Quiz de stats

Introduction

Ceci est un document collaboratif. Tout le monde s'est vu attribuer une question au hasard, mais n'hésitez pas à contribuer sur d'autres questions, notamment si vous repérez des erreurs.

Afin d'effectuer un suivi de la complétion du document, mettez en regard des questions une coche lorsque la réponse est correcte et complète ✓. Si vous repérez une erreur, indiquez le avec ✗ et proposez une correction ou a minima expliquez ce qui vous semble faux.

Essayons de finaliser toutes les questions d'ici le **dimanche 6 novembre** au soir pour avoir le temps de réviser (le quiz a lieu le 9).

Pour retrouver le code LateX associé à un symbole vous pouvez le dessiner à la souris sur ce site : (merci Willie pour le lien !)
http ://detexify.kirelabs.org/classify.html

Général

Question 1 ✓

Que vaut $Cov(X + \mu)$ pour tout $\mu \in \mathbb{R}^p$ déterministe, et tout vecteur aléatoire $X \in \mathbb{R}^p$?

$$Cov(X + \mu) = Cov(X)$$

Preuve :

$$\begin{aligned} Cov(X + \mu) &= E[(X + \mu - E(X + \mu))(X + \mu - E(X + \mu))^\top] \\ &= E[(X + \mu - \mu - E(X))(X + \mu - \mu - E(X))^\top] \\ &= E[(X - E(X))(X - E(X))^\top] \end{aligned}$$

Question 2 ✓

Que vaut $Cov(AX)$, pour toute matrice $A \in \mathbb{R}^{m \times p}$ et tout vecteur aléatoire $X \in \mathbb{R}^p$?

$$Cov(AX) = ACov(X)A^\top$$

Preuve :

$$\begin{aligned} Cov(AX) &= E[(AX - E(AX))(AX - E(AX))^\top] \\ &= E[(AX - AE(X))(AX - AE(X))^\top] \\ &= E[A(X - E(X))((X - E(X))A)^\top] \\ &= AE[(X - E(X))(X - E(X))^\top A^\top] \end{aligned}$$

Question 3

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

Un modèle naturel pour un lancer de dé (à 6 faces) est une loi multinomiale de paramètres $n = 1$ et $p \in [0,1]^6$ avec $\sum p_i = 1$.

$$\mathbb{P}(X_i = i) = p_i, \forall i = \{1, ..., 6\}$$

Généralisation : soient les variables $X_i, i = \{1, ..., 6\}$ correspondantes aux probabilités $p_i, i = \{1, ..., 6\}$ Avec les conditions : $\sum n_i = n$ et $\sum p_i = 1$
La fonction de probabilité s'écrit alors :

$$\mathbb{P}(X_1 = n_1, ..., X_6 = n_6) = \frac{n!}{n_1!...n_6!} p_1^{n_1}...p_6^{n_6}$$

Question 4 ✓ : merci de sélectionner une seule méthode pour plus de lisibilité ...

Que vaut le biais de $\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2$ *(*$\overline{y}_n$ *est la moyenne empirique) pour des* y_i *i.i.d, gaussiens, centrés et de variance* σ^2 *?*

la variance empirique est $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$

(✗Remarque : la variance empirique est $S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$)

$$E(\overline{X}_n^2) = E((\frac{1}{n}\sum_{i=1}^{n} X_i)^2) = \frac{1}{n^2}E(\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} X_i X_j)$$

les X_1,...X_n *sont idd* donc

$$E(X_i^2) = E(X^2)$$

et par indépendance

$$\forall i \neq j : \quad E(X_i X_j) = E(X_i)E(X_j) = E(X)^2$$

Ainsi :

$$E(\overline{X}_n^2) = \frac{1}{n}E(X^2) + \frac{(n-1)E(X)^2}{n}$$

Par ailleurs :

$$E(\frac{1}{n}\sum_{i=1}^{n} X_i^2) = E(X^2)$$

D'où en reprenant la formule initiale :

$$E(S_n) = E(X^2) - \frac{1}{n}E(X^2) - \frac{(n-1)}{n}E(X)^2$$

$$E(S_n) = \frac{(n-1)}{n} * \sigma^2$$

donc le biais est :

$$Bias = \frac{(n-1)}{n} * \sigma^2 - \sigma^2$$

$$= -\frac{1}{n}\sigma^2$$

=> Autre méthode :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

est l'estimateur de σ^2 donc, le biais vaut :

$$E(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2) - \sigma^2$$

$$= E[\frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2 * \frac{1}{n}\sum_{i=1}^{n} y_i * \overline{y} + \frac{1}{n}\sum_{i=1}^{n}(\overline{y})^2] - \sigma^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(y_i^2) - 2 * (\frac{1}{n}\sum_{i=1}^{n} y_i)^2 + (\overline{y})^2 - \sigma^2$$

or

$$E(y_i^2) = Var(y_i) + E(y_i)^2 = Var(y_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} Var(y_i) - E(\overline{y}^2) - \sigma^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} Var(y_i) - Var(\overline{y}) + E(\overline{y}) - \sigma^2$$

$$= \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

Question 5

On suppose que l'on observe $y_1, \ldots, y_n$, des variables réelles i.i.d., gaussiennes, centrées et de variance σ^2. Quel est le risque quadratique de l'estimateur $\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2$ de σ^2 ($\overline{y}_n$ est la moyenne empirique) ?

Le risque quadratique d'un estimateur est égal à : R = Variance + $Biais^2$ de l'estimateur.

Nous avons le biais dans la question précédente. Calculons la variance de notre estimateur :

Soit $(X_1, X_2, ..., X_n)$ de loi normale centrée réduite N(0,1).

On a alors $\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$ suit une loi du χ^2 à (n-1) degrés de liberté (n degrés de libertés d'après de nombreux site.

Or on connait la variance de la loi du χ^2. On a donc :
$\text{Var}(\sum_{i=1}^{n}(x_i - \overline{x}_n)^2) = 2(\text{n-1})$.
Or ici, Yi suit une loi N(0,σ^2) donc $X_i = \dfrac{Yi}{\sigma}$
On remplace dans l'expression, on obtient :
$\text{Var}(\sum_{i=1}^{n}(x_i - \overline{x}_n)^2) = \text{Var}(\sum_{i=1}^{n}(\dfrac{yi}{\sigma} - \dfrac{\overline{y}_n}{\sigma})^2)$
$= \text{Var}\ (\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2)$
$= \dfrac{1}{\sigma^4}\ \text{Var}\ (\sum_{i=1}^{n}(y_i - \overline{y}_n)^2)$
Donc $\text{Var}\ (\dfrac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2) = \dfrac{2(n-1)\sigma^4}{n^2}$
Donc, finalement le risque de notre estimateur est égale à
$\text{R} = \dfrac{2(n-1)\sigma^4}{n^2} + \dfrac{\sigma^4}{n^2}$(biais question précédente au carré)
$\text{R} = \dfrac{\sigma^4(2n-1)}{n^2}$

Question 6 ✓

Quelle est la projection du vecteur $\mathbf{y} \in \mathbb{R}^n$ *sur* $\text{Vect}(1_n)$*, avec* $1_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ *?*

$$\frac{< y, 1_n > 1_n}{< 1_n, 1_n >} = \frac{1_n \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} 1} = \bar{y} 1_n$$

Question 7 ✓

Quels sont les vecteurs $\mathbf{y} \in \mathbb{R}^n$ *tels que* $\text{var}_n(\mathbf{y}) = 0$ *(*var_n *est la variance empirique) ?*

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = 0 \iff \forall i, y_i = \bar{y}_n$$

Donc tous les vecteurs constants.

Moindres carrés unidimensionnels : on observe $\mathbf{y} = (y_1, \ldots, y_n)^\top$ et $\mathbf{x} = (x_1, \ldots, x_n)^\top$

Question 8 ✗

La fonction $(\theta_0, \theta_1) \to \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2$ *est-elle convexe ou concave ?*

(a) f est convexe ssi H, la matrice Hessienne (la matrice des dérivées partielles secondes) de f est semi-définie positive, ssi les valeurs propres de H sont positives ou nulles, ssi det(H) >= 0 et tr(H) >= 0

(b) C'est une fonction convexe car c'est polynomial de degré deux en θ_0 et en θ_1 et les monômes de plus hauts degrés en theta sont à coefficients positifs.

Dans cette question, il faut calculer les dérivées partielles de f en fonction de θ_0, puis de θ_1, puis des deux, construire la matrice Hessienne (composée des dérivées secondes sur θ_0 et θ_1). Le déterminant de la Hessienne est :

$$det(H) = n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2$$

Pour prouver la convexité de la fonction, il suffit de montrer que le déterminant est positif (c'est Faux : pour une matrice de dimension 2, il faut aussi que la trace soit positive...) !

Il suffit d'utiliser l'inégalité de Cauchy Schwarz pour prouver que le determinant est positif :

$$\left(\sum_{i=1}^{n} x_i\right)^2 \leq \left(\sum_{i=1}^{n} 1^2\right)\left(\sum_{i=1}^{n} x_i^2\right)$$

Donc :

$$\left(\sum_{i=1}^{n} x_i\right)^2 \leq n * \left(\sum_{i=1}^{n} x_i^2\right)$$

Finalement :

$$0 \leq det(H) => convexe$$

Question 9 ✓

Donner la formule $(\hat{\theta}_0, \hat{\theta}_1)$ *des estimateurs des moindres carrés où* $\hat{\theta}_0$ *correspond au coefficient des constantes et* $\hat{\theta}_1$ *correspond à l'influence de* $\mathbf{x}$ *sur* $\mathbf{y}$*. On les exprimera en fonction des* $x_i, y_i, \overline{x}_n, \overline{y}_n$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} = \frac{\overline{x}\ \overline{y} - \overline{xy}}{\overline{x}^2 - \overline{x^2}} = \frac{cov(x,y)}{var(x)}$$

$$\hat{\theta}_0 = \overline{y}_n - \hat{\theta}_1 \overline{x}_n$$

Moindres carrés : $\mathbf{y} = (y_1, \ldots, y_n)^\top$ et $X \in \mathbb{R}^{n \times p}$

Question 10

Écrire un pseudo-code de descente de gradient pour résoudre le problème des moindres carrés.
Solution de Joseph himself (cf. Mooc Big Data)

Données initiales : $x^{(0)}$, nombre maximum d'itérations T, critère d'arrêt ϵ, pas α

for t in 0 : T do :

$$x^{(t+1)} \leftarrow x^{(t)} - < \alpha, \nabla f(x^{(t)}) >$$

STOP si critère d'arrêt inférieur à ϵ

Plusieurs critères d'arrêt possibles, par exemple :

$$x^{(t+1)} - x^{(t)} < \epsilon$$

Résultat : un point $x^{(t_{arret})}$ proche du minimum de la fonction f.

En python :

```
tolerance = 1e-6
stop_condition = tolerance + 1
max_iter = 10000
m = 0
step = 1e-1
n = array_datas.shape[0]
theta_gradient = np.zeros(7)
previous_theta_sum = 0.0
while stop_condition > tolerance and m < max_iter:
    prediction = np.dot(array_x,
        theta_gradient).reshape(n, 1)
    error = prediction - array_y
    gradient = np.multiply(array_x, error)
    theta_gradient -= step * np.mean(gradient, axis=0)
    stop_condition = np.sum(theta_gradient) -
                previous_theta_sum
```

```
previous_theta_sum = np.sum(theta_gradient)
m += 1
```

Question 11 ✓

Pour une matrice $X \in \mathbb{R}^{n \times p}$*, que vaut* $\mathrm{Ker}(X^\top X)$ *?*

$$Ker(X^\top X) = Ker(X)$$

Preuve :

(i) Montrons que $Ker(X) \subset Ker(X^\top X)$:

$\forall a \in Ker(X) : Xa = 0$ (0 étant un vecteur nul de taille p)

$\Rightarrow X^\top Xa = 0$ (par multiplication à gauche par $X^\top$)

$\Rightarrow a \in Ker(X^\top X)$

$\Rightarrow Ker(X) \subset Ker(X^\top X)$

(ii) Montrons que $Ker(X^\top X) \subset Ker(X)$:

$\forall a \in Ker(X^\top X) : X^\top Xa = 0$ (0 étant un vecteur nul de taille p)

$\Rightarrow a^\top X^\top Xa = 0$ (par multiplication à gauche par $a^\top$)

$\Rightarrow \|Xa\| = 0 \Rightarrow Xa = 0$ (vecteur nul de taille p)

$\Rightarrow a \in Ker(X) \Rightarrow Ker(X) \subset Ker(X^\top X)$

D'après (i) et (ii) on a bien l'égalité des deux ensembles.

Question 12 ✓

Si la matrice $X \in \mathbb{R}^{n \times p}$ est de plein rang, donner une formule exacte de l'estimateur des moindres carrés.

$$\hat{\theta} = (X^\top X)^{-1} X^\top y$$

Preuve :

Partant de :

$$\nabla f(\theta) = X^\top X\theta - X^\top y$$

On obtient (en appliquant la CPO en $\hat{\theta}$) : (i)

$$\nabla f(\hat{\theta}) = 0 \Leftrightarrow X^\top X\hat{\theta} = X^\top y$$

Or si X est de plein rang, alors $X^\top X$ est inversible.
Démonstration :

$$X^\top X \text{inversible} \Leftrightarrow Ker(X^\top X) = \{0\}$$

En utilisant le résultat de la question 11 :

$$\Leftrightarrow Ker(X) = \{0\}$$

$$\Leftrightarrow X \text{ est de plein rang.}$$

On peut donc appliquer une multiplication matricielle par la gauche de $(X^\top X)^{-1}$. Ainsi, en repartant de (i) on a :

$$X^\top X\hat{\theta} = X^\top y \Leftrightarrow (X^\top X)^{-1} X^\top X\hat{\theta} = (X^\top X)^{-1} X^\top y$$

$$\Leftrightarrow \hat{\theta} = (X^\top X)^{-1} X^\top y$$

Question 13 ✓

Si la matrice $X \in \mathbb{R}^{n \times p}$ est de plein rang, donner la matrice de covariance de l'estimateur des moindres carrés (dans l'hypothèse d'un bruit $\epsilon = \mathbf{y} - X\theta^*$ centré et de matrice de covariance $\sigma^2 \mathrm{Id}_n$).

$$Cov(\hat{\theta}) = \sigma^2 (X^\top X)^{-1}$$

Preuve :

$$\begin{aligned}
var(\hat{\theta}) &= var((X^\top X)^{-1} X^\top y) \\
&= var((X^\top X)^{-1} X^\top (\epsilon + X\theta^*) \\
&= var((X^\top X)^{-1} X^\top \epsilon + (X^\top X)^{-1} X^\top X\theta^*) \\
&= var((X^\top X)^{-1} X^\top \epsilon + \theta^*) \\
&= var((X^\top X)^{-1} X^\top \epsilon) \\
&= (X^\top X)^{-1} X^\top var(\epsilon)((X^\top X)^{-1} X^\top)^\top \\
&= (X^\top X)^{-1} X^\top var(\epsilon) X((X^\top X)^{-1})^\top
\end{aligned}$$

or $(X^\top X)^{-1}$ symétrique donc $((X^\top X)^{-1})^\top = (X^\top X)^{-1}$

$$\begin{aligned}
&= (X^\top X)^{-1} X^\top X var(\epsilon)(X^\top X)^{-1} \\
&= \sigma^2 (X^\top X)^{-1}
\end{aligned}$$

Autre preuve (celle de la slide 39/51 du Chapitre 3) :

$$V = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\hat{\theta} - \mathbb{E}[\hat{\theta}])^\top] = \mathbb{E}[(\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*])^\top]$$

$$V = \mathbb{E}[((X^\top X)^{-1} X^\top (X\theta^* + \epsilon) - \theta^*)((X^\top X)^{-1} X^\top (X\theta^* + \epsilon) - \theta^*)^\top]$$

$$V = \mathbb{E}[((X^\top X)^{-1}X^\top \epsilon))((X^\top X)^{-1}X^\top \epsilon))^\top]$$
$$V = (X^\top X)^{-1}X^\top \mathbb{E}[\epsilon\epsilon^\top]X(X^\top X)^{-1}$$
$$V = (X^\top X)^{-1}X^\top (\sigma^2 Id_n)X(X^\top X)^{-1}$$
$$V = \sigma^2(X^\top X)^{-1}$$

Question 14 ✓

Donner la formulation de la pseudo inverse si la SVD de X peut s'écrire $X = \sum_{i=1}^{r} s_i \mathbf{v}_i \mathbf{u}_i^\top$.

$$X^+ = \sum_{i=1}^{r} \frac{1}{s_i} \mathbf{u_i} \mathbf{v_i}^\top$$

Preuve :

La matrice X peut s'ecrire de la manière suivante :

$$X = U\Sigma V^\top$$
$$XX^+ = Id$$
$$U\Sigma V^\top X^+ = Id$$
$$U^\top U\Sigma V^\top X^+ = U^\top$$
$$\Sigma V^\top X^+ = U^\top$$

car U est une matrice orthogonale

$$\Sigma^{-1}\Sigma V^\top X^+ = \Sigma^{-1}U^\top$$
$$VV^\top X^+ = V\Sigma^{-1}U^\top$$

comme V est une matrice orthogonale

$$X^+ = \sum_{i=1}^{r} \frac{1}{s_i} \mathbf{u_i} \mathbf{v_i}^\top$$

Question 15

Donner une formule explicite du problème $\arg\min_\theta \frac{1}{2}(\mathbf{y}-X\theta)^\top\Omega(\mathbf{y}-X\theta)$ *pour une matrice* $\Omega = \mathrm{diag}(w_1,\ldots,w_n)$ *définie positive.*

Faire les changements de variable :

$$X' = \sqrt{\Omega}X = diag(\sqrt{w_1},...,\sqrt{w_n})X$$

$$Y' = \sqrt{\Omega}Y = diag(\sqrt{w_1},...,\sqrt{w_n})Y$$

Puis résoudre les moindres carrés comme dans le cours :

$$\begin{aligned}\widehat{Y'} &= X'(X'^\top X')^{-1}X'^\top Y' \\ \Leftrightarrow \sqrt{\Omega}\widehat{Y} &= \sqrt{\Omega}X(X^\top\sqrt{\Omega}^\top\sqrt{\Omega}X)^{-1}X^\top\sqrt{\Omega}^\top\sqrt{\Omega}Y \\ \Leftrightarrow \widehat{Y} &= X(X^\top\Omega X)^{-1}X^\top\Omega Y = X\hat{\theta}\end{aligned}$$

Ok pour changement de variable, ensuite :

$$\begin{aligned}\hat{\theta} &\in \arg\min_{\theta\in\Theta}\frac{1}{2}(\mathbf{y}-X\theta)^\top\Omega(\mathbf{y}-X\theta) \\ \Leftrightarrow \hat{\theta} &\in \arg\min_{\theta\in\Theta}\frac{1}{2}(\mathbf{y}'-X'\theta)^\top(\mathbf{y}'-X'\theta)\end{aligned}$$

Une solution est donnée (en cas d'inversibilité) par :

$$\begin{aligned}\hat{\theta} &= (X'^\top X')^{-1}X'^\top\mathbf{y}' \\ &= (X^\top\sqrt{\Omega}^\top\sqrt{\Omega}X)^{-1}X^\top\sqrt{\Omega}^\top\sqrt{\Omega}Y \\ &= (X^\top\Omega X)^{-1}X^\top\Omega Y\end{aligned}$$

Si l'on ne peut présupposer de l'invertibilité, une solution est donnée par :

$$\begin{aligned}\hat{\theta} &= X'^{+}\mathbf{y}' \\ &= \sum_{i=1}^{r} \frac{1}{s'_i}\mathbf{u'_i}\mathbf{v'_i}^{\top}\mathbf{y}'\end{aligned}$$

On a $X' = \sqrt{\Omega}X = \sqrt{\Omega}USV^T$, or $\sqrt{\Omega}U$ n'est pas orthogonale en général, donc il n'y a à priori pas de lien évident entre la SVD de X' et celle de X, idem pour la pseudo-inverse.

Ridge

On note $\hat{\theta} = \arg\min_{\theta} \frac{1}{2}\|\mathbf{y} - X\theta\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2$ l'estimateur ridge.

Question 16

Donner une formule explicite pour l'estimateur Ridge en fonction de y et λ quand $X = \mathrm{Id}_n$.

$$\hat{\theta}_\lambda^{rdg} = \frac{y}{\lambda + 1}$$

Preuve :

$$\hat{\theta}_\lambda^{rdg} = (\lambda Id_p + X^\top X)^{-1} X^\top y$$

Si $X = Id_n$, on a :

$$\begin{aligned}\hat{\theta}_\lambda^{rdg} &= (Id_p + \lambda Id_p)^{-1} Id_p y \\ \hat{\theta}_\lambda^{rdg} &= (\lambda + 1)^{-1} y \\ \hat{\theta}_\lambda^{rdg} &= \frac{y}{\lambda + 1}\end{aligned}$$

Question 17

Donner une formule explicite pour l'estimateur Ridge en fonction de X, y et λ.

$$\hat{\theta}_\lambda^{rdg} = (\lambda Id_p + X^\top X)^{-1} X^\top y$$

Preuve :

On part du gradient :

$$\nabla f(\theta) = X^\top (X\theta - y) + \lambda\theta$$

Par la CDO en $\hat{\theta}_\lambda^{rdg}$:

$$\nabla f(\theta) = 0 \Leftrightarrow (\lambda Id_p + X^\top X)\hat{\theta}_\lambda^{rdg} = X^\top y$$
$$\Leftrightarrow \hat{\theta}_\lambda^{rdg} = (\lambda Id_p + X^\top X)^{-1} X^\top y$$

Question 18 ✗ : on pourrait avoir juste la réponse quand même svp ?

Donner la variance de l'estimateur Ridge sous l'hypothèse que le bruit $\mathbf{y} - X\theta^*$ est centré et de variance $\sigma^2 \mathrm{Id}_n$.
$Var(\theta)_\lambda^{rdg} = E((\widehat{\Theta_\lambda}^{rdg} - E(\widehat{\Theta_\lambda}^{rdg}))(\widehat{\Theta_\lambda}^{rdg} - E(\widehat{\Theta_\lambda}^{rdg})^\top))$

$$Var(\hat{\theta}_\lambda^{Ridge}) = \sum_{i=1}^{rang(X)} \frac{s_i^2 \sigma^2}{(s_i^2 + \lambda)^2} \mathbf{v}_i \mathbf{v}_i^T$$

Question 19

Donner une formule explicite pour l'estimateur Ridge généralisé : $\hat{\theta} = \arg\min_\theta \frac{1}{2}\|\mathbf{y} - X\theta\|_2^2 + \frac{\lambda}{2}\|D\theta\|_2^2$, en fonction de X, y,

$D \in \mathbb{R}^{p \times p}$ *et* λ.

$$\hat{\theta} = (X^\top X + \lambda D^\top D)^{-1} X^\top y$$

Preuve :

Soit,

$$f(\theta) = \frac{1}{2}\|\mathbf{y} - X\theta\|_2^2 + \frac{\lambda}{2}\|D\theta\|_2^2$$

Alors,

$$\nabla f(\theta) = X^t(X\theta - y) + \lambda D^\top D\theta$$

En appliquant la CTO en $\hat{\theta}$:

$$\nabla f(\hat{\theta}) = 0 \Leftrightarrow (X^\top X + \lambda D^t D)\hat{\theta} = X^\top y$$

Sous réserve que $(X^\top X + \lambda D^\top D)$ soit inversible (peut on le prouver ?)

Lasso

Question 20

Calculer $\eta_\lambda(Z) = \arg\min_{x \in \mathbb{R}} x \mapsto \frac{1}{2}(z-x)^2 + \lambda|x|$ *en fonction du signe de* x *et de la partie positive* $(\cdot)_+$

$$\eta_\lambda(Z) = sign(z)(|z| - \lambda)_+$$

Preuve :

Calculer le sous gradient de $J(x, \lambda, z) = \frac{1}{2}(z-x)^2 + \lambda|x|$
le sous gradient est appelé SG
$SG = -(z-x) + \lambda.F(x, \lambda, z)$

$F(x, \lambda, z) = sign(x)$ pour la fonction $sign : \mathbb{R} \to \mathcal{P}(\mathbb{R})$ voir slide 16/66 du cours lasso puis on en déduit (facilement ?) la forme explicite du slide 17/66

le cas $x = 0$ donne en ce qui concerne le sous gradient que

$$0 \in -z + x + \lambda[-1; +1]$$

ce qui revient à $|z| \leq \lambda$ ($\lambda \in \mathbb{R}^+$)

la formulation de la question est bizarre d'autant qu'au slide 17/66 x n'apparait pas ... du coup quid de la parte positive et du signe de "x"

Question 21 ✓

Donner en tout point la sous-différentielle de la fonction réelle $x \mapsto (x)_+ = \max(x, 0)$.

Notons f cette fonction. C'est une fonction convexe, elle a donc une sous-différentielle non-vide. Cet ensemble est de la forme :

$$\partial f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x > 0 \\ [0; 1] & \text{si } x = 0 \end{cases}$$

Question 22

*Donner l'étape de mise à jour principale en descente par coordonnée pour résoudre le problème de l'*Elastic Net* :*
$\hat{\theta}_\lambda = \arg\min_{\theta \in \mathbb{R}^p} \left[\frac{1}{2} \|\mathbf{y} - X\theta\|_2^2 + \lambda \left(\alpha \|\theta\|_1 + (1 - \alpha) \frac{\|\theta\|_2^2}{2} \right) \right]$.

L'étape de mise à jour principale est la suivante :

$$\theta_j^{(k+1)} \leftarrow \eta_{ST,\frac{\alpha\lambda}{\|\mathbf{x}_j\|^2+\lambda(1-\alpha)}} \left(\frac{\mathbf{x}_j^\top r^{int}}{\|\mathbf{x}_j\|^2 + \lambda\,(1-\alpha)} \right)$$

Remarque : on devrait obtenir

$$\theta_j^{(k+1)} \leftarrow \eta_{ST,\alpha,\lambda/\|\mathbf{x}_j\|^2}(\mathbf{x}_j^\top r^{int}/\|\mathbf{x}_j\|^2)$$

Où :

$$\eta_{ST,\alpha,\lambda} = \frac{sign(z)}{1+\lambda(1-\alpha)}(|z| - \lambda\alpha)_+$$

Question 23

*Donner l'étape de mise à jour principale en descente par coordonnée pour résoudre le problème du *Lasso Positif* :*
$\hat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}_+^p} \frac{1}{2}\|\mathbf{y} - X\theta\|_2^2 + \lambda\|\theta\|_1$.

L'étape de mise à jour principale est la suivante :

$$\theta_j^{(k+1)} \leftarrow \eta_{ST+,\lambda/\|\mathbf{x}_j\|^2}(\mathbf{x}_j^\top r^{int}/\|\mathbf{x}_j\|^2)$$

Où $\eta_{ST+,\lambda} = (z-\lambda)_+$

Preuve :

Voir slide 56 du cours sur le Lasso qui précise les calculs intermédiaires de la descente par coordonnée pour le Lasso.
Pour l'adapter au Lasso positif il suffit d'expliciter :

$$\eta_{ST+,\lambda}(z) = \arg\min_{t\in\mathbb{R}_+} \frac{1}{2}(z-t)^2 + \lambda t$$

On note $f(t) = \frac{1}{2}(z-t)^2 + \lambda t$.

$$f(t) = \frac{z^2}{2} + \frac{t^2}{2} - t(z-\lambda)$$

Donc :

$$f'(t) = t - (z-\lambda)$$

La dérivée de f est strictement croissante sur $\mathbb{R}$ et nulle en $(z-\lambda)$. Le minimum de f sur $\mathbb{R}_+$ est donc atteint en $t = (z-\lambda)$ si c'est positif, ou en $t = 0$ sinon. D'où $\eta_{ST+,\lambda} = (z-\lambda)_+$

Question 24

On suppose que l'on dispose d'un solveur $\text{Lasso}(X, y, \lambda)$ *qui résout le problème du Lasso* $\hat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\theta\|_2^2 + \lambda\|\theta\|_1$. *En utilisant ce solveur comment résoudre le problème suivant :* $\hat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\theta\|_2^2 + \lambda\sum_{j=1}^{p} w_j|\theta_j|$, *pour des* $w_j \geq 0$ *?*

Réponse : On definit A $=diag(\omega_1, ..., \omega_p)$, on pose $\gamma = A\theta$ ainsi $\gamma_i = w_i\theta_i \quad \forall i$.

On souhaite resoudre :

$$\hat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\theta\|_2^2 + \lambda\sum_{j=1}^{p} w_j|\theta_j|, \text{pour des } w_j \geq 0$$

ce qui est équivalent à résoudre :

$$\hat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - XA^{-1}\gamma\|_2^2 + \lambda\sum_{j=1}^{p} |\gamma_j| \quad (\star)$$

On note $\tilde{X} = XA^{-1}$

$$(\star) \Leftrightarrow \hat{\theta}_\lambda = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \tilde{X}\gamma\|_2^2 + \lambda \sum_{j=1}^{p} |\gamma_j|$$

ce qui correspond à un problème lasso classique et peut être resolu grâce au solver avec $\tilde{X}$ et γ, on obtient ainsi les $\hat{\gamma}_\lambda$ et on peut en déduire les $\hat{\theta}_\lambda = A^{-1} * \hat{\gamma}_\lambda$

ACP/SVD

Question 25

$$\textit{Que vaut} \begin{cases} \max\limits_{u \in \mathbb{R}^n, v \in \mathbb{R}^p} u^\top X v \\ \textit{s.c. } \|u\|_2^2 = 1 \textit{ et } \|v\|_2^2 = 1 \end{cases} \quad ?$$

slide 6/33 mod lin biais variance .
le max est la plus grande valeur singulière de X
Ecriture du Lagrangien lié au problème de maximisation :

$$L(u, v) = u^\top X v - \lambda_1(\|u\|^2 - 1) - \lambda_2(\|v\|^2 - 1)$$

Conditions du premier ordre :

$$\begin{cases} \nabla_u L = Xv - 2\lambda_1 u = 0 \\ \nabla_v L = X^\top u - 2\lambda_2 u = 0 \end{cases} \iff \begin{cases} Xv = 2\lambda_1 u \\ X^\top u = 2\lambda_2 v \end{cases} \implies \begin{cases} X^\top X v = \alpha v \\ X X^\top u = \alpha u \end{cases}$$

Avec
$\alpha = 4\lambda_1\lambda_2$, d'après le théorème Spectrale (4/33), on a v et u les vecteurs propres associés à $X^\top X$ et $XX^\top$ ayant les mêmes valeurs propres.

On peut alors décomposer l'expression à maximiser de la façon suivante :
$u^\top Xv = diag(s_1, ..., s_{min(n,p)})$ avec $s_1 \geqslant s_2 \geqslant ... \geqslant s_{min(n,p)} \geqslant 0$

En respectant les contraintes, il existe un s_1 tel que l'on peut écrire l'égalité suivante :

$$s_1 = \begin{cases} \max\limits_{u \in \mathbb{R}^n, v \in \mathbb{R}^p} u^\top Xv \\ \text{s.c. } \|u\|_2^2 = 1 \text{ et } \|v\|_2^2 = 1 \end{cases}$$

Test

Question 26

Pour des $X_1, \ldots, X_n$ identiquement distribuées à valeur dans $\{0, 1\}$, décrire une procédure de test de l'hypothèse $p = P(X_1 = 1) = 1/2$ contre son contraire.

Diapo ICTests, voir entre 6 et 10/27

Proposition de solution :
Soit $X = 1/n * \sum_{i=1}^{n} X_i$, X suit ainsi une loi de Bernoulli $\mathcal{B}(p, p(1-p))$ On cherche à décrire une procédure de test pour le paramètre p = 1/2. On teste l'hypothèse de test H_0 vs H_1 : $\begin{cases} H_0 : p = \frac{1}{2} \\ H_1 : p \neq \frac{1}{2} \end{cases}$
Grâce au TCL, on peut montrer que :

$$\sqrt{n} \frac{X - p}{\sqrt{p(1-p)}} \sim \mathcal{N}(0, 1)$$

Ainsi, on peut dire, au risque α et pour ϵ_k les quantiles d'une loi

normale centrée réduite, qu'on accepte l'hypothèse H_0 si :

$$\mathbb{P}(\epsilon_{\alpha/2} < \sqrt{n}\frac{X-p}{\sqrt{p(1-p)}} < \epsilon_{1-\alpha/2}) > 1-\alpha$$

On la refuse sinon.
On peut aussi trouver un intervalle de confiance au risque 5% pour le paramètre p. il suffit d'encadrer le paramètre p mais je n'ai malheureusement pas eu le courage de le faire.

Question 27

*Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d selon des lois gaussiennes de moyenne (inconnue) μ et de variance connue σ^2, *i.e.,* $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Décrire une procédure de test de l'hypothèse $\mu = 1$ contre son contraire.*

On a : $X_1, \ldots, X_n$ des VAs iid qui suivent une loi $\mathrm{N}(\mu, \sigma^2)$.
Procédure de test de l'hypothése μ = 1 contre son contraire La statistique de test choisie est : $T_n(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$
La loi de la statistique de test est : $\mathrm{E}[T_n] = \mu$ et $\mathrm{Var}[T_n] = \frac{\sigma^2}{n}$
Sous H_0, la loi de la statistique de test T_n est donc : $\mathrm{E}[T_n] = 1$ et $\mathrm{Var}[T_n] = \frac{\sigma^2}{n}$.

Sous H_0 et d'après le théorème Central Limite, on a :
$\mathrm{Z} = \sqrt{\frac{T_n - 1}{\sigma}} \sim N(0,1)$.

Détermination d'un IC de niveau 1-α pour l'acceptation de T_n :
$\mathrm{P}(t_{\frac{\alpha}{2}} \leq z \leq t_{1-\frac{\alpha}{2}}) = 1$ - $\alpha \Leftrightarrow$
Diapo ICTests, voir 10/27

Autre solution qui me paraît correcte :
L'intervalle de confiance au risque α pour μ est donné par :

$$\left[\bar{x} - t_{1-\alpha/2}^{n-1}\sqrt{\frac{S}{n}}, \bar{x} + t_{1-\alpha/2}^{n-1}\sqrt{\frac{S}{n}}\right]$$

Avec $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
et $S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Preuve :

$$T_0 = \frac{\bar{x} - \mu}{\sqrt{S}/\sqrt{n}}$$

Remarque : On peut voir cette expression comme la TCL avec l'estimateur de la variance non biaisée

T_0 suit une loi de Student de paramètre $n-1$. Donc, on a :

$$\mathbb{P}(t_{\alpha/2}^{n-1} < T_0 < t_{1-\alpha/2}^{n-1}) = 1 - \alpha$$

En encadrant μ on retombe sur la solution. Ainsi, si μ appartient à l'intervalle, on valide l'hypothèse au risque 5%.

Question 28

*Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et distribuées selon des lois gaussiennes de moyenne (inconnue) μ et de variances connues σ_i^2, *i.e.,* $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$. Décrire une procédure de test de l'hypothèse $\mu = 1$ contre son contraire.*

Diapo ICTests, voir 10/27

Proposition de correction :

On propose de réduire les variables. On effectue le changement de variable suivant :

$Y_i = X_i \ / \ \sigma_i$

Dès lors Y_i est équivalent à une loi normale de moyenne μ et de variance 1.

On peut reprendre la démarche du 27

Bootstrap

Question 29

*Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d selon des lois gaussiennes de moyenne (inconnue) μ et de variance connue σ^2, *i.e.,* $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Écrire un pseudo code de bootstrap pour le test sur la moyenne $\mu = 1$.*

Notons $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} Xi$ l'estimateur sans biais de l'espérance μ des X_i.

Notons $\hat{R}_n = \sqrt{n}(\bar{X}_n - \mu)$ la racine (convergente) de notre estimateur.

Réalisons un test de niveau $1 - \alpha$, α étant donc le risque de première espèce de notre test, en posant :
H_0 : L'espérance μ vaut 1.
H_1 : L'espérance $\mu \neq 1$.

Sous l'hypothèse H_0, on a :
$\hat{R}_n = \sqrt{n}(\bar{X}_n - 1)$

Nous allons comparer la valeur de $\bar{X}_n$ sur l'échantillon intial par rapport à l'intervalle de confiance obtenu par méthode Bootstrap.

Bootstrap :
for b in B :
On créé un nouvel échantillon : tirage avec remise : $(X_i)^{*b}_{i\in[1,n]}$
On calcule la racine bootstrap de l'échantillon : $\hat{R}^{*b}_n = \sqrt{n}(\bar{X}^{*b}_n - 1)$

Et en notant $\hat{\xi}^{(bb)}_{B,\alpha}$ le α-quantile de de l'estimation de la loi de distribution de $\hat{R}^*_n$).

On obtient l'intervalle de confiance suivant :

$$I_{(1-\alpha)}(\bar{X}_n)^{(bb)} = \left[1 + \frac{\hat{\xi}^{(bb)}_{\alpha/2}}{\sqrt{n}}, 1 + \frac{\hat{\xi}^{(bb)}_{1-\alpha/2}}{\sqrt{n}}\right]$$

Ainsi on accepte H_0 avec un risque de première espèce α si $\bar{X}_n \in \left[1 + \frac{\hat{\xi}^{(bb)}_{\alpha/2}}{\sqrt{n}}, 1 + \frac{\hat{\xi}^{(bb)}_{1-\alpha/2}}{\sqrt{n}}\right]$.

Question 30 ✗

Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d et $w_1, \ldots, w_n$ une suite de variables i.i.d. de moyenne 1 et de variance 1. A l'aide des X_i et des w_i construire un intervalle de confiance à 99% pour la quantité $\mathbb{P}(X_1 \geq 10)$.

Je propose un intervalle type percentile bootstrap (plus simple de s'en souvenir comparé aux autres)

$[q_{\alpha/2}; q_{1-\alpha/2}]$ est l'intervalle des quantiles associés à ... il faut supposer n assez grand

Proposition de réponse (Basic bootstrap, reste à prouver la convergence de la racine avec les w_i) : On va chercher

à estimer la fonction de répartition F_X de X par une méthode Bootstrap.
Notons $\hat{F}_n(x)$ notre estimateur de F_X en x, défini comme suit :

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}$$

Notons $R_n(x)$ une racine de notre estimateur (dont nous admettrons la convergence étant donné que je ne sais pas comment la prouver, voir page 41/53 du cours Bootstrap, merci d'avance à celui qui pourrait aider à le prouver) :

$$R_n(x) = \sqrt{n}\left(\frac{\hat{F}_n(x) - F_X(x)}{\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}}\right)$$

Pour chacun des échantillons X^{*k} obtenus par méthode bootstrap, définissons la racine bootstrap suivante (dont nous admettrons aussi qu'elle mime le comportement de notre racine intiale, attention, je ne suis pas sûr que cette racine bootstrap soit la bonne) :

$$R_n^{*k}(x) = \sqrt{n}\left(\frac{\hat{F}_n^{*k}(x) - \hat{F}_n(x)}{\sqrt{\hat{F}_n^{*k}(x)(1-\hat{F}_n^{*k}(x))}}\right)$$

Il ne reste plus qu'à construire l'intervalle de confiance à partir de l'estimation de la densité de notre racine.

En notant $R_n^*(x)$ la distribution des $R_n^{*k}(x)$.
Et en notant $\hat{\xi}_{B,\alpha}(x)$ le α-quantile de $R_n^*(x)$ (obtenue suite à un nombre B de rééchantillonnages bootstrap).

On obtient l'intervalle de confiance suivant :

$I_{(1-\alpha)}(F_X(x)) =$
$\left[\hat{F}_n(x) - \frac{\hat{\xi}_{1-\alpha/2}(x)\sqrt{\hat{F}_n^{*k}(x)(1-\hat{F}_n^{*k}(x))}}{\sqrt{n}}, \hat{F}_n(x) - \frac{\hat{\xi}_{\alpha/2}(x)\sqrt{\hat{F}_n^{*k}(x)(1-\hat{F}_n^{*k}(x))}}{\sqrt{n}}\right]$

D'où : $I_{(0.99)}(F_X(10)) =$
$\left[\hat{F}_n(10) - \frac{\hat{\xi}_{0.995}(10)\sqrt{\hat{F}_n^{*k}(10)(1-\hat{F}_n^{*k}(10))}}{\sqrt{n}}, \hat{F}_n(10) - \frac{\hat{\xi}_{0.005}(10)\sqrt{\hat{F}_n^{*k}(10)(1-\hat{F}_n^{*k}(10))}}{\sqrt{n}}\right]$

Finalement, on en déduit par simple transformation $P(X \geq 10)$ (étant donné que $P(X \geq 10) = P(X > 10)$ (cas continu) $= 1 - P(X \leq 10)$).

Je pense que les w_i peuvent servir pour prouver la convergence de la racine (p31/53, independant bootstrap).

Question 31

Proposer une procédure bootstrap pour estimer l'écart quadratique moyen de la méthode des moindres carrées dans le cas d'une régression linéaire.

We start with a little Bootstrap of the residuals
From the sample $(Y1, X1), ..., (Yn, Xn)$ ...

we compute the θ of the (multi)-linear regression such that

$Y = X\theta + \epsilon$

by the (now) usual multi linear regression process.

Then we bootstrap the ϵ into ϵ^* and we compute the $Y^* = X\theta + \epsilon^*$, the bootstraped responses,
we compute the θ^* i.e. the bootstraped (multi)-linear regression operator such that

$Y^* = X\theta^* + \epsilon^{**}$
by multi linear regression.
Then we compute $\frac{sum(\epsilon^{**2})}{n - rg(X)}$

autre solution : il faut plutôt faire un bootstrap des résidus et calculer la MSE (risque quadratique) pour les coefficients bootstrapés, voir le pseudo code suivant :

```
X = matrice(n,p)
y = X*Theta + epsilon (on a un modele MCO)

for b in B :
        epsilon_boot <- epsilon(rand(n))
        y_boot <- X*Theta + epsilon_boot
    theta_chap[b] <- inverse(Xt*X)*Xt*y_boot

MSE <- (E(theta_chap) - theta)**2 + var(theta_chap)
```