Mathématiques pour l'intelligence artificielle

UE MIA, M2 IAAA, AMU, 2022-2023

Thomas Schatz

Mardi 13 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

Algèbre linéaire

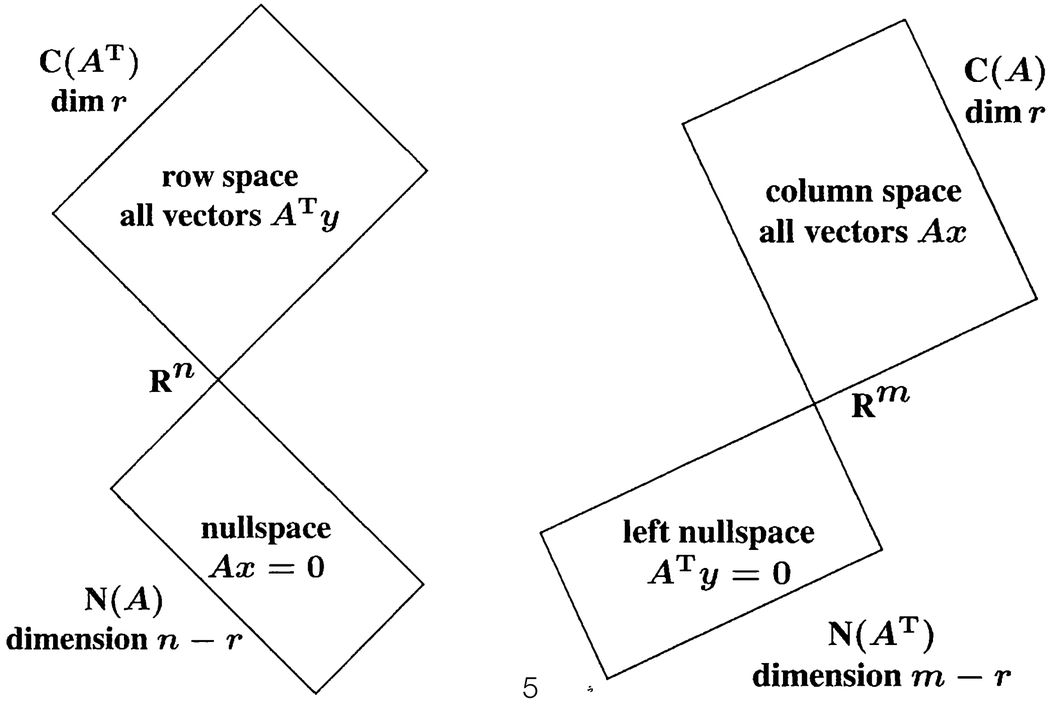
1. Espaces vectoriels et fonctions linéaires
2. Matrices
3. Angles et orthogonalité
4. Structure des applications linéaires et matrices
5. Espaces de matrices et normes
6. Algèbre linéaire numérique

Algèbre linéaire

1. Espaces vectoriels et fonctions linéaires
2. Matrices
3. Angles et orthogonalité
4. Structure des applications linéaires et matrices
5. Espaces de matrices et applications linéaires

Espaces associés à une matrice

$$AV = U\Sigma \qquad A \begin{bmatrix} v_1 \,..\, v_r \,..\, v_n \end{bmatrix} = \begin{bmatrix} u_1 \,..\, u_r \,..\, u_m \end{bmatrix} \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]$$



Exercice manipulation algébrique produit scalaire

Intérêt ?

Soit E un espace vectoriel de dimension finie n muni d'un produit scalaire $(\, . \mid . \,)$ et $V = (v_1, \ldots, v_n)$ une base orthonormale de E. Alors, pour tout $v \in E$,

$$v = \sum_{i=1}^{n} (v \mid v_i) v_i.$$

- step 1a. $\tilde{q}_1 := a_1$
- step 1b. $q_1 := \tilde{q}_1 / \|\tilde{q}_1\|$ (normalize)
- step 2a. $\tilde{q}_2 := a_2 - (q_1^T a_2) q_1$ (remove q_1 component from a_2)
- step 2b. $q_2 := \tilde{q}_2 / \|\tilde{q}_2\|$ (normalize)
- step 3a. $\tilde{q}_3 := a_3 - (q_1^T a_3) q_1 - (q_2^T a_3) q_2$ (remove q_1, q_2 components)
- step 3b. $q_3 := \tilde{q}_3 / \|\tilde{q}_3\|$ (normalize)
- etc.

Théorème spectral

Si S est une matrice symétrique, réelle de taille $m \times m$, alors il existe une matrice orthogonale réelle Q de taille $m \times m$ et une matrice diagonale réelle Λ de taille $m \times m$ telles que $S = Q\Lambda Q^T$.

Une matrice symétrique réelle est dite définie positive, noté $S \succ 0$ ssi pour toute matrice colonne u, $u^T S u > 0$.

Une matrice symétrique réelle est dite semi-définie positive, $S \succeq 0$, ssi pour toute matrice colonne u, $u^T S u \geq 0$.

Relation d'ordre sur les matrices symétriques réelles:

$S_1 \prec S_2$ ssi $S_2 - S_1 \succ 0$

et

$S_1 \preceq S_2$ ssi $S_2 - S_1 \succeq 0$

Diagonalisation

Soit A une matrice à coefficients réels de taille $m \times m$. $\lambda \in \mathbf{C}$ est une valeur propre de A ssi il existe $v \in \mathbf{C}^m$, non-nul, tel que $Av = \lambda v$, i.e. l'image de v par A est dans la même direction que v. Un tel v est appelé un vecteur propre de A associé à la valeur propre λ.

