Mathématiques pour l'intelligence artificielle

UE MIA, M2 IAAA, AMU, 2022-2023

Thomas Schatz

Mardi 20 septembre 2022

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage

9. Putting it all together

Probabilités

1. Revue du calcul probabiliste, sur la base des transparents de Ding & Khani, Stanford CS229, April 2022 (revus , réordonnés et augmentés)

2. Quelques points plus avancés (intro théorie de la mesure)

Moments

Moment d’ordre k

$$E[X^k]$$

Moment centré d’ordre k

$$\mu = E[X]$$
$$E[(X - \mu)^k]$$

Multivariate Gaussian

The multivariate Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^n$:

$$p(x; \mu, \Sigma) = \frac{1}{det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

The univariate Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$, $X \in \mathbb{R}$ is just the special case of the multivariate Gaussian when $n = 1$.

$$p(x; \mu, \sigma^2) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Notice that if $\Sigma \in \mathbb{R}^{1\times 1}$, then $\Sigma = Var[X_1] = \sigma^2$, and so

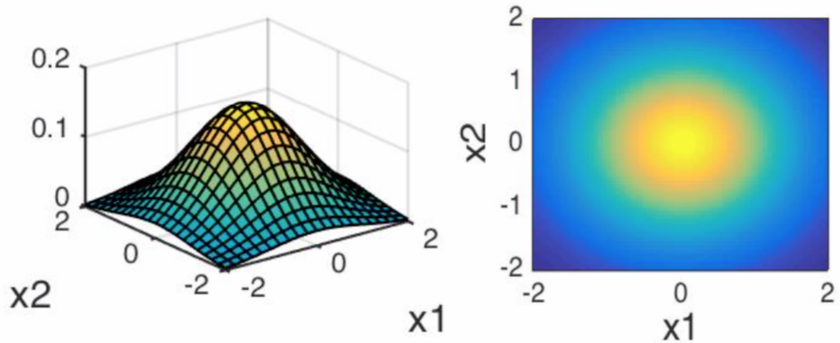
- $\Sigma^{-1} = \frac{1}{\sigma^2}$
- $det(\Sigma)^{\frac{1}{2}} = \sigma$

Visualizations of MV Gaussians

Effect of changing variance

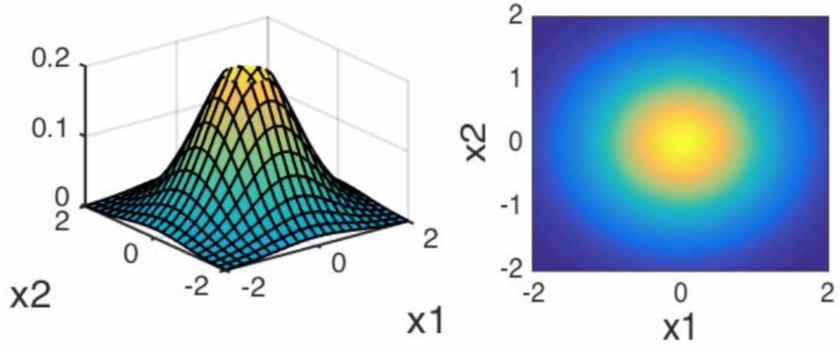
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



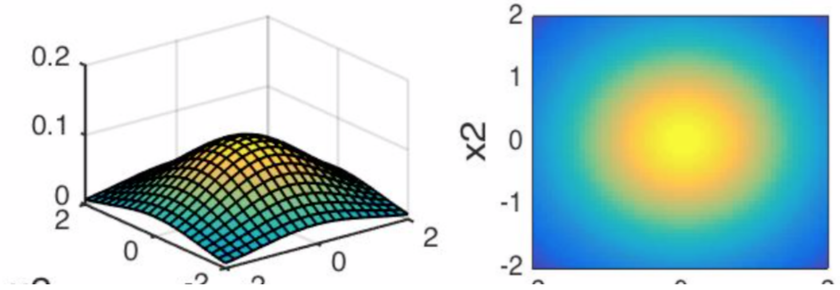
$$\Sigma = \begin{matrix} 0.7 & 0 \\ 0 & 0.7 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 1.5 & 0 \\ 0 & 1.5 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$

