Mathématiques pour l'intelligence artificielle

M2 informatique parcours IAAA, Aix-Marseille Université

Thomas Schatz
Jeudi 14 septembre 2023

Moments

“Bilinéarité” de la variance

$$\operatorname{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n}\sum_{i=1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j)$$

Bilinéarité de la covariance

$$\operatorname{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{Cov}(X_i, Y_j)$$

Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion

Probabilités

1. Notions de base

2. Calculs d'espérance et de variance

3. Loi gaussienne multivariée et autres distributions

Example Distributions

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, ..., n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, ...$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, ...$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ for all $x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

[2] Table reproduced from Maleki & Do's review handout by Koochak & Irvin

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
- If $X_i \perp X_j$ for all i, j, then $\Sigma = diag(Var[X_1], ..., Var[X_n])$

Multivariate Gaussian

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

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

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

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

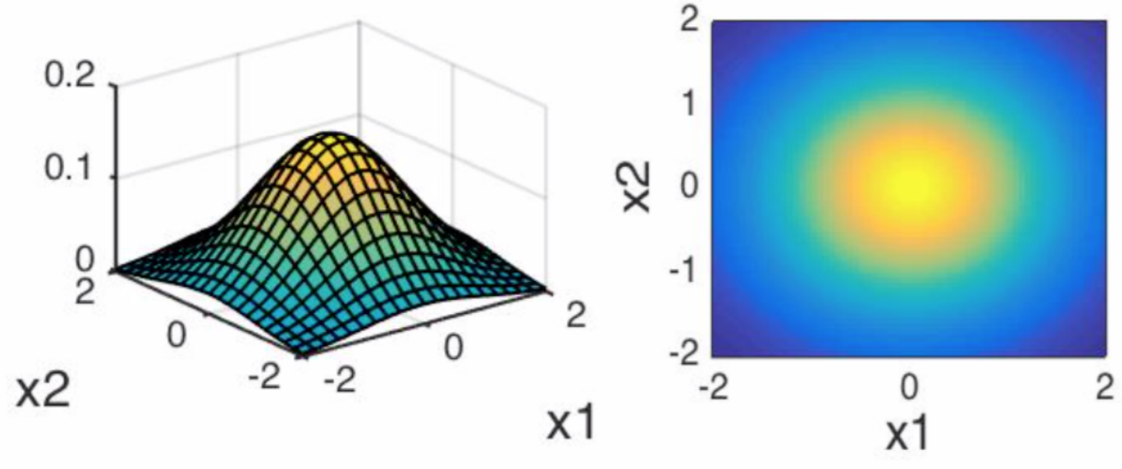
Some Nice Properties of MV Gaussians

- 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.

