MS BGD
MDI 720/721 : Statistiques

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Enseignants

- **Joseph Salmon** :
  - Précédemment : Paris Diderot-Paris 7, Duke University, Télécom ParisTech
  - Spécialités : statistiques en grande dimension, optimisation, agrégation, traitement des images
  - Email :*joseph.salmon@telecom-paristech.fr*
  - Bureau : E410
- **François Portier** :
  - Précédemment : Université de Rennes 1, Université catholique de Louvain, Télécom ParisTech
  - Spécialités : régression parcimonieuse, bootstrap, estimation semi-paramétrique, méthodes à noyaux
  - Email : *fportier@enst.fr*
  - Bureau : E 302

Calendrier de validation

- Devoir maison 1 : **5% note finale**
  - Date : séance du 03/10 (sujet déjà disponible) à rendre avant 23h59 le mercredi 04/10
  - Validation par pairs : à rendre pour le dimanche 08/10
- Devoir maison 2 : **20% note finale**
  - Date : sujet mis à disposition le 10/10 et à rendre avant 23h59 le dimanche 15/10
  - Validation par pairs : à rendre pour le vendredi 13/10
- Devoir maison 3 : **20% note finale**
  - Date : sujet mis à disposition le 20/10 et à rendre avant 23h59 le dimanche 22/10 (zéro passé cette limite)
  - Validation par pairs : à rendre pour le vendredi 27/10
- Un devoir final : **55% note finale**
  - Date : 25/10
  - Format : partie Quiz (1h30), partie numérique (1h30) qui pourra être rendue jusqu'à 23h59 le 25/10.

ATTENTION : le travail rendu doit être personnel ! ! !

Notation des parties numériques

Pour chaque rendu, note totale sur **20**
Détails des points :

- Qualité des réponses aux questions **15** pts
- Qualité de rédaction, de présentation et d'orthographe **2** pts
- Indentation, Style PEP8, commentaires dans le code **2** pts
- absence de bug **1** pt

Retard : 0 pour tout retard, sauf excuse validée par l'administration

Consigne supplémentaire : un fichier unique au format **.ipynb**

Aucun travail par mail accepté !

Bonus

1 pt supplémentaire sur **la note finale** pour toute contribution à l'amélioration des cours (présentations, codes, etc.)

Contraintes :

- seule la première amélioration reçue est "rémunérée"
- déposer un fichier **.txt** par proposition dans la partie du site pédagogique intitulée "Propositions d'améliorations"
- détailler précisément (ligne de code, page des présentations, etc.) l'amélioration proposée, ce qu'elle corrige et/ou améliore
- pour les fautes d'orthographe : proposer au minimum 5 corrections par contribution

Plan du cours

Séance 1. Joseph Salmon (13/09) : Introduction

Séance 2. Joseph Salmon (20/09) : Modèle linéaire (p<n)

Séance 3. Joseph Salmon (27/09) : Modèle linéaire (suite)

Séance 4. A. G./ E.N. / F.P. / J. S. (03/10) : **TP #1 : noté**

Séance 5. Joseph Salmon (4/10) : SVD / IC

Séance 6. A. G. / E.N. / F.P. / J. S. (10/10) : **TP #2 : noté**

Séance 7. François Portier (11/10) : IC / Bootstrap

Séance 8. Joseph Salmon (13/10) : Tests, Ridge

Séance 9. Joseph Salmon (17/10) : Sélection de variables / Lasso

Séance 10. Joseph Salmon (18/10) : GLM / Régression logistique

Séance 11. A.G. / E.N. / F.P. / J.S. (20/10) : **TP #3 : noté**

Séance 12. F.P. / J.S. (25/10) : **Validation finale**

Prérequis - à revoir seul

- Bases de **probabilités** : probabilité, densité, espérance, loi des grands nombres, lois gaussiennes, théorème central limite
  Lecture : Foata et Fuchs (1996)
- Bases de l'**optimisation** : fonctions convexes, condition du premier ordre, descente de gradient, méthode de Newton
  Lecture : Boyd et Vandenberghe (2004), Bertsekas (1999)

Prérequis - à revoir seul

- Bases de **probabilités** : probabilité, densité, espérance, loi des grands nombres, lois gaussiennes, théorème central limite
  Lecture : Foata et Fuchs (1996)
- Bases de l'**optimisation** : fonctions convexes, condition du premier ordre, descente de gradient, méthode de Newton
  Lecture : Boyd et Vandenberghe (2004), Bertsekas (1999)
- Bases de l'**algèbre (bi-)linéaire** : espaces vectoriels, normes, produit scalaire, matrices, déterminants, diagonalisation
  Lecture : Horn et Johnson (1994)

Prérequis - à revoir seul

- Bases de **probabilités** : probabilité, densité, espérance, loi des grands nombres, lois gaussiennes, théorème central limite
  Lecture : Foata et Fuchs (1996)
- Bases de l'**optimisation** : fonctions convexes, condition du premier ordre, descente de gradient, méthode de Newton
  Lecture : Boyd et Vandenberghe (2004), Bertsekas (1999)
- Bases de l'**algèbre (bi-)linéaire** : espaces vectoriels, normes, produit scalaire, matrices, déterminants, diagonalisation
  Lecture : Horn et Johnson (1994)
- Bases de l'**algèbre linéaire numérique** : résolution de système, factorisation de matrices, conditionnement, etc.
  Lecture : Golub et VanLoan (2013), Applied Numerical Computing par L. Vandenberghe

Prérequis - à revoir seul

- Bases de **probabilités** : probabilité, densité, espérance, loi des grands nombres, lois gaussiennes, théorème central limite
  Lecture : Foata et Fuchs (1996)
- Bases de l'**optimisation** : fonctions convexes, condition du premier ordre, descente de gradient, méthode de Newton
  Lecture : Boyd et Vandenberghe (2004), Bertsekas (1999)
- Bases de l'**algèbre (bi-)linéaire** : espaces vectoriels, normes, produit scalaire, matrices, déterminants, diagonalisation
  Lecture : Horn et Johnson (1994)
- Bases de l'**algèbre linéaire numérique** : résolution de système, factorisation de matrices, conditionnement, etc.
  Lecture : Golub et VanLoan (2013), Applied Numerical Computing par L. Vandenberghe

Aspects algorithmiques : quelques conseils

Installation Python : **Conda / Anaconda** (tous OS)

Rem: sur ce point je ne peux rien pour vous : entraidez-vous !

Outils :

- Rendus **Jupyter / IPython Notebook**
- Projets plus importants : **IPython** + éditeur de texte avancé ;

e.g., **Atom**, **Sublime Text**, **PyCharm**, etc.

- **Python**, **Scipy**, **Numpy** : Reproducible Data Analysis in Jupyter (Tutos de Jake Vanderplas) : OBLIGATOIRE !

  http://perso.telecom-paristech.fr/~gramfort/liesse_python/
- **Pandas** : https://github.com/jorisvandenbossche/pandas-tutorial
- **scikit-learn** : http://scikit-learn.org/stable/tutorial/index.html

Rem: en TP, prenez vos portables si vous préférez garder votre environnement (packages, versions, etc.)
Rem: je suis passé à Python 3 cette année... vous aussi ?

Conseils généraux pour l'année

- Utilisez un système de versionnement de fichiers pour vos travaux en groupe : **Git** (*e.g.*, **Bitbucket**, **Github**, etc.)
- Adoptez des règles d'écriture de code et tenez-y vous ! Exemple : **PEP8** pour Python (utiliser **AutoPEP8**)
- Utilisez **Markdown** (.md) (markdown-preview-plus avec **Atom**), *e.g.*, pour les parties rédigées / compte-rendus

"A (wo)man must have a code."
– Bunk

- Apprenez de bons exemples (ouvrez les codes sources !) : https://github.com/scikit-learn/, http://jakevdp.github.io/, etc.

Plan

Cadre statistique standard

On notera $\mathbb{P}, \mathbb{E}$ pour probabilité et l'espérance

- On observe des réalisations $(y_1, \ldots, y_n)$ de variables aléatoires inconnues (éventuellement vectorielles)
- On suppose ici que les variables sont indépendantes et identiquement distribuées (*i.i.d.*) selon une loi $\mathbb{P}_Y$

Rem: on note souvent Y une variable aléatoire et y une réalisation

Estimation

Comment apprendre certaines caractéristiques de $\mathbb{P}_Y$ seulement à partir des observations $(y_1, \ldots, y_n)$?

Prédiction

On se prépare à observer y_{n+1} : comment approcher y_{n+1}, quantifier une incertitude sur cette grandeur, etc. ?

Vocabulaire

- Observations $\mathbf{y} = y_{1:n} = (y_1, \ldots, y_n)$: **échantillon** de **taille** n
- Grandeurs **théoriques** : dépendant de la loi $\mathbb{P}_Y$, **inconnue**, et qui génère les observations
  Exemple : l'espérance $\mathbb{E}(Y)$ ou la variance $\mathrm{Var}(Y)$ de Y
- Grandeurs **empiriques** : calculées à partir des observations y_i
  Exemple : la moyenne empirique $\bar{y}_n = \dfrac{1}{n}\sum_{i=1}^{n} y_i$
- Objectif général : apprendre les caractéristiques théoriques de $\mathbb{P}_Y$ à partir de résumés empiriques.

Rem: les grandeurs théoriques dépendent de $\mathbb{P}_Y$ alors que les grandeurs empiriques dépendent de $\mathbb{P}_n := \dfrac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$ (ici δ_{y_i} est la mesure de Dirac au point y_i)

Sommaire

Modèle statistique : contexte

Rappel

- On observe des réalisations $(y_1, \ldots, y_n)$ de variables aléatoires inconnues (éventuellement vectorielles)
- On suppose ici que les variables sont indépendantes et identiquement distribuées (*i.i.d.*) selon une loi $\mathbb{P}_Y$

- Selon la situation, la loi $\mathbb{P}_Y$ a certaines caractéristiques.
  Exemple : "Pile ou face" : on sait que $\mathbb{P}_Y = \text{Bernoulli}(\theta)$ pour un certain $\theta \in [0, 1]$ inconnu

- Reformulation : on dispose d'une **famille de lois candidates**, (parfois naturelle) pour $\mathbb{P}_Y$
  Exemple : la famille des lois de Bernoulli

Exo: Quel est un modèle naturel pour "un lancer de dé" ?

Modèle statistique

- La loi cible $\mathbb{P}_Y$ est indexée par un **paramètre** $\theta \in \Theta$:
  $\mathbb{P}_Y = \mathbb{P}_\theta$ pour un θ inconnu, et Θ est l'ensemble d'indexation

  Exemple : "Pile ou face", $\theta \in \Theta = [0, 1]$ et $\mathbb{P}_\theta = \text{Bernoulli}(\theta)$

Définition

Un **modèle statistique** est une famille de lois

$$\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \Theta\}$$

indexée par un ensemble de paramètres Θ.

Exo: Proposer un modèle Θ pour le "lancer de dé".

Modèle statistique paramétrique

Définition

Un **modèle paramétrique** est une famille de lois $\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ indexée par un nombre fini p de paramètres : $\Theta \subset \mathbb{R}^p$. On note aussi $\mathbb{E}_\theta$ l'espérance associée.

Rem: le modèle est indexé par un nombre ou un vecteur réel ; p est la dimension du modèle

Exemple :

- Modèle de Bernoulli (ou "Pile ou face") : $\Theta = [0, 1]$.
- Modèle gaussien : $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times \mathbb{R}_+^*$.

Rem: le modèle est dit **non-paramétrique** s'il n'est pas indexable par un paramètre de dimension finie, *e.g.*, $\{f : \int f = 1, \text{ et } f \geqslant 0\}$

Rem: dans le cadre **fréquentiste**, on suppose qu'il existe un vrai paramètre inconnu, tel que $\mathbb{P}_Y = \mathbb{P}_\theta$

Estimateur

- *Objectif :* estimer une quantité $g = g(\theta)$ qui ne dépend que de la loi $\mathbb{P}_\theta$ des observations. g est une constante inconnue **déterministe** *i.e.*, non aléatoire.

  Exemple : espérance, quantile, variance, écart-type, etc.

- *Intuition :* un **estimateur** $\hat{g}$ est calculé à partir de l'échantillon $(y_1, \ldots, y_n)$, dans le but d'approcher $g(\theta)$.

Définition

Un **estimateur** $\hat{g}$ de g est une fonction (mesurable) des observations :

$$\hat{g} : (y_1, \ldots, y_n) \mapsto \hat{g}(y_1, \ldots, y_n)$$

Rem: un estimateur est parfois aussi appelé une **statistique**

Rem: en pratique l'estimateur doit être calculable efficacement

Sommaire

Propriétés d'un estimateur : le biais

Définition

Le **biais** d'un estimateur $\hat{g}$ est l'espérance de son écart au paramètre :

$$\text{Biais}(\hat{g}, g) = \mathbb{E}_\theta(\hat{g}(y_1, \ldots, y_n)) - g(\theta) \qquad \text{(dépend de } \theta\text{)}$$

Définition

Un estimateur $\hat{g}$ de g est dit **non biaisé** (ou **sans biais**) si :

$$\forall \theta \in \Theta, \quad \mathbb{E}(\hat{g}(y_1, \ldots, y_n)) = g(\theta)$$

Rem: le biais mesure l'erreur systématique d'un estimateur

Estimateur sans biais de l'espérance

- L'espérance 'théorique' dépend de la loi $\mathbb{P}_\theta$
- On cherche ici à estimer $g(\theta) = \mathbb{E}(Y) (= \mathbb{E}_\theta(Y))$

Théorème

Sous l'hypothèse que l'échantillon est *i.i.d.* , la moyenne empirique $\overline{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ est un estimateur sans biais de l'espérance $\mathbb{E}(Y)$

Démonstration :

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n y_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(y_i) = \mathbb{E}(Y)$$

car $\mathbb{E}(y_i) = \mathbb{E}(Y)$ (caractère *i.i.d.* des y_i)

Rem: $\hat{g}(y_1, \ldots, y_n) = y_1$ est un estimateur sans biais de l'espérance

Estimateur sans biais de la variance

- La variance 'théorique' dépend de la loi $\mathbb{P}_\theta$
- On cherche ici à estimer $g(\theta) = \mathrm{Var}(Y) (= \mathrm{Var}_\theta(Y))$

Théorème

L'estimateur $\hat{g}(y_1, \ldots, y_n) = \dfrac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y}_n)^2$ est un estimateur sans biais de la variance $\mathrm{Var}(Y)$

Rem: l'estimateur $\dfrac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y}_n)^2$ est lui biaisé

Exo: Vérifier cette propriété par le calcul

Propriétés d'un estimateur : la variance

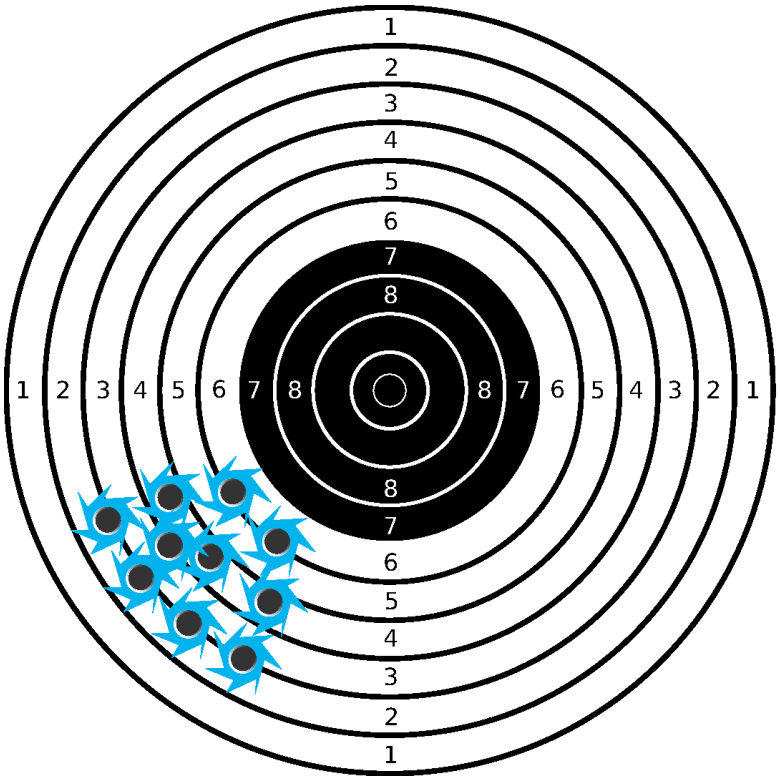
Définition

La **variance** d'un estimateur $\hat{g}$ est sa variance théorique :

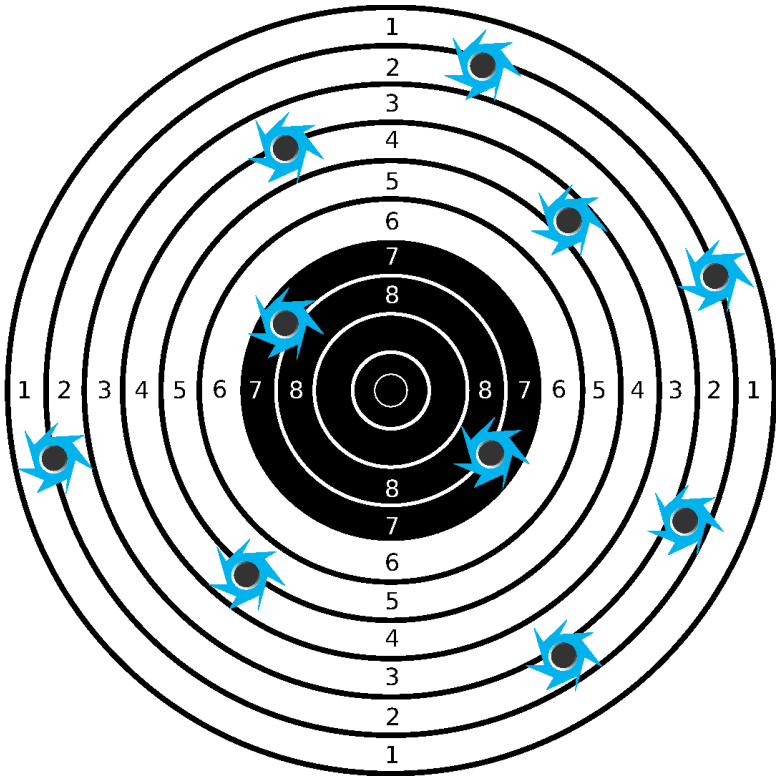
$$\mathrm{Var}(\hat{g}) = \mathrm{Var}(\hat{g}(y_1, \ldots, y_n)) = \mathbb{E}_\theta(\hat{g} - \mathbb{E}_\theta(\hat{g}))^2 \quad \text{(dépend de } \theta\text{)}$$

Rem: la variance mesure la dispersion autour de l'espérance

Biais ou variance ?