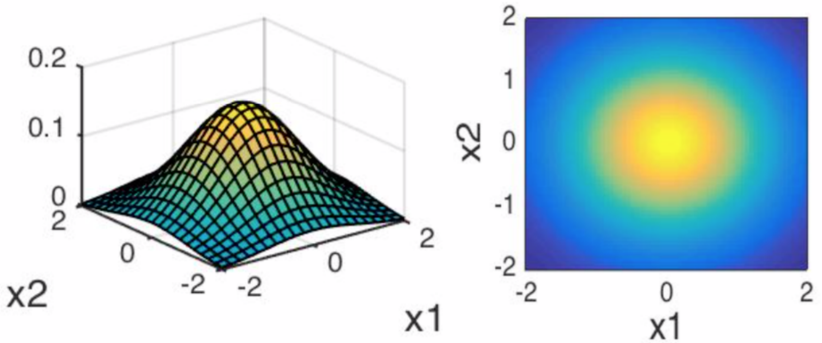


Visualizations of MV Gaussians

If $Var[X_1] \neq Var[X_2]$:

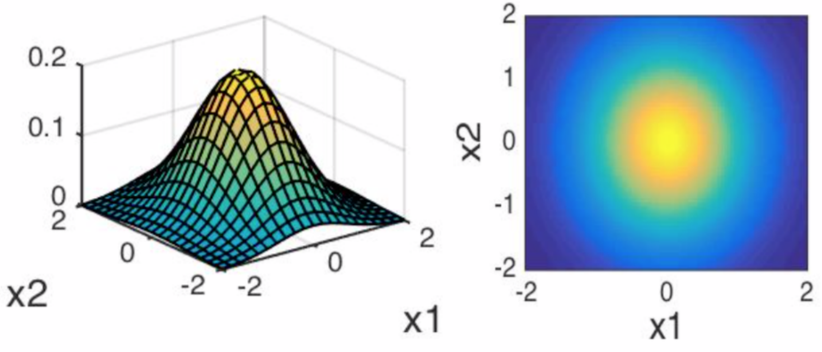
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



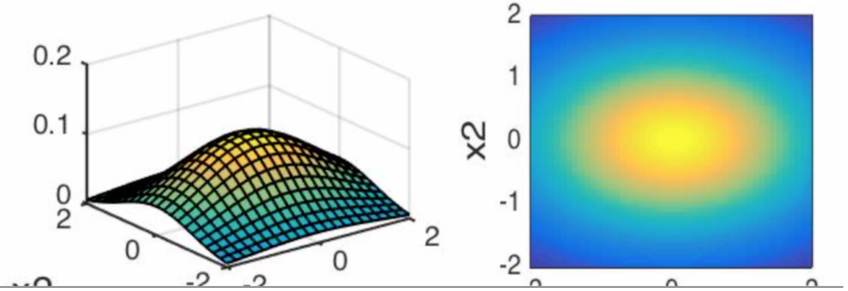
$$\Sigma = \begin{matrix} 0.6 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 2 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$

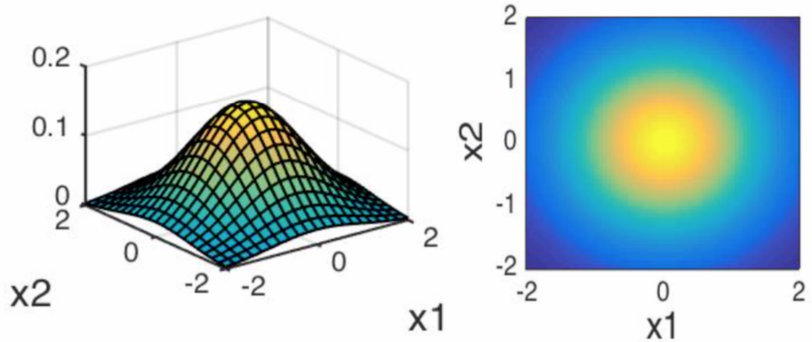


Visualizations of MV Gaussians

If X_1 and X_2 are positively correlated:

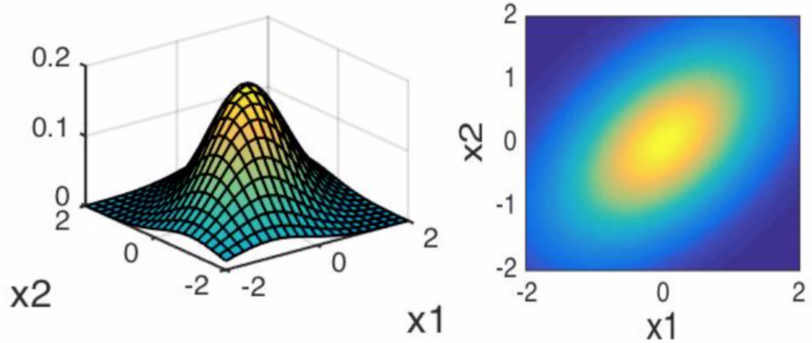
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



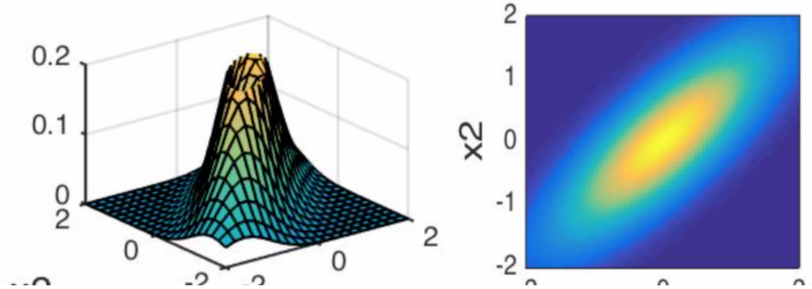
$$\Sigma = \begin{matrix} 1 & 0.5 \\ 0.5 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 1 & 0.8 \\ 0.8 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$

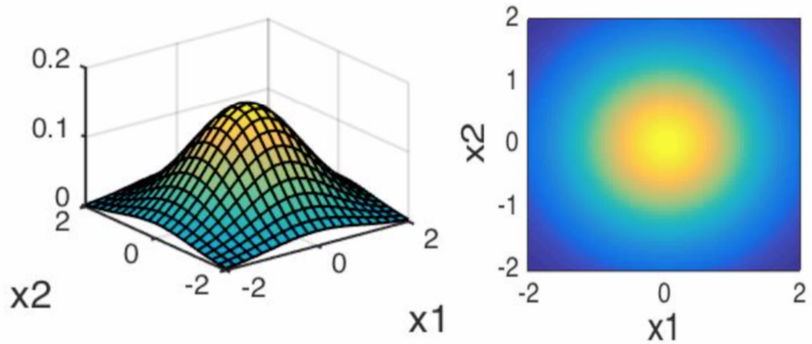


Visualizations of MV Gaussians

If X_1 and X_2 are negatively correlated:

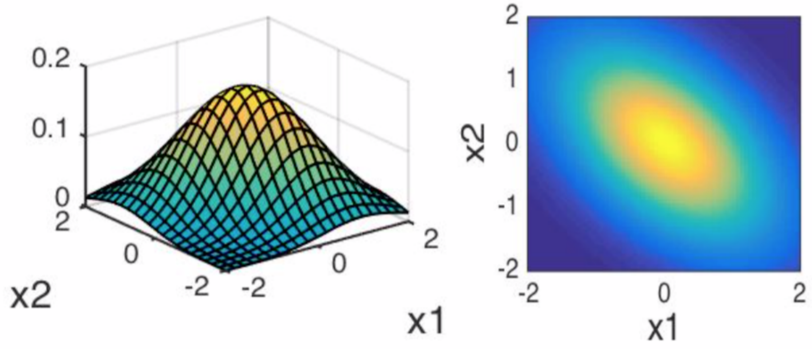
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