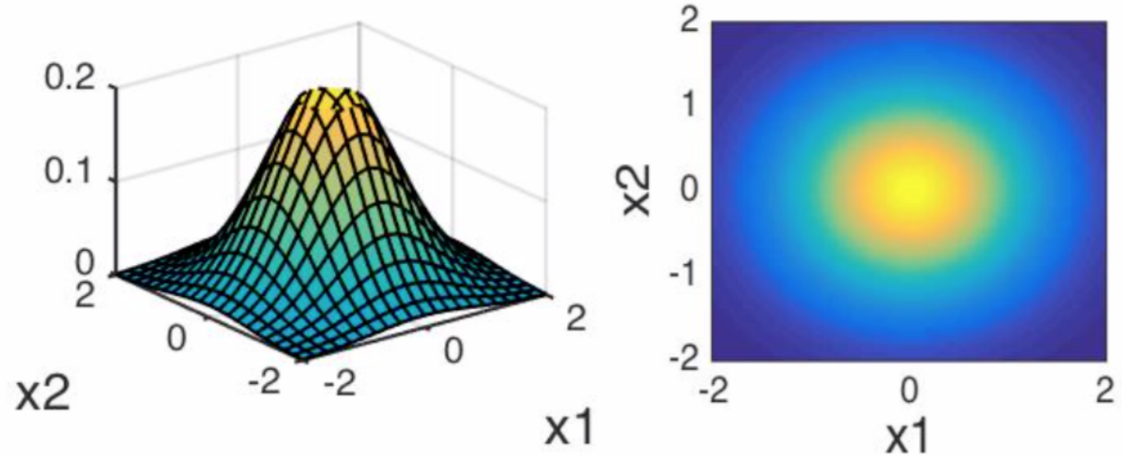
Visualizations of MV Gaussians

Effect of changing variance

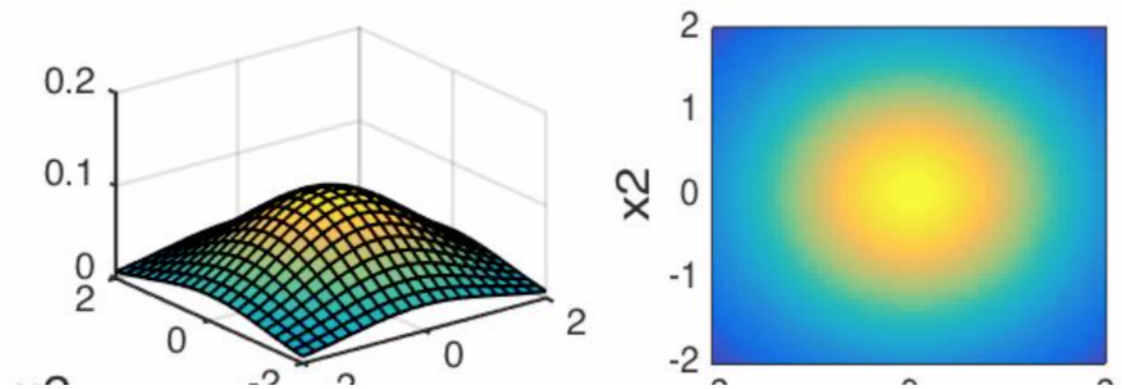
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 0.7 & 0 \\ 0 & 0.7 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$



$$\Sigma = \begin{matrix} 1.5 & 0 \\ 0 & 1.5 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$