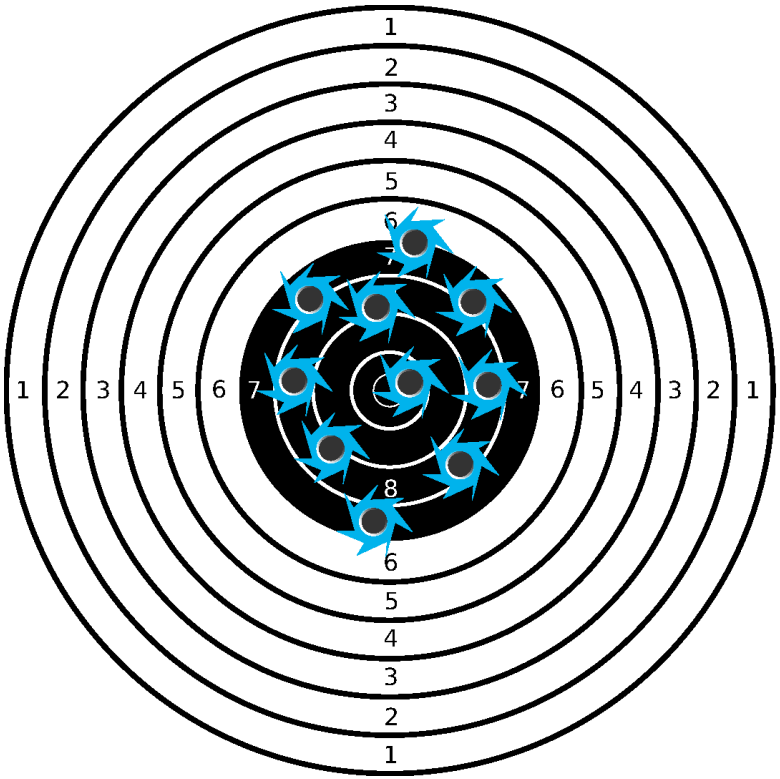


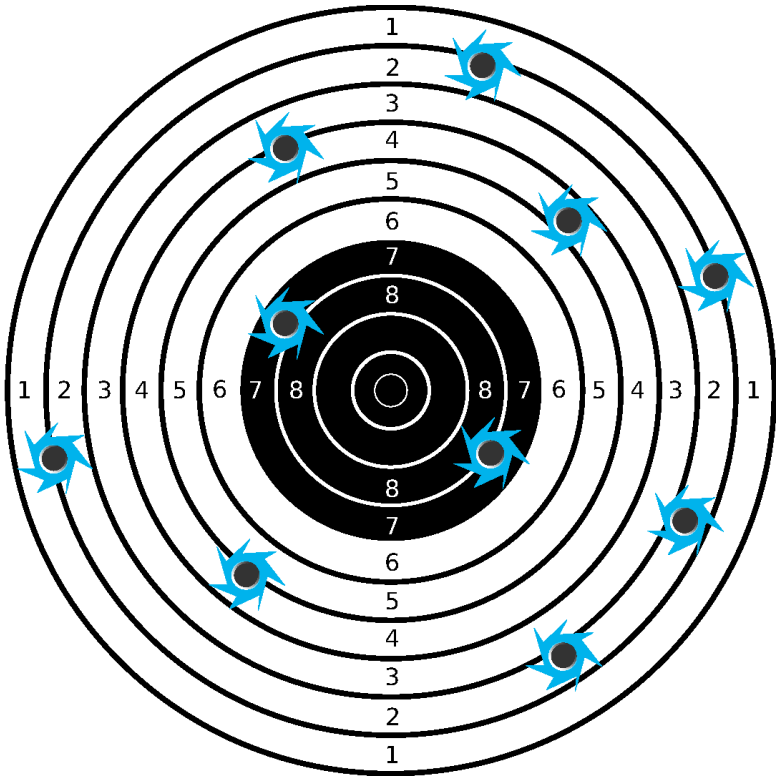
Erreurs systématiques



Erreurs stochastiques

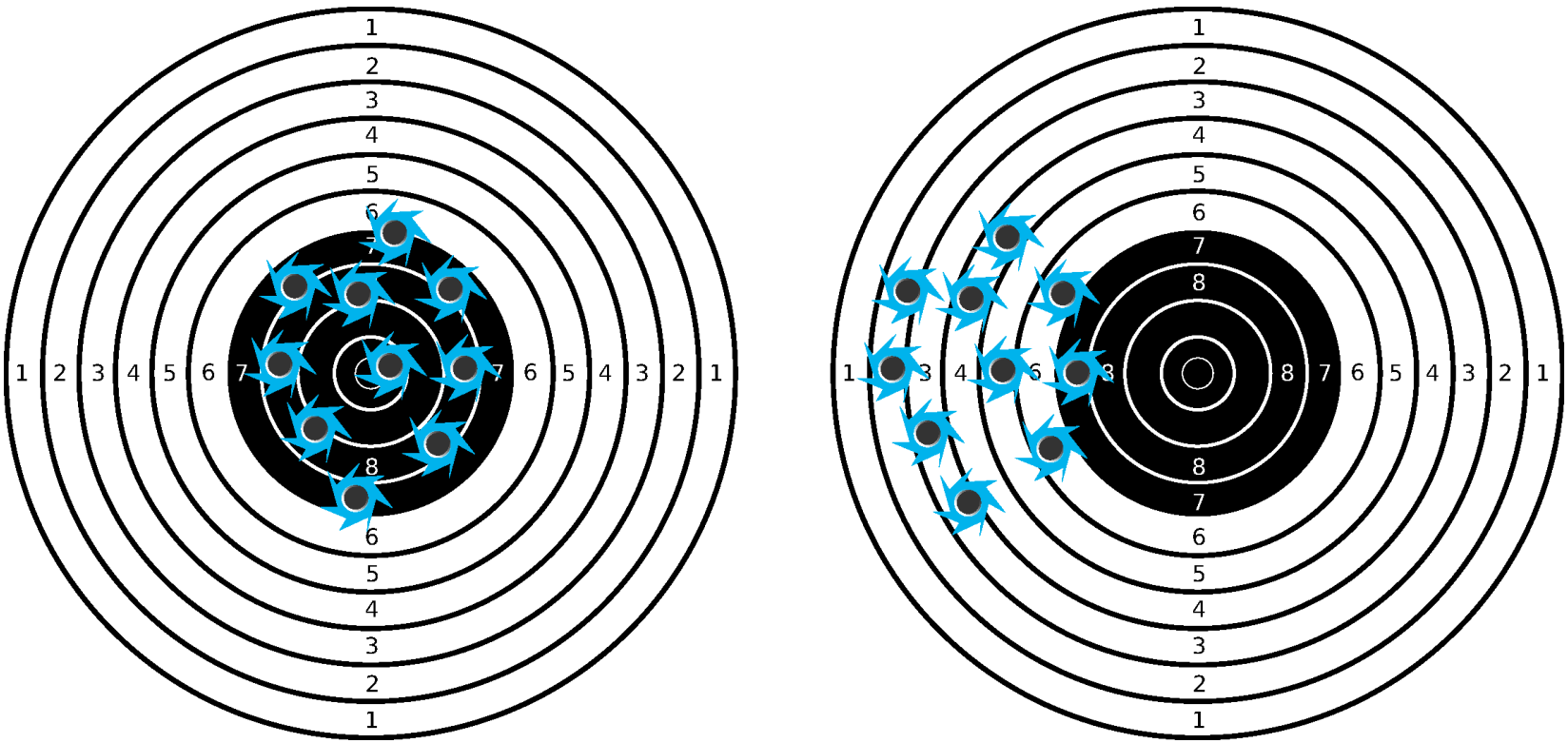
Biais ou variance ?





- Si $\hat{g}_0$ et $\hat{g}_1$ sont sans biais, on préfère avoir une faible variance

Biais ou variance ?



- Si $\hat{g}_0$ et $\hat{g}_1$ ont la même variance, on préfère un biais faible

Risque quadratique / compromis biais-variance

Définition

Le **risque quadratique** d'un estimateur $\hat{g}$ est l'espérance de son erreur au carré :

$$R(\hat{g}) = \mathbb{E}\left[(\hat{g} - g)^2\right]$$

Règle de choix : prendre l'estimateur dont le risque est le plus petit

Théorème : décomposition biais / variance

$$\text{Risque}(\hat{g}) = \text{Variance}(\hat{g}) + \big(\text{Biais}(\hat{g})\big)^2$$

Démonstration : faire apparaître le biais $B = \mathbb{E}[\hat{g}] - g$; développer

$$\begin{aligned} R(\hat{g}) &= \mathbb{E}\left[(\hat{g} - \mathbb{E}(\hat{g}) + B)^2\right] \\ &= \mathbb{E}\left[(\hat{g} - \mathbb{E}(\hat{g}))^2 + B^2 + 2B(\hat{g} - \mathbb{E}(\hat{g}))\right] \\ &= \mathrm{Var}(\hat{g}) + B^2 + 2B \underbrace{\mathbb{E}\left[\hat{g} - \mathbb{E}(\hat{g})\right]}_{=0} = \mathrm{Var}(\hat{g}) + B^2 \end{aligned}$$

Sommaire

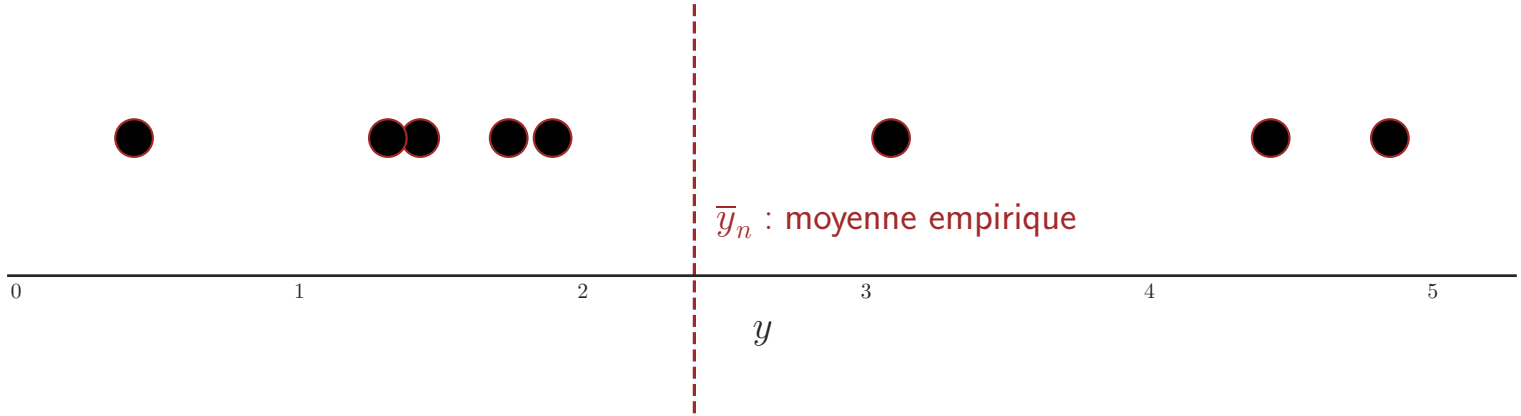
Statistique exploratoire et descriptive

- Première analyse sans hypothèse sur la loi $\mathbb{P}_Y$.
- Analyse qualitative du jeu de données / échantillon
- Visualisation du jeu de données / échantillon

Rappel : **statistique** $=$ **estimateur**, c'est une fonction (mesurable) des observations $(y_1, \ldots, y_n)$ (et qu'on espère être une fonction calculable des observations $(y_1, \ldots, y_n)$!)

Rem: les enjeux computationnels seront a prendre en compte dans la plupart de vos applications pratiques

Moyenne (arithmétique)



Définition

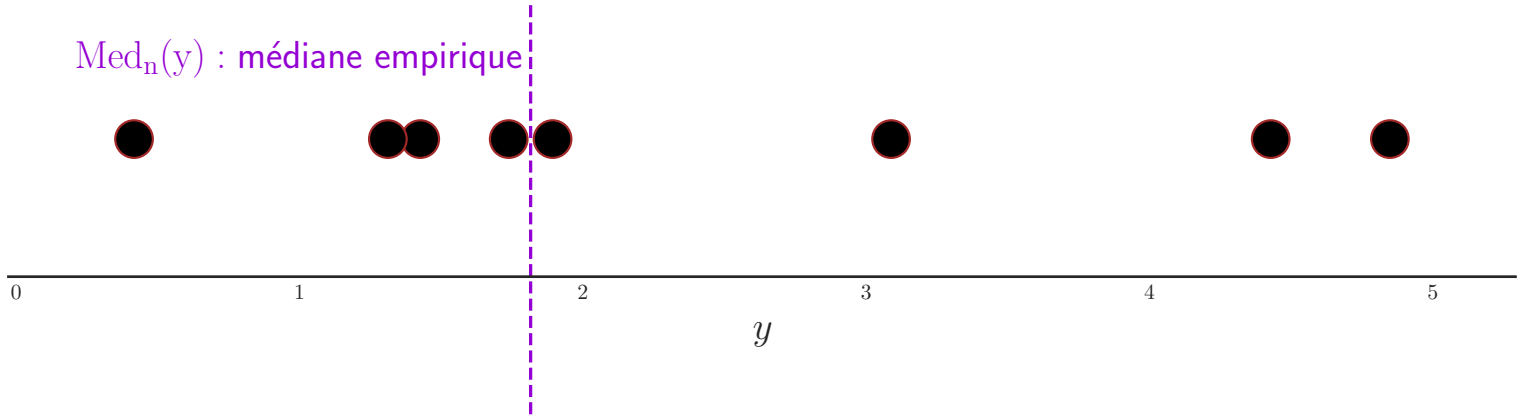
Moyenne (arithmétique) : $\overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$

Si $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i$ (produit scalaire) et $\mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$:

$$\overline{y}_n = \left\langle \mathbf{y}, \frac{\mathbf{1}_n}{n} \right\rangle$$

Exo: Le vecteur $\overline{y}_n \mathbf{1}_n$ est la projection de $\mathbf{y}$ sur l'espace $\mathrm{vect}(\mathbf{1}_n)$

Médiane



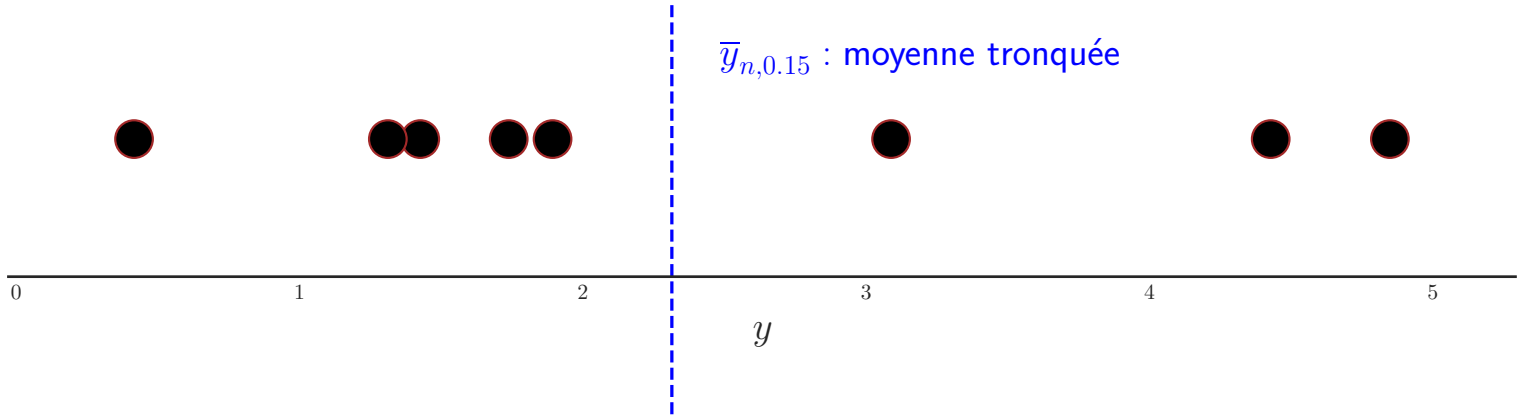
On ordonne les y_i dans l'ordre croissant : $y_{(1)} \leqslant y_{(2)} \leqslant \ldots \leqslant y_{(n)}$

Définition

$$\textbf{Médiane} : \mathrm{Med}_n(\mathbf{y}) = \begin{cases} \dfrac{y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}}{2}, & \text{si } n \text{ est pair} \\ y_{(\frac{n+1}{2})}, & \text{si } n \text{ est impair} \end{cases}$$

Rem: la définition d'une médiane est non-unique, et peut être parfois ambiguë...