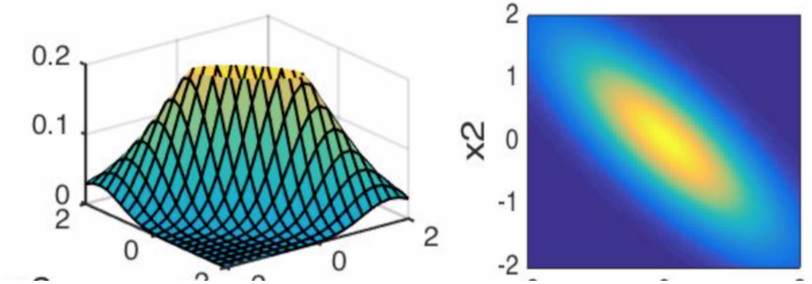
$$\Sigma = \begin{matrix} 1 & -0.5 \\ -0.5 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 1 & -0.8 \\ -0.8 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



Some Nice Properties of MV Gaussians

- Marginals and conditionals of a joint Gaussian are Gaussian
- A d-dimensional Gaussian $X \in \mathcal{N}(\mu, \Sigma = diag(\sigma_1^2, ..., \sigma_n^2))$ is equivalent to a collection of d **independent** Gaussians $X_i \in \mathcal{N}(\mu_i, \sigma_i^2)$. This results in isocontours aligned with the coordinate axes.
- In general, the isocontours of a MV Gaussian are n-dimensional ellipsoids with principal axes in the directions of the eigenvectors of covariance matrix Σ (remember, Σ is PSD, so all n eigenvectors are non-negative). The axes' relative lengths depend on the eigenvalues of Σ.

Multivariate Gaussian

Définition générale

$$\mathbf{X} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma}) \quad \iff \quad \text{there exist } \mu \in \mathbb{R}^k, \boldsymbol{A} \in \mathbb{R}^{k \times \ell} \text{ such that } \mathbf{X} = \boldsymbol{A}\mathbf{Z} + \mu \text{ for } Z_n \sim \mathcal{N}(0,1), \text{i.i.d.}$$

Distributions conditionnelles

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$

$$p(\mathbf{x}_1 \mid \mathbf{x}_2 = \mathbf{a}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}), with$$

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{+}(\mathbf{a} - \boldsymbol{\mu}_2)$$

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{+}\boldsymbol{\Sigma}_{21}$$

Distributions marginales ?

Probabilités

1. Revue du calcul probabiliste, sur la base des transparents de Ding & Khani, Stanford CS229, April 2022 (revus , réordonnés et augmentés)

2. Quelques points plus avancés (intro théorie de la mesure)

Théorie de la mesure ?

- Quel intérêt : traitement rigoureux, unifié et général de la théorie des probabilité

  - Unifié : e.g. PMF et PDF deviennent deux cas particuliers d’un concept plus général, pas de différence à faire entre somme discrètes et intégrales continues

  - Général : variables mixtes (ni continues ni discrètes), variables plus compliquées (e.g. signal), espace de probabilité aussi complexes que nécessaires, théorèmes de convergence plus simples et plus forts qu’avec l’intégrale de Riemann (utile en statistique)

Théorie de la mesure ?