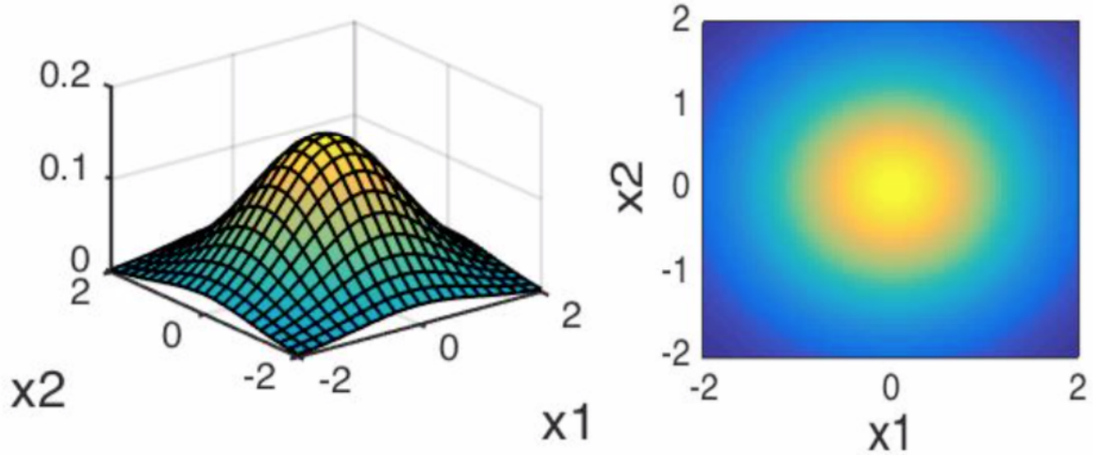


Visualizations of MV Gaussians

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

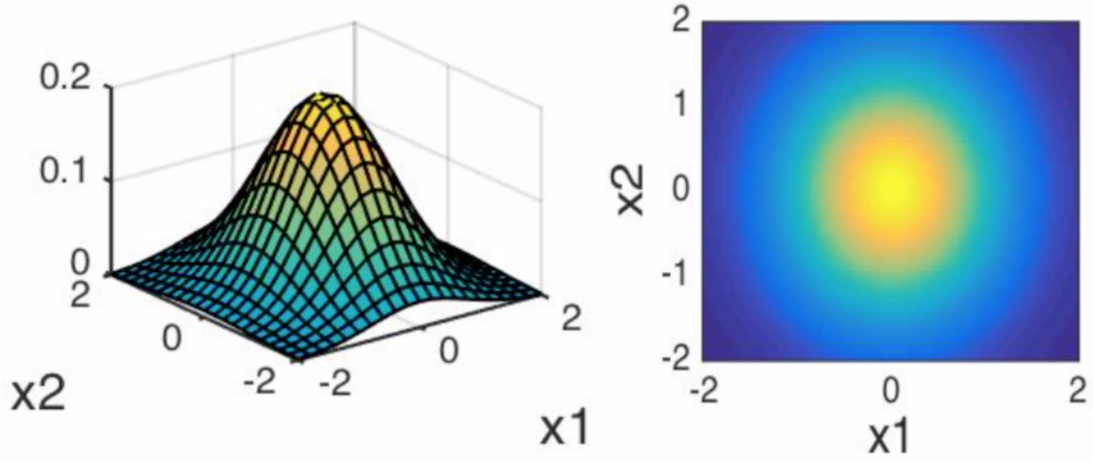
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



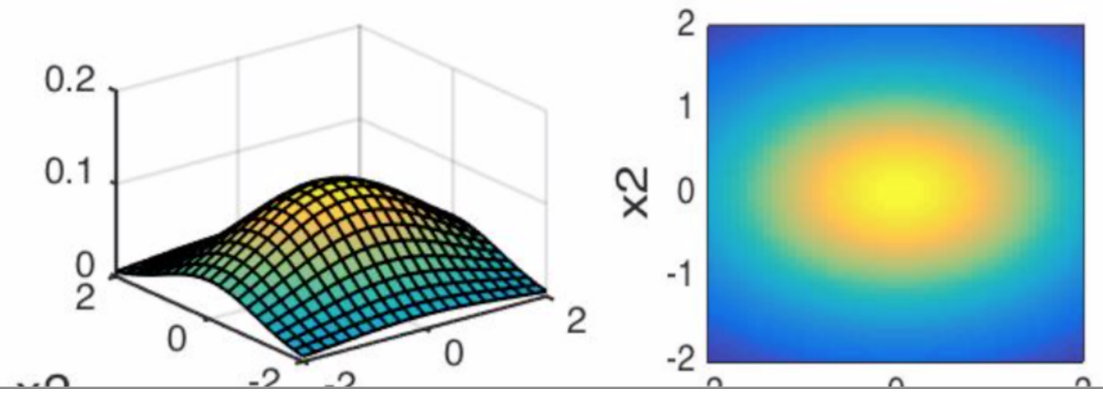
$$\Sigma = \begin{matrix} 0.6 & 0 \\ 0 & 1 \end{matrix}$$

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



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

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

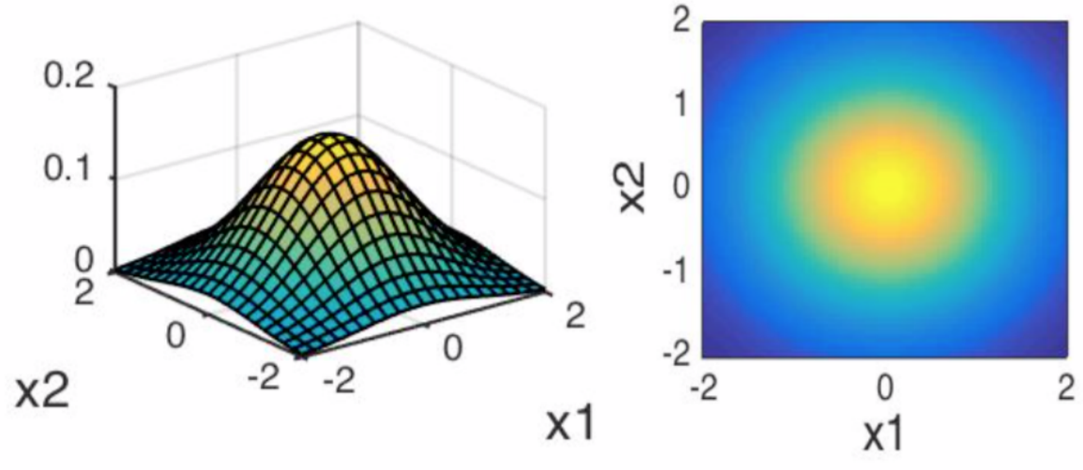


Visualizations of MV Gaussians

If X_1 and X_2 are positively correlated:

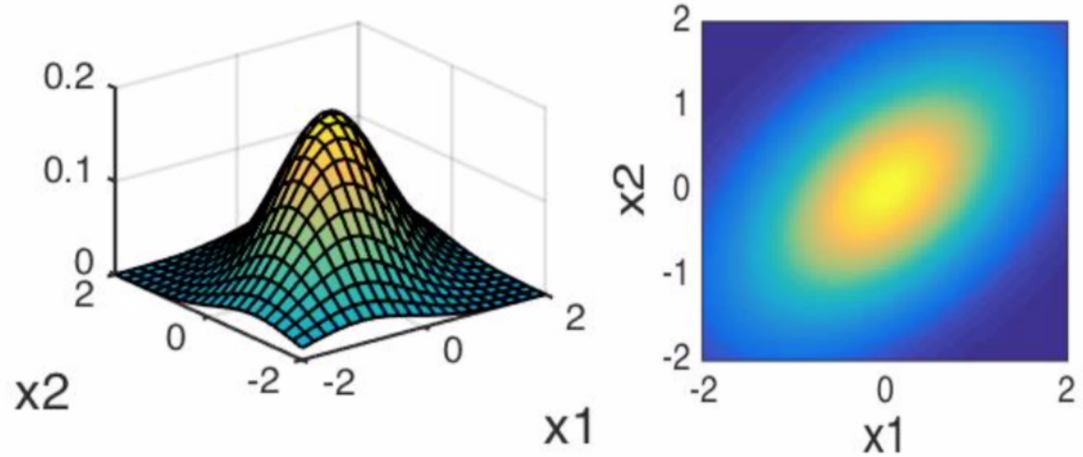
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



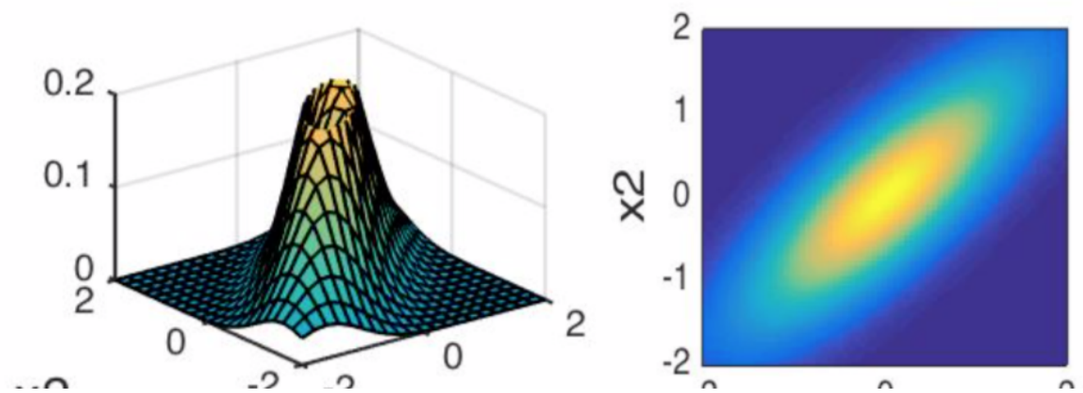
$$\Sigma = \begin{matrix} 1 & 0.5 \\ 0.5 & 1 \end{matrix}$$

