Huitième cours

- DM1 à rendre aujourd'hui (en main propre ou par email: thomas.schatz@univ-amu.fr

- DM2 en ligne, à rendre pour le 27 Octobre

- DM3 -> bonus, à rendre pour le 27 Octobre, en ligne la semaine prochaine

- Scribes : Aymene-Mohammed Bouayed (rendu = fichier .tex)

- Aujourd'hui

    - Probabilités et statistique

Random Vectors

Given n RV's $X_1, ..., X_n$, we can define a random vector X s.t.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Note: all the notions of joint PDF/CDF will apply to X.

Given $g : \mathbb{R}^n \to \mathbb{R}^m$, we have:

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}, \mathbb{E}[g(X)] = \begin{bmatrix} \mathbb{E}[g_1(X)] \\ \mathbb{E}[g_2(X)] \\ \vdots \\ \mathbb{E}[g_m(X)] \end{bmatrix}.$$

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij-th entry contains the covariance between X_i and X_j.

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \dots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \dots & Cov[X_n, X_n] \end{bmatrix}$$

applying linearity of expectation and the fact that $Cov[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$, we obtain

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

Properties:

- Σ is symmetric and PSD Proof ?
- If $X_i \perp X_j$ for all i, j, then $\Sigma = diag(Var[X_1], ..., Var[X_n])$

Multivariate Gaussian

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

If Sigma is not degenerate

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

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

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

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

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

Distributions conditionnelles

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

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

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

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

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

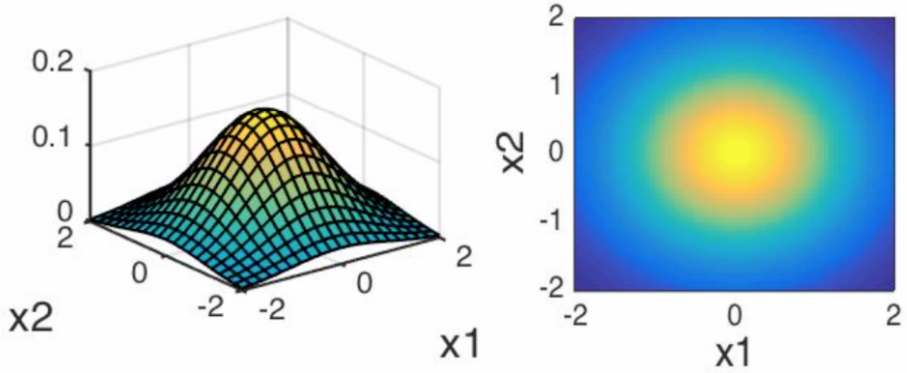
Distributions marginales ?

Visualizations of MV Gaussians

Effect of changing variance

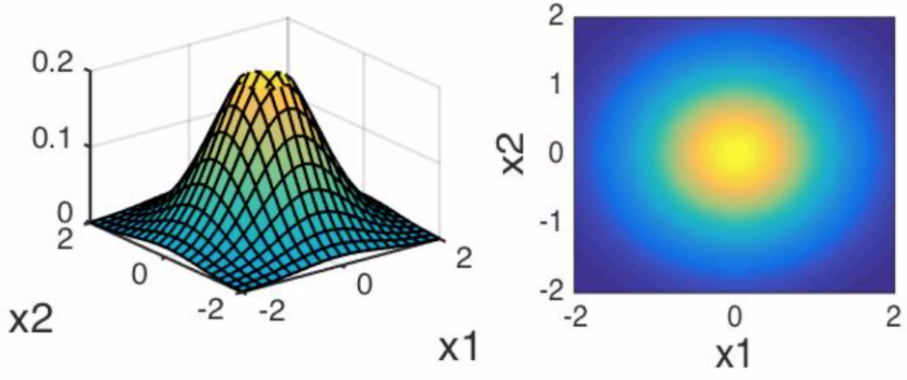
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



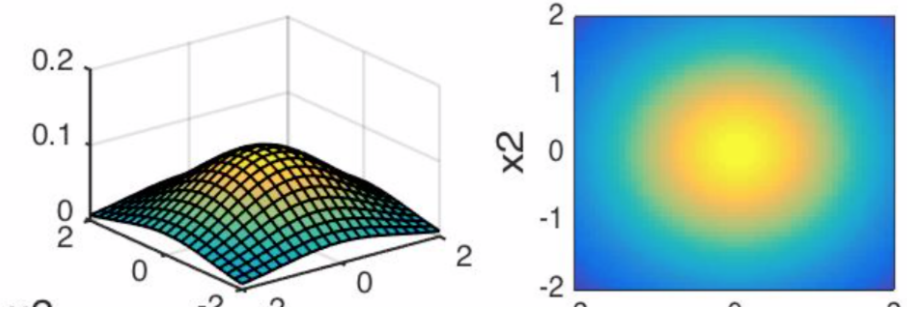
$$\Sigma = \begin{matrix} 0.7 & 0 \\ 0 & 0.7 \end{matrix}$$

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



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

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

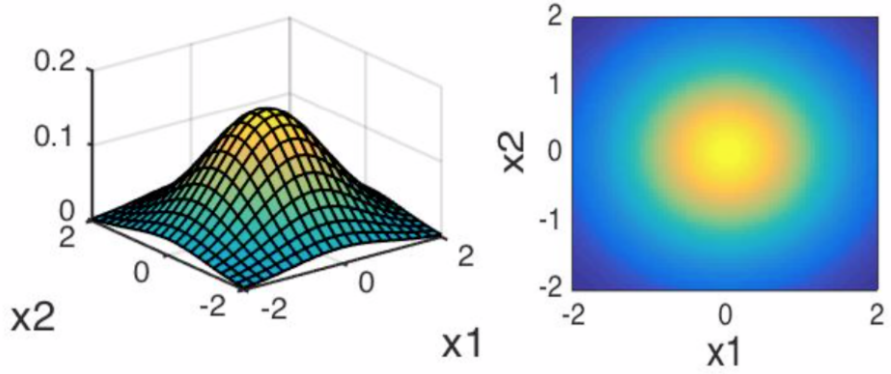


Visualizations of MV Gaussians

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

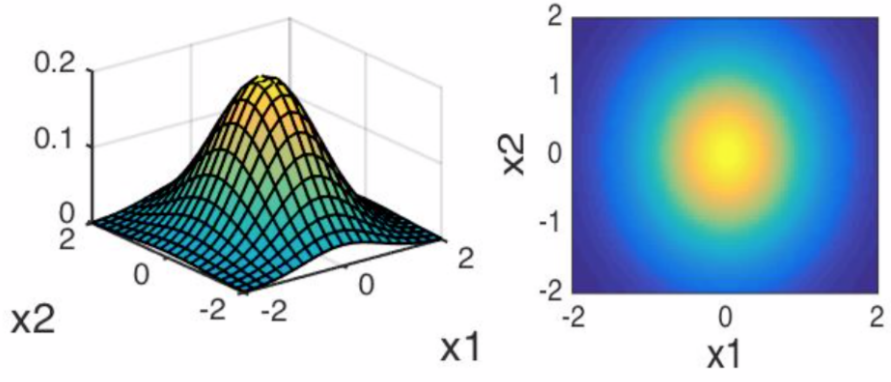
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



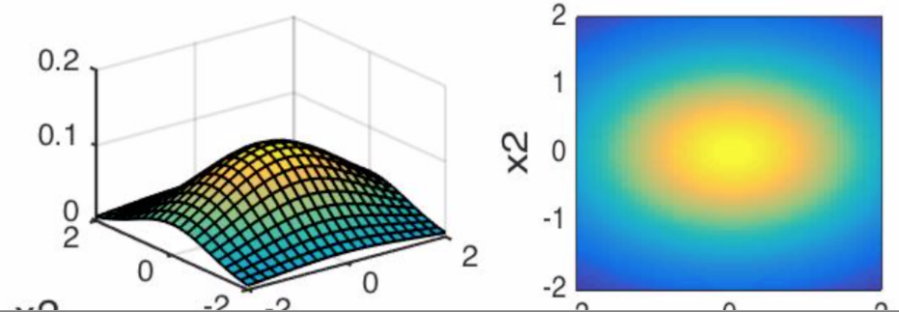
$$\Sigma = \begin{matrix} 0.6 & 0 \\ 0 & 1 \end{matrix}$$

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



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

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

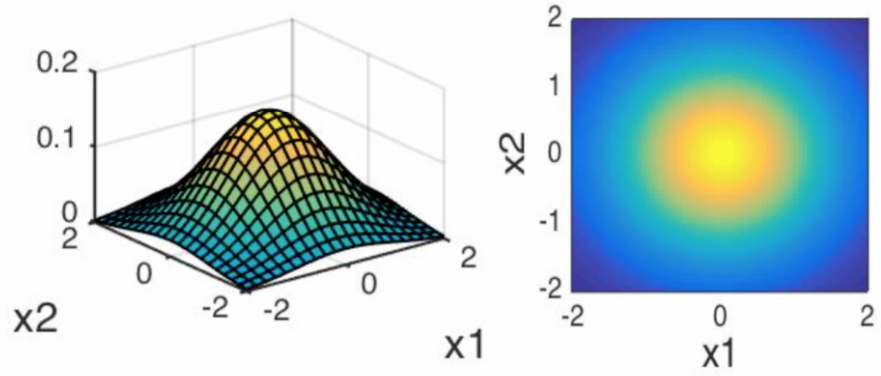


Visualizations of MV Gaussians

If X_1 and X_2 are positively correlated:

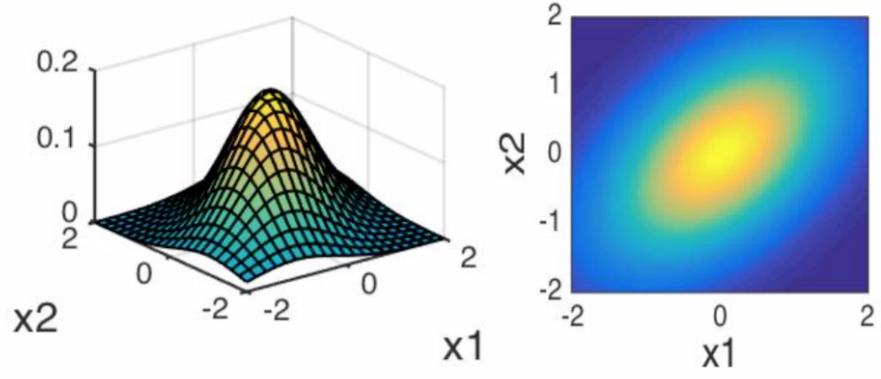
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



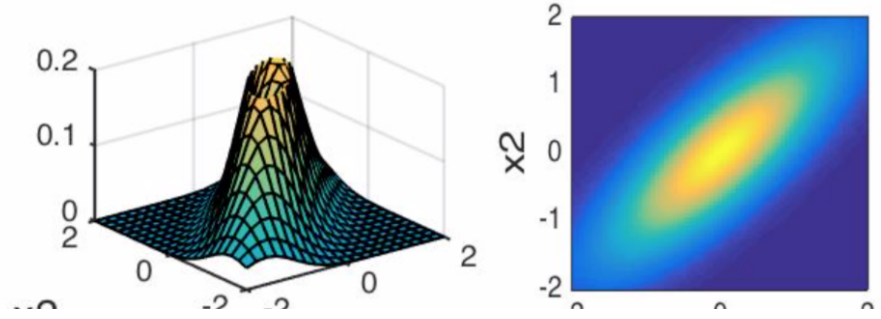
$$\Sigma = \begin{matrix} 1 & 0.5 \\ 0.5 & 1 \end{matrix}$$

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



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

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

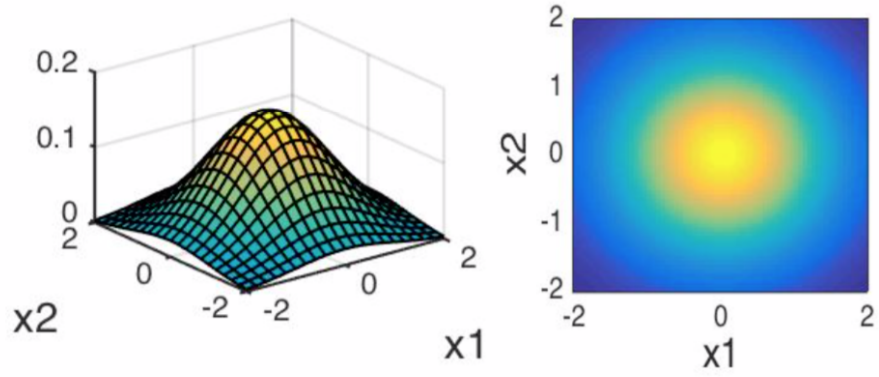


Visualizations of MV Gaussians

If X_1 and X_2 are negatively correlated:

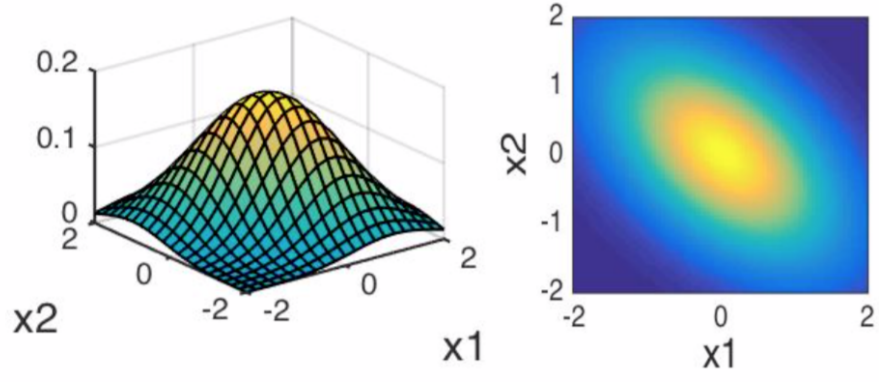
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



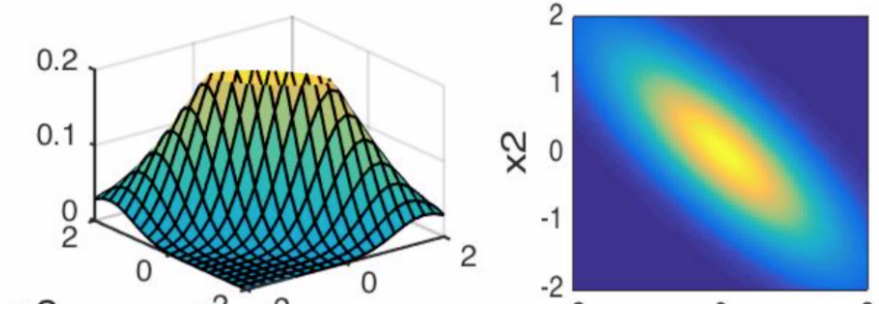
$$\Sigma = \begin{matrix} 1 & -0.5 \\ -0.5 & 1 \end{matrix}$$

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



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

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



Elements of Probability

Sample Space Ω

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

Event $A \subseteq \Omega$

$\{HH, HT\}$, Ω

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

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

Plan du cours

1. Introduction générale

2. Preuves (revue)

3. Algèbre linéaire (revue)

4. Optimisation (revue)

5. Optimisation sous contraintes

6. Probabilités (revue)

7. Statistique

8. Théorie de l’apprentissage

9. Optimisation pour l’apprentissage

Modes de convergence

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

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 $\Sigma = \text{Var}(X_1)$. Then

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

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 $c \neq 0$.

Plan du cours

1. Introduction générale

2. Preuves (revue)

3. Algèbre linéaire (revue)

4. Optimisation (revue)

5. Optimisation sous contraintes

6. Probabilités (revue)

7. Statistique

8. Théorie de l’apprentissage

9. Optimisation pour l’apprentissage

Introduction

Statistique “Bayésienne” et “fréquentiste”

Type de problèmes

- Estimation ponctuelle
- Estimation par intervalle
- Test statistique

Outils pour le contrôle qualité en statistique

- Estimateur ponctuel
    - biais, variance, risque, consistence, vitesse de convergence (asymptotique ou non)
        - optimalité : admissibilité, UMVUE, minimax, minimum Bayes risk
- Estimateur par intervalle
    - probabilité de couverture
- Test
    - correction, pouvoir statistique

Crédits

Certains théorèmes de Jun Shao, Mathematical Statistics

Transparents sur les probabilités repris et modifié de

- Taide Ding, Fereshte Khani, Stanford CS229 probability theory review, April 2020