Moyenne tronquée



Pour un paramètre α (*e.g.*, $\alpha = 15\%$), on calcule la moyenne en enlevant les $\alpha\%$ plus grandes et plus petites valeurs

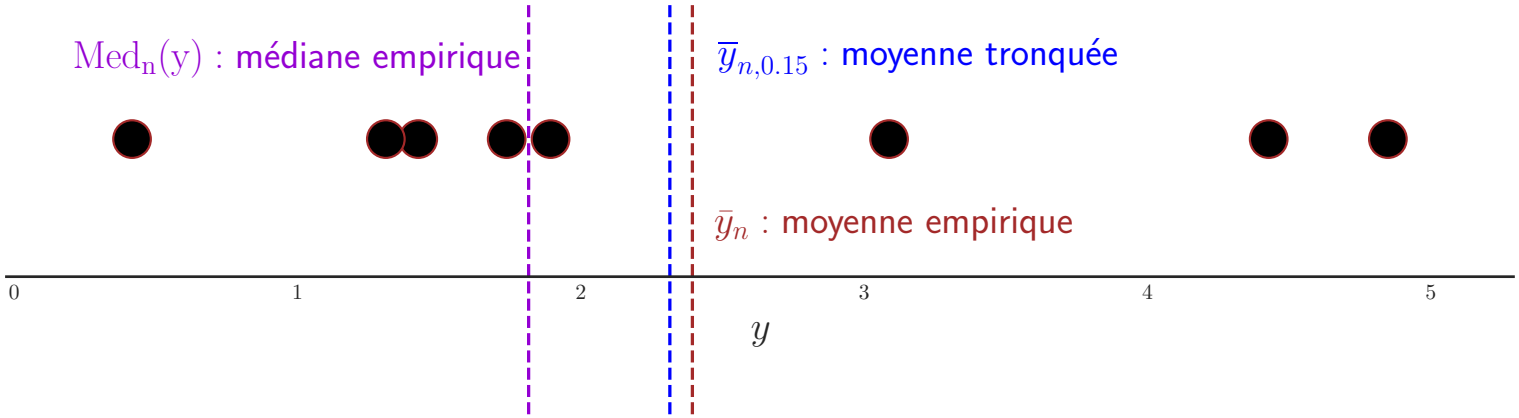
Définition

Moyenne tronquée (à l'ordre α) : $\overline{y}_{n,\alpha} = \overline{z}_n$

où $\mathbf{z} = (y_{(\lfloor \alpha n \rfloor)}, \ldots, y_{(\lfloor (1-\alpha) n \rfloor)})$ est l'échantillon α-tronqué

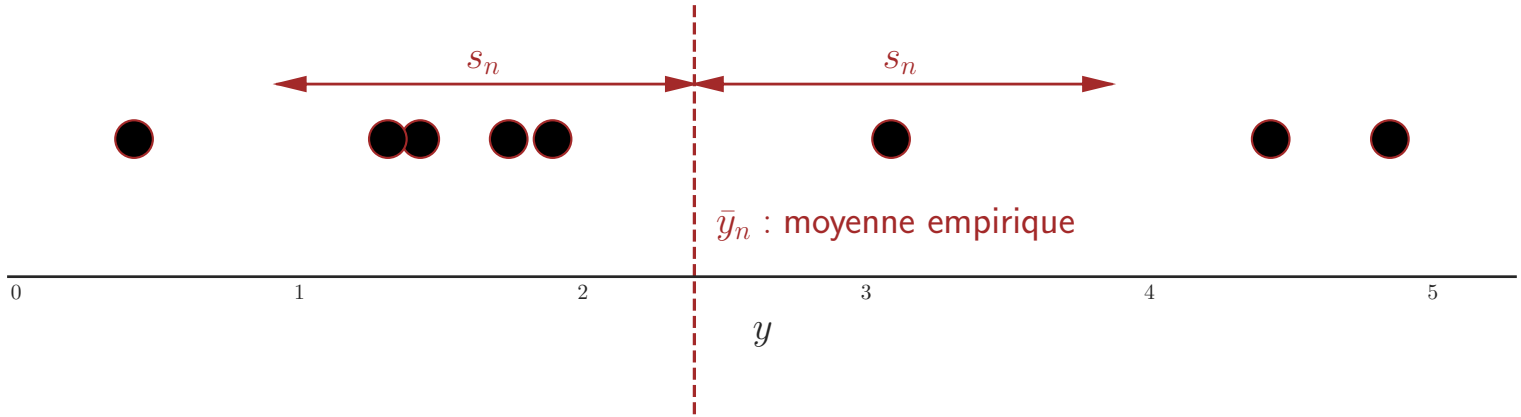
Rem: $\lfloor u \rfloor$ est le nombre entier tel que $\lfloor u \rfloor - 1 < u \leqslant \lfloor u \rfloor$

Moyenne *vs* médiane



- Les trois statistiques ne coïncident pas
- Moyennes tronquées et médianes sont robustes aux points atypiques (🇬🇧 : *outliers*), la moyenne non !

Dispersion : variance / écart-type



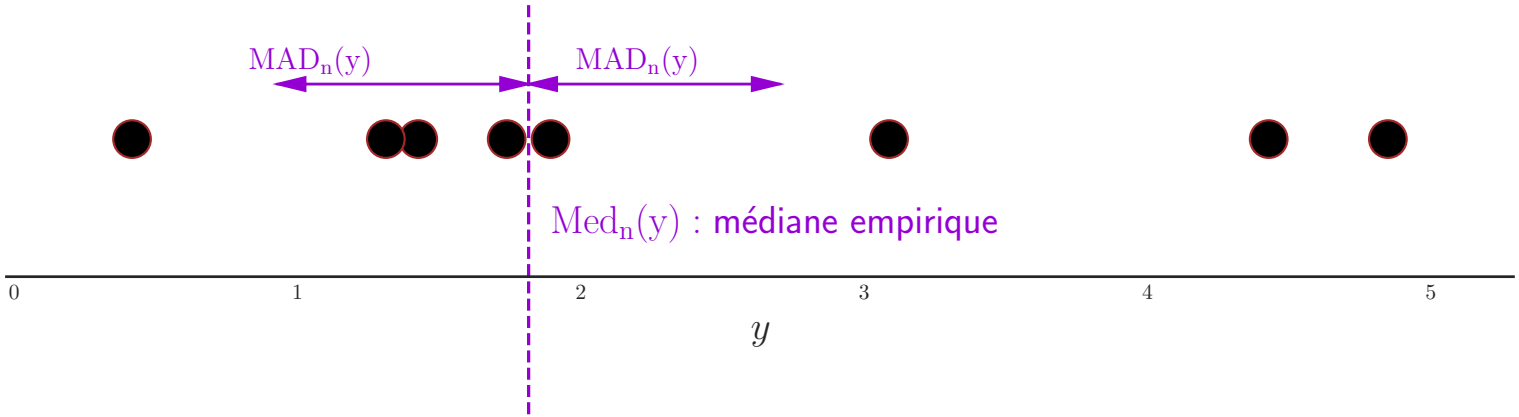
Définitions

Variance : $$\mathrm{var}_n(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2 = \frac{1}{n}\left\|\mathbf{y} - \overline{y}_n\mathbf{1}_n\right\|^2$$

Écart-type : $$s_n(\mathbf{y}) = \sqrt{\mathrm{var}_n(\mathbf{y})} \quad (\text{où } \|\mathbf{z}\|^2 = \sum_{i=1}^{n} z_i^2)$$

Exo: Quels sont les vecteurs $\mathbf{y} \in \mathbb{R}^n$ tels que $\mathrm{var}_n(\mathbf{y}) = 0$?

Dispersion : MAD

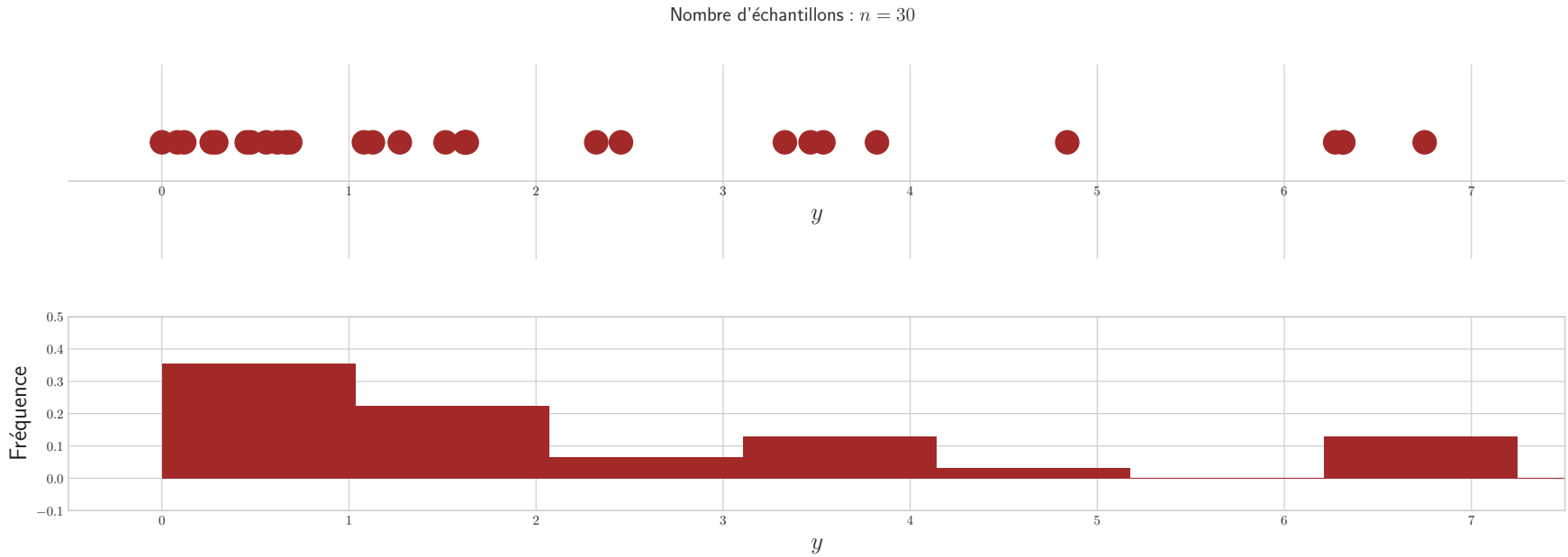


Définition

Déviation médiane absolue (🇬🇧 : *Mean Absolute Deviation*) :

$$\mathrm{MAD}_n(\mathbf{y}) = \mathrm{Med}_n\left(|\mathrm{Med}_n(\mathbf{y}) - \mathbf{y}|\right)$$

Estimation de la densité : histogramme

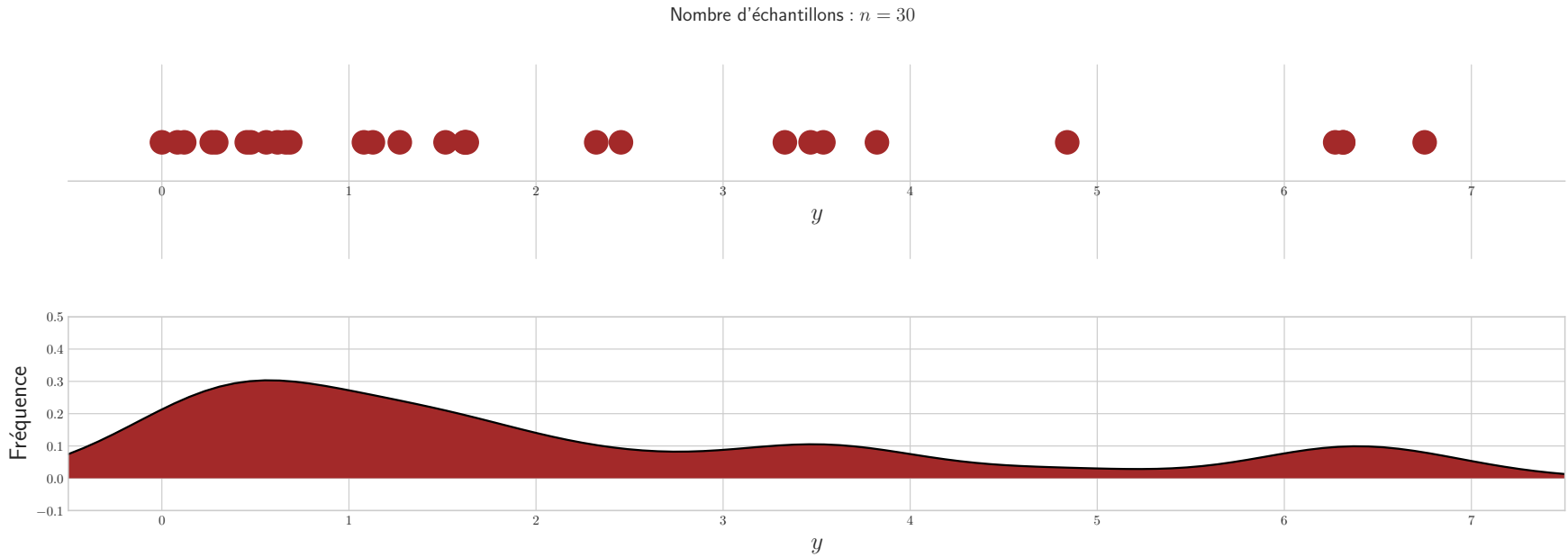


L'**histogramme** est une approximation de la **densité** par une fonction constante par morceaux

Rem: les « cases » (🇬🇧 : *bins*) ont une aire proportionnelle au nombre de données qu'elles contiennent

Rem: en Python, on compte le nombre ou la proportion de données par case, *e.g.*, avec `normed=False(True)` dans la fonction `hist`

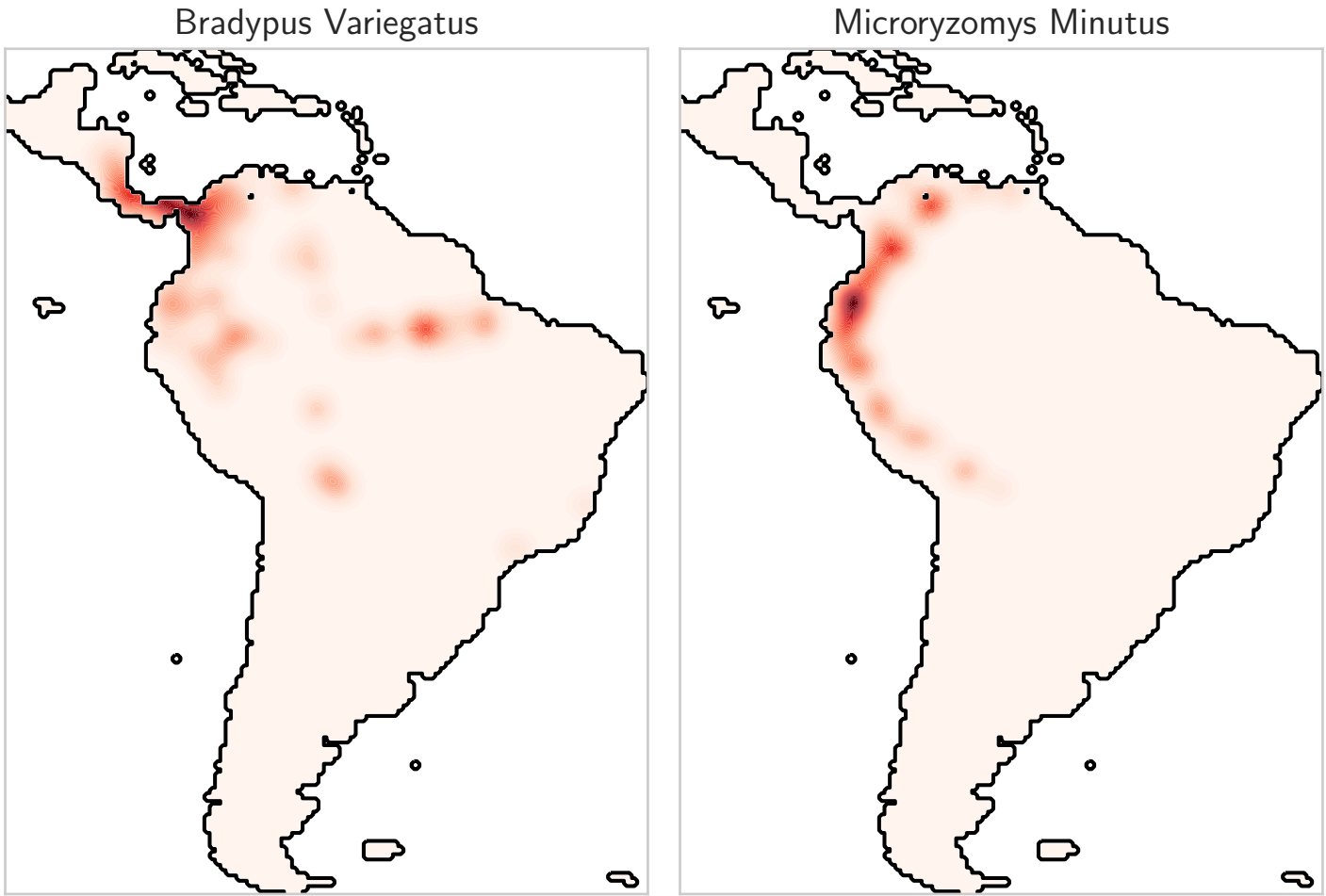
Estimation de la densité : méthode à noyau



- Méthode à noyau (🇬🇧 : *Kernel Density Estimation*, *KDE*) : approche non-paramétrique estimant la densité par une fonction continue – généralisation de l'histogramme

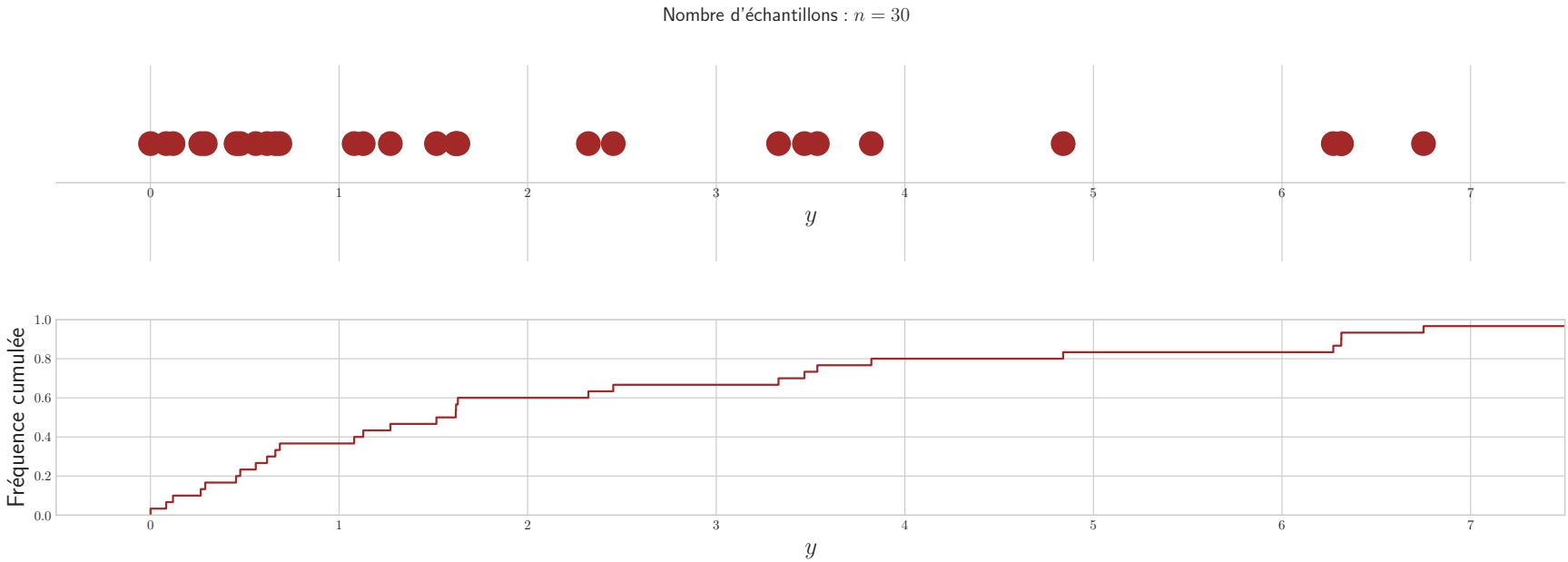
Pour plus de détails voir le livre Silverman (1986)

Densité bi-dimensionnelle (spatiale)



cf. http://scikit-learn.org/stable/_downloads/plot_species_kde.py

Fonction de répartition



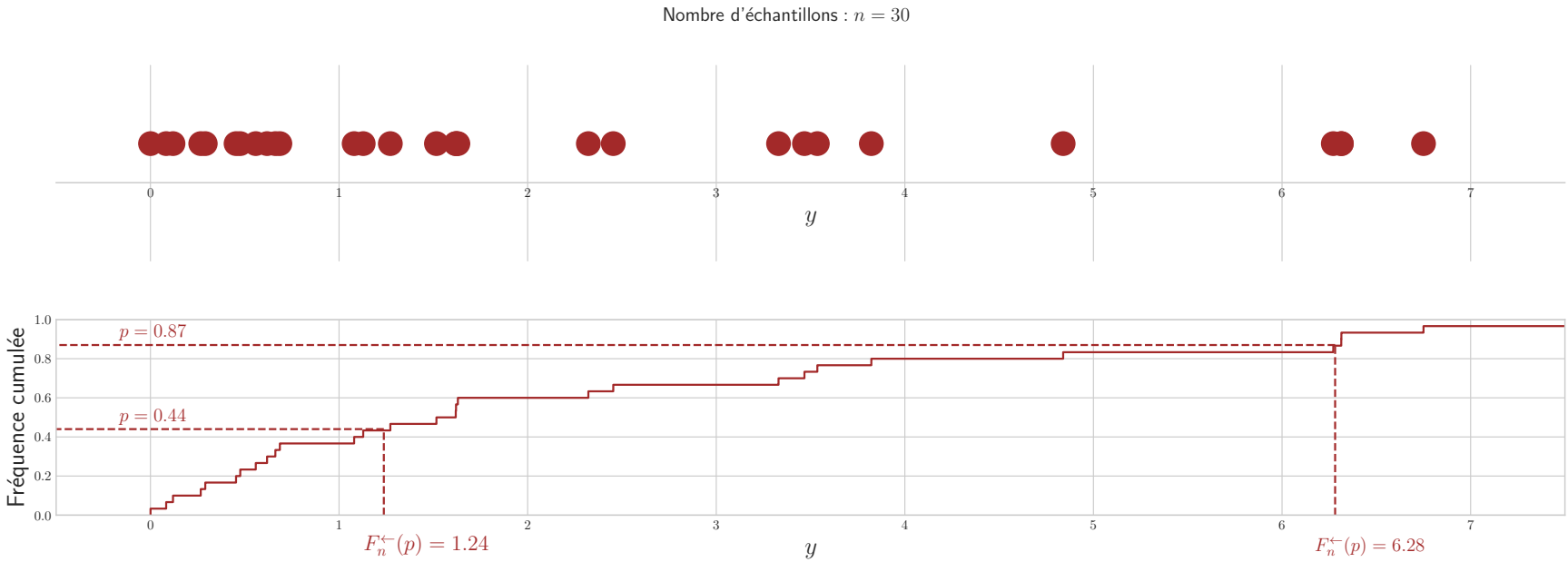
Définition : fonction de répartition

Théorique : $$F(u) = \mathbb{P}(Y \leqslant u) = \int_{-\infty}^{u} f_Y(x)dx$$

Empirique : $$F_n(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i \leqslant u\}}$$

Interprétation : proportion d'observations sous un certain niveau

Fonction quantile



Définition

Pour $p \in]0, 1]$,

Quantile théorique (d'ordre p) : $F^{\leftarrow}(p) = \inf\{u \in \mathbb{R} : F(u) \geqslant p\}$

Quantile empirique (d'ordre p) : $F_n^{\leftarrow}(p) = y_{(\lfloor np \rfloor)}$

Rem: c'est l'inverse (généralisée) de la fonction de répartition

Sommaire

Covariances et corrélations empiriques

Covariance empirique

Pour deux échantillons $\mathbf{x}$ et $\mathbf{y}$ de moyennes et variances empiriques $\overline{x}_n$, $\overline{y}_n$ et $\mathrm{var}_n(\mathbf{x})$, $\mathrm{var}_n(\mathbf{y})$:

$$\mathrm{cov}_n(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) \quad \text{c'est-à-dire}$$

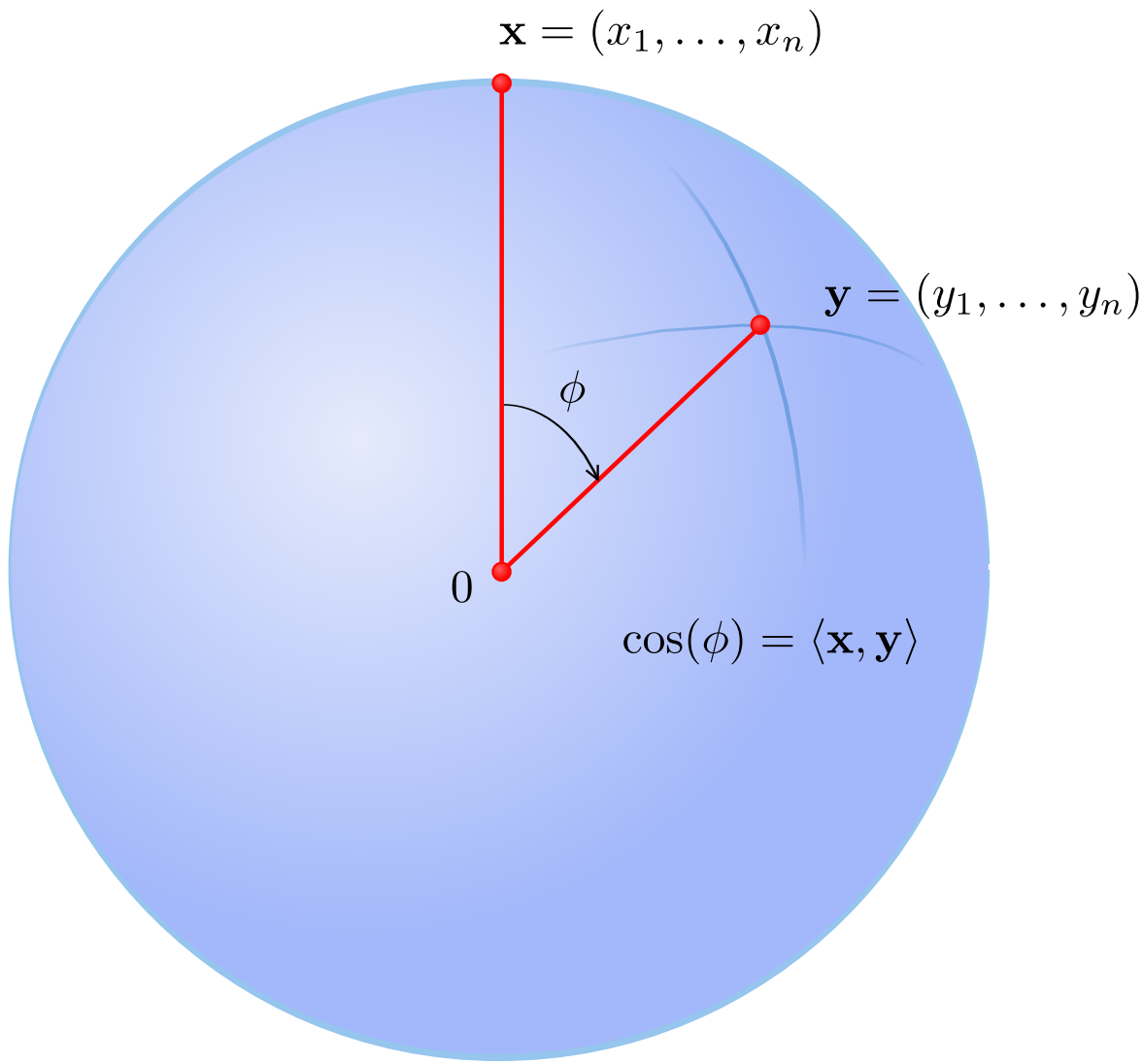
$$\mathrm{cov}_n(\mathbf{x},\mathbf{y}) = \frac{1}{n}\langle \mathbf{x} - \overline{x}_n\mathbf{1}_n, \mathbf{y} - \overline{y}_n\mathbf{1}_n\rangle$$

Corrélation empirique

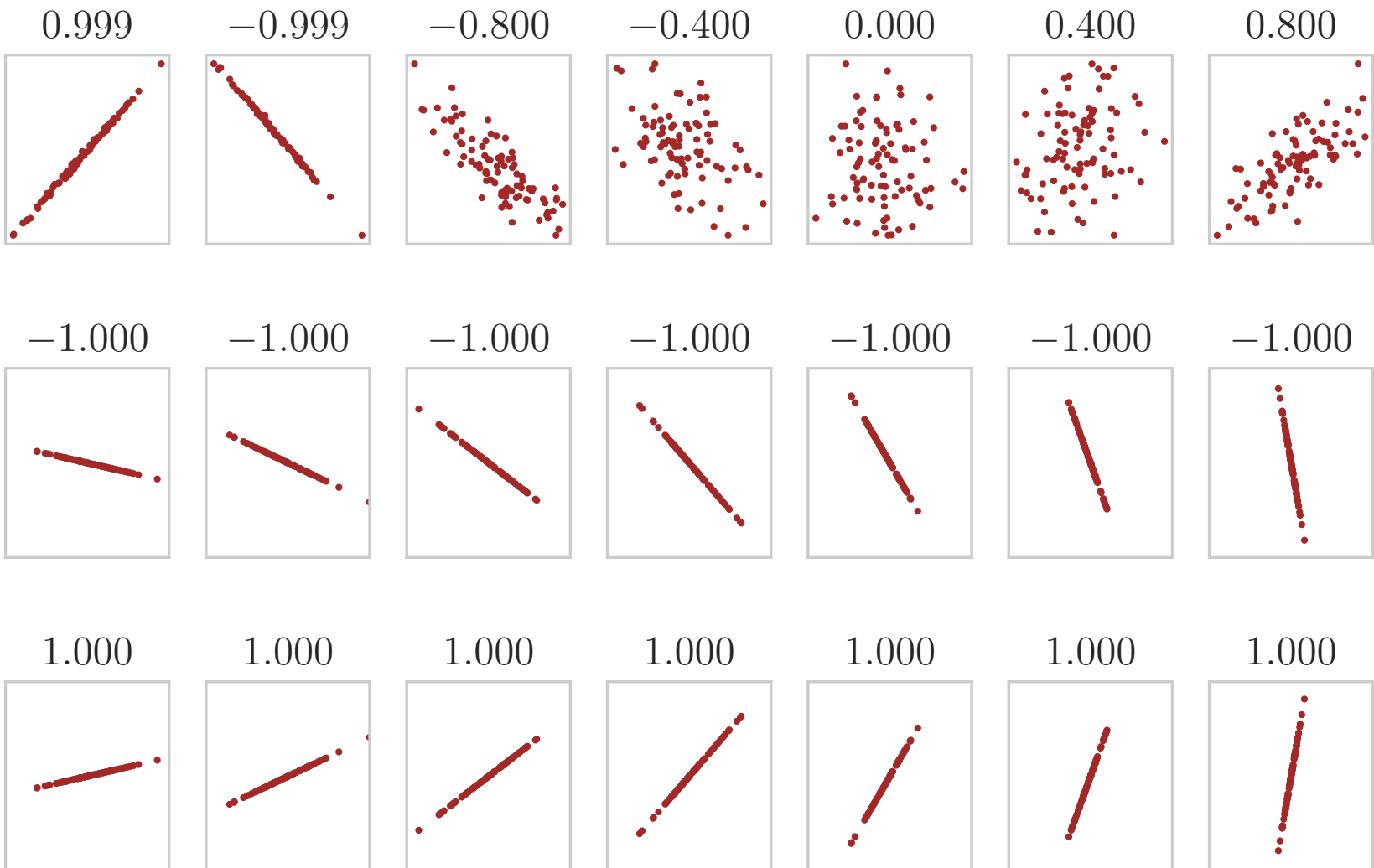
$$\rho = \mathrm{corr}_n(\mathbf{x},\mathbf{y}) = \frac{\mathrm{cov}_n(\mathbf{x},\mathbf{y})}{\sqrt{\mathrm{var}_n(\mathbf{x})}\sqrt{\mathrm{var}_n(\mathbf{y})}}, \quad \text{c'est-à-dire}$$

$$\rho = \frac{\langle \mathbf{x} - \overline{x}_n\mathbf{1}_n, \mathbf{y} - \overline{y}_n\mathbf{1}_n\rangle}{\|\mathbf{x} - \overline{x}_n\|\ \|\mathbf{y} - \overline{y}_n\|} = \cos(\mathbf{x} - \overline{x}_n\mathbf{1}_n, \mathbf{y} - \overline{y}_n\mathbf{1}_n)$$

Interprétation pour $n = 3$ et $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$

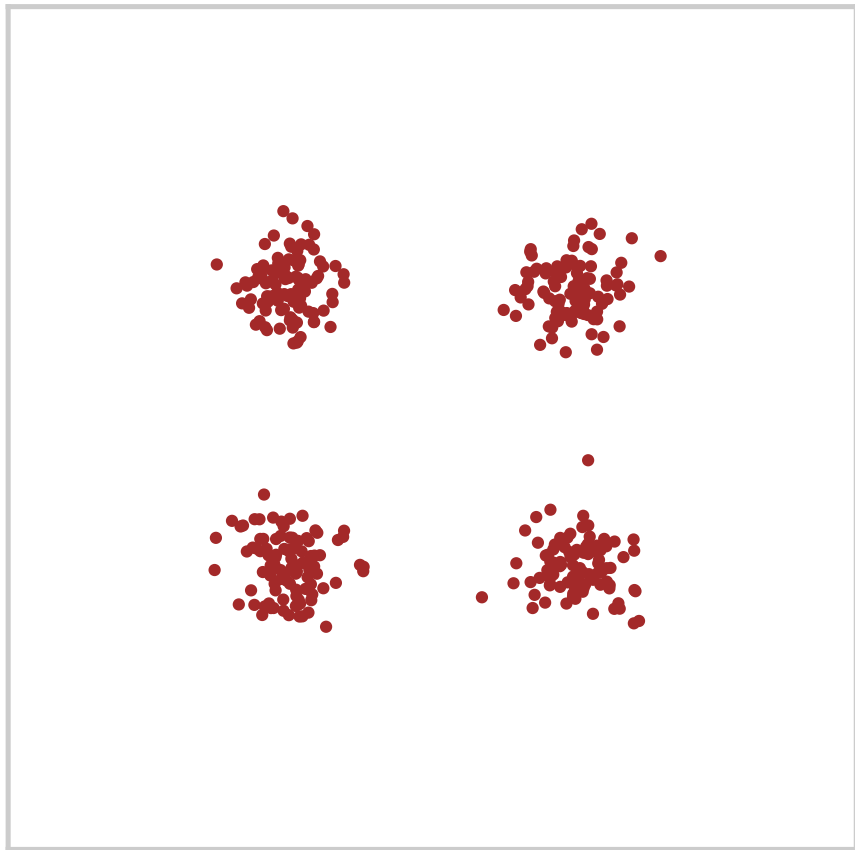


Exemples de corrélations



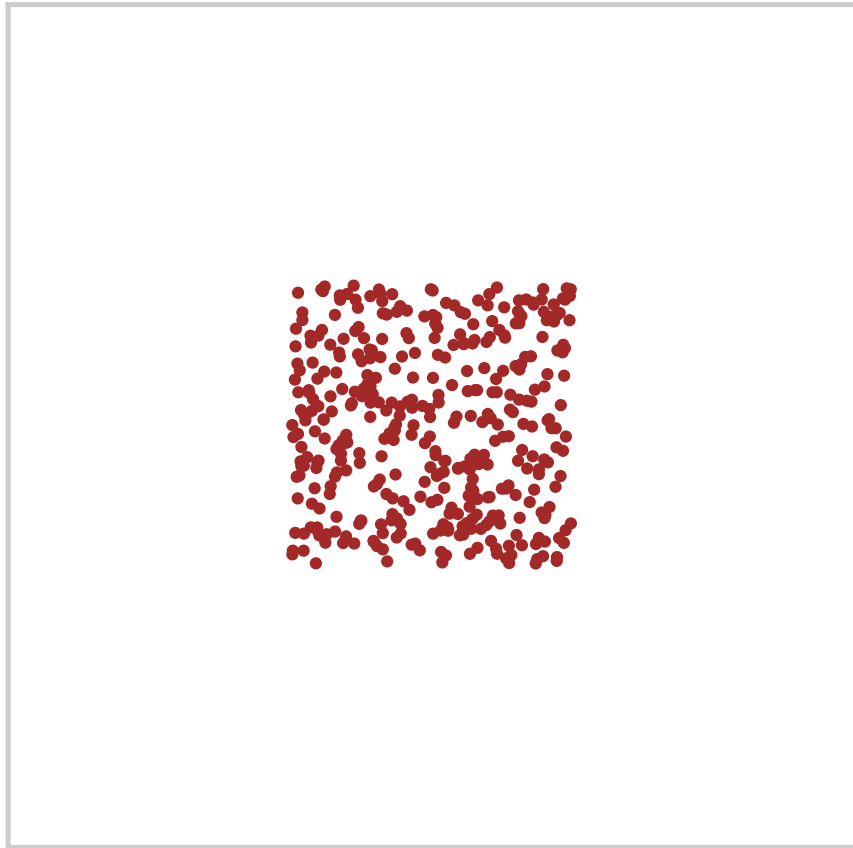
Exemples de corrélations proches de zéro

Corrélation $= -0.021$



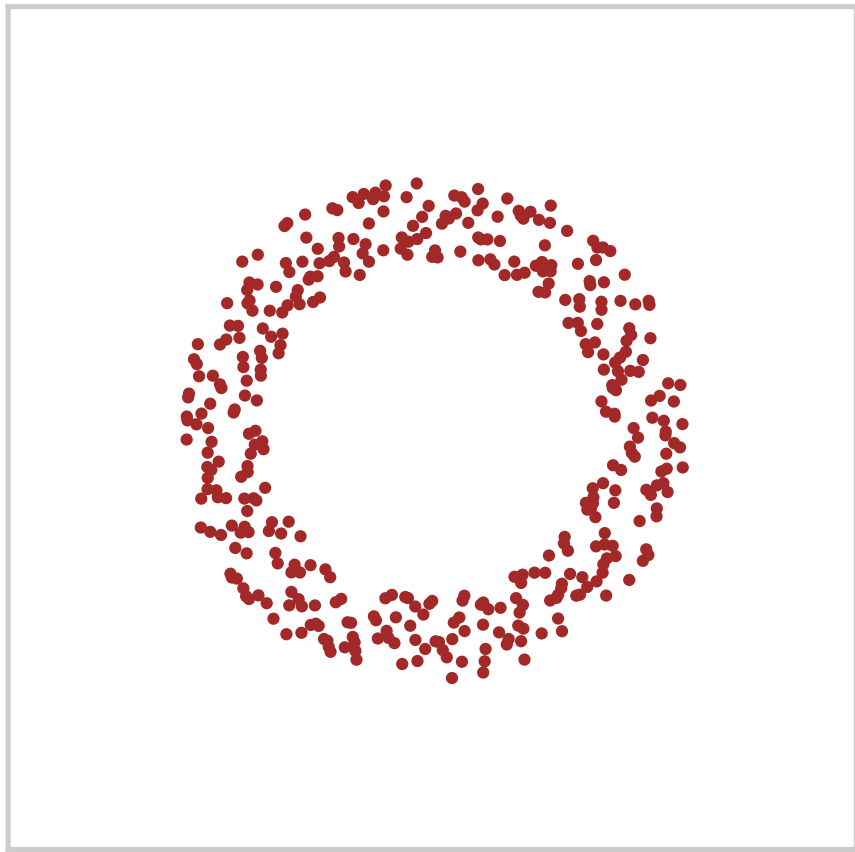
Exemples de corrélations proches de zéro

Corrélation $= 0.007$

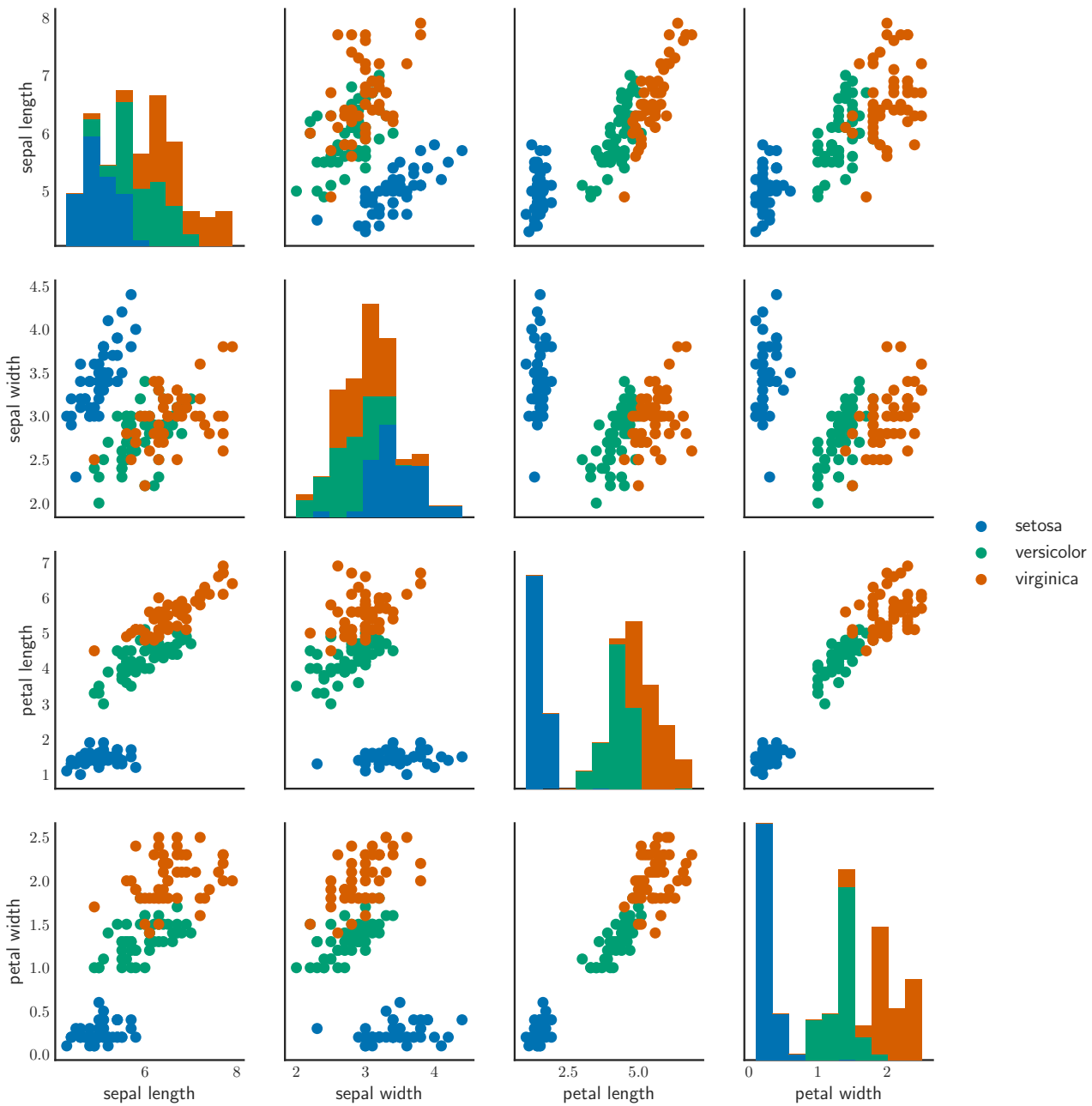


Exemples de corrélations proches de zéro

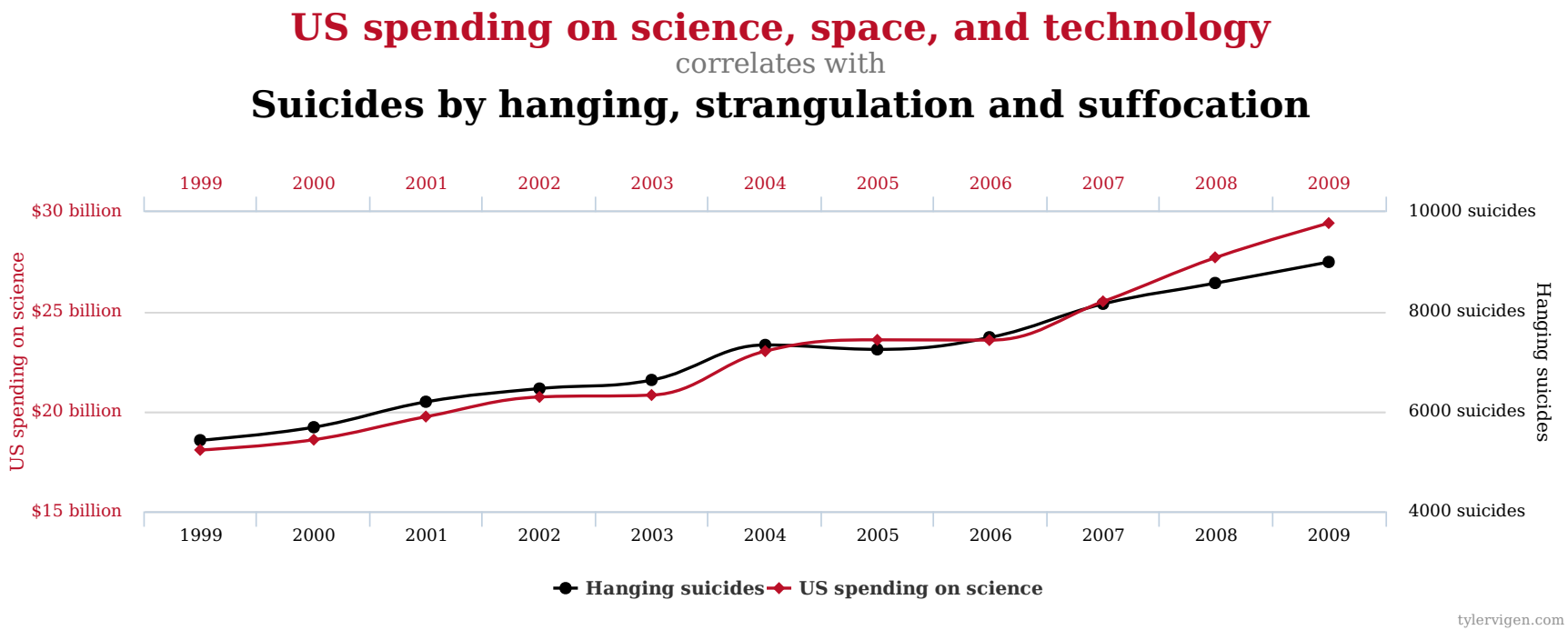
Corrélation $= 0.011$



Nuages de points / Scatter plot / PairGrid



Covariance $\neq$ causalité



Corrélation : 0.9979

cf. http://www.tylervigen.com/spurious-correlations

Sommaire

Covariance d'un couple de *v.a.*

Soient X et Y des variables aléatoires réelles de carré intégrable.

Définition

La **covariance** de X et Y est la moyenne des fluctuations jointes :

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big]$$

Propriété : la covariance est bilinéaire, pour tous $\alpha, \beta \in \mathbb{R}$ et toutes variables aléatoires réelles X_1, X_2, Y_1, Y_2 on a

$$\begin{aligned}\mathrm{Cov}(\alpha X_1 + \beta X_2, Y_1) &= \alpha\,\mathrm{Cov}(X_1, Y_1) + \beta\,\mathrm{Cov}(X_2, Y_1)\\ \mathrm{Cov}(X_1, \alpha Y_1 + \beta Y_2) &= \alpha\,\mathrm{Cov}(X_1, Y_1) + \beta\,\mathrm{Cov}(X_1, Y_2)\end{aligned}$$

Rappel : inégalité de Cauchy–Schwarz dans ce cadre

$$|\,\mathrm{Cov}(X,Y)| \leqslant \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$$

Matrice de covariance d'un vecteur aléatoire

Notation : $X = (X_1, \dots, X_p)^\top$ est vecteur aléatoire t.q.
$\forall j \in [\![1, p]\!], \mathbb{E}(X_j^2) < +\infty$ et $\sigma_{i,j} = \mathrm{cov}(X_i, X_j)$ $(\sigma_{i,i} = \mathrm{var}(X_i))$

Définition

La **matrice de covariance** du vecteur X est la matrice $\mathrm{Cov}(X)$, de taille $p \times p$, formée par les $\sigma_{i,j}$ (i^{e} ligne, j^{e} colonne). Ainsi,

$$\mathrm{Cov}(X) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \dots & \mathrm{cov}(X_1,X_p) \\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{cov}(X_p,X_1) & \dots & & \mathrm{var}(X_p) \end{pmatrix} \in \mathbb{R}^{p\times p}$$

Version condensée : $\mathrm{Cov}(X) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(X - \mathbb{E}(X)\big)^\top\Big]$

Exo: Montrer que pour μ déterministe $\mathrm{Cov}(X + \mu) = \mathrm{Cov}(X)$

Quelques propriétés de la covariance

- Une matrice de covariance est symétrique :

$$\mathrm{Cov}(X) = \mathrm{Cov}(X)^\top \Leftrightarrow \forall (i,j) \in [\![1,p]\!]^2,\ \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$$

- Une matrice de covariance est (semi-définie) positive :

$$\forall u \in \mathbb{R}^p,\ u^\top \mathrm{Cov}(X) u \geqslant 0$$

Démonstration :

$$u^\top \mathrm{Cov}(X) u = \sum_{i=1}^p \sum_{j=1}^p u_i u_j \mathrm{Cov}(X_i, X_j) = \underbrace{\mathrm{Cov}\left(\sum_{i=1}^p u_i X_i, \sum_{j=1}^p u_j X_j\right)}_{=\mathrm{Var}\left(\sum_{j=1}^p u_j X_j\right) \geqslant 0}$$

Exo: $\mathrm{Cov}(AX) = A\,\mathrm{Cov}(X) A^\top$, pour toute matrice $A \in \mathbb{R}^{m \times p}$

La décomposition spectrale

Théorème spectral

Une matrice symétrique $S \in \mathbb{R}^{n\times n}$ est diagonalisable en base orthonormée, *i.e.*, il existe $\lambda_1 \geqslant \cdots \geqslant \lambda_n$ et une matrice orthogonale $U \in \mathbb{R}^{n\times n}$ telle que :

$$S = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^\top \text{ ou } SU = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

Rappel : une matrice orthogonale $U \in \mathbb{R}^n$ est une matrice telle que $U^\top U = UU^\top = \mathrm{Id}_n$ ou $\forall (i,j) \in [\![1,n]\!], \mathbf{u}_i^\top \mathbf{u}_j = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}$

Rem: si l'on écrit $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ cela signifie que :

$$S = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \quad \text{et} \quad \forall i \in [\![1,n]\!],\ S\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

Vocabulaire :

- les λ_i sont les **valeurs propres** de S (🇬🇧 : *eigenvalues*)
- les $\mathbf{u}_i$ sont les **vecteurs propres** de S (🇬🇧 : *eigenvectors*)

La décomposition spectrale : exemple

$$A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 2 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 \\ 2 & 0 & 2 & 1 \end{pmatrix} = UDU^\top$$

avec

$$D = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -3 \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{\sqrt{2}}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{\sqrt{2}}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{\sqrt{2}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{2}}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

La décomposition spectrale : numérique

```
import numpy as np
from scipy.linalg import toeplitz
from numpy.linalg import eigh

A = toeplitz([1, 2, 0, 2])
[Dint, Uint] = eigh(A)
# use eigh not eig for symmetric matrices

idx = Dint.argsort()[::-1]
D = Dint[idx]
U = Uint[:, idx]

print(np.allclose(U.dot(np.diag(D)).dot(U.T), A))
```

Sommaire

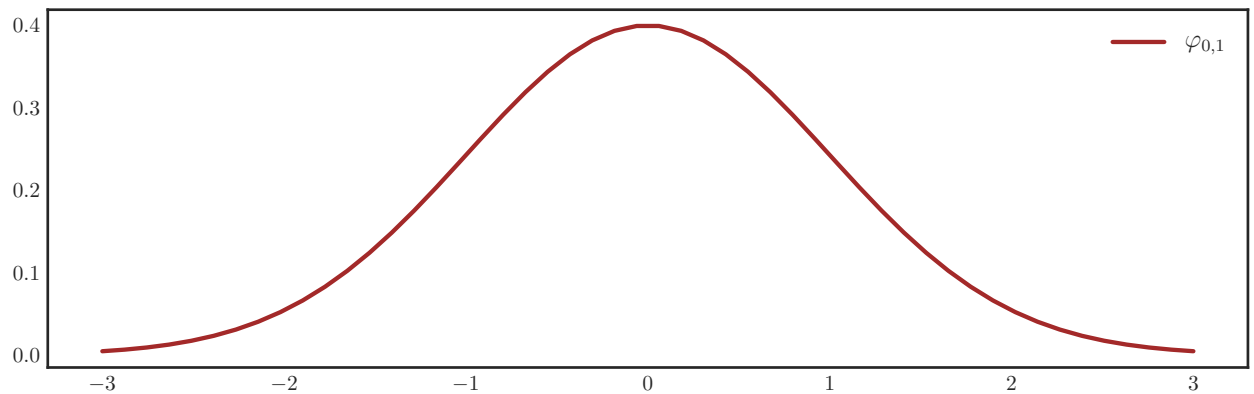
Loi normale unidimensionnelle

- Une v.a. réelle X suit une « **loi normale** standard » (ou « **loi gaussienne** » ou « loi de Laplace-Gauss ») si sa densité vaut

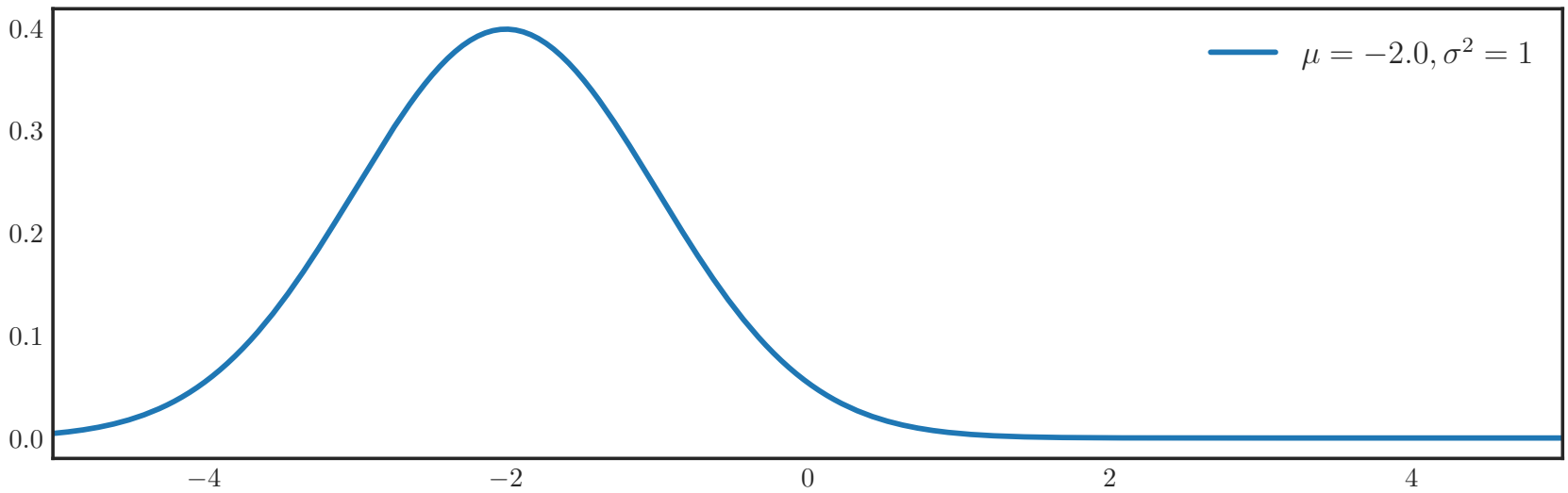
$$\varphi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

  On note $X \sim \mathcal{N}(0,1)$.
- Une v.a. Y suit une loi normale de paramètres μ et σ^2 si $Y = \mu + \sqrt{\sigma^2} X$, où $X \sim \mathcal{N}(0,1)$, et on note $Y \sim \mathcal{N}(\mu, \sigma^2)$

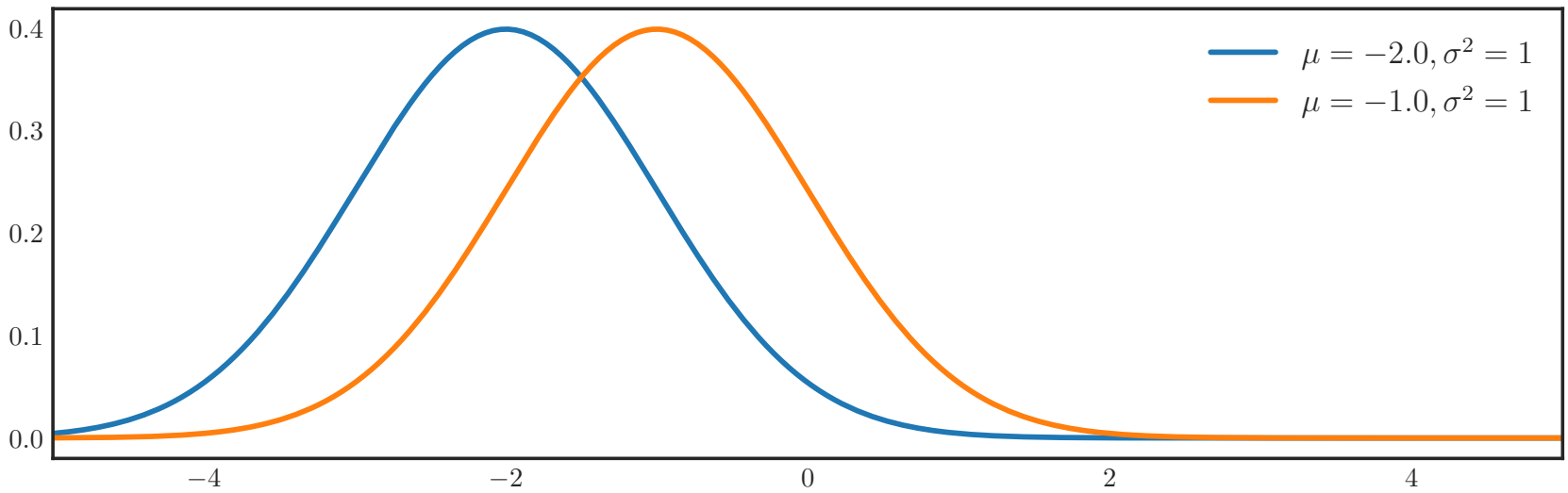
Densité : $\boxed{\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$



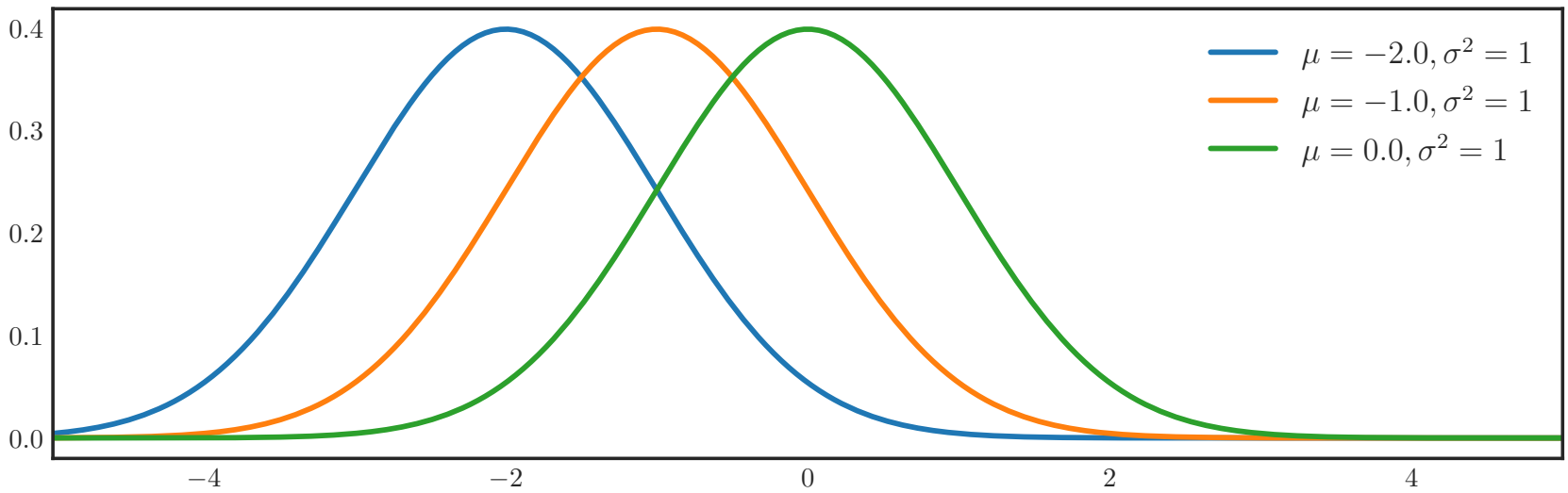
Exemple : variation sur μ



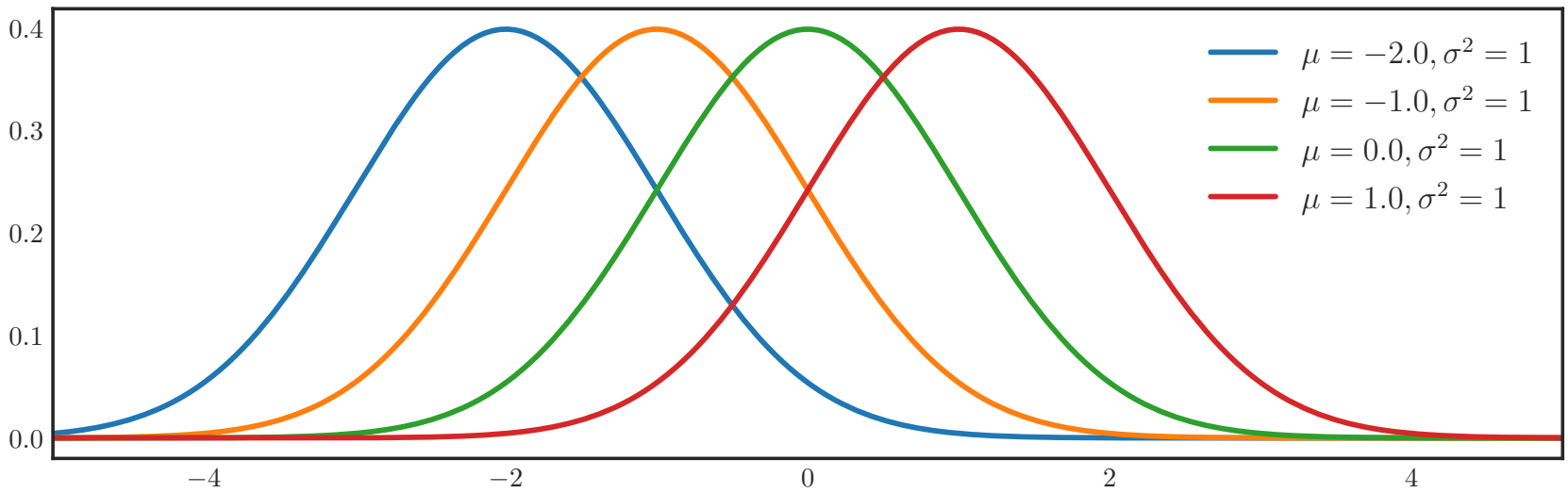
Exemple : variation sur μ



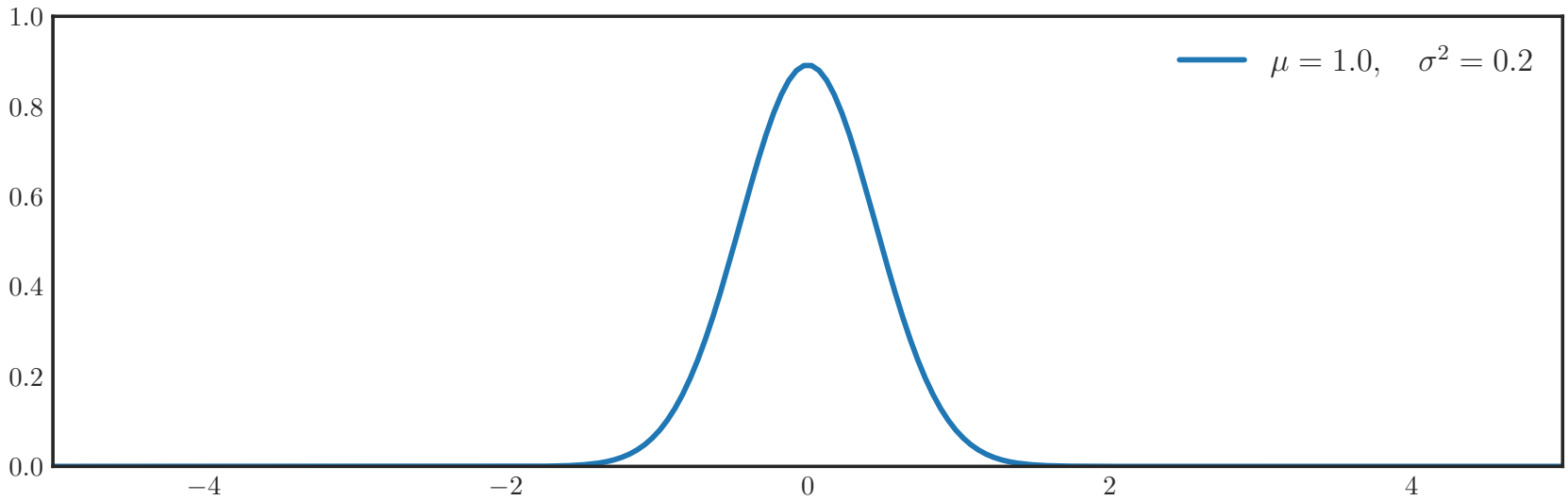
Exemple : variation sur μ



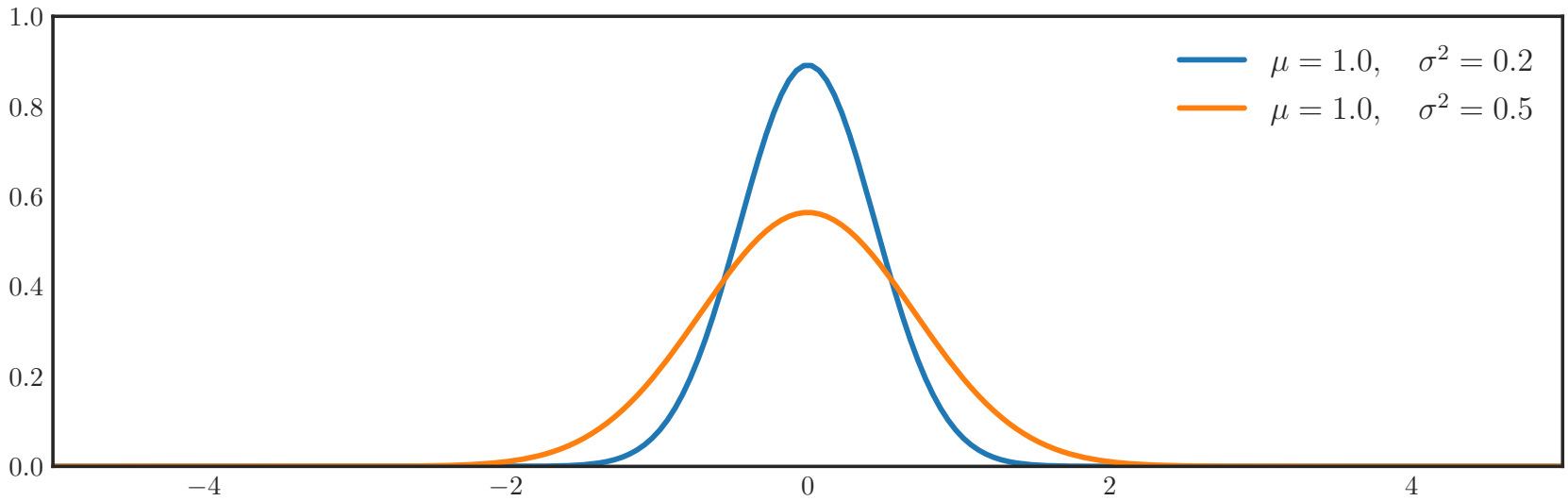
Exemple : variation sur μ



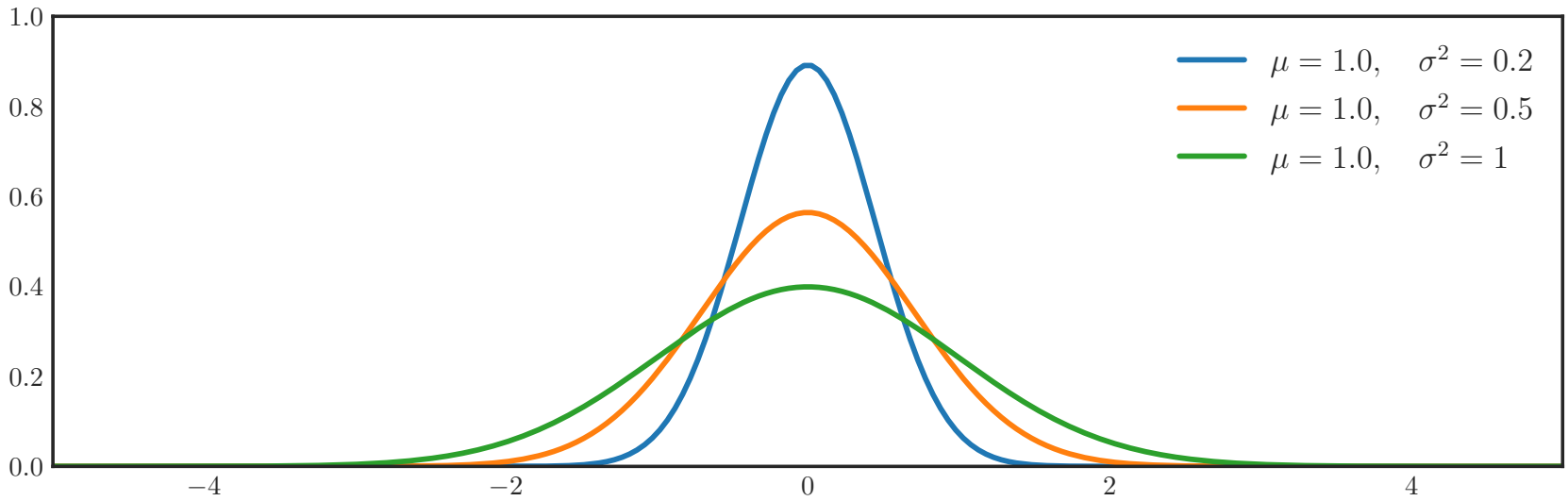
Exemple : variation sur σ



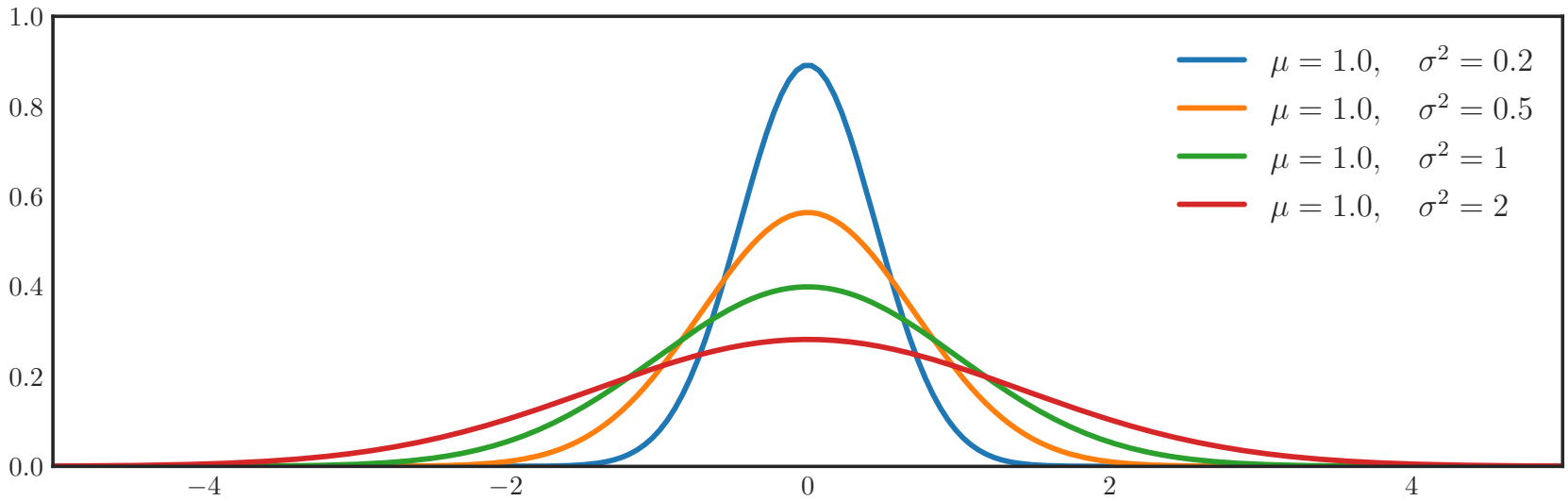
Exemple : variation sur σ



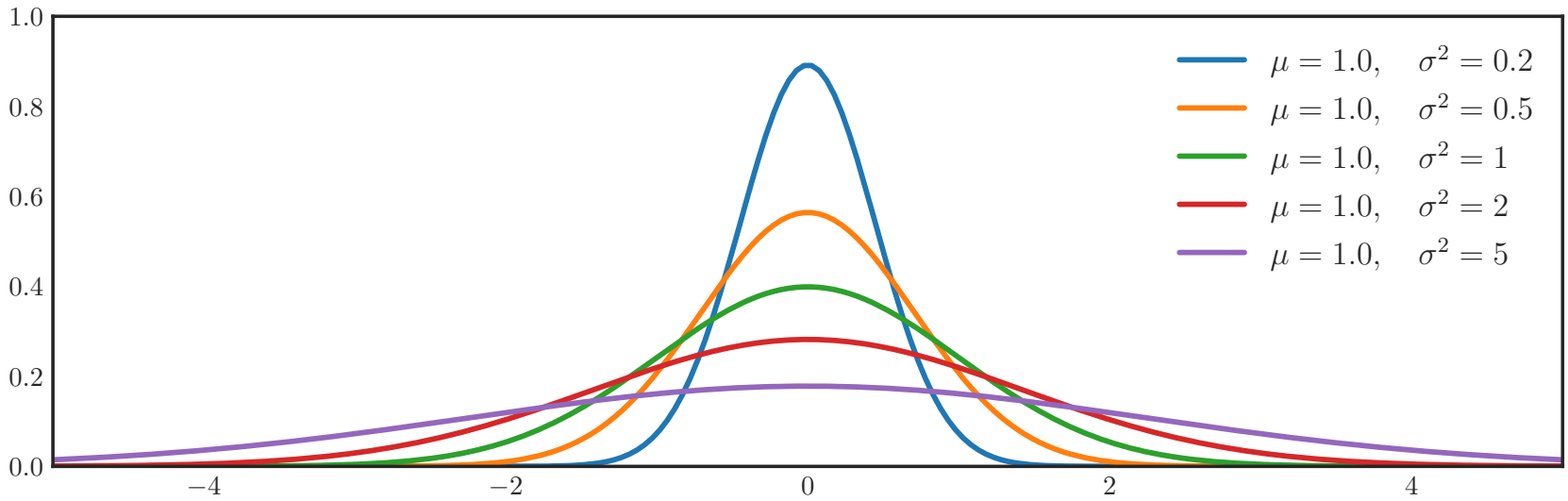
Exemple : variation sur σ



Exemple : variation sur σ



Exemple : variation sur σ



Vecteurs Gaussiens

En dimension p, les lois gaussiennes ont des densités de la forme :

$$\boxed{\varphi_{\mu,\Sigma}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}}.$$

La fonction $\varphi_{\mu,\Sigma}$ est gouvernée par deux paramètres :

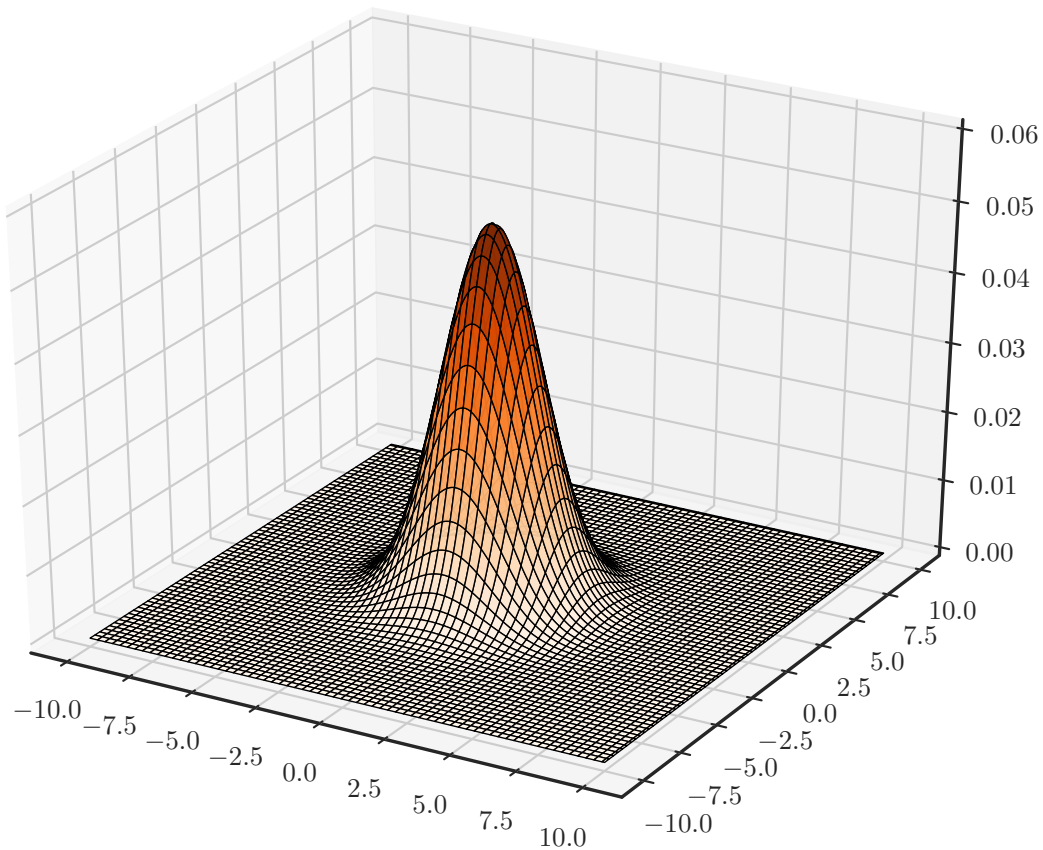
- le vecteur d'espérance $\mu \in \mathbb{R}^p$
- la matrice de covariance $\Sigma \in \mathbb{R}^{p\times p}$

Notation : lorsque le vecteur aléatoire X suit une loi normale d'espérance μ et de covariance Σ, on note $X \sim \mathcal{N}(\mu, \Sigma)$ qu'on suppose définie positive

Rem: $|\Sigma| = \det(\Sigma)$ est le produit des valeurs propres de Σ. On parle de cas dégénéré quand $\det(\Sigma) = 0$

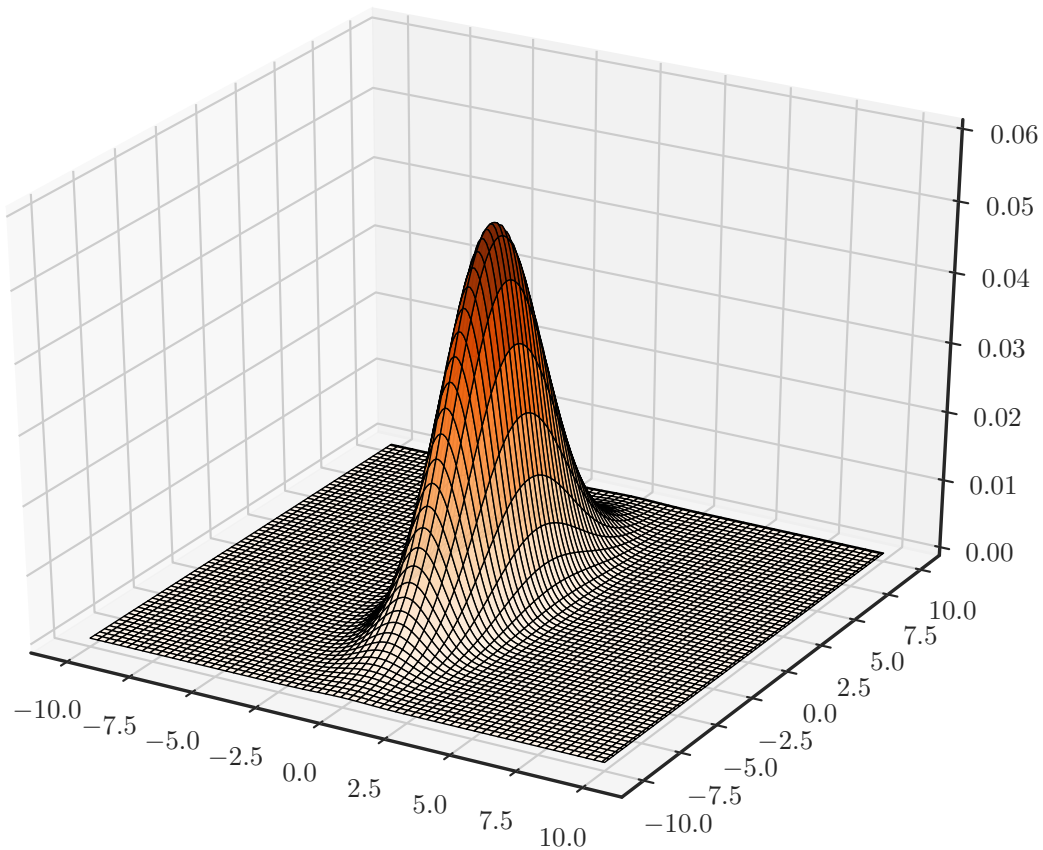
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 3 & \\ & 3 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix},$$



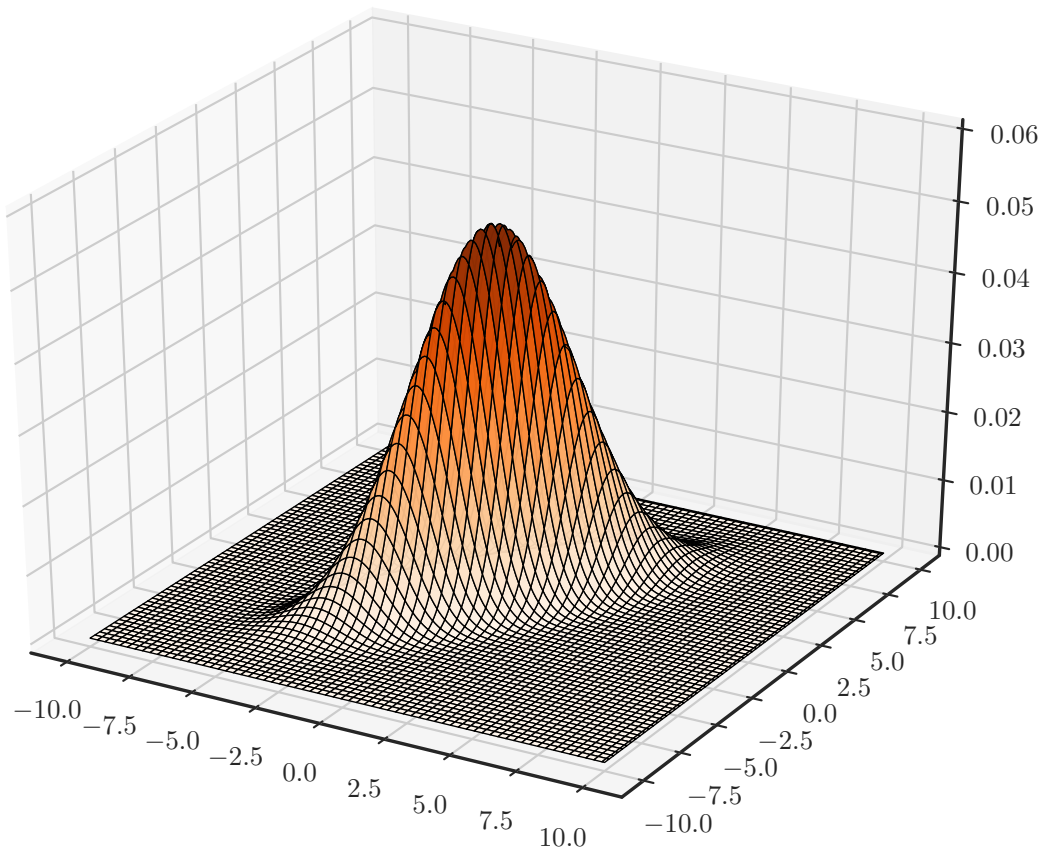
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = \qquad 0$$



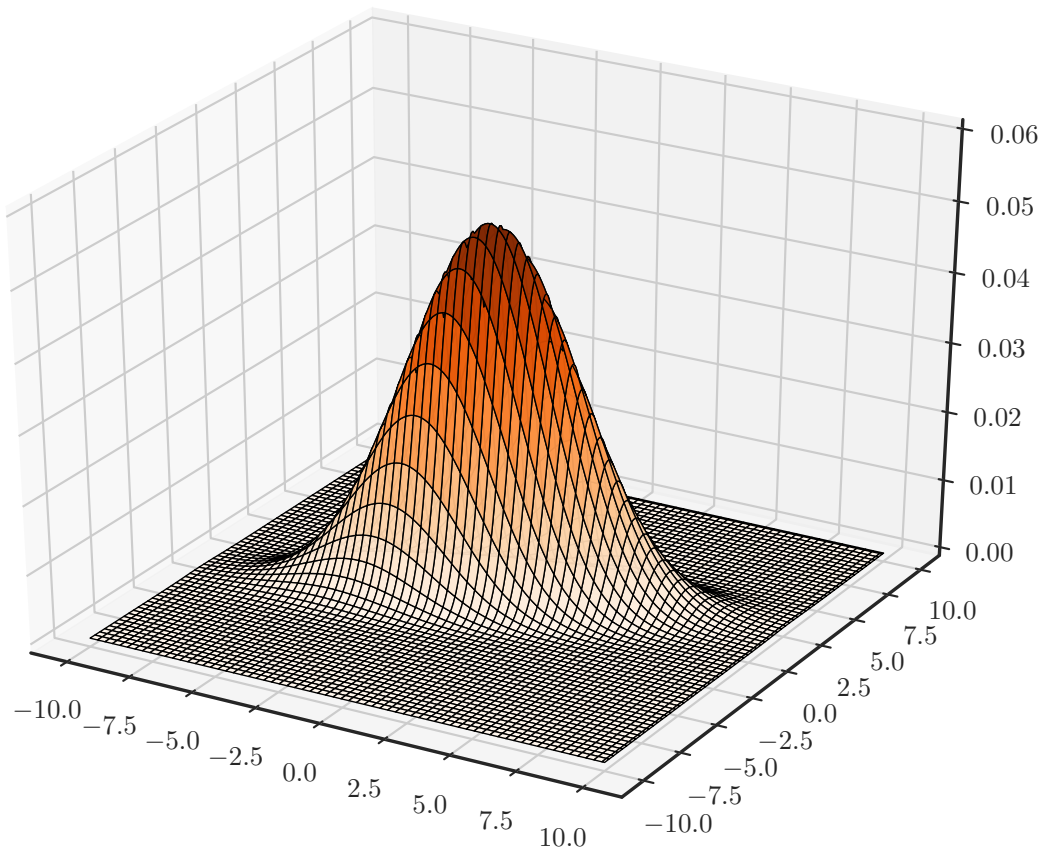
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 1 \cdot \pi/5$$



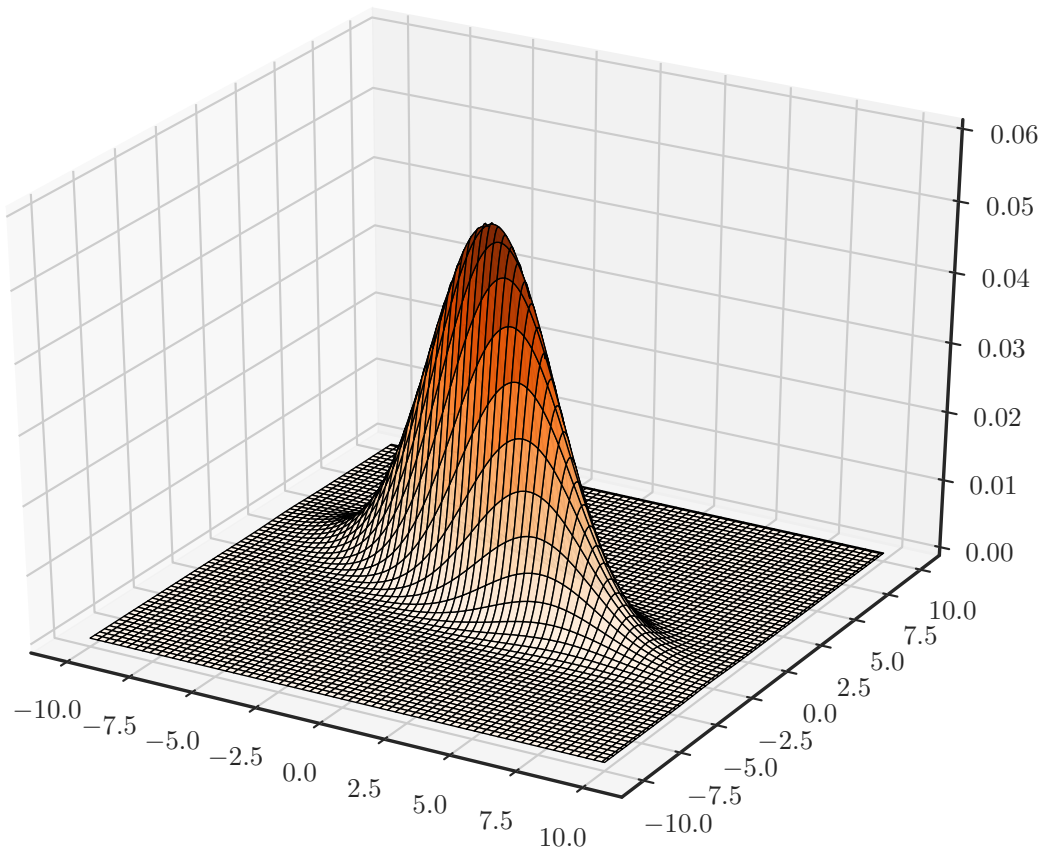
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 2 \cdot \pi/5$$



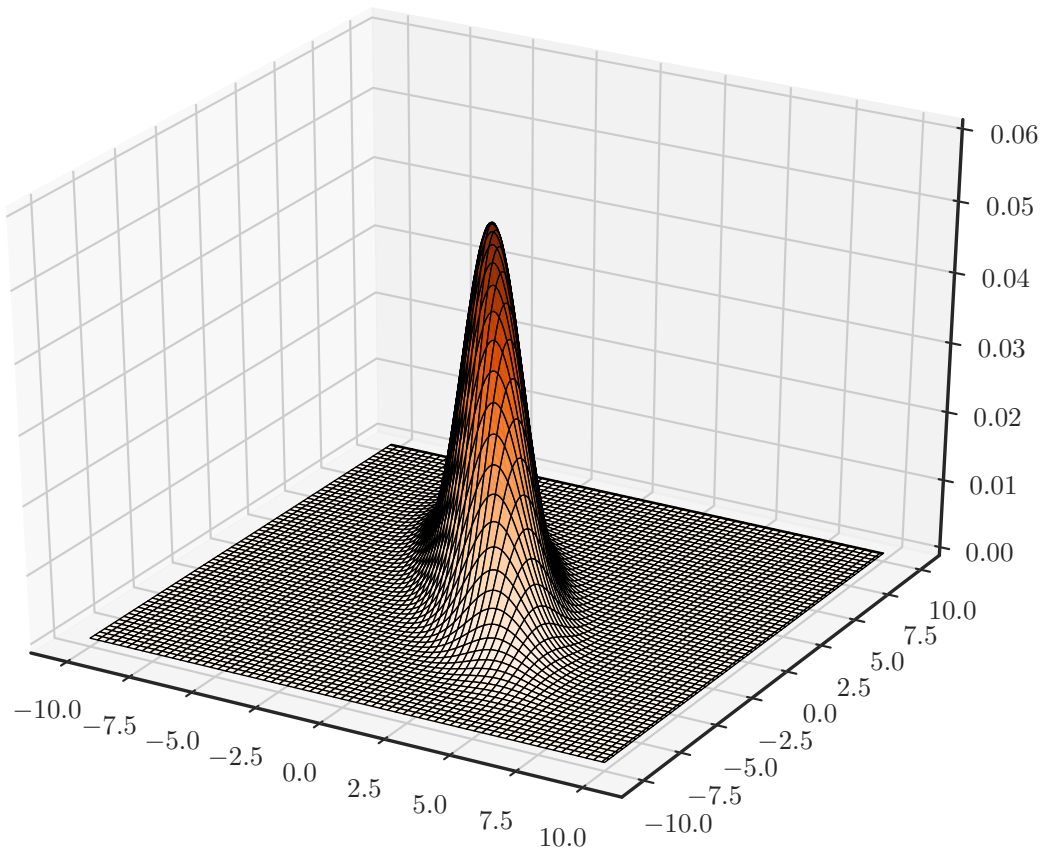
Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 3 \cdot \pi/5$$



Exemple 2D

$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & \\ & 9 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \theta = 4 \cdot \pi/5$$



Propriétés des vecteurs gaussiens

Proposition

Si X est un vecteur gaussien de $\mathbb{R}^p$, et si A est une matrice de $\mathbb{R}^{m\times p}$ et que b est un vecteur de $\mathbb{R}^m$ alors $Y = AX + b$ est un vecteur gaussien de $\mathbb{R}^m$

Construction

Soit $X \in \mathbb{R}^p$ un vecteur gaussien centré-réduit $X \sim \mathcal{N}(0, \mathrm{Id}_p)$. Supposons que l'on connaisse $L \in \mathbb{R}^{p\times p}$ telle que $LL^\top = \Sigma$, alors pour tout $\mu \in \mathbb{R}^p$, $Y = \mu + LX \sim \mathcal{N}(\mu, \Sigma)$

Démonstration :
$\mathrm{Cov}(Y) = \mathrm{Cov}(LX) = L\,\mathrm{Cov}(X)L^\top = L\,\mathrm{Id}_p\,L^\top = \Sigma$

Rem: L peut être obtenue par la factorisation de Cholesky

Factorisation de Cholesky

Théorème

Toute matrice symétrique définie positive $\Sigma \in \mathbb{R}^{p \times p}$ peut s'écrire $\Sigma = LL^\top$ pour une matrice L triangulaire inférieure

$$L = \begin{bmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ L_{p1} & L_{p2} & \cdots & L_{pp} \end{bmatrix}$$

Rem: on peut imposer que les éléments diagonaux de la matrice L soient tous positifs ; la factorisation correspondante est alors unique

Rem: numériquement L est obtenue par la méthode du pivot de Gauss, *e.g.*, avec `numpy.linalg.cholesky`

Bibliographie

DataScience :

- Blog + videos de Jake Vanderplas : http://jakevdp.github.io/, http://jakevdp.github.io/blog/2017/03/03/reproducible-data-analysis-in-jupyter/
- VanderPlas (2016), Müller et Guido (2016) : statistiques/apprentissage avec `Python`
- Exemples d'application de `scikit-learn` : http://www.baglom.com/b/10-scikit-learn-case-studies-examples-tutorials-cm572/?utm_content=bufferbde5d&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Math :

- Hastie *et al.* (2009) : *Elements of Statistical Learning*
- James *et al.* (2013) : *An introduction to statistical learning* (version simplifiée du précédent)
- Tsybakov (2006) cours de "Statistique appliquée"
- Delyon (2015) cours de Régression

Références I

- Bertsekas, D. P. (1999).
  Nonlinear programming.
  Athena Scientific.
- Boyd, S. and Vandenberghe, L. (2004).
  Convex optimization.
  Cambridge University Press, Cambridge.
- Delyon, B. (2015).
  Régression.
  https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf.
- Foata, D. and Fuchs, A. (1996).
  Calcul des probabilités : cours et exercices corrigés.
  Masson.

Références II

- Golub, G. H. and van Loan, C. F. (2013).
  Matrix computations.
  Johns Hopkins University Press, Baltimore, MD, fourth edition.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009).
  The elements of statistical learning.
  Springer Series in Statistics. Springer, New York, second edition.
  http://www-stat.stanford.edu/~tibs/ElemStatLearn/.
- Horn, R. A. and Johnson, C. R. (1994).
  Topics in matrix analysis.
  Cambridge University Press, Cambridge.
  Corrected reprint of the 1991 original.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
  An introduction to statistical learning, volume 6.
  Springer.

Références III

- Müller, A. C. and Guido, S. (2016).
  Introduction to Machine Learning with Python : A Guide for Data Scientists.
  O'Reilly Media, early access edition.
- Silverman, B. W. (1986).
  Density estimation for statistics and data analysis.
  Monographs on Statistics and Applied Probability. Chapman & Hall, London.
- Tsybakov, A. B. (2006).
  Statistique appliquée.
  http://josephsalmon.eu/enseignement/ENSAE/StatAppli_tsybakov.pdf.
- VanderPlas, J. (2016).
  Python Data Science Handbook.
  O'Reilly Media.