A moi
A toi
A lui
A nous non
que moi

# TD d'analyse de données. 2

1. Soient $X_1, X_2, \ldots, X_n$, $n$ variables aléatoires gaussiennes indépendantes, et des nombres positifs $\lambda_1, \lambda_2, \ldots, \lambda_n$ dont la somme est égale à 1. Comparez l'espérance mathématique et la variance de la variable aléatoire $X = \lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_n X_n$, et de la loi mélange $Y$, dont la fonction de répartition s'écrit $F = \lambda_1 F_{X_1} + \lambda_2 F_{X_2} + \ldots + \lambda_n F_{X_n}$. Que peut-on dire d'un mélange de lois uniformes ?

2. On rappelle qu'un test du Khi deux teste les fréquences d'une variable aléatoire discrète. Ce test est lié directement à la loi du Khi deux, définie pour un paramètre $r$ (Loi du Khi deux à $r$ degrés de liberté, notée en abrégé $\chi^2_r$) comme la loi de probabilité d'une somme de $r$ variables indépendantes, chaque terme étant le carré d'une variable gaussienne réduite.
   Si on note $\pi_1$, $\pi_2$ ... , $\pi_m$ les fréquences (probabilités) théoriques testées, $N$ la taille de l'échantillon, et $N_1$, $N_2$ ... , $N_m$ les fréquences absolues (effectifs) observées, l'écart entre la loi théorique testée et la loi de l'échantillon s'écrit : $D_N = \sum_{k=1}^{m} \frac{(N_k - N\,\pi_k)^2}{N\,\pi_k}$
   Comme pour tous les tests d'hypothèses la justification est liée à une propriété asymptotique. Dans le cas présent, que la loi de $D_N$ tend, quand $N$ augmente, vers une loi du Khi deux à $m - 1$ degrés de liberté. Sur un effectif de 100 observations, on observe les 3 valeurs 0, 1, et – 1 avec les fréquences respectives de 45, 22 et 33. On souhaite tester au risque de 5% les probabilités de ½, ¼ et ¼. Acceptera-t-on dans ce cas cette hypothèse, sachant qu'une loi du Khi deux à 2 degrés de liberté coïncide avec une loi exponentielle exp (½) ? (log (0.05) = – 3)

3. Si $X$ est un tableau de données ayant $n$ lignes et $m$ colonnes, on définit sa variance par les relations (avec les notations du TD 1) $V = \frac{1}{n}\, {}^tY\, Y = \frac{1}{n}\, {}^tX\, X \;-\; {}^t\hat{\mu}\; \hat{\mu}$. Les valeurs principales de l'ACP sont les valeurs propres de $V$, et ses axes principaux sont les vecteurs propres associés.

4. Recherche de l'axe principal : retrouvez que le projecteur orthogonal sur $u$ (représenté comme un vecteur colonne) a pour matrice dans la base canonique de $\mathrm{R}^n$ : $P = u\,{}^tu$ L'axe principal est le vecteur $u$ qui minimise l'écart quadratique moyen des colonnes de la matrice $X$ (ou $Y$) à leurs projections. Exprimez cet écart, pour un vecteur $u$ unitaire, et écrivez le problème comme une maximisation. Après avoir exprimé la fonction à maximiser sous forme matricielle, en déduire que $u$ doit être un vecteur propre pour la plus grande valeur propre de la matrice $W = {}^tY\ Y$.

5. On considère la matrice $X$ associée à 3 caractères, chacun ayant 10 valeurs d'échantillon.

   La transposée de $X$ s'écrit : $\begin{bmatrix} 12 & 13 & 11 & 10 & 14 & 9 & 10 & 12 & 11 & 13 \\ 2 & 3 & 4 & 1 & 4 & 2 & 3 & 1 & 2 & 4 \\ 21 & 20 & 19 & 20 & 21 & 19 & 19 & 21 & 21 & 19 \end{bmatrix}$

   Ecrivez les matrices $Y$ et $V$. Quels sont les axes principaux ?

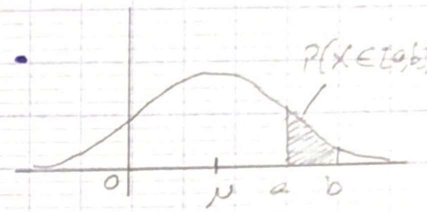
6. On considère une matrice de données $X$ telle que ${}^tX = \begin{bmatrix} 1 & 2 & 3 & 4 \\ -1 & 2 & 3 & -4 \end{bmatrix}$. Ecrivez la matrice de covariance associée. Déterminez les axes principaux, et représentez les projections des données sur le premier axe principal.

# Analyse de données : TD2

## Exo 1

Loi mélange (mélange gaussien)

- $f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}$

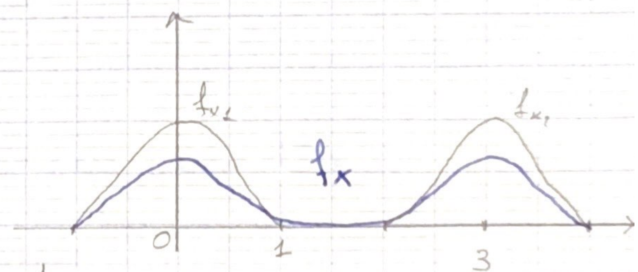



- $X_0 : N(0,1)$ $\quad \mu=0, \sigma=1$
  $X : N(\mu,\sigma)$

$\underline{X = X_0 \cdot \sigma + \mu}$

$f_X = \lambda_1 f_{X_1} + \lambda_2 f_{X_2}$

$\begin{cases} \lambda_1 = \lambda_2 = \frac{1}{2} \\ X_1 : N(0,1) \\ X_2 : N(3,1) \end{cases}$

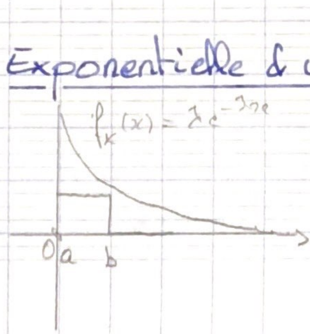



- $\int_{\mathbb{R}} f_X(x)dx = 1 = \lambda_1 \underbrace{\int_{\mathbb{R}} f_{X_1}(x)dx}_{=1} + \lambda_2 \underbrace{\int_{\mathbb{R}} f_{X_2}(x)dx}_{=1}$ -> 50% du temps la valeur sera dans $0 \pm \sigma_1$ -> les autres 50% elle sera dans $3 \pm \sigma_2$

  $= \underline{\lambda_1 + \lambda_2} = 1$ -> Fonctionne parce que $\underline{\sum \lambda_i = 1}$

- $F_X = \sum_i^n \lambda_i F_{X_i} \Rightarrow F_X' = \sum_i^n \lambda_i F_{X_i}'$

### Exponentielle & uniforme



$E(X) = \int_{\mathbb{R}} x f_X(x)\, dx$

↳ $E(X) = \lambda_1 E(X_1) + \dots + \lambda_n E(X_n)$
-> $V_{tot} = \Sigma V_i$ si les $V_i$ sont indépendantes
-> $\Sigma$ var alea gaussienne -> still gauss

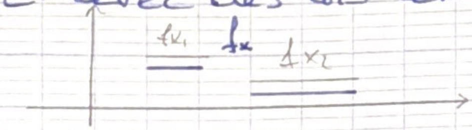
$$Y = \frac{1}{2} X_1 + \frac{1}{2} X_2 \ : N(\mu, \sigma)$$

$$\mu = \frac{1}{2} \times 0 + \frac{1}{2} \times 3 = \frac{3}{2}$$

$$\sigma^2 = \left(\frac{1}{2}\right)^2 \times 1 + \left(\frac{1}{2}\right)^2 \times 1 = \frac{1}{2} \quad \to \quad \begin{cases} N(0,1) \\ N(3,1) \end{cases}$$

$$\hookrightarrow = E((\lambda X)^2) - E(\lambda X)^2 = \lambda^2 (E(X^2) - E(X)^2)$$
$$= \lambda^2 V(X)$$
$$\hookrightarrow \sigma^2 = \Sigma \lambda_i^2 V(X_i)$$

*Et avec des lois continues ?




ça fait des fonctions en escalier.

Exo 2

$$D \underset{H_1}{\overset{H_0}{\lessgtr}} \to P(D > \eta \mid H_0) = \lambda$$

| | | Théorique $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ |
|---|---|---|
| $X = 0$ | 45 | (50) |
| 1 | 22 | (25) |
| $-1$ | 33 | (25) |

$H_0 : \begin{cases} P(0) = \frac{1}{2} \\ P(1) = P(-1) = \frac{1}{4} \end{cases}$ $n = 100$

$$D = \frac{(45-50)^2}{50} + \frac{(22-25)^2}{25} + \frac{(33-25)^2}{25}$$

$$= \frac{25}{50} + \frac{18}{50} + \frac{128}{50} = \frac{171}{50} = 3,42$$

$D \sim Z \chi^2_{2}$ → ddl

$\alpha \to \eta \;/\; \underbrace{P(Z > \eta)}_{1 - F_Z(\eta)} = \alpha$

Pour $Z$ :

$F_Z(x) = 1 - e^{-\frac{1}{2}x}$

$1 - F_Z(\eta) = e^{-\frac{1}{2}\eta} = \alpha$

Si $\alpha = 0,05$, $\eta = -2\log(\alpha)$
$= 6$

Calcul de la valeur qu'on avait dans le tableau en SE

$D < 6 \Rightarrow$ On ne rejette pas $H_0$

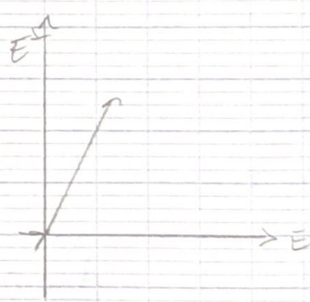
## Exo 4

$u$ : vecteur colonne

$p$ : projecteur $\perp$ sur $u$ $(n \times 1)$

$P = uu^t$

$\begin{cases} p(x) = x & \text{si } x \in E \\ p(x) = 0 & \text{si } x \in E^\perp \end{cases}$




$x$ vecteur colonne

① Si $x = \lambda u$

$Px = uu^t \lambda u$ vecteur colonne !

$= \lambda u u^t u$ scalaire somme des carrés des composantes de $x$

$\|u\|^2 = 1$

② Si $x \perp u$ : $u^t x = 0$

→ $u$ et $x$ vecteur colonne, pour le produit scalaire => on prend $u^t$ sinon on peut pas faire de produit scalaire (impossible entre vecteur colonnes)

$Px = uu^t x = 0$

Exo6

$A = \begin{bmatrix} 1 & -1 \\ 2 & 2 \\ 3 & 3 \\ 4 & -4 \end{bmatrix}$

B matrice centrée

V matrice de covariance $\frac{1}{n} B^t B$

Axes principaux (ACP)

B matrice centré

-> on fait la moyenne des colonnes

$C_1 = \frac{1+2+3+4}{4} = 2{,}5$

$C_2 = \frac{5-5}{4} = 0$

$B = \begin{bmatrix} -1{,}5 & -1 \\ -0{,}5 & 2 \\ 0{,}5 & 3 \\ 1{,}5 & -4 \end{bmatrix}$

V matrice de covarience

$B^t B = \begin{bmatrix} -1{,}5 & -0{,}5 & 0{,}5 & 1{,}5 \\ -1 & 2 & 3 & -4 \end{bmatrix} \times \begin{bmatrix} -1{,}5 & -1 \\ -0{,}5 & 2 \\ 0{,}5 & 3 \\ 1{,}5 & -4 \end{bmatrix} = \begin{bmatrix} 5 & -4 \\ -4 & 30 \end{bmatrix} \to \begin{bmatrix} 5-x & -4 \\ -4 & 30-x \end{bmatrix}$

$\text{Det} = (5-x)(30-x) - (-4)^2$

$\text{Det} = 0 \iff (5-x)(30-x) - (-4)^2 = 0$

$150 - 5x - 30x + x^2 - 16 = 0$

$x^2 - 35x + 134 = 0$

$\Delta = 35^2 - 4 \times 134 = 689$

$\lambda_1 = \frac{+35 - \sqrt{689}}{2}$

$\lambda_2 = \frac{+35 + \sqrt{689}}{2} \simeq 30$

↳ le + intéressant

Axe principal

$5x - 4y = \lambda_2 x$

car l'Axe principal est associé à la + grande des racines

↳ c'est une droite, inutile de prendre l'autre ligne car ça donnerait juste une autre équation décrivant la m̃ droite et surtout on s'interesse qu'au $\lambda i$ le plus grand

# Analyse de données : TD2

## Exo 6

$$\begin{cases} \boxed{5x - 4y = \lambda_2 x} \\ -4x + 30y = \lambda_2 x \end{cases}$$

On cherche à avoir la distance à 0 avec la + grande et on devrait avoir réussi