Diagonalisation

suppose $v_1, \ldots, v_n$ is a *linearly independent* set of eigenvectors of $A \in \mathbf{R}^{n \times n}$:

$$Av_i = \lambda_i v_i, \quad i = 1, \ldots, n$$

express as

$$A \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

define $T = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}$ and $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, so

$$AT = T\Lambda$$

and finally

$$T^{-1}AT = \Lambda$$

Forme canonique de Jordan

any matrix $A \in \mathbf{R}^{n \times n}$ can be put in *Jordan canonical form* by a similarity transformation, $i.e.$

$$T^{-1}AT = J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_q \end{bmatrix}$$

where

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbf{C}^{n_i \times n_i}$$

is called a *Jordan block* of size n_i with eigenvalue λ_i (so $n = \sum_{i=1}^{q} n_i$)

Déterminant et trace

A matrice carrée $n \times n$

$$\det(A) = \sum_{\sigma \in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma_i} \right),$$

sgn(σ) est la parité du nombre d'élement dans une décomposition de σ en une séquence de transpositions (échange de deux éléments).

$$\det(AB) = \det(A)\det(B)$$

$$\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{i,i} \qquad \mathrm{Tr(AB)} = \mathrm{Tr(BA)}$$

$$\mathrm{Tr}(A_1 A_2 \dots A_k) = \mathrm{Tr}(A_2 A_3 \dots A_k A_1) = \cdots = \mathrm{Tr}(A_k A_1 A_2 \dots A_{k-1})$$

Théorème de Cayley-Hamilton

Cayley-Hamilton theorem: for any $A \in \mathbf{R}^{n \times n}$ we have $\mathcal{X}(A) = 0$, where $\mathcal{X}(s) = \det(sI - A)$

Corollaire : pour tout entier naturel p, $A^p \in \mathrm{Vect}(I, A, A^2, \ldots, A^{n-1})$

Algèbre linéaire

1. Espaces vectoriels et fonctions linéaires
2. Matrices
3. Angles et orthogonalité
4. Structure des applications linéaires et matrices
5. Espaces de matrices et normes
6. Algèbre linéaire numérique

Espaces de fonctions linéaires

Espace de matrice, muni de l'addition et la multiplication matricielle et de la multiplication par un scalaire:

$$\mathcal{M}_{m,n}(\mathbf{R})$$

Espace de fonctions linéaires muni de l'addition et de la composition de fonctions linéaires et de la multiplication par un scalaire:

$$\mathcal{L}(E, F)$$

Norme induite par la norme 2 pour une matrice

$$\|A\|_2 := \sup_{x \in \mathbf{R}^n,\ x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \in \mathbf{R}^n,\ \|x\|_2 = 1} \|Ax\|_2$$

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 et théorèmes de convergence)

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

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

Conditional Probability and Bayes' Rule

For any events A, B such that $P(B) \neq 0$, we define:

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$

Let's apply conditional probability to obtain **Bayes' Rule**!

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

$$= \boxed{\frac{P(B)P(A \mid B)}{P(A)}}$$

Conditioned Bayes' Rule: given events A, B, C,

$$P(A \mid B, C) = \frac{P(B \mid A, C)P(A \mid C)}{P(B \mid C)}$$

See Appendix for proof :)

Law of Total Probability

Let $B_1, ..., B_n$ be n disjoint events whose union is the entire sample space. Then, for any event A,

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A \mid B_i) P(B_i) \end{aligned}$$

We can then write Bayes' Rule as:

$$\begin{aligned} P(B_k \mid A) &= \frac{P(B_k) P(A \mid B_k)}{P(A)} \\ &= \boxed{\frac{P(B_k) P(A \mid B_k)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}} \end{aligned}$$

Example

Treasure chest **A** holds 100 gold coins. Treasure chest **B** holds 60 gold and 40 silver coins.
Choose a treasure chest uniformly at random, and pick a coin from that chest uniformly at random. If the coin is gold, then what is the probability that you chose chest **A**? [1]
Solution:

[1] Question based on slides by Koochak & Irvin

Example

Treasure chest **A** holds 100 gold coins. Treasure chest **B** holds 60 gold and 40 silver coins.
Choose a treasure chest uniformly at random, and pick a coin from that chest uniformly at random. If the coin is gold, then what is the probability that you chose chest **A**? [1]
Solution:

$$
\begin{aligned}
P(A \mid G) &= \frac{P(A)P(G \mid A)}{P(A)P(G \mid A) + P(B)P(G \mid B)} \\
&= \frac{0.5 \times 1}{0.5 \times 1 + 0.5 \times 0.6} \\
&= \boxed{0.625}
\end{aligned}
$$

[1] Question based on slides by Koochak & Irvin

Chain Rule

For any n events $A_1, ..., A_n$, the joint probability can be expressed as a product of conditionals:

$$P(A_1 \cap A_2 \cap ... \cap A_n)$$
$$= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_2 \cap A_1)...P(A_n \mid A_{n-1} \cap A_{n-2} \cap ... \cap A_1)$$

Independence

Events A, B are independent if

$$P(AB) = P(A)P(B)$$

We denote this as $A \perp B$. From this, we know that if $A \perp B$,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Implication: If two events are independent, observing one event does not change the probability that the other event occurs.
In general: events $A_1, ..., A_n$ are **mutually independent** if

$$P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$$

for any subset $S \subseteq \{1, ..., n\}$.

Random Variables

real-valued

- A **random variable** X maps outcomes to real values.
- X takes on values in $Val(X) \subseteq \mathbb{R}$.
- $X = k$ is the event that random variable X takes on value k.

Discrete RVs:

- $Val(X)$ is a set
- $P(X = k)$ can be nonzero

Continuous RVs:

- $Val(X)$ is a range
- $P(X = k) = 0$ for all k. $P(a \leq X \leq b)$ can be nonzero.