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



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

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

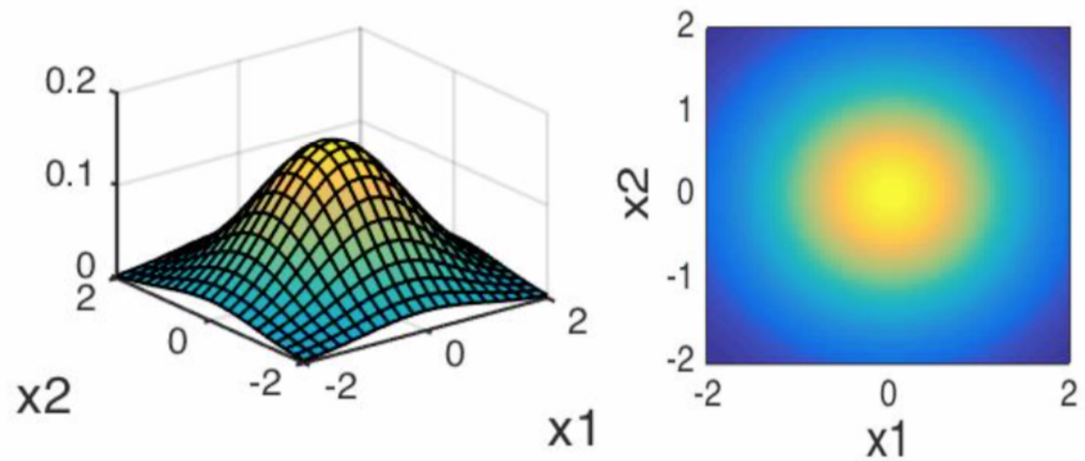


Visualizations of MV Gaussians

If X_1 and X_2 are negatively correlated:

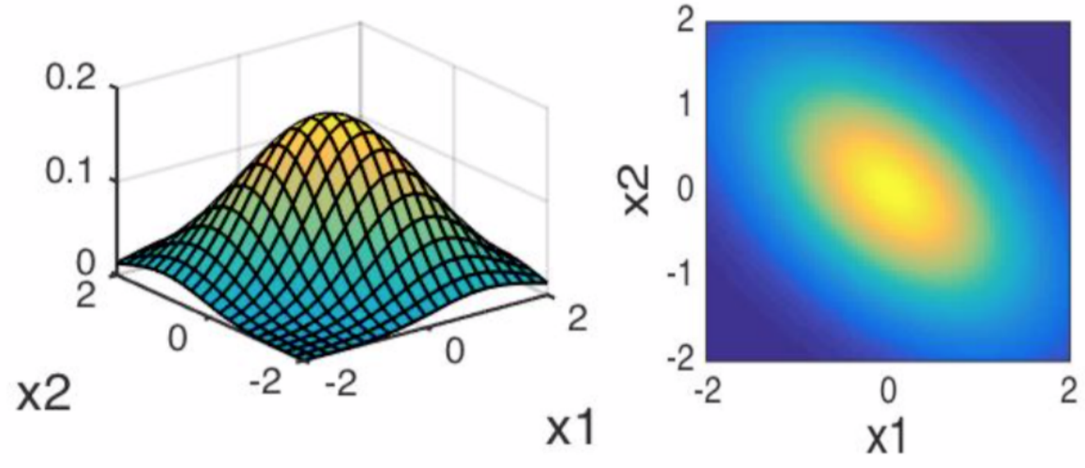
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



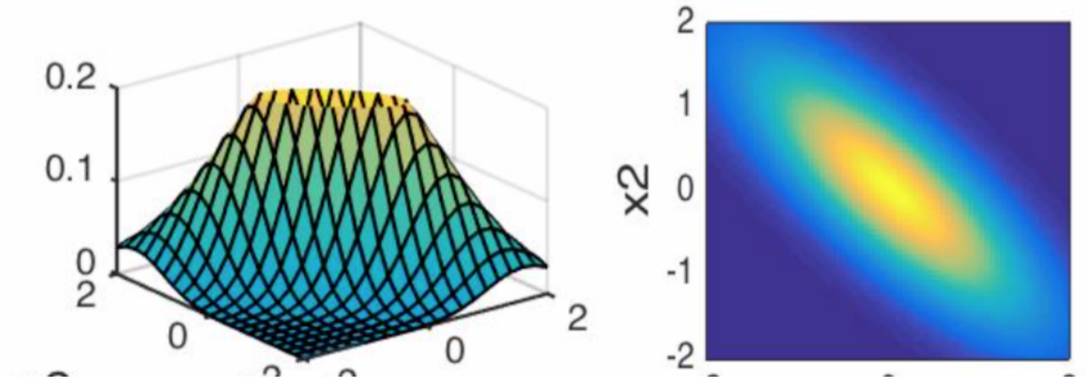
$$\Sigma = \begin{matrix} 1 & -0.5 \\ -0.5 & 1 \end{matrix}$$

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



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

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



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 ?