- Difficulté :
  - Plus abstrait
  - Le problème de la mesurabilité des fonctions

Elements of Probability

Sample Space Ω

$\{HH, HT, TH, TT\}$

Event $A \subseteq \Omega$

$\{HH, HT\}, \Omega$

Event Space $\mathcal{F}$

Probability Measure $P : \mathcal{F} \to \mathbb{R}$

$P(A) \geq 0 \quad \forall A \in \mathcal{F}$

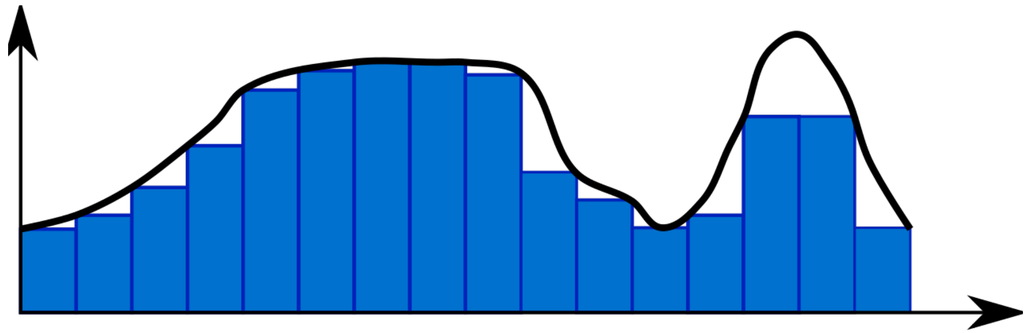
$P(\Omega) = 1$

countable

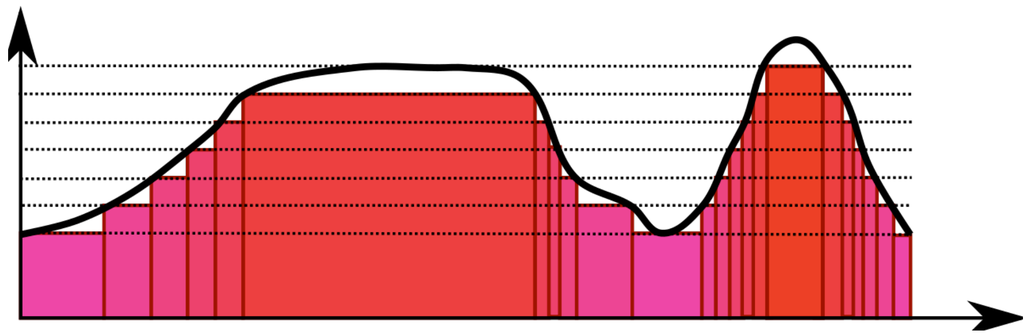
If $A_1, A_2, \ldots$ disjoint set of events ($A_i \cap A_j = \emptyset$ when $i \neq j$), then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

Espérance d'une variable aléatoire: intégrale de Lebesgue



$$\int_a^b f(x)dx = \lim_{n\to+\infty} \sum_{i=0}^{n-1} \left(\frac{b-a}{n}\right) f\left(a + i\frac{b-a}{n}\right)$$



$$\int_X f d\mu = \sup_{g\leq f, g \text{ étagée}} g$$

$$\int_X f d\mu = \sum_{k=1}^{K} a_k \mu(A_k)$$

$$\exists K \in \mathbf{N},\ \exists (A_1, A_2, \ldots, A_K) \in P(X),\ \exists (a_1, a_2, \ldots, a_K) \in \mathbf{R}^K,\ g = \sum_{k=1}^{K} a_k \mathbf{1}_{A_k}$$

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage

9. Putting it all together

Statistique

- Introduction générale
- Statistique non asymptotique partie 1: biais, variance et risque
- Statistique asymptotique
- Statistique non asymptotique partie 2: inégalités de concentration

Introduction

- Problème de l’induction, inséparable de la démarche scientifique dans son ensemble
- Bayésien vs fréquentiste
- Probabilistic vs non-probabilistic models/algorithms

Introduction

Types de problèmes statistiques

$$\begin{aligned}&\text{Modèle statistique} \quad \mathcal{P}\\ &\text{Distribution} \quad P \in \mathcal{P}\\ &\text{Observations} \quad O \sim P\end{aligned}$$

- Estimation ponctuelle

$$\hat{f}(O) \approx f(P)$$

- Estimation par intervalle

$$R(O) \subset \mathrm{Im} f, \; P(f(P) \in R(O)) \approx 1 - \alpha$$

- Test statistique

$$H_0 \quad \hat{T}(O) \in \mathbf{R} \quad p = P(|\hat{T}| \geq |\hat{T}(O)|| \; H_0)$$

Statistique non asymptotique partie 1: biais, variance et risque

$$\begin{aligned}\text{Modèle statistique} &\quad \mathcal{P} \\ \text{Distribution} &\quad P \in \mathcal{P} \\ \text{Observations} &\quad O \sim P\end{aligned}$$

- Estimation ponctuelle
  - biais $b_P(\hat{f}) = E_{O\sim P}[\hat{f}(O)] - f(P) \quad \|b_P(\hat{f})\|_2$
  - variance $\mathrm{Var}_P(\hat{f}) = \mathrm{Var}_{O\sim P}[\hat{f}(O)]$
  - risque $\ell : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}$

$$R_P(\hat{f}) = E_{O\sim P}[\ell(f(P), \hat{f}(O))]$$

Outils pour la construction d'estimateurs ponctuels

- Statistique suffisante

  Definition 2.4 (Sufficiency). Let X be a sample from an unknown population $P \in \mathcal{P}$, where $\mathcal{P}$ is a family of populations. A statistic $T(X)$ is said to be *sufficient* for $P \in \mathcal{P}$ (or for $\theta \in \Theta$ when $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a parametric family) if and only if the conditional distribution of X given T is *known* (does not depend on P or θ). ∎

- Maximum de vraisemblance

Statistique

- Introduction générale
- Statistique non asymptotique partie 1: biais, variance et risque
- Statistique asymptotique
- Statistique non asymptotique partie 2: inégalités de concentration

Modes de convergence

- Convergence presque sûre ou presque partout
- Convergence en probabilité
- Convergence en loi
- Convergence en moyenne quadratique

Loi forte des grands nombres

$X_1, X_2, \ldots$ variables aléatoires i.i.d.

(ii) (The SLLN). A necessary and sufficient condition for the existence of a constant c for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} c \tag{1.81}$$

is that $E|X_1| < \infty$, in which case $c = EX_1$

Théorème de la limite centrale

(Multivariate CLT). Let $X_1, ..., X_n$ be i.i.d. random k-vectors with a finite $\Sigma = \mathrm{Var}(X_1)$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - EX_1) \to_d N_k(0, \Sigma). \quad \blacksquare$$

Transformations continues et théorème de Slutsky

Theorem 1.10. Let $X, X_1, X_2, ...$ be random k-vectors defined on a probability space and g be a measurable function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^l, \mathcal{B}^l)$. Suppose that g is continuous a.s. P_X. Then
(i) $X_n \to_{a.s.} X$ implies $g(X_n) \to_{a.s.} g(X)$;
(ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$;
(iii) $X_n \to_d X$ implies $g(X_n) \to_d g(X)$. ∎

Theorem 1.11 (Slutsky's theorem). Let $X, X_1, X_2, ...,\ Y_1, Y_2, ...$ be random variables on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_p c$, where c is a fixed real number. Then
(i) $X_n + Y_n \to_d X + c$;
(ii) $Y_n X_n \to_d cX$;
(iii) $X_n / Y_n \to_d X/c$ if $c \neq 0$.

Statistique

- Introduction générale
- Statistique non asymptotique partie 1: biais, variance et risque
- Statistique asymptotique
- Statistique non asymptotique partie 2: inégalités de concentration

Statistique non asymptotique partie 2: inégalités de concentration

- Hoeffding's bound
- Application : intervalles de confiances non-asymptotiques (cf. DM2)