Probability Mass Function (PMF)

Given a **discrete** RV X, a PMF maps values of X to probabilities.

$$p_X(x) := P(X = x)$$

For a valid PMF, $\sum_{x \in Val(x)} p_X(x) = 1$.

Cumulative Distribution Function (CDF)

A CDF maps a continuous RV to a probability (i.e. $\mathbb{R} \to [0, 1]$)

$$F_X(x) := P(X \leq x)$$

A CDF must fulfill the following:

- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- If $a \leq b$, then $F_X(a) \leq F_X(b)$ (i.e. CDF must be nondecreasing)

Also note: $P(a \leq X \leq b) = F_X(b) - F_X(a)$.

Probability Density Function (PDF)

PDF of a continuous RV is simply the derivative of the CDF.

$$f_X(x) := \frac{dF_X(x)}{dx}$$

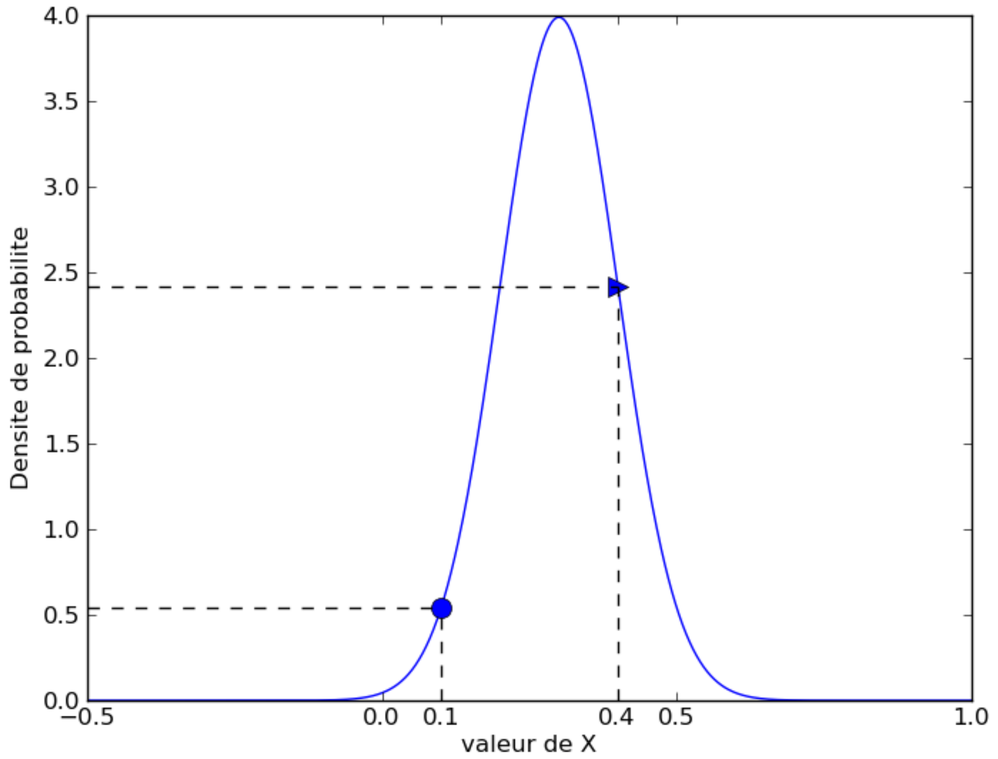
Thus,

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$$

A valid PDF must be such that

- for all real numbers x, $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

On a représente sur le graphe ci-dessous la densité de probabilité d'une variable aléatoire réelle X.

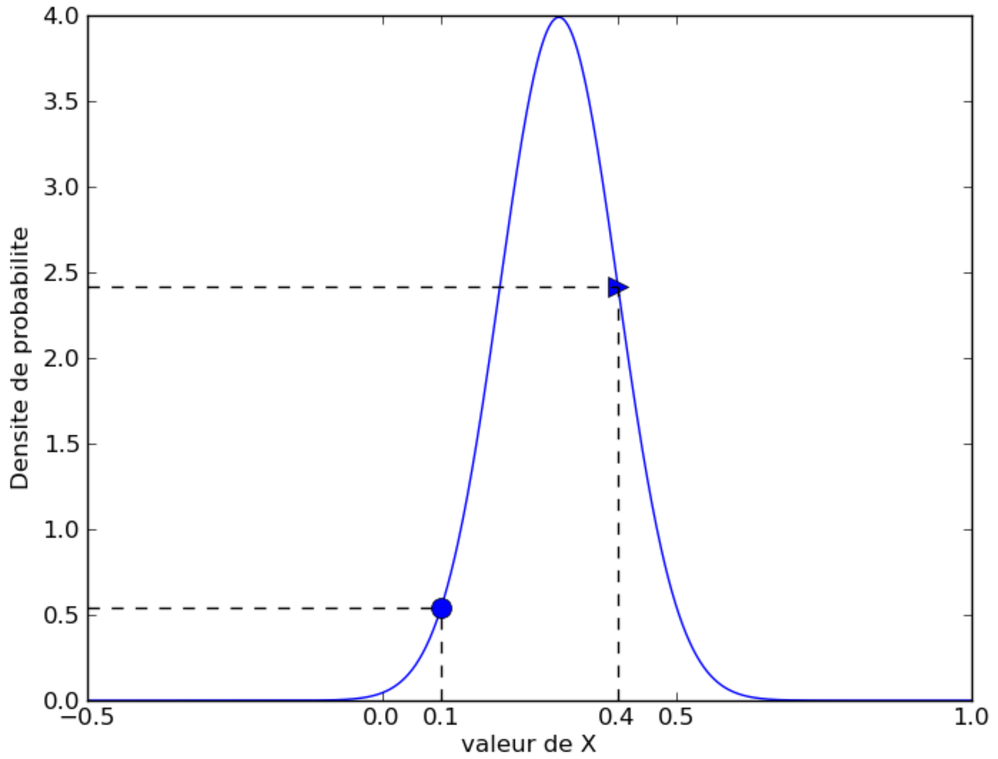


1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous parait-il normal ? Justifiez votre réponse.