Probabilités

1. Notions de base

2. Calculs d’espérance et de variance

3. Loi gaussienne multivariée et autres distributions

Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion

Statistique

1. Cadre général
2. Statistique non-asymptotique: biais, variance, risque et concentration
3. Statistique asymptotique

Cadre général

Cadre général

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

- Estimation ponctuelle

$$\begin{aligned}\text{Fonctionnelle} \quad & f : \mathcal{P} \to E \\ \text{Estimateur} \quad & \hat{f} : \mathcal{O} \to E \\ \text{But} \quad & \hat{f}(O) \approx f(P)\end{aligned}$$

- Estimation par intervalle, tests d'hypothèses

Cadre général

- Problème de l'induction, inséparable de la démarche scientifique dans son ensemble

- Bayésien vs fréquentiste et autres ("best systems" etc.)

- Question d'**interprétation** de la modélisation probabiliste dans un cadre applicatif donné (que représentent les probabilités considérées ?)

    - Une méthode formelle n'est donc pas intrinsèquement "Bayésienne" ou "fréquentiste", tout dépend de son utilisation

    - Référence sur le sujet : https://plato.stanford.edu/entries/probability-interpret/

Statistique

1. Cadre général
2. Statistique non-asymptotique: biais, variance, risque et concentration
3. Statistique asymptotique

Qualité statistique d'un estimateur ponctuel

- biais $b_P(\hat{f}) = E_{O\sim P}[\hat{f}(O)] - f(P)$ $\|b_P(\hat{f})\|_2$
- variance $\mathrm{Var}_P(\hat{f}) = \mathrm{Var}_{O\sim P}[\hat{f}(O)]$
- risque $R_P(\hat{f}) = E_{O\sim P}[\ell(f(P), \hat{f}(O))]$

ℓ Fonction de coût

Décomposition biais-variance du risque dans le cas $\ell : x, y \mapsto \|x - y\|_2^2$

$$R_P(\hat{f}) = \mathrm{Var}_P(\hat{f}) + \|b_P(\hat{f})\|_2^2$$

Comment obtenir un estimateur ?

Cas paramétrique: estimateur du maximum de vraisemblance

$$\text{Modèle statistique} \quad \mathcal{P}$$
$$\text{Distribution} \quad P \in \mathcal{P}$$
$$\text{Observations} \quad O \sim P$$

$$\text{Fonctionnelle} \ f : \mathcal{P} \to E$$
$$\text{Estimateur} \quad \hat{f} : \mathcal{O} \to E$$

Cas paramétrique :

$$\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$$
$$f(P_\theta) = \theta$$

$$\hat{f} : O \mapsto \arg\max_{\theta \in \Theta}(p_\theta(O))$$

Statistique

1. Cadre général
2. Statistique non-asymptotique: biais, variance, risque et concentration
3. Statistique asymptotique

Modes de convergence

- Convergence en loi $X_n \to_d X$ si et seulement si

$$\lim_{n \to +\infty} E[f(X_n)] = E[f(X)]$$

pour toute fonction f à valeurs réelles, continue et bornée

Définition équivalente pour des variables aléatoire réelles:

$$\lim_{n \to +\infty} F_{X_n}(x) = F_X(x) \text{ en tout point } X \text{ où } F_X \text{ est continue}$$

Modes de convergence

- Convergence presque sûre ou presque partout $X_n \to_{a.s.} X$

$$P\left(\lim_{n\to+\infty} X_n = X\right) = 1$$

- Convergence en probabilité $X_n \to_p X$

$$\forall \epsilon > 0, \lim_{n\to+\infty} P(|X_n - X| > \epsilon) = 0$$

- Convergence en moyenne quadratique $X_n \to_{\mathcal{L}_2} X$

$$\lim_{n\to+\infty} E[|X_n - X|^2] = 0$$

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 $c \neq 0$.

Delta method

Soit X_1, X_2, ... and Y des vecteurs aléatoires de dimension k, tels que :

$$a_n(X_n - c) \to_d Y,$$

pour $c \in \mathbf{R}^k$ et (a_n) une suite de nombres positifs tendant vers $+\infty$. Alors pour toute fonction $g : \mathbf{R}^k \to \mathbf{R}$ différentiable en c, on a:

$$a_n[g(X_n) - g(c)] \to_d [\nabla g(c)]^T Y.$$

Statistique

1. Cadre général
2. Statistique non-asymptotique: biais, variance, risque et concentration
3. Statistique asymptotique

Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion