Cours M1 - Probabilité et Statistique

A. Decelle, A. Bonin

19 décembre 2019

Table des matières

1 Introduction 2

2 Plan 2

3 Les bases en proba 2
- 3.1 Les opérations ensemblistes . . . 2
- 3.2 Loi de probabilité . . . 3
- 3.3 Distributions discrètes . . . 3
  - 3.3.1 Loi de Bernoulli . . . 3
  - 3.3.2 Loi Binomiale . . . 5
- 3.4 Loi de Poisson et évènements rares . . . 6
- 3.5 Distributions continues . . . 7
  - 3.5.1 Loi uniforme . . . 8
  - 3.5.2 Loi normale (Gaussienne) . . . 8
  - 3.5.3 Remarque sur les lois continues . . . 10
- 3.6 La loi faible des grands nombres (Théorème de Khintchin) . . . 10
- 3.7 Le théorème central limite . . . 11

4 Analyse en composante principale (ACP / PCA) 14
- 4.1 Introduction . . . 14
- 4.2 Formulation du maximum de variance . . . 15
- 4.3 Formulation du maximum de variance . . . 17
- 4.4 ACP et décomposition en valeurs singulière (SVD) . . . 18

5 Inférence Bayésienne 18
- 5.1 Les probabilités conditionnelles . . . 18
- 5.2 Problème direct . . . 19
- 5.3 Problème inverse . . . 19
- 5.4 Théorème de Bayes : données et modèles . . . 20

Enseignant : Aurélien decelle (Email : aurelien dot decelle at lri dot fr) Pour les TP : Python3 avec numpy, matplotlib et scikit-learn, jupyter-notebook (ou jupyter-lab)

1 Introduction

Proba : étude mathématique des phénomènes caractérisés par le hasard et l'incertitude (= Phénomène stochastique).
Statistique : Recueillir, traiter et interpréter un ensemble de données.

2 Plan

1. Probabilité & variable aléatoire (base et théorie des probabilités)
2. Variables indépendantes (Lois classique, combinatoire, etc)
3. Variables en interactions : plus proche des données réelles
4. Estimateur d'observables (moyenne, variante), corrélations (voir le site des corrélations : tylervigen)

3 Les bases en proba

En probabilité on décrit des objets dont le comportement n'est pas déterministe. L'objet fondamental est la variable aléatoire. En généaral, on note une variable aléatoire (va) X, qui suit une loi de probabilité (i.e. on ne connait pas le résultat de sa réalisation à l'avance). On appelle x une réalisation de X. Dans la suite du cours on utilisera x à la fois pour la variable aléatoire et sa réalisation quand le contexte permet de ne pas se tromper, c'est aussi plus simple pour les notations.
On notera l'ensemble des états possibles que peut prendre une v.a. par : Ω

Exemples

— un jeu de pile ou face : $\Omega = \{\ll pile\gg, \ll face\gg\}$
— un dé à 6 faces : $\Omega = \{1, 2, 3, 4, 5, 6\}$

On note $|\Omega|$ Le cardinal de Ω, soit le nombre d'éléments possibles.
On note x une réalisation possible d'une va. Dans les deux exemples précédents, on a respectivement 2 et 6 pour la valeur du cardinale.
Un "événement" est une partie de Ω (un sous ensemble). Exemple pour pile ou face : $\{pile\}, \{face\}, \{pile, face\}, \emptyset$
La réalisation d'un événement A de Ω se formule ainsi : soit w une réalisation, A se réalise $\Leftrightarrow w \in A$.
Exemple : si A "le résultat est pair" $= \{4, 5, 6\} \rightarrow$ L'événement se réalise si $x = 2$, 4 ou 6
Un événement est donc un ensemble de réalisations, il suffit que la va soit égale à l'une des réalisations pour réaliser l'événement.

3.1 Les opérations ensemblistes

— **Complémentaire** : Le complémentaire de A est $\overline{A} \Leftrightarrow \{x \in \Omega, x \notin A\}$. Si les deux ensembles sont disjoints, $A \cap \overline{A} = \emptyset$ (leur intersection est nulle)

$$A \cup \overline{A} = \Omega$$

— **Union** : L'union $A \cup B$:={ $x \in \Omega,\ \ x \in A \bigvee x \in B$ }
— **Intersection** : L'intersection $A \cap B := \{x \in \Omega, x \in A \ \wedge \ x \in B\}$
— **Inclusion** $A \subset B := (x \in A \Rightarrow x \in B)$

3.2 Loi de probabilité

La loi de probabilité est une fonction mesurant la « chance » ou la « fréquence » de réalisation d'un événement de Ω. Elle est définie telle que $P : A \rightarrow [0,1]$: à une sous-partie de Ω, on associe une valeur entre 0 et 1 représentant la fréquence d'observer ces évènements.
La loi doit satisfaire 3 axiomes pour être bien définie :

$$- \ (1) \quad \forall A' \in A : 0 \leq P(A') \leq 1$$

Tous les événements possibles ont une valeur associée entre 0 et 1

$$- \ (2) \quad P(\Omega) = 1$$

La probabilité d'apparition de l'ensemble des événements est 1 (La somme des probabilités de tous les événements de l'univers vaut 1)

$$- \ (3) \quad P(\underset{i \in \mathbb{N}}{U} A_i) = \sum_{i \in \mathbb{N}} P(A_i) \text{ si les ensembles sont disjoints.}$$

La probabilité d'un ou plusieurs événements disjoints est la somme de leurs probabilités respectives.

3.3 Distributions discrètes

3.3.1 Loi de Bernoulli

Les réalisations sont binaires (pile ou face) :

$$\Omega = \{0,1\}$$

On paramètrise la distribution à l'aide de : p_0 et p_1 :

$$p(X = 0) = p_0$$

$$p(X = 1) = p_1$$

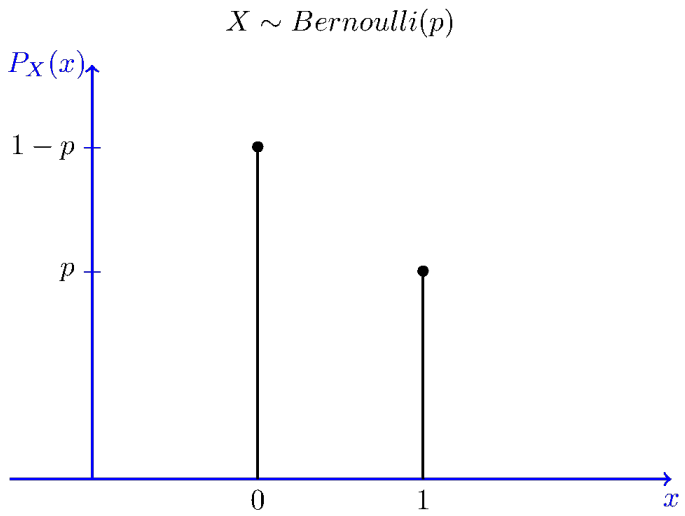
$$p_0 + p_= 1$$

$$0 \leq p_{0,1} \leq 1$$

Cette égalité permet d'écrire seulement :

$$p_0 = 1 - p_1$$

On représente les distributions par un graphe : une barre pour chaque valeur réalisable x et la barre monte jusqu'à la valeur $P(X = x)$.



On note $X \sim B(p_1)$ pour dire « X suit une loi de Bernoulli de paramètre p_1 ».

Moyenne — Définition de la valeur moyenne : $m = \mathbb{E}(x) = \sum_{x\in\Omega} xp(X = x)$ que l'on écrira plus simplement $\sum_{x\in\Omega} xp(x)$. On trouve ici $m = 0p_0 + 1p_1 = p_1$

Variance — Elle est définie par : $\sigma^2 = \mathbb{E}((x-m)^2) = \sum_{x\in\Omega} p(x)(x-m)^2$. Pour Bernoulli, on trouve $\sigma^2 = (0-m)^2p_0 + (1-m)^2p_1 = m^2(1-p_1) + (1-m)^2p_1 = (1-p_1)m^2$. On note $variance = \sigma^2$. La variance mesure l'étalement des échantillons autour de la moyenne.

La somme de deux variables aléatoires — A partir de la distribution de Bernoulli, il est possible de se poser la question suivante. Si j'ai deux variables aléatoires indépendantes X_1 et X_2, chacune distribuée selon $\sim B(p_1)$. Peut-on caractériser la distribution de la variable aléatoire $Y = X_1 + X_2$? Tout d'abord, on peut constate que l'ensemble des valeurs possibles pour Y est $\Omega_Y = \{0, 1, 2\}$. Ensuite, il suffit de compter pour obtenir la loi $p(Y)$. La distribution de Y peut s'écrire

$$p(Y = y) = \sum_{x_1=0,1,x2=0,1} p(x_1, x_2)\delta_{y,x1+x2} = \sum_{x_1=0,1} p(x_1, y - x1)$$

où δ est le delta de Kronecker : $\delta_{ij} = \begin{cases} 1 \; si \; i = j \\ 0 \; sinon \end{cases}$

autrement dit, la distribution de Y consiste à regarder tous les cas pour lesquelles la variable Y est égale à une certaine valeur y et les sommer entre eux. Comme X_1 et X_2 sont indépendant, on a $p(x_1, x_2) = p(x_1)p(x_2)$. On peut donc calculer $p(y)$

$$p(Y=0) = p(X_1=0)p(x_2=0) = (1-p_1)^2$$
$$p(Y=1) = 2p_1(1-p_1)$$
$$p(Y=2) = p_1^2$$

On observe bien que la somme des p(y=i) pour i allant de 0 à 2 donne 1 (propriété d'une loi de probabilité).

Séquence de variables de Bernoulli — Soient $x_i \sim B(p_i)$ pour $i = 1,...,N$. Quelle est la probabilité de $p(x_1, x_2, ...x_N)$?

$$p(x_i) = \begin{cases} = p_i & \text{si } x_i = 1 \\ 1-p_i & \text{si } x_i = 0 \end{cases} = p_i^{x_i}(1-p_i)^{1-x_i}$$

donc

$$p(x_1, x_2, ...x_N) = \prod_{i=1}^{N} p_i^{x_i} * (1-p_i)^{1-x_i}$$

3.3.2 Loi Binomiale

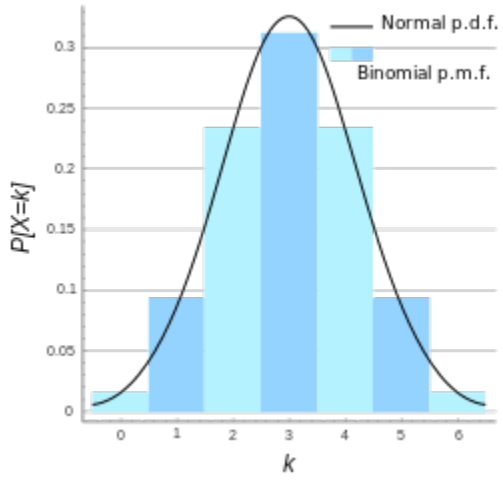
Si on prend le cas $p_i = p$, on peut alors caractériser la distribution de probabilité de la somme de N variables de Bernoulli. La distribution prend la forme suivante

$$p(\sum_i = k) = (\#\text{ de configurations possibles donc la somme donne } k \text{ par } N) p^k (1-p)^{N-k}$$

Le facteur combinatoire peut se calculer et se note $C_N^k \ \frac{N!}{k!(N-k)!}$. On obtient alors

$$p(k) = C_N^k p^k (1-p)^{N-k}$$

Le graphe typique de la distribution binomiale est : (les valeurs d'abscisse et d'ordonnée sont un exemple)



3.4 Loi de Poisson et évènements rares

On va finir sur une dernière distribution qui peut s'obtenir de la façon suivante. Imaginons que l'on soit intéressé à compter des évènements rares pouvant apparaître avec une certaine probabilité. On va noter

p_0 (densité par unité de temps) probabilité que rien ne se passe
p_1 (densité par unité de temps) probabilité d'un évènement rare

On va donc vouloir savoir, à un temps t quelle est la probabilité d'avoir vu $N(t)$ évènements rares. Pour obtenir cette quantité il faut commencer par discrétiser le temps. On découpe la trame temporelle en petits intervalles δ, au sein de chacun de ces petits intervalles, on a alors la probabilité $p_1 = \delta\lambda$ où δ est la taille de la fenêtre temporelle et λ le taux d'apparition de l'évènement rare par unité de temps.
On va maintenant chercher la forme de la distributino $p(N(t)$ donnant la probabilité du nombre d'évènements rares au temps t selon les hypothèses suivantes

1. p est petit devant ?
2. $N = t/\delta$ (le nombre d'intervalles) est grand (t est grand devant δ

On voit que la distributino de $N(t)$ est donnée par une loi binomiale de probabilité $p = \lambda\delta$ pour un nombre $N = t/\delta$ de lancer. On s'intéresse maintenant au régime où un évènement est rare : $\delta \to 0$. On obtient

$$p(N(t) = k) = C_N^k (\delta\lambda)^k (1 - \delta\lambda)^{N-k} = \frac{(t/\delta)!}{k!(t/\delta - k)!}(1 - \delta\lambda)^{t/\delta - k}$$

En utilisant l'approximation de Stirling pour la fonction factorielle et en faisant tendre $\delta \to 0$ on obtient

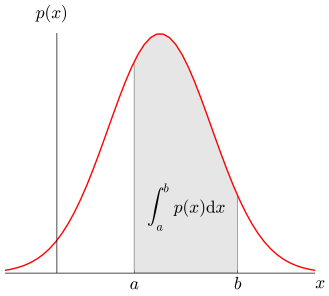


FIGURE 1 – Allure d'une densité de proababilté. La fréquence d'un évènement, par exemple $x \in [a, b]$ est maintenant mesuré par l'aire sous la courbe

.

$$p(N(t)) = \frac{e^{-\lambda t}\lambda t^{N(t)}}{N(t)!}$$

qui est la loi de Poisson.
Moyenne $m = \lambda t$ Variance $\sigma^2 = \lambda t$

3.5 Distributions continues

Lorsque l'on considère uniquement les distributions discrètes l'interprétation des lois de probabilités peut au moins se faire simplement : la fonction de probabilité peut se comprendre comme la fréquence à laquelle apparaît un évènement dans le cas d'un tirage infini. Par ailleurs la normalisation se fait simplement en sommant tous les éléments. Pour les v.a. continues la différence est qu'on va s'appuyer cette fois-ci sur la densité de probabilté. Une conséquence directe est que l'observation d'une valeur isolée est associé à une fréquence nul puisqu'il faut multiplié la densité de probabilité par un intervalle pour en obtenir la fréquence d'un évènement.
On notera comme précédemment $p(x)$ la densité de probabilité. On peut également la représenter par un graphe :
Puisque $p(x)$ représente la densité de probabilité, il n'est pas interdit d'avoir des valeurs supérieur à 1, par contre tout intégration sur un intervalle doit effectivement être plus petit que 1 puisqu'il représente la fréquence d'un évènement.
La normalisation de ces densité de probabilité (axiome 2) est maintenant obtenu par

$$\int_{\mathcal{R}} p(x)dx = 1$$

Pour une loi continue, la moyenne et la variance sont données par :

1. $m = \int_{\mathcal{R}} xp(x)dx$
2. $\sigma^2 = \int_{\mathcal{R}} (x-m)^2 p(x)dx$

3.5.1 Loi uniforme

La loi uniforme est caractérisée par une densité de probabilité constante : en d'autre terme, si je fixe un intervalle δx , la fréquence de réalisation ne dépend pas de "où" est mon intervalle. On peut donc écrire

$$p(x) = \begin{cases} \frac{1}{b-a} \text{ si } x \in [a,b] \\ 0 \text{ sinon} \end{cases}$$

On considère fréquemment le cas $a = 0$ et $b = 1$.

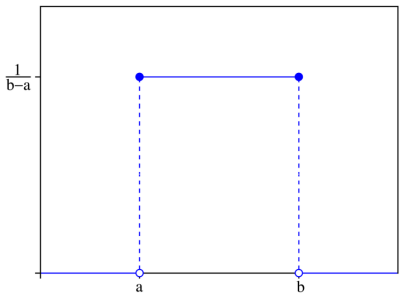


FIGURE 2 – Densité de probabilté de la loi uniforme

Cette loi est importante pour générer des événements aléatoires de toute sorte. Par exemple, si on sait générer des nombres aléatoires selon la loi uniforme dans $[0,1]$, il est très facile de générer n'importe quel loi discrète simplement en utilisant la cumulative.

3.5.2 Loi normale (Gaussienne)

La loi normale correspond à la "fameuse" courbe en cloche. En fait, elle correspond à une densité de probabilité centrée autour d'une valeur moyenne, et qui a pratiquement toute son aire concentré sur 4 fois la variance. La densité est donnée par

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-m)^2}{2\sigma^2}\right)$$

où m représente la valeur moyenne et σ^2 la variance. L'allure de la densité de probabilité est donnée ci-dessous :
De nouveau, ici il est très clair que le paramètre σ caractérise sur quel interval autour de la moyenne m la probabilité de x reste grande. On voit sur le graphe que dans un intervale de $\pm\sigma$ autour de la valeur moyenne, on a 95% de l'aire sous la courbe.
Il peut-être également intéressant de la représenter en log-lin, afin d'observer le comportement quadratique de l'argument de l'exponentiel

Le log linéaire montre qu'on descend vite en proba (valeurs négatives sont très petites une fois passées en exponentiel).

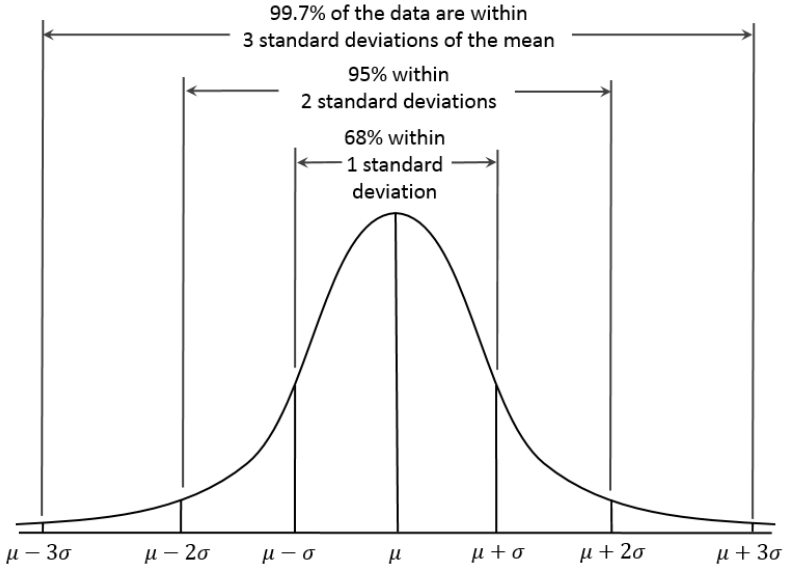


FIGURE 3 – Plot linéaire de la loi gaussienne.

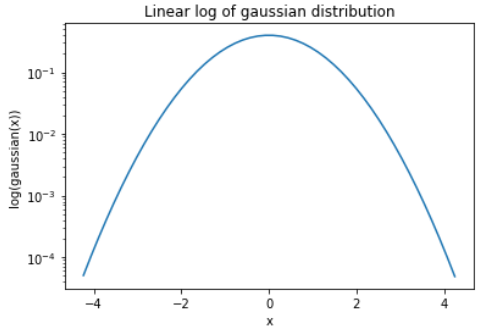


FIGURE 4 – Plot log-linéaire de la loi gaussienne.

Notation — on note $X \sim \mathcal{N}(m, \sigma)$ si la VA X est distribuée selon une loi gaussienne de paramètres m et σ. (On note le N en cursive).

3.5.3 Remarque sur les lois continues

On a souvent l'habitude de dire, pour une loi discrète (par exemple Bernoulli) que la probabilité d'une séquence d'observation (par exemple 0110001) est donnée par $p_0^4 p_1^3$. On peut faire la même chose pour une loi continue mais pour rendre les choses plus claires il peut-être utile de "discrétiser" la loi continue. Si on image qu'on discrétise la loi normale de la façon suivante : on définit un interval δx, et pour toute valeur $x \in [i\delta x, (i+1)\delta x]$, avec $i \in \mathbb{Z}$, on a la probabilité $p_d(x) = \int_{i\delta x}^{(i+1)\delta x} p(x)$. Alors, il devient clair commen associé maintenant la probabilité d'une séquence de variables gaussiennes $\{x_i\}_{i=1,..,N}$ sur cette loi discrétisée :

$$pr(\{x_i\}) = \prod_{i=1}^{N} p_d(x_i)$$

Le problème de cette notation est que la probabilité de la séquence dépend de la discrétisation utilisée. On généralise en générale en utilise la fonction de densité de la loi continue considérée.

3.6 La loi faible des grands nombres (Théorème de Khintchin)

Ici, on va rappeler quelques théorèmes centrales permettant de montrer la convergence d'une séquence de variables aléatoires vers la valeur moyenne ainsi que le rôle joué par la variaance.

Soit $\{x_i\}_{i=1,...,N}$ une séquence de VA indépendantes ayant chacun pour moyenne m et pour variance σ^2, tels que m et σ^2 sont finis.

Alors :

$$\forall \epsilon > 0, \lim_{N \to \infty} proba(|\frac{\sum_{i=1}^{N} x_i}{N} - m| \geq \epsilon) = 0 \tag{1}$$

Quand on somme N réalisations (N valeurs de VA) et qu'on les divise par N, alors cet object se rapproche de plus en plus de m lorsque N devient grand. La démonstration utilise le théorème de Tchebychev.

$$\forall \alpha > 0, proba(|x - m| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2} \tag{2}$$

Ce théorème nous dit que, la probabilité de se trouver à une distance α de la valeur moyenne est bornée par la variance de la distribution. Pour la démonstration on aura besoin de ...

Inégalité de Markov :

$$\forall a > 0, X > 0 \quad \text{une VA positive de moyenne } m, \quad Proba(X \geq a) \leq \frac{m}{a} \tag{3}$$

Preuve de 3 — $proba(x \geq a) = \sum_{x \geq a} p(x) \leq \sum_{x \geq 0} p(x)$ *Car* $a > 0$ *donc la seconde somme a plus de termes tout positifs*

$= \sum_{x \geq a} a * p(x) \leq \sum_{x \geq 0} x * p(x)$ *Car* x ¿= a ¿= 0

$a * proba(X \geq a) \leq m$ *On extrait la constante a du premier terme, le second est la définition de la moyenne*

$proba(X \geq a) \leq \frac{m}{a}$ *On retrouve l'inégalité de markov 3*

Preuve de 2 — On utilise 3 avec $Y = (X - m)^2$

$proba(Y \geq \alpha^2) \leq \frac{\mathbb{E}[Y]}{\alpha^2}$ $\mathbb{E}[Y]$ *l'espérance de Y*

$proba((x - m)^2 \geq \alpha^2) \leq \frac{\sigma^2}{\alpha^2}$ *Réécriture de Y par sa definition, Espérance étant la moyenne on retrouve la définition de la variance quand on l'applique à Y.*

$proba(|x - m| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$ *On retrouve 2 en retirant les carrés à chaque membre de* $(x - m)^2 \geq \alpha^2$

Preuve de 1 — On utilise $X = \frac{\sum_i x_i}{N}$

$$proba(|\frac{\sum x_i}{N} - m| \geq \alpha) \leq \frac{\mathbb{E}[(\frac{\sum x_i}{N} - m)^2]}{\alpha^2} = \frac{(\frac{\sum x_i}{N} - m)^2}{N * \alpha^2} \rightarrow \frac{(\mathrm{m} - m)^2}{N * \alpha^2} \rightarrow \frac{0}{\infty} \rightarrow 0 \; (N \rightarrow \infty)$$

3.7 Le théorème central limite

Soit $\{x_i\}_{i_1,\ldots,N}$ une séquence de VA indépendantes de moyenne m et de variance σ^2.

On montre que $S_N = \sum_{i_1}^{N} x_i$

$$\mathbb{E}[S_N] = \sum_{i=1}^{N} \mathbb{E}[x_i] = \sum_{i=1}^{N} m = Nm$$

$\mathbb{E}[(S_N - \mathbb{E}[S_N])^2] = \mathbb{E}[S_N^2] - 2\mathbb{E}[S_N]^2 + \mathbb{E}[S_N]^2$ *(*$2\mathbb{E}[S_N]^2$ *vient de* $\mathbb{E}[2 * S_N \mathbb{E}[S_N]] = 2 * \mathbb{E}[S_N]\mathbb{E}[S_N]$*)*

$$= \mathbb{E}[S_N^2] - \mathbb{E}[S_N]^2 (2\mathbb{E}[S_N]^2 + \mathbb{E}[S_N]^2 = \mathbb{E}[S_N]^2)$$

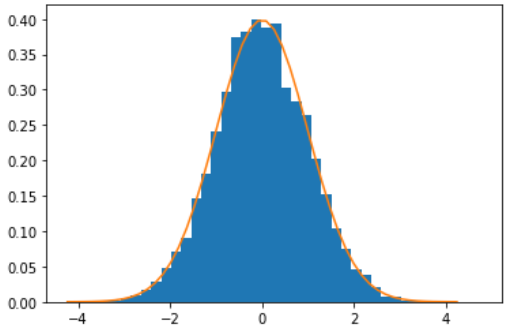
$$\mathbb{E}[S_N^2] = \mathbb{E}[(\sum_{i=1}^{N} x_i)^2] = \mathbb{E}[\sum_{i=1}^{N} x_i^2 + 2\sum_{i<j} x_i x_j] = \sum_{i=1}^{N} \mathbb{E}[x_i^2] + 2\sum_{i<j} \mathbb{E}[x_i]\mathbb{E}[x_j]$$

(Décomposition de $\mathbb{E}[a+b]$ *en* $\mathbb{E}[a] + \mathbb{E}[b]$ *permise car* $\mathbb{E}$ *est une application linéaire)*

$\mathbb{E}[\mathrm{x_i}]\mathbb{E}[\mathrm{x_j}] = \mathrm{m}^2$ car les variables ont la même moyenne

Que vaut $\mathbb{E}[x_i^2]$***?***

$$\mathbb{E}[(x_i - m)^2] = \mathbb{E}[x_i^2 - m^2] = \sigma^2 <=> \mathbb{E}[x_i^2] = \sigma^2 + m^2$$



$\mathbb{E}[(S_N - \mathbb{E}[S_N])^2] = N*\sigma^2 + N*m^2 + N(N-1)m^2 - N^2m^2$ *(Décomposition du carré)*
$= N*\sigma^2 + Nm^2 + N^2m^2 - Nm^2 - N^2m^2$ *(Distribution de* $N(N-1)m^2$*)*
$= N\sigma^2$ *(Suppression des termes qui s'éliminent)*

On définit : $Z_N = \frac{S_N - Nm}{\sqrt{N\sigma^2}} = \sqrt{N}(\frac{\frac{S_N}{N} - m}{\sigma})$

$\lim_{N\to\infty} proba(Z_N = x) = \mathcal{N}(0,1)$ *Quand N tend vers l'infini, la distribution de* Z_N *devient une loi Gaussienne centrée (c'est à dire vers 0) et normée (c'est à dire* σ *= 1).*

Exemple du théorème en pratique

1. Générer N variables aléatoires entre 0 et 1
2. Calculer leur somme
3. Répéter l'opération M fois et créer un histogramme des Z_N
4. Afficher la courbe de gauss (m, σ)

Figure **1** : Superposition de l'histogramme des moyennes et de la courbe de gauss (N = M = 10000, m = 0.5, σ = 0.5)

Malgré le caractère aléatoire uniforme des VA, la somme des variables indépendantes se comportent comme sous une loi normale.

2. Estimateurs et observables (lien entre proba et stats) Estimateur de la moyenne :

En général, on évalue empiriquement que la moyenne d'une séquence est donnée par $\hat{m} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Question : à quel point $\widehat{m_N}$ est-il proche de m ?

On calcule la variance entre $\widehat{m_N}$ et m :

$\mathbb{E}[(\widehat{m_N} - m)^2]$ où m est la réelle moyenne des éléments de la séquence.

On trouve $\mathbb{E}[(\widehat{m_N} - m)^2] = \frac{\sigma^2}{N}$

$$\widehat{m_N} \simeq m \pm \frac{\sigma}{\sqrt{N}}$$

$$\lim_{N\to\infty} \widehat{m_N} = m$$

Estimateur de la variance :

La variance requiert la moyenne de la distribution. Si seulement $\widehat{m_N}$ est connue, l'estimation est biaisée car $\mathbb{E}[\widehat{\sigma^2}] = \frac{N-1}{N}\sigma^2 = (1-\frac{1}{N})\sigma^2$. L'estimateur est d'autant erroné que l'échantillon est petit.

Pour corriger ce biais on divise non pas par N mais par N – 1.

$$\widehat{\sigma_{N-1}}^2 = \frac{\sum_{i=1}^{N}(x_i - \widehat{m_N})^2}{N-1}$$

Cette correction est utile seulement sur les très petits échantillons. Car $\frac{\widehat{\sigma_N}^2}{\widehat{\sigma_{N-1}}^2}$ $\to 1$ $(N \to \infty)$.

4 Analyse en composante principale (ACP / PCA)

4.1 Introduction

L'analyse en composante principale trouve son intérêt lorsque l'on traite des jeux de données possédant :

1. Beaucoup de données
2. Des données en grandes dimensions (Malédiction de la dimensionnalité)

L'ACP permet d'extraire de l'information rapidement en se basant sur la "forme" du jeu de données. En particulier, si le jeu de données vit dans un espace beaucoup plus petit, où en présence de bruits, elle permet d'obtenir une idée des dimensinos ayant une importance particulière. Il faut toutefois faire attention aux "outliers" lorsque le jeu de données n'est pas très grands.

a) Nombre de "vraies" dimensions par rapport aux données — un exemple bête :

1. x_1 : nombre d'accidents
2. x_2 : nombre d'écoles fermées
3. x_3 : nombre d'explosion de canalisations
4. x_4 : nombre d'épisodes neigeux

Ces quatre features peuvent être correlées par la seule variable de **Température**.

b) La malédiction de la dimensionnalité **Exemple :** Considérons un jeu de données d'images de 100x100 pixels en Noir/Blanc. Sur un tel exemple, le nombre d'images possibles est de 2^{10^4}=>. En réalité, notre jeu de données sera beaucoup plus petit car il encode des corrélations entre les pixels et considérer simplement une énumération de toutes les images possibles revient à rajouter beaucoup de bruits.

Un outil pour analyser les données : l'ACP L'ACP donne les directions où les données varient le plus. De façon équivalente, elle minimise l'erreur de reconstruction (elle donne les k meilleures directions pour reconstruire le plus fidèlement le jeu de données). L'ACP va en général trouver de novuelles directions potentiellement intéressantes pour comprendre le jeu de données, ces directions seront en général une composition des dimensions originales du jeu de données. Regardons comment l'ACP fonction sur un exemple simple en 2 dimensions. On considère un ensemble de points $\overrightarrow{x}^{(i)}$ dans l'espace à deux dimensions. L'ACP va trouver la direction de vecteur unitaire $\overrightarrow{u}_1$ le long de laquelle les données varient beaucoup.

Le vecteur $\overrightarrow{u}_1$ est défini comme la direction où les données varient le plus (maximum de variance), et où la distance entre les points et la courbe est la moins grande (minimisation de l'erreur de reconstruction).

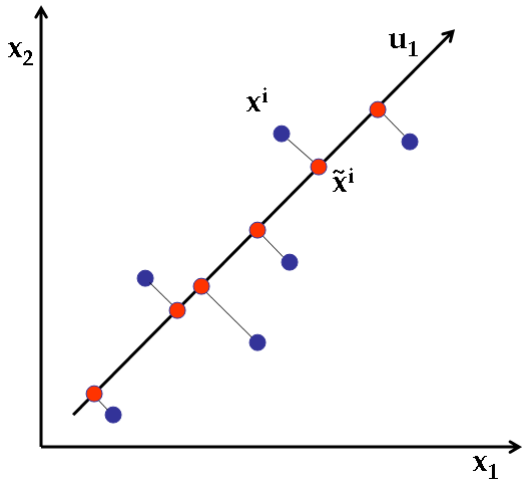


FIGURE 5 – Exemple sur un jeu de données simple. En bleu les données originales, en rouge leur projection le long de la direction trouvée par l'ACP.

4.2 Formulation du maximum de variance

On considère un ensemble d'observation $\{\overrightarrow{x_n}\}$ où $n = 1,...,N$, $\overrightarrow{x_i} \in \mathbb{R}^{DxN}$, N données de dimensions D. On cherche à projeter les données dans un sous espace de dimension M < D tout en maximisant la variance des données projetées.

On commence à $M = 1$, le résultat pour une dimension quelconque pouvant se généraliser à partir de là. On définit une direction quelconque par $\overrightarrow{u_1}$, un vecteur unitaire

$$\left\|\overrightarrow{u_1}\right\|^2 = \sum_{i=1}^{D} x_i^2 = \overrightarrow{u_1}^T \cdot \overrightarrow{u_1} = 1$$

Définition :

1. La moyenne des données $\overrightarrow{m} = \frac{1}{N}\sum_{n=1}^{N} \overrightarrow{x_n}$
2. La moyenne le long de $\overrightarrow{u_1}$: $\overrightarrow{u_1}^T \cdot \overrightarrow{m} = \frac{1}{N}\sum_{n=1}^{N} \overrightarrow{u_1}^T \cdot \overrightarrow{x_n}$ (car le vecteur est unitaire, on obtient un produit de la norme de m par le cosinus de l'angle pour un produit scalaire)
3. La variance le long de $\overrightarrow{u_1}$: $\frac{1}{N}\sum_{n=1}^{N} (\overrightarrow{u_1}^T \cdot \overrightarrow{x_n} - \overrightarrow{u_1}^T \cdot \overrightarrow{m})^2$ *(où $\overrightarrow{u_1}^T \cdot \overrightarrow{x_n}$ est la projection de $\overrightarrow{x_n}$ le long de $\overrightarrow{u_1}$et le deuxième terme étant la moyenne le long de $\overrightarrow{u_1}$)*

En développant la variance projettée le long du $\overrightarrow{u_1}$, on obtient le résultat suivant :

$$\frac{1}{N}\sum_{n=1}^{N}(\overrightarrow{u_1}^T\cdot\overrightarrow{x_n}-\overrightarrow{u_1}^T\cdot\overrightarrow{m})^2=\frac{1}{N}\sum_{n=1}^{N}(\overrightarrow{u_1}^T\cdot\overrightarrow{x_n}-\overrightarrow{u_1}^T\cdot\overrightarrow{m})(\overrightarrow{u_1}^T\cdot\overrightarrow{x_n}-\overrightarrow{u_1}^T\cdot\overrightarrow{m})$$
$$=\frac{1}{N}\sum_{n=1}^{N}\overrightarrow{u_1}^T(\overrightarrow{x_n}-\overrightarrow{m})(\overrightarrow{x_n}^T-\overrightarrow{m}^T)\overrightarrow{u_1}$$
$$=\overrightarrow{u_1}^T S\overrightarrow{u_1}$$

où on a définit

$$S=\frac{1}{N}\sum_{n=1}^{N}(\overrightarrow{x_n}-\overrightarrow{m})(\overrightarrow{x_n}^T-\overrightarrow{m}^T)$$

la matrice de covariance du jeu de données. S est symétrique et définie positive.

1. Symétrique : $S=S^T$
2. Définie semi-positive : $\forall\overrightarrow{x}!=\overrightarrow{0},\overrightarrow{x}^T S\overrightarrow{x}\geq 0$

On voit que le problème se réduit à trouver $\overrightarrow{u_1}$ tel que $\overrightarrow{u_1}^T S\ \overrightarrow{u_1}$ soit le plus grand possible. On va décomposer S selon ses vecteurs propres. On dit que $\overrightarrow{u}$ est un vecteur propre de S si $S\overrightarrow{u}=\lambda\overrightarrow{u}$ où λ est un scalaire différent de zéro (le vecteur conserve sa direction). A l'aide de la matrice de passage V composée de tous les vecteurs propres :

$$V=\begin{pmatrix}\\ \overrightarrow{v_1},\overrightarrow{v_2}...\overrightarrow{v_D}\\ \\ \end{pmatrix}$$

on peut diagonaliser la matrice S de la façon suivante :

$$S=VDV^T$$

où

— V est une matrice orthogonale ($VV^T=I_D$ la matric identité dans $\mathbb{R}^{DxD}$), V est une matrice de changement de base.
— D est une matrice diagonale (seuls les éléments diagonaux sont non nuls) contenant les variances le long des directions données par V.

On peut maintenant réécrire le problème sous la forme

$$\overrightarrow{u_1}^T S\overrightarrow{u_1}=(\overrightarrow{u_1}^T V)D(V^T\overrightarrow{u_1})$$

On peut définir $\overrightarrow{w_1}=V^T\overrightarrow{u_1}$ correspondnat au vecteur $\overrightarrow{u_1}$ réécrit dans la nouvelle base. Cette transformation conserve la norme et donc

$$\overrightarrow{w_1}^T\overrightarrow{w_1}=(\overrightarrow{u_1}^T V)(V^T\overrightarrow{u_1})=\overrightarrow{u_1}^T\overrightarrow{u_1}=1$$

Une fois que l'on écrit tout dans la base des vecteurs propres, on obitnet que

$$\overrightarrow{w_1}^T D \overrightarrow{w_1} = \sum_{i=1}^{D} \lambda_i (w_{1i})^2$$

On cherche $\max\{\sum_i \lambda_i (w_{1i})^2\}$ par rapport aux composantes du vecteur $\overrightarrow{w_1}$ avec la contrainte que $\sum_i w_{1i}^2 = 1$.
Solution : Soit $i^* = argmax\{\lambda_i\}$

$\begin{cases} {w_{1i^*}}^2 = 1 \\ {w_{1i}}^2 = 0, \forall i \neq i^* \end{cases}$ <=> $\overrightarrow{w_1}$ est le vecteur propre associé à la plus grande valeur propre.
L'ACP consiste à trouver les M vecteurs propres de S qui correspondant aux plus grandes valeurs propres.

4.3 Formulation du maximum de variance

On peut arriver à la même soution en posant le problème d'une façon différente : en voulant minimiser l'erreur de reconstruction (L_2 norme) si l'on projette les données dans un sous-espace de dimension $M < D$. Commençons par définir une nouvelle base de $\mathbb{R}^D; \{\overrightarrow{u_i}\} \in \mathbb{R}^D; i = 1, ..., D$

$$\begin{bmatrix} \|\overrightarrow{u_i}\|^2 = 1 \\ \overrightarrow{u_i}^T \cdot \overrightarrow{u_j} = \begin{cases} 1 \; si \; i = j \\ 0 \text{ sinon} \end{cases} \end{bmatrix}$$

On peut donc réécrire les données sur cette nouvelle base :

$$\overrightarrow{x_n} = \sum_{i=1}^{D} \alpha_{ni} \overrightarrow{u_i} = \sum_{i=1}^{D} (\overrightarrow{x_n}^T \cdot \overrightarrow{u_i}) \overrightarrow{u_i}$$

où les α correspondent aux projections sur les nouveaux vecteurs de base : $\alpha_{ni} = (\overrightarrow{x_n}^T \overrightarrow{u_i})$. On peut maitenant projetter les données sur les M premiers vecteurs $\{\overrightarrow{u_i}\}$

$$\overrightarrow{\widetilde{x_n}} = \sum_{i=1}^{M} Z_{ni} \overrightarrow{u_i} + \sum_{i=M+1}^{D} b_i \overrightarrow{u_i}$$

où le premier terme correspond à la projection et le second à un décalage systèmatique (il ne dépend pas de la donnée considérée). La quantité que l'on souhaite minimiser ici est alors l'écart entre les points et les points projetés :

$$\min\{Z_{ni}, b_i\} J \text{ tel que } J = \frac{1}{N} \sum_{n=1}^{N} \| \overrightarrow{x_n} - \overrightarrow{\widetilde{x_n}} \|^2$$

La solution ici pour minimiser J sera donner en prenant il faut prendre les M vecteurs propres de S associés aux M plus grandes valeurs propres.

4.4 ACP et décomposition en valeurs singulière (SVD)

On peut réécrire le jeu de données en utilisant la base des vecteur propres de la matrice de covariance S. Considérons d'abord l'ensemble des données $X \in \mathbb{R}^{DxN}$ (D dimensions, N données), et recentrons les autour de la valeur moyenne :

$$X^c = X - \overrightarrow{m}$$

ici on a utilisé un leger abus de notation pour dire que l'on va soustraire la valeur moyenne pour chacune des données. Remarquons tout d'abord qu'il est très simple de calculer la matrice de covariance de la façon suivante : $S = \frac{1}{N} X^c (X^c)^T$. On peut maintenant montrer que l'on peut écrire la matrice X^c de la façon suivante :

$$X^c = V \Sigma U^T \text{ où } \begin{cases} V \in \mathbb{R}^{DxD} \text{ est la matrice de passage de } S \\ \Sigma \in \mathbb{R}^{DxD} \text{ l'écart type le long des directions données par} V \ (\sqrt{\sigma}) \\ U \in \mathbb{R}^{NxD} \text{ les nouvelles coordonnées} \end{cases}$$

Si on veut utiliser seulement M dimensions pour modéliser les données, il suffit de n'utiliser que les k vecteur propres nécessaire. On définit donc $V^{(k)} \in \mathbb{R}^{Dxk}$ la réduction aux k plus grands vecteurs propres, $\Sigma^{(k)} \in \mathbb{R}^{kxk}$, la matrice Σ correspondant aux k plus grandes valeurs propres et $U^{(k)} \in \mathbb{R}^{Nxk}$ la matrice des nouvelles cooordonnées à laquelle on a gardé que les coordonnées le long des k plus grands vecteurs propres.

5 Inférence Bayésienne

L'inférence peut être défini comme la déduction des observables sur un jeu de données à l'aide d'un modèle. Le terme de "bayésienne" fait ici référence au théorème de Bayes.

5.1 Les probabilités conditionnelles

Une distribution jointe sur x et y se note $p(x,y)$. On utilisant la notation suivante $p(x|y)$ pour indiquer la probabilité d'observer x sachant qu'on a observé y.
Exemple : x : Il pleut ? ; y : la saison. $p(x|y)$, avec y =automne : représente la probabilité qu'il pleuve en automne.
Par définition, on peut aussi obtenir que : $p(x,y) = p(x|y)p(y) = p(y|y)p(y)$. On peut en déduire le **théorème de Bayes** :

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

On va maintenant regarder comment le théorème de Bayes nous permet d'inférer (de déduire) les paramètre d'un modèle. Pour cela on va illustrer le

mécanisme par une exemple sur lequel on va d'abord résouder le problème direct : sachant les paramètres du modèle, comment estimer la statistique et ensuite on passera au problème inverse : sachant un modèle et un jeu de données, comment trouver les paramètres du modèle les mieux adaptés.

5.2 Problème direct

Soit une boite qui contient K boules : N noires et B=K-N blanches. On prend une boule au hasard, et on la remet. On répète M fois. On se demande la probabilité d'obtenir N boules noire :

$$\text{proba}\,(\textit{observer N boules noiresparmintirages})\ C_M^n\left(\frac{N}{K}\right)^n(1-\ \frac{N}{K})^{\ M-n}$$

qui correspond à la loi binomiale. On notera le paramètre de la loi ici $f_n = \frac{N}{K}$. On a modélisé le problème et on en déduit une loi de probabilité et donc on peut prédire la statistique des différents observables (moyenne, variance, ...).
A présent, on peut s'intéresser au problème inverse.

5.3 Problème inverse

A partir d'une modélisation, et de données (expériences réalisées), on cherche les paramètres de la modélisation. On va considérer 11 boîtes possibles, contenant chacune 10 boules. On les étiquette par $u \in \{0,\ 1,\ \ldots,\ 10\}$: représentant le nombre de boules noires dans la boîte u. On va maintenant réaliser l'expérience suivante : on choisit une boîte au hasard, on tire M boules (avec remise). On observe n boules noires et $M - n$ boules blanches. Le but ici est donc d'inférer **quelle boîte a été utilisée**. On cherche donc le u qui a la plus grande chance (ou la plus grande probabilité) d'être la boite utilisée. C'est-à-dire : $p(u|n,M)$: la probabilité que la boite utilisée était celle avec l'étiquette u boules noires, sachant l'expérience (on a effectué M tirages).
A partir du théorème de Bayes, on réécrit le problème :

$$p\,(u|n,M) = \frac{p\,(n|u,M)\,p(u|M)}{p(n|M)}$$

Ici : $p\,(u|M) = p(u)$. On pense que la boîte est prise au hasard (notre hypothèse) et on en déduit que $p\,(u) = \frac{1}{\#boites} = \frac{1}{11}$. Mais on pourrait aussi choisir $p(u)$ en fonction des connaissances a priori. Le terme au dénominateur peut aussi s'écrire

$$p\,(n|M) = \sum_u p\,(n|u,M)\,p\,(u)$$

,
Et, sachant que la loi $p\,(n|u,M)$ est décrit par une loi binomiale de paramètre $f_u = u/10$ on obtient :

$$p(u|n,M) = \frac{1}{11}\frac{1}{p(n|M)} C_M^n \, f_u^n (1-f_u)^{M-n}$$

Figure 1
Si on souhaite maintenant trouver la boîte la plus probable, il faut chercher le maximum de $p(u|n,\ M)$, noté u^*. Dans ce cas simple on aura $u^* \sim \frac{n}{M}$ car plus le nombre de tirage est grand, plus la moyenne empirique de l'expérience s'approche de u.

5.4 Théorème de Bayes : données et modèles

Considérons maintenant la situation suivante. On observe un certain nombre de données $\{\vec{x}_i\}$ et on pense qu'elles sont bien modélisés par une distribution $p\left(\left\{\overrightarrow{x_i}\right\}_i|\theta\right)$, où θ correspond donc au paramètre de la modélisation. On peut donc écrire à l'aide du théorème de Bayes

$$p\left(\theta|\left\{\overrightarrow{x_i}\right\}\right) = \frac{p\left(\left\{\overrightarrow{x_i}\right\}|\theta\right)p(\theta)}{\sum_\theta p\left(\left\{\overrightarrow{x_i}\right\}|\ \theta\right)p(\theta)}$$