Soit x une réalisation de X.

2) Quelle est la probabilité d'avoir $x = 0.1$? Quelle est la probabilité d'avoir $x = 0.4$? Est il plus probable d'observer $x = 0.4$ ou $x = 0.1$? A quel point (approximativement) ?

On a représente sur le graphe ci-dessous la densité de probabilité d'une variable aléatoire réelle X.



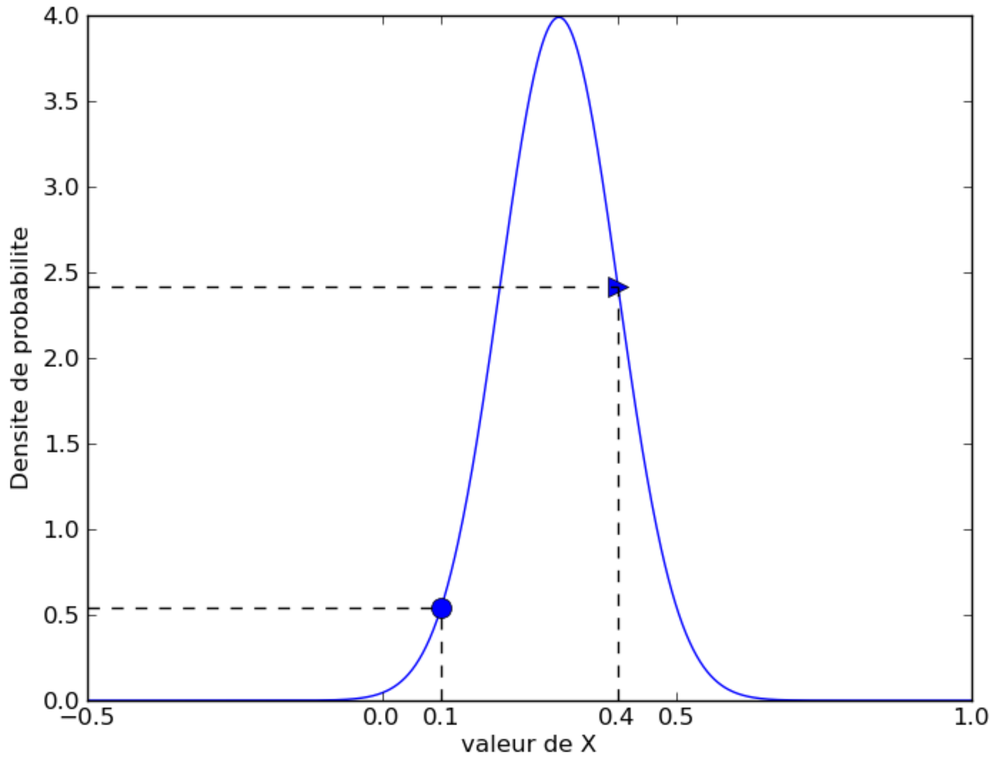
1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous parait-il normal ? Justifiez votre réponse.

L'aire sous la courbe doit être égale à 1 mais la densité en un point particulier peut être supérieure et même arbitrairement grande.

Soit x une réalisation de X.

2) Quelle est la probabilité d'avoir $x = 0.1$? Quelle est la probabilité d'avoir $x = 0.4$? Est il plus probable d'observer $x = 0.4$ ou $x = 0.1$? A quel point (approximativement) ?

On a représente sur le graphe ci-dessous la densité de probabilité d'une variable aléatoire réelle X.



1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous parait-il normal ? Justifiez votre réponse.

L'aire sous la courbe doit être égale à 1 mais la densité en un point particulier peut être supérieure et même arbitrairement grande.

Soit x une réalisation de X.

2) Quelle est la probabilité d'avoir $x = 0.1$? Quelle est la probabilité d'avoir $x = 0.4$? Est il plus probable d'observer $x = 0.4$ ou $x = 0.1$? A quel point (approximativement) ?

La probabilité que x soit 0.1 ou 0.4 est 0. Par contre, il est environ 5 fois plus probable d'observer $X = 0.4$ que $x = 0.1$.

Expectation

Let g be an arbitrary real-valued function.

- If X is a discrete RV with PMF p_X:

$$\mathbb{E}[g(X)] := \sum_{x \in Val(X)} g(x) p_X(x)$$

- If X is a continuous RV with PDF f_X:

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Intuitively, expectation is a weighted average of the values of $g(x)$, weighted by the probability of x.

Properties of Expectation

For any constant $a \in \mathbb{R}$ and arbitrary real function f:

- $\mathbb{E}[a] = a$
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$

Linearity of Expectation

Given n real-valued functions $f_1(X), ..., f_n(X)$,

$$\mathbb{E}[\sum_{i=1}^{n} f_i(X)] = \sum_{i=1}^{n} \mathbb{E}[f_i(X)]$$

Variance

The **variance** of a RV X measures how concentrated the distribution of X is around its mean.

$$\begin{aligned} Var(X) &:= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

Interpretation: $Var(X)$ is the expected deviation of X from $\mathbb{E}[X]$.
Properties: For any constant $a \in \mathbb{R}$, real-valued function $f(X)$

- $Var[a] = 0$
- $Var[af(X)] = a^2 Var[f(X)]$

Joint and Marginal Distributions

- **Joint PMF** for discrete RV's X, Y:

$$p_{XY}(x, y) = P(X = x, Y = y)$$

  Note that $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} p_{XY}(x, y) = 1$
- **Marginal PMF** of X, given joint PMF of X, Y:

$$p_X(x) = \sum_y p_{XY}(x, y)$$

- **Joint PDF** for continuous X, Y:

$$f_{XY}(x, y) = \frac{\delta^2 F_{XY}(x, y)}{\delta x \delta y}$$

  Note that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$
- **Marginal PDF** of X, given joint PDF of X, Y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

Joint and Marginal Distributions for Multiple RVs

- **Joint PMF** for discrete RV's $X_1, ..., X_n$:

$$p(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n)$$

  Note that $\sum_{x_1} \sum_{x_2} ... \sum_{x_n} p(x_1, ..., x_n) = 1$
- **Marginal PMF** of X_1, given joint PMF of $X_1, ..., X_n$:

$$p_{X_1}(x_1) = \sum_{x_2} ... \sum_{x_n} p(x_1, ..., x_n)$$

- **Joint PDF** for continuous RV's $X_1, ..., X_n$:

$$f(x_1, ..., x_n) = \frac{\delta^n F(x_1, ...x_n)}{\delta x_1 \delta x_2 ... \delta x_n}$$

  Note that $\int_{x_1} \int_{x_2} ... \int_{x_n} f(x_1, ..., x_n) dx_1 ... dx_n = 1$
- **Marginal PDF** of X_1, given joint PDF of $X_1, ..., X_n$:

$$f_{X_1}(x_1) = \int_{x_2} ... \int_{x_n} f(x_1, ..., x_n) dx_2 ... dx_n$$

Expectation for multiple random variables

Given two RV's X, Y and a function $g : \mathbb{R}^2 \to \mathbb{R}$ of X, Y,

- for discrete X, Y:

$$\mathbb{E}[g(X, Y)] := \sum_{x \in Val(x)} \sum_{y \in Val(y)} g(x, y) p_{XY}(x, y)$$

- for continuous X, Y:

$$\mathbb{E}[g(X, Y)] := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy$$

These definitions can be extended to multiple random variables in the same way as in the previous slide. For example, for n continuous RV's $X_1, .., X_n$ and function $g : \mathbb{R}^n \to \mathbb{R}$:

$$\mathbb{E}[g(X)] = \int \int ... \int g(x_1, ..., x_n) f_{X_1,...,X_n}(x_1, ..., x_n) dx_1, ..., dx_n$$

Conditional distributions for RVs

Works the same way with *RV*'s as with events:

- For discrete X, Y:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

- For continuous X, Y:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

- In general, for continuous $X_1, ..., X_n$:

$$f_{X_1|X_2,...,X_n}(x_1|x_2, ..., x_n) = \frac{f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)}{f_{X_2,...,X_n}(x_2, ..., x_n)}$$

Bayes' Rule for RVs

Also works the same way for *RV*'s as with events:

- For discrete X, Y:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y' \in Val(Y)} p_{X|Y}(x|y')p_Y(y')}$$

- For continuous X, Y:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y')dy'}$$

Chain Rule for RVs

Also works the same way as with events:

$$\begin{aligned} f(x_1, x_2, ..., x_n) &= f(x_1) f(x_2|x_1) ... f(x_n|x_1, x_2, ..., x_{n-1}) \\ &= f(x_1) \prod_{i=2}^{n} f(x_i|x_1, ..., x_{i-1}) \end{aligned}$$

Independence for RVs

- For $X \perp Y$ to hold, it must be that $F_{XY}(x, y) = F_X(x)F_Y(y)$ **FOR ALL VALUES** of x, y.

- Since $f_{Y|X}(y|x) = f_Y(y)$ if $X \perp Y$, chain rule for mutually independent $X_1, ..., X_n$ is:

$$f(x_1, ..., x_n) = f(x_1)f(x_2)...f(x_n) = \prod_{i=1}^{n} f(x_i)$$

(very important assumption for a Naive Bayes classifier!)

Graphical model

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_{\pi_i})$$

Properties of Expectation

Law of Total Expectation
Given two RVs X, Y:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

N.B. $\mathbb{E}[X \mid Y] = \sum_{x \in Val(x)} x p_{X|Y}(x|y)$ is a function of Y.
See Appendix for details :)

Example of Law of Total Expectation

El Goog sources two batteries, A and B, for its phone. A phone with battery A runs on average 12 hours on a single charge, but only 8 hours on average with battery B. El Goog puts battery A in 80% of its phones and battery B in the rest. If you buy a phone from El Goog, how many hours do you expect it to run on a single charge?

Example of Law of Total Expectation

El Goog sources two batteries, A and B, for its phone. A phone with battery A runs on average 12 hours on a single charge, but only 8 hours on average with battery B. El Goog puts battery A in 80% of its phones and battery B in the rest. If you buy a phone from El Goog, how many hours do you expect it to run on a single charge?

Solution: Let L be the time your phone runs on a single charge. We know the following:

- $p_X(A) = 0.8$, $p_X(B) = 0.2$,
- $\mathbb{E}[L \mid A] = 12$, $\mathbb{E}[L \mid B] = 8$.

Then, by Law of Total Expectation,

$$\begin{aligned}\mathbb{E}[L] = \mathbb{E}[\mathbb{E}[L \mid X]] &= \sum_{X \in \{A,B\}} \mathbb{E}[L \mid X] p_X(X) \\ &= \mathbb{E}[L \mid A] p_X(A) + \mathbb{E}[L \mid B] p_X(B) \\ &= 12 \times 0.8 + 8 \times 0.2 = \boxed{11.2}\end{aligned}$$

Covariance

Intuitively: measures how much one RV's value tends to move with another RV's value. For RV's X, Y:

$$\begin{aligned} Cov[X, Y] &:= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

- If $Cov[X, Y] < 0$, then X and Y are negatively correlated
- If $Cov[X, Y] > 0$, then X and Y are positively correlated
- If $Cov[X, Y] = 0$, then X and Y are uncorrelated

Properties Involving Covariance

- If $X \perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Thus,

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

  This is unidirectional! $Cov[X, Y] = 0$ **does not imply** $X \perp Y$
- **Variance of two variables:**

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

  i.e. if $X \perp Y$, $Var[X + Y] = Var[X] + Var[Y]$.
- **Special Case:**

$$Cov[X, X] = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = Var[X]$$

Variance of a sum

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{Cov}(X_i, X_j)$$

Exercice : espérance et variance de la moyenne de n variables aléatoires i.id. ?

Variance of a sum

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{Cov}(X_i, X_j)$$

Exercice : espérance et variance de la moyenne de n variables aléatoires i.id. ?

Exercice : On a une procédure aléatoire pour entrainer un classificateur binaire, dont on obtient n échantillons (n classifieurs). Pour tester la qualité de la procédure d'entrainement, on a une procédure aléatoire de test qui produit une erreur de classification et qu'on applique m fois sur chacun des n classificateurs entrainé . Comment mesurer la variance de l'erreur de classification (par exemple pour savoir si elle est significativement en dessous du hasard) à partir des erreurs de classifications $(e_{i,j})_{1 \leq i \leq n, 1 \leq j \leq m}$?

Variance of a sum

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) + 2 \sum_{1 \le i < j \le n} \operatorname{Cov}(X_i, X_j)$$

Exercice : espérance et variance de la moyenne de n variables aléatoires i.id. ?

Law of total variance

$$\operatorname{Var}(Y) = \operatorname{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}(\operatorname{E}[Y \mid X]).$$

Exercice : On a une procédure aléatoire pour entrainer un classificateur binaire, dont on obtient n échantillons (n classifieurs). Pour tester la qualité de la procédure d'entrainement, on a une procédure aléatoire de test qui produit une erreur de classification et qu'on applique m fois sur chacun des n classificateurs entrainé . Comment mesurer la variance de l'erreur de classification (par exemple pour savoir si elle est significativement en dessous du hasard) à partir des erreurs de classifications $(e_{i,j})_{1 \le i \le n, 1 \le j \le m}$?

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

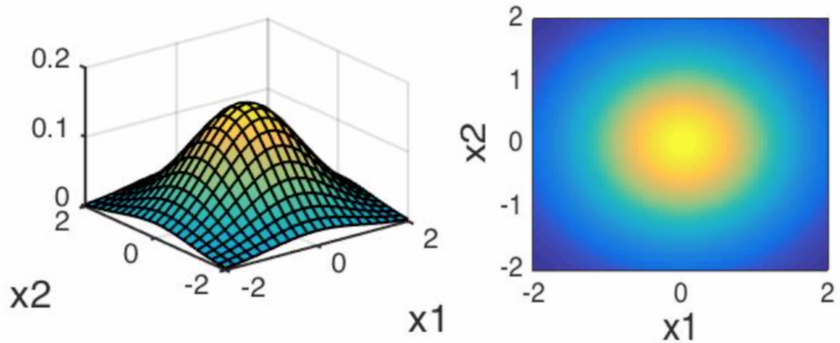
- Marginals and conditionals of a joint Gaussian are Gaussian
- A d-dimensional Gaussian $X \in \mathcal{N}(\mu, \Sigma = diag(\sigma_1^2, ..., \sigma_n^2))$ is equivalent to a collection of d **independent** Gaussians $X_i \in \mathcal{N}(\mu_i, \sigma_i^2)$. This results in isocontours aligned with 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

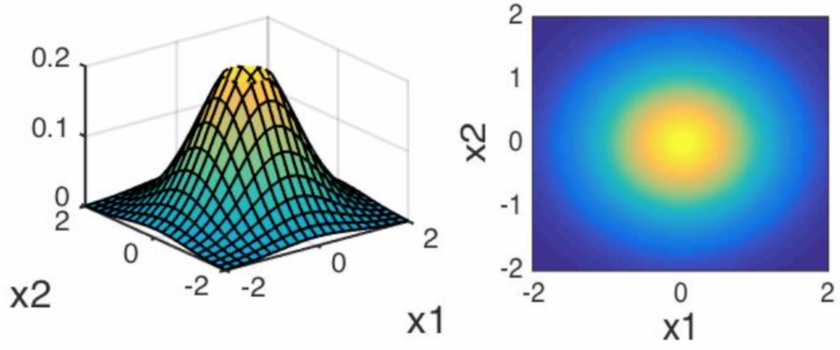
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



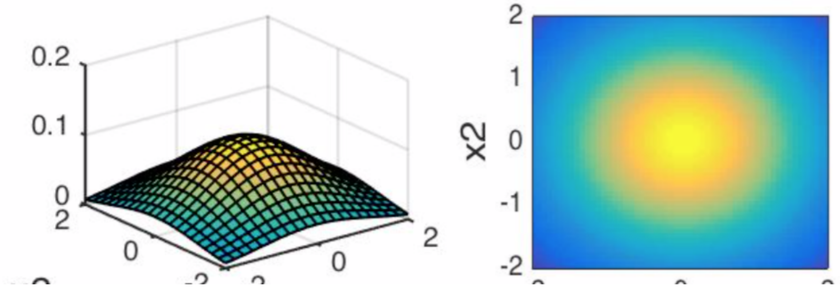
$$\Sigma = \begin{matrix} 0.7 & 0 \\ 0 & 0.7 \end{matrix}$$

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



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

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

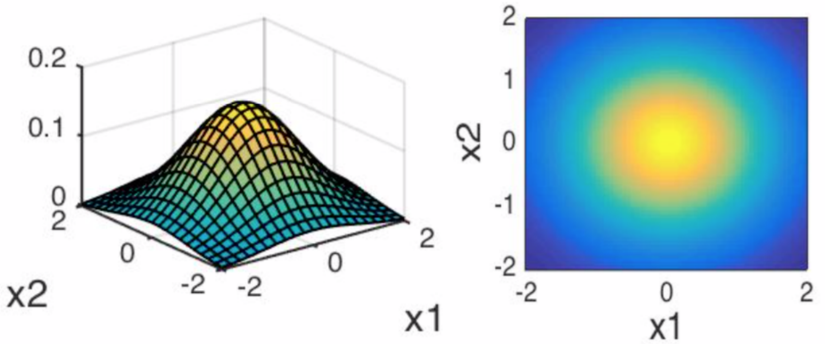


Visualizations of MV Gaussians

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

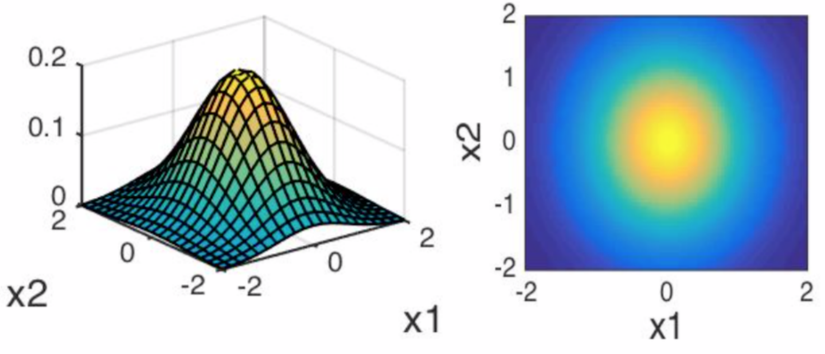
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



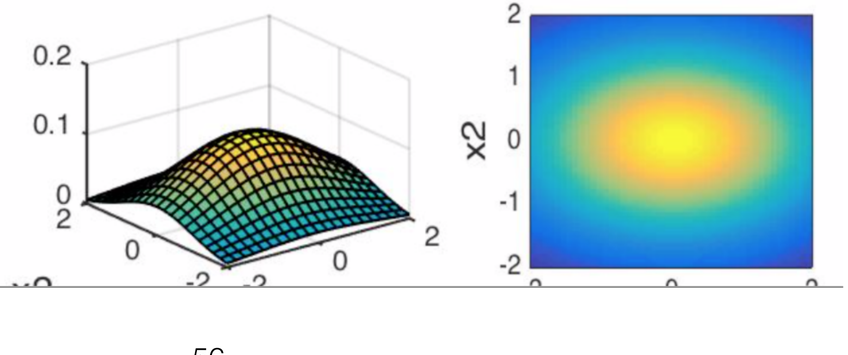
$$\Sigma = \begin{matrix} 0.6 & 0 \\ 0 & 1 \end{matrix}$$

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



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

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

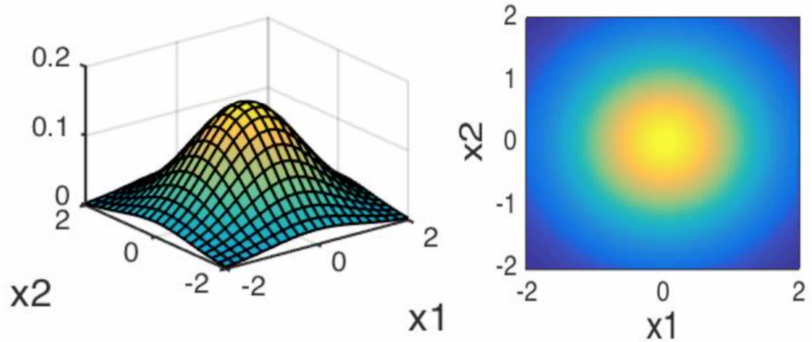


Visualizations of MV Gaussians

If X_1 and X_2 are positively correlated:

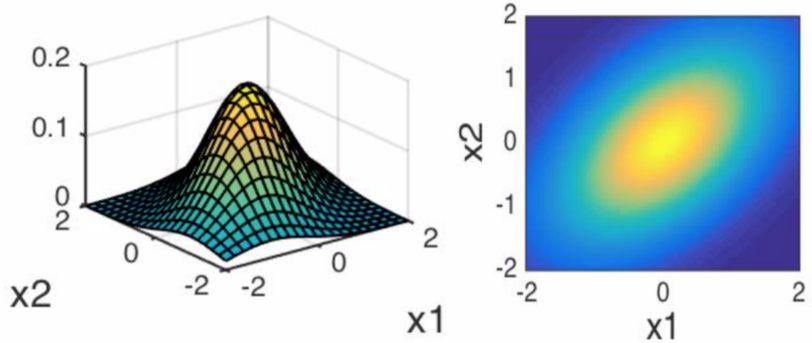
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



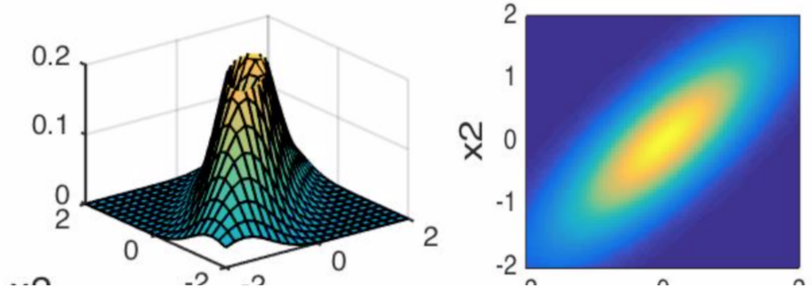
$$\Sigma = \begin{matrix} 1 & 0.5 \\ 0.5 & 1 \end{matrix}$$

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



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

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

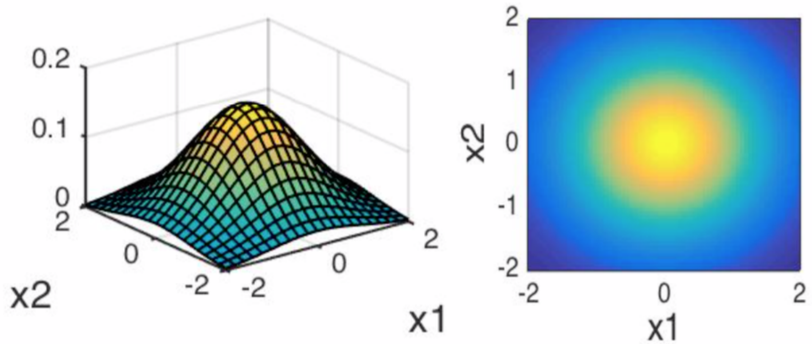


Visualizations of MV Gaussians

If X_1 and X_2 are negatively correlated:

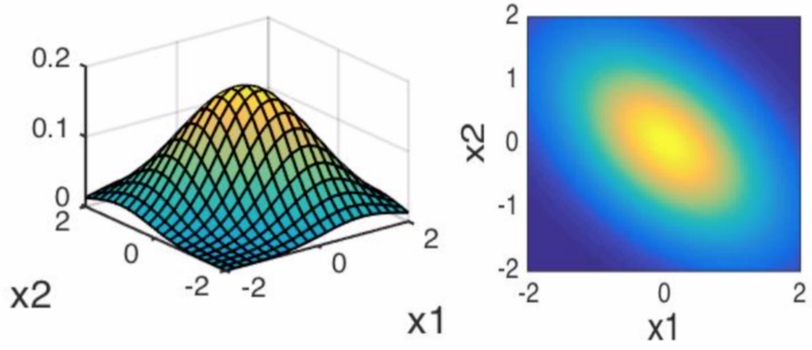
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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



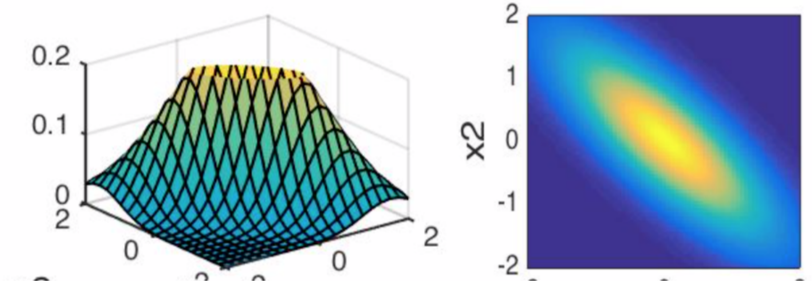
$$\Sigma = \begin{matrix} 1 & -0.5 \\ -0.5 & 1 \end{matrix}$$

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



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

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



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 ?

Appendix: More on Total Expectation

Why is $\mathbb{E}[X|Y]$ a function of Y? Consider the following:

- $\mathbb{E}[X|Y=y]$ is a scalar that only depends on y.
- Thus, $\mathbb{E}[X|Y]$ is a random variable that only depends on Y. Specifically, $\mathbb{E}[X|Y]$ is a function of Y mapping $Val(Y)$ to the real numbers.

An example: Consider RV X such that

$$X = Y^2 + \epsilon$$

such that $\epsilon \sim \mathcal{N}(0,1)$ is a standard Gaussian. Then,

- $\mathbb{E}[X|Y] = Y^2$
- $\mathbb{E}[X|Y=y] = y^2$

Appendix: More on Total Expectation

A derivation of Law of Total Expectation for discrete X, Y:[3]

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[\sum_x xP(X = x \mid Y)] \tag{1}$$

$$= \sum_y \sum_x xP(X = x \mid Y)P(Y = y) \tag{2}$$

$$= \sum_y \sum_x xP(X = x, Y = y) \tag{3}$$

$$= \sum_x x \sum_y P(X = x, Y = y) \tag{4}$$

$$= \sum_x xP(X = x) = \boxed{\mathbb{E}[X]} \tag{5}$$

where (1), (2), and (5) result from the definition of expectation, (3) results from the definition of cond. prob., and (5) results from marginalizing out Y.

[3] from slides by Koochak & Irvin

Appendix: A proof of Conditioned Bayes Rule

Repeatedly applying the definition of conditional probability, we have: [4]

$$\begin{aligned}
\frac{P(b|a,c)P(a|c)}{P(b|c)} &= \frac{P(b,a,c)}{P(a,c)} \cdot \frac{P(a|c)}{P(b|c)} \\
&= \frac{P(b,a,c)}{P(a,c)} \cdot \frac{P(a,c)}{P(b|c)P(c)} \\
&= \frac{P(b,a,c)}{P(b|c)P(c)} \\
&= \frac{P(b,a,c)}{P(b,c)} \\
&= P(a|b,c)
\end{aligned}$$

[4] from slides by Koochak & Irvin