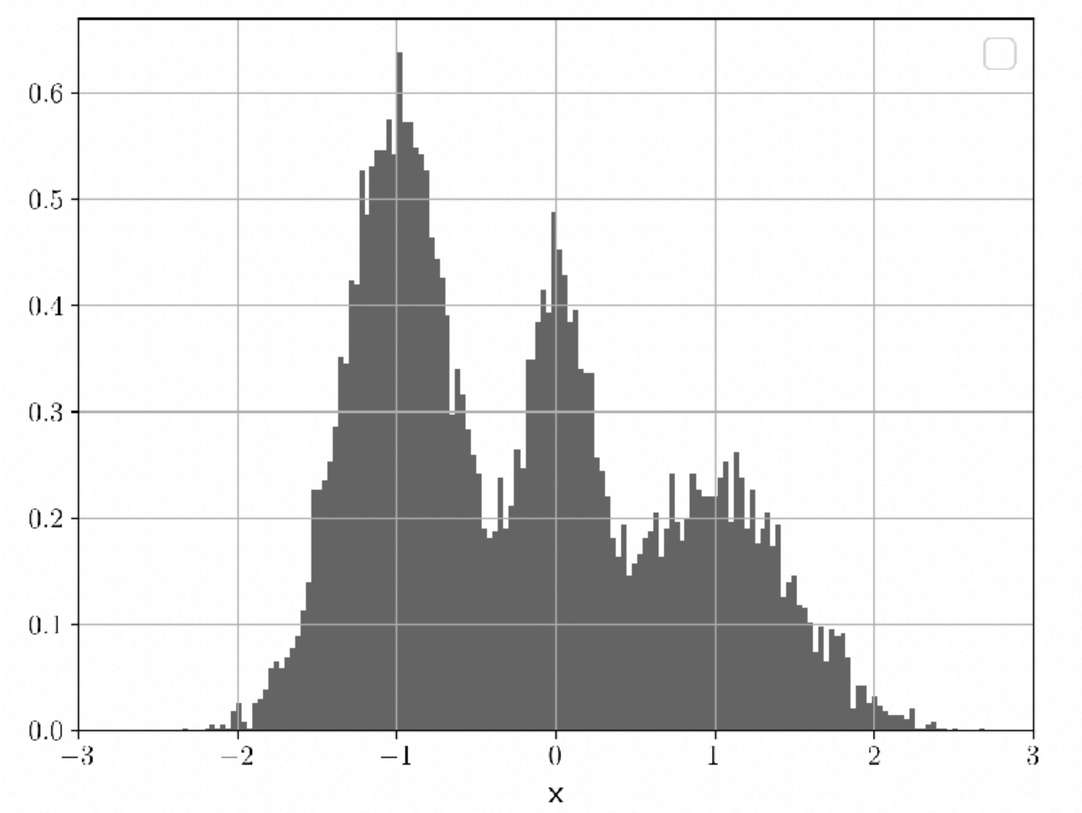
$$= \int_{-\infty}^{x} \left(\pi_1 f_{\mathcal{N}(-1,0.5^2)}(u) + \pi_2 f_{\mathcal{N}(1,0.5^2)}(u)\right) \mathrm{d}u$$

$\mathbf{X}$ est un mélange de Gaussiennes:

$$\mathbf{X} \sim \pi_1 \mathcal{N}(-1, 0.5^2) + \pi_2 \mathcal{N}(1, 0.5^2)$$

Exemple en 1D

On observe une variable (1D) dont l'histogramme est:



Quel modèle paramétrique est adapté pour ces données ?

Visuellement, un GMM(3) devrait bien modéliser ces données

On suppose donc $X_1, \ldots, X_n \sim \text{GMM}(3)$

Formellement: $X \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2)$ avec K = 3

Comment identifier les paramètres π_k, μ_k, σ_k^2 ?

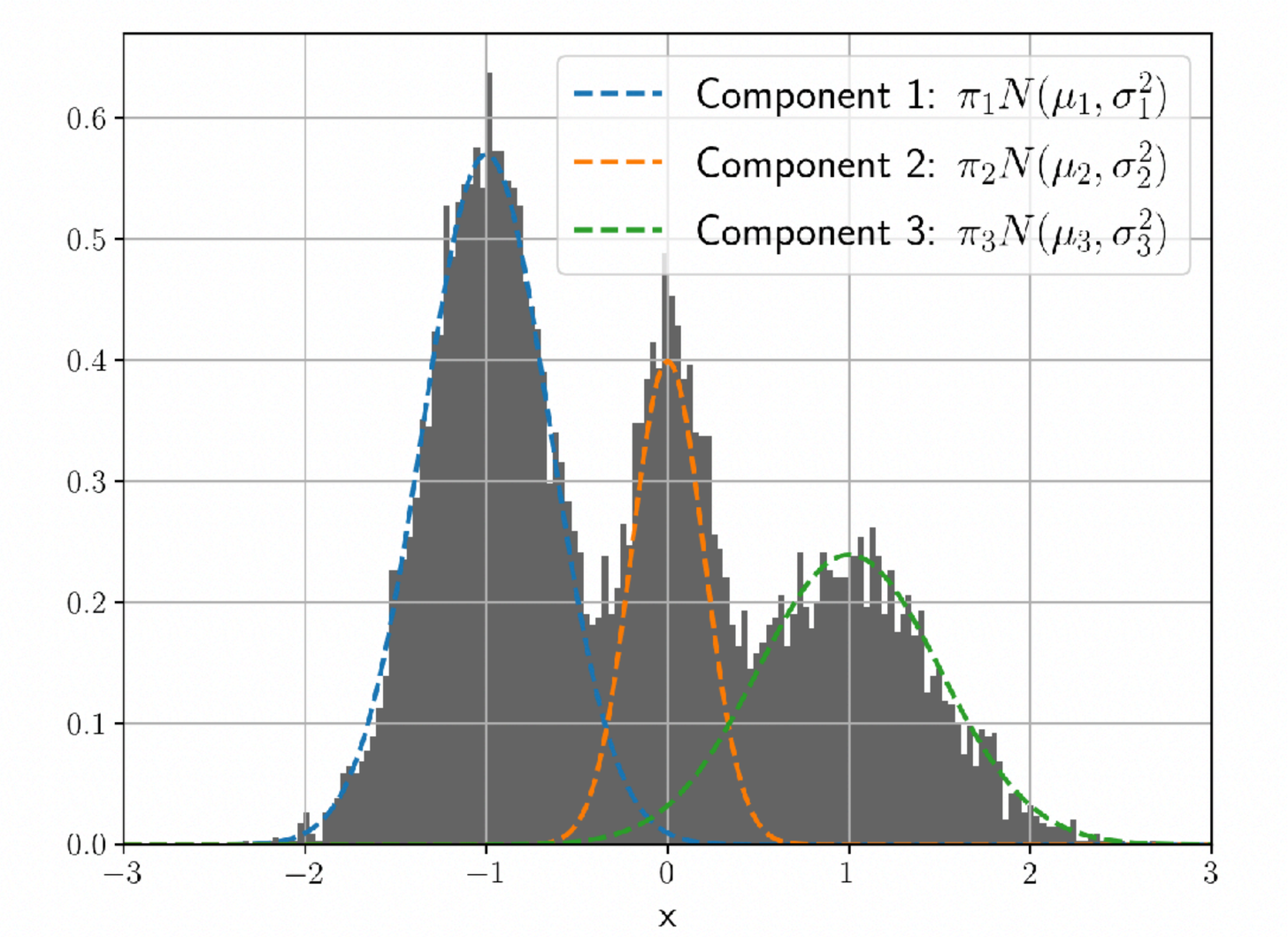
On maximise sa log-vraisemblance:

$$\max_{\substack{\pi,\mu,\sigma^2 \\ \sum_{k=1}^{K} \pi_k = 1}} \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f_{\mathcal{N}(\mu_k, \sigma_k^2)}(X_i) \right)$$

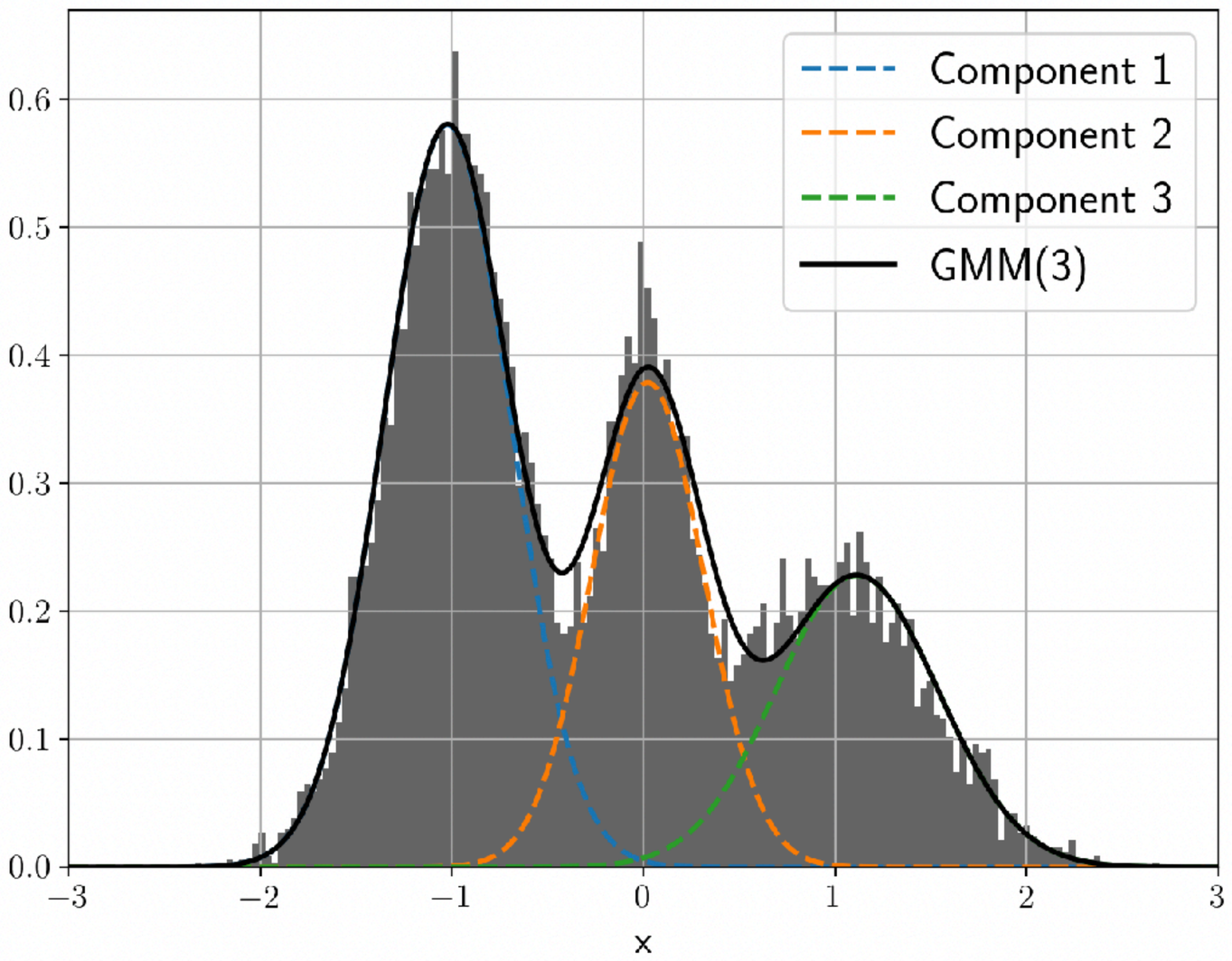
Numériquement, on obtient:

$$\pi = [0.48, 0.28, 0.24]$$
$$\mu = [-1.02, 0.02, 1.1]$$
$$\sigma^2 = [0.1, 0.08, 0.17]$$

Composantes du modèle GMM:

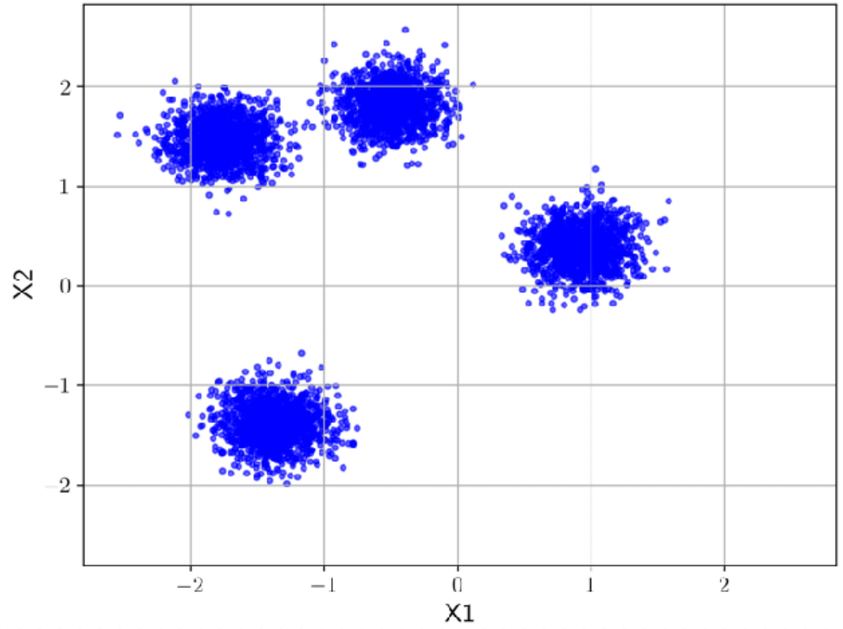


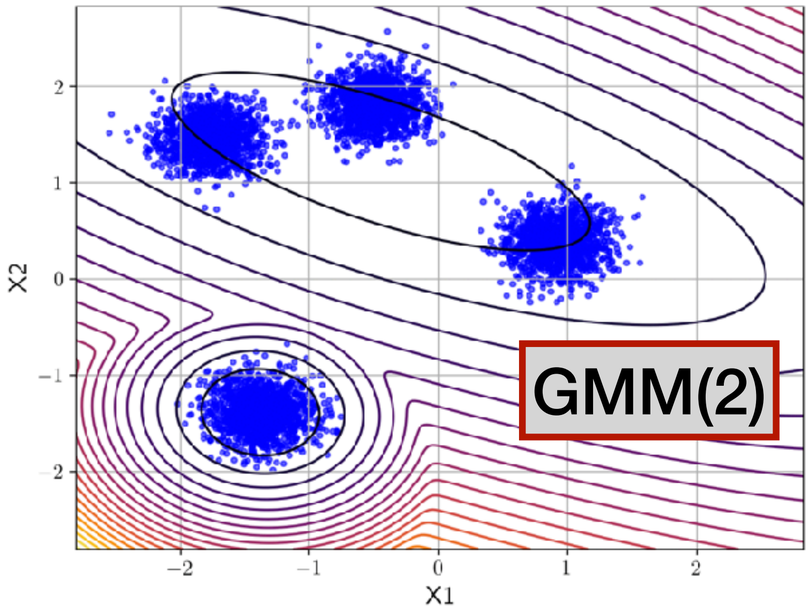
Densité du modèle GMM (somme des composantes)

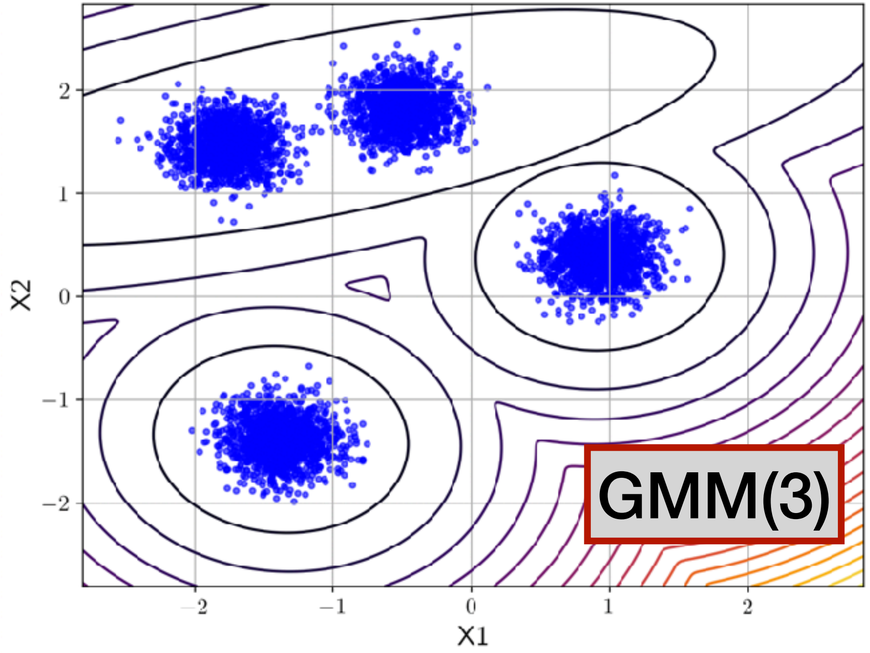


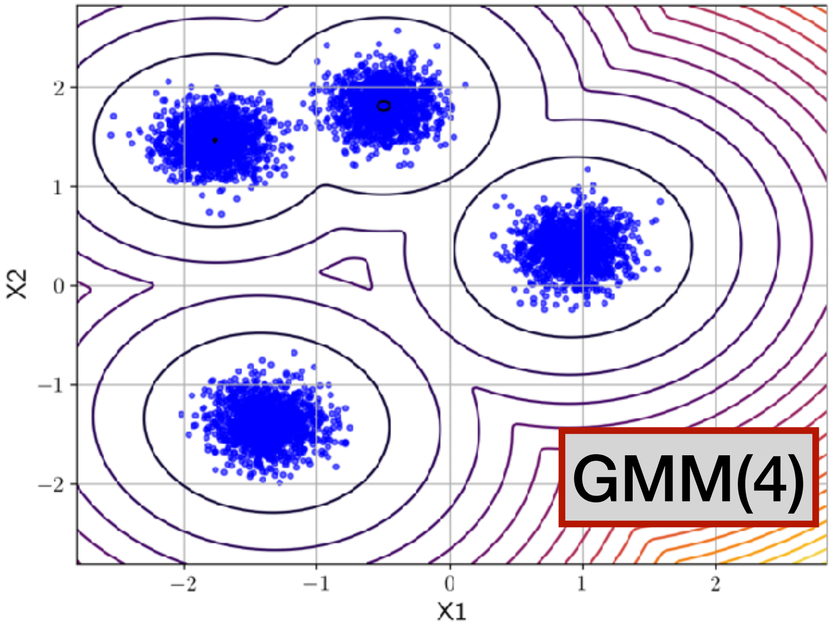
2. Mélange de Gaussiennes (GMM)

Exemple en 2D









INSEA

$$\text{Soit } \mathbf{X}_1, \ldots, \mathbf{X}_n \text{ des observations i.i.d } \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

La vraisemblance n'est pas facile à maximiser directement:

$$\log f(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k) \right)$$

On l'augmente en utilisant la **variable latente Z**

On peut montrer que: $$\log f(\mathbf{X}, Z|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}(Z_i = k)\left(\log \pi_k + \log \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k)\right)$$

Mais on n'observe pas les variables Z_i, on "marginalise" en intégrant la variable Z:

$$\mathbb{E}_{Z|\mathbf{X}}[\log f(X, Z|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{Z|\mathbf{X}}[\mathbb{1}(Z_i = k)\left(\log \pi_k + \log \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k)\right)]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{Z|\mathbf{X}}[\mathbb{1}(Z_i = k)]\left(\log \pi_k + \log \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k)\right)$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}\left(\log \pi_k + \log \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k)\right)$$

$$\gamma_{ik} = \mathbb{P}(Z_i = k|\mathbf{X} = \mathbf{x}_i) = \frac{f_{\mathbf{X}|Z=k}(\mathbf{x}_i)\mathbb{P}(Z = k)}{f_{\mathbf{X}}(\mathbf{x}_i)} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i|\mu_j, \Sigma_j)}$$

INSEA

Si on suppose que ces probabilités sont connues, alors on peut facilement maximiser la nouvelle vraisemblance pondérée par les γ_{ik}:

$$\varphi(\boldsymbol{pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \stackrel{\text{def}}{=} \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} \left(\log \pi_k + \log \mathcal{N}(x_i | \mu_k, \Sigma_k) \right)$$

On obtient: $\hat{\pi_k} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ik}$ $\hat{\mu_k} = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}}$ $\hat{\Sigma_k} = \frac{\sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^\top}{\sum_{i=1}^{N} \gamma_{ik}}$ (A)

Mais les γ_{ik} dépendent des paramètres: $\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j)}$ (B)

Idée: un algorithme alternatif

0. Initialiser $\pi^{(0)}, \mu^{(0)}, \Sigma^{(0)}$

Pour chaque itération t:

1. Calculer $\gamma^{(t)}$ avec (B) avec et les $\pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}$ *Expectation* step
2. Mettre à jour $\pi^{(t+1)}, \mu^{(t+1)}, \Sigma^{(t+1)}$ avec (A) et $\gamma^{(t)}$ *Maximization* step

L'algorithme E-M

0. Initialiser $\pi^{(0)}, \mu^{(0)}, \Sigma^{(0)}$

Pour chaque itération t:

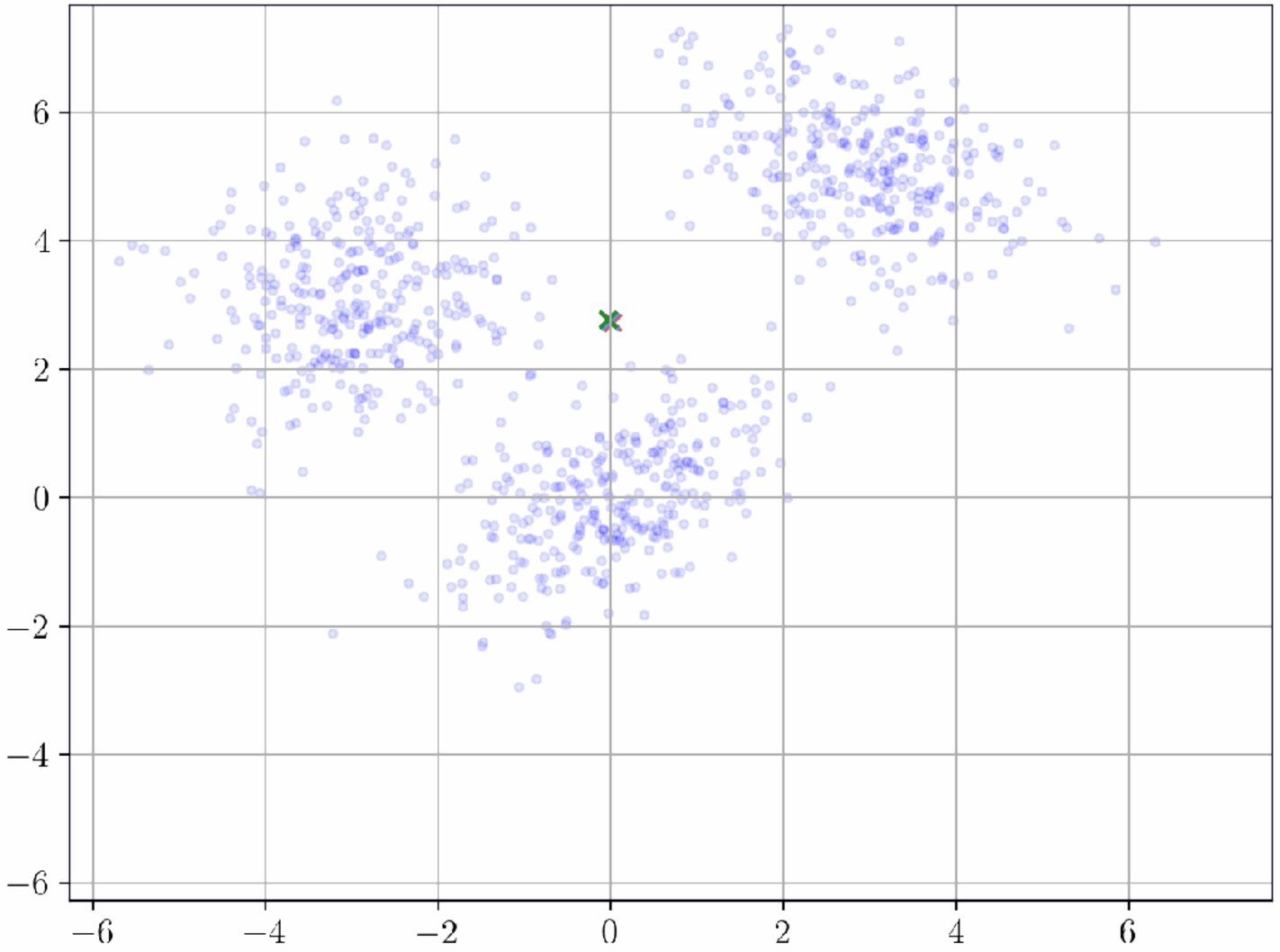
1. Calculer $\gamma^{(t)}$ avec (B) avec et les $\pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}$ — *Expectation* step
2. Mettre à jour $\pi^{(t+1)}, \mu^{(t+1)}, \Sigma^{(t+1)}$ avec (A) et $\gamma^{(t)}$ — *Maximization* step

Remarques

1. La vraisemblance du modèle GMM($\pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}$) croit en fonction de t.
2. La vraisemblance du modèle n'est pas convexe: pas de garantie d'un maximum global
3. La vraisemblance du modèle n'est pas convexe: le résultat E-M dépend de l'initialisation

INSEA

Algorithme E-M in action



INSEA

On souhaite identifier des groupes différents dans des données **sans labels:**

On suppose qu'il y a K labels possibles et on modélise les données par un GMM(K):

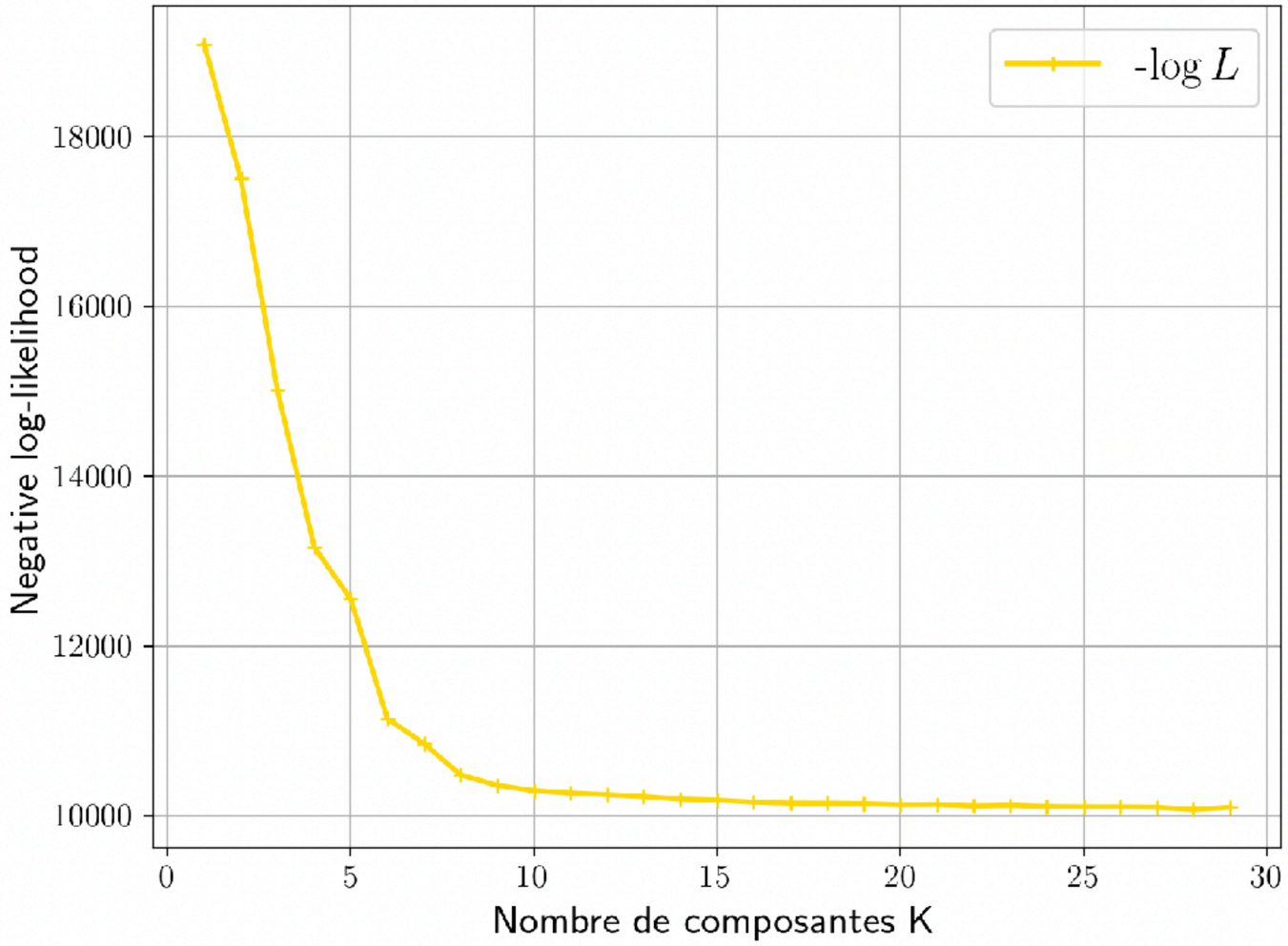
$$\text{Soit } \mathbf{X} \sim \sum_{k=1}^{K} \hat{\pi}_k \mathcal{N}(\hat{\mu}_k, \widehat{\Sigma_k})$$

Où les paramètres ont été estimés sur n observations i.i.d $\mathbf{x}_1, \ldots, \mathbf{x}_n$

Comment prédire la classe d'un nouveau point $\mathbf{x}_{n+1}$?

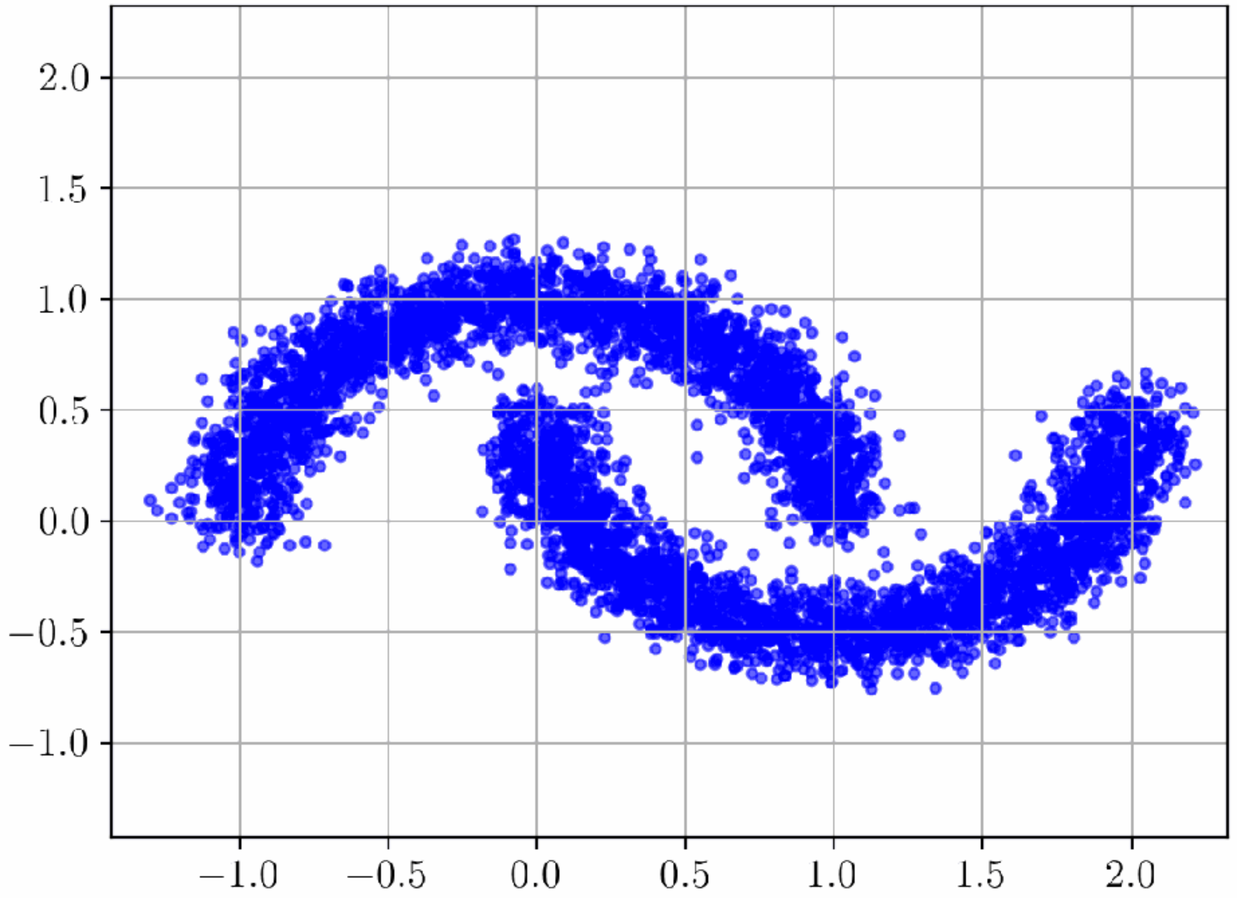
Choix du nombre de composantes

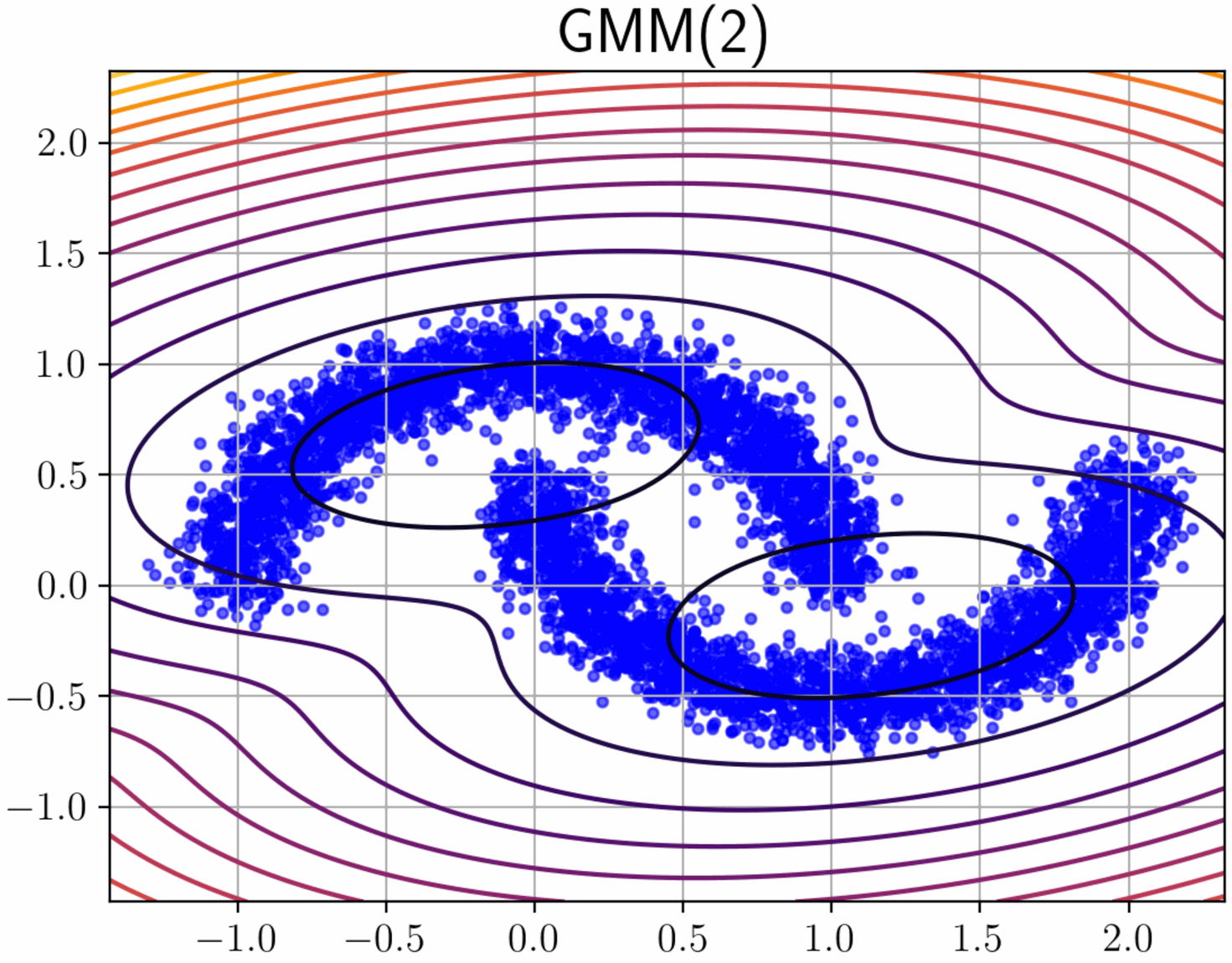
Quel est l'effet du nombre de composantes sur la vraisemblance du modèle ?

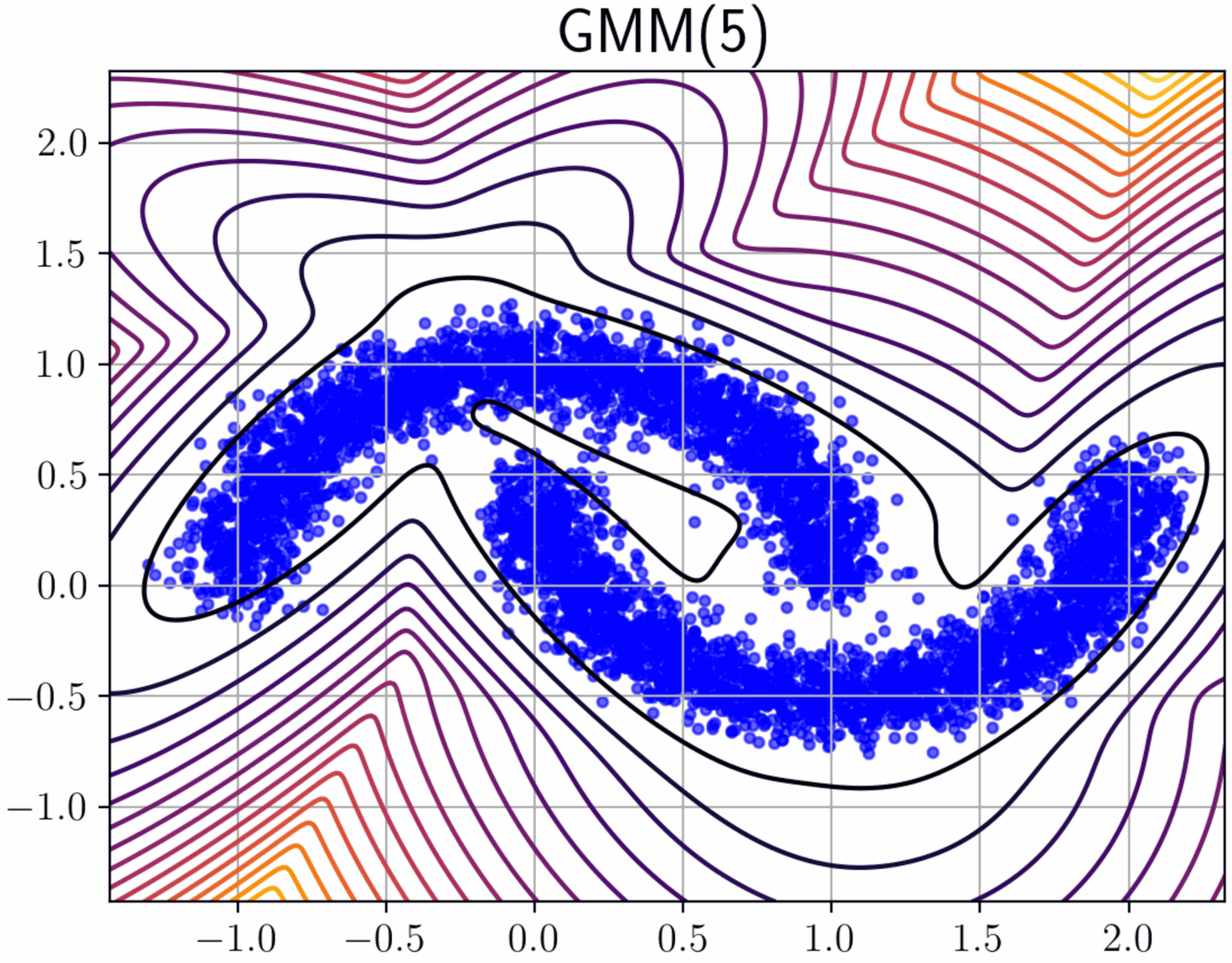


1. Plus K est grand, plus le modèle est riche
2. .. mais on risque d'overfit les données

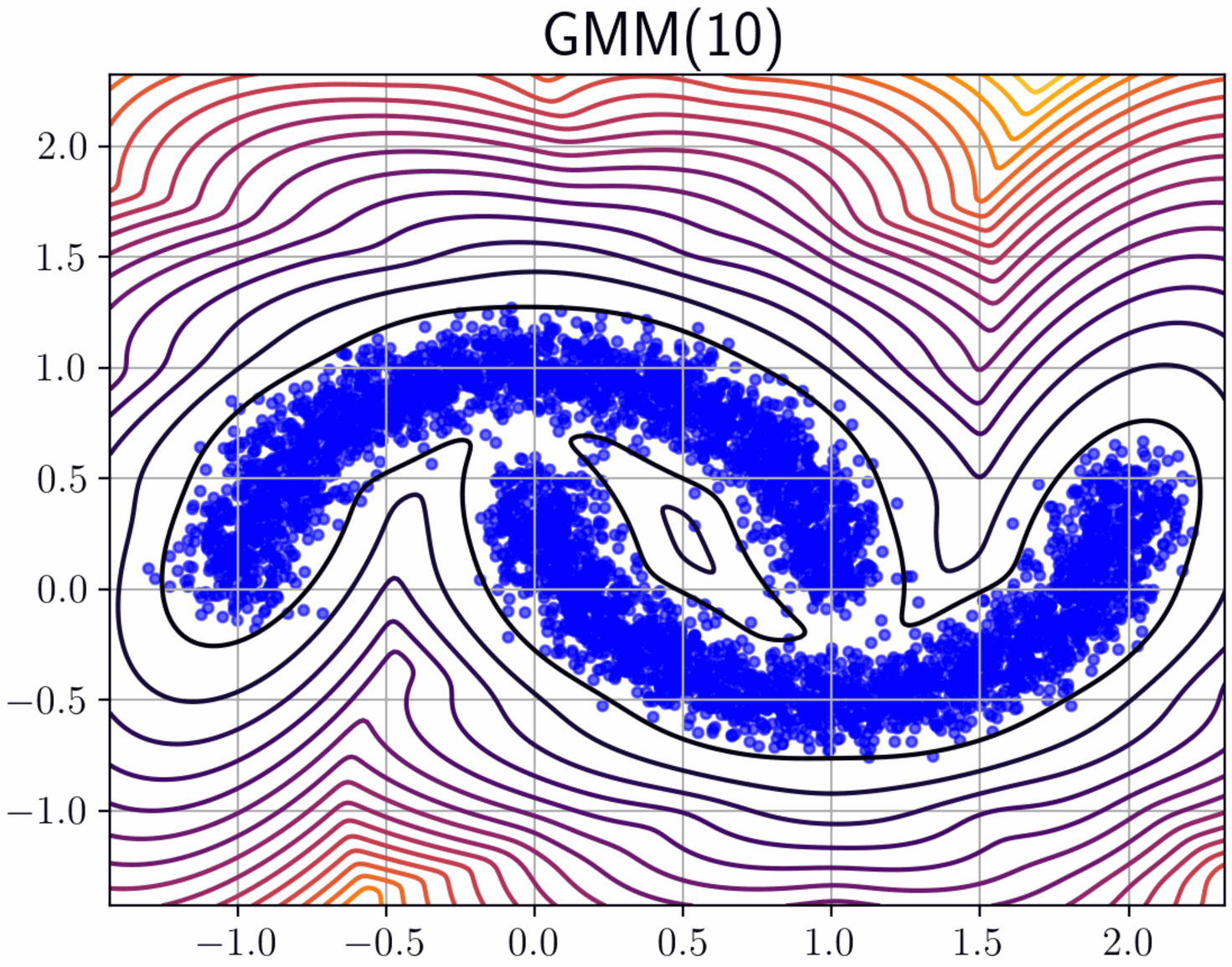
Quel est l'effet du nombre de composantes sur le modèle ?




GMM(2)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


GMM(5)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

Exemple en 2D

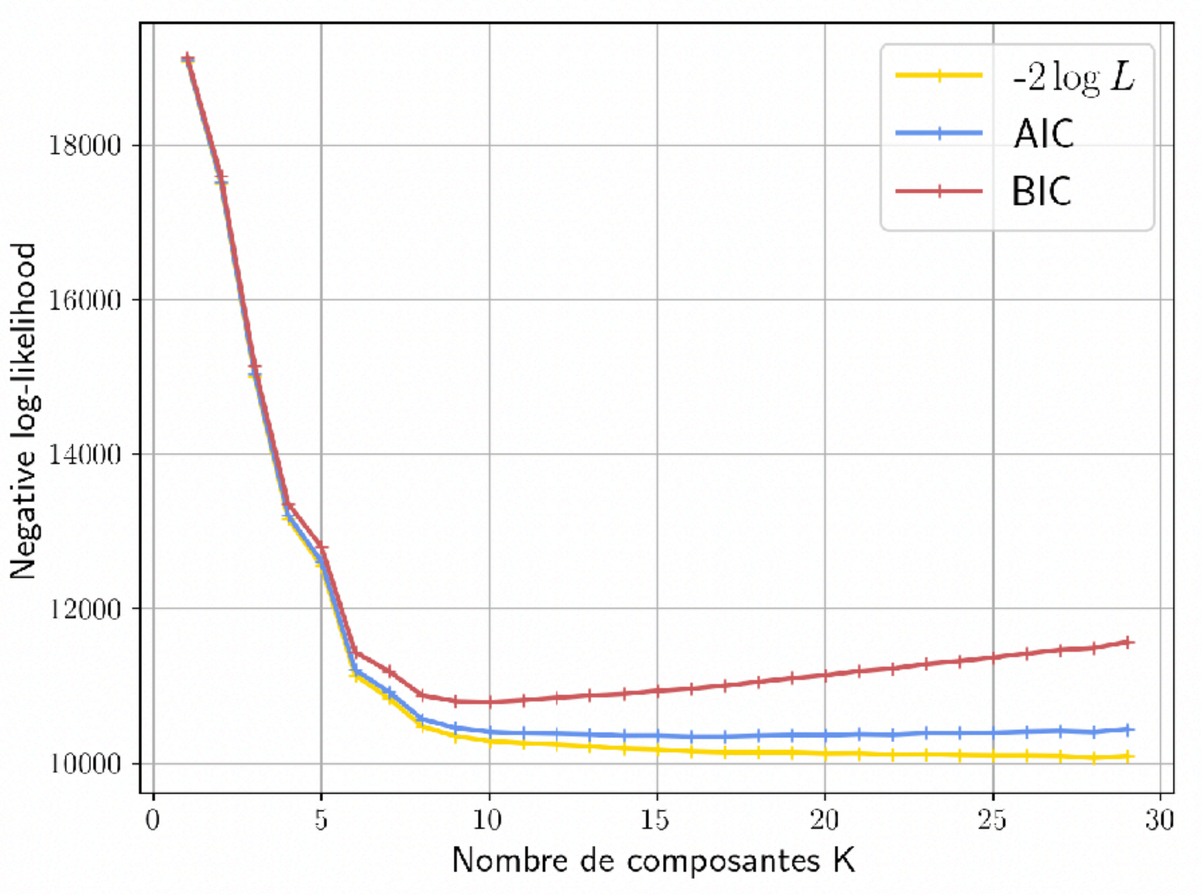


Choix du nombre de composantes

Principe du *rasoir d'Ockham: "The simplest explanation is usually the best one"*

Un modèle simple = modèle avec peu de paramètres à estimer

On minimise un trade-off entre la vraisemblance et la complexité



"Akaike Information Criterion"

$$\mathrm{AIC} = -2\log(L) + 2K$$

"Bayesian Information Criterion"

$$\mathrm{BIC} = -2\log(L) + \log(n)K$$

Clustering (apprentissage non supervisé)

Soit $\mathbf{X}$ un vecteur aléatoire dans $\mathbb{R}^d$ et Y une variable aléatoire dans $\{0, 1, \ldots, K-1\}$.

On observe $\mathbf{x}_1, \ldots, \mathbf{x}_n$ supposés i.i.d. Les Y ne sont pas observés. K peut éventuellement être connu à l'avance.

On note $\pi_k = \mathbb{P}(Y = k)$. On suppose que l'on peut modéliser les lois conditionelles: $\mathbb{P}(\mathbf{X} = \mathbf{x}|Y = k) = f_{a_k}(\mathbf{x})$

Alors, on considérant M composantes, par la formule des probabilités totales:

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \sum_{k=0}^{M-1} \pi_k f_{a_k}(\mathbf{x})$$

$\mathbf{X}$ est un modèle de mélange de la loi à densité f_a

Ses paramètres $(a_0, \ldots, a_{M-1})$ peuvent être estimés avec l'algorithme E-M.

La fonction de prédiction est donnée par:

$$g : \mathbf{x} \mapsto \underset{j \in \{0,\ldots,M-1\}}{\operatorname{argmax}} \log\left(f_{a_j}(\mathbf{x}) p_j\right)$$