Université de Montpellier **HMMA307**
2019/2020 **Joseph Salmon**

TD N° 2 : Optimisation sous contraintes

Exercice 1. Résoudre le problème suivant :

$$\min_{(x_1,x_2,x_3)\in\mathbb{R}^3} \tfrac{1}{2}\left[(x_1-2)^2+(x_2-2)^2+(x_3-2)^2\right] \quad (1)$$
$$\text{s.c.} \quad x_1+x_2+x_3=1$$

Correction:

$$\mathcal{L}(x,\lambda)=\tfrac{1}{2}\left[(x_1-2)^2+(x_2-2)^2+(x_3-2)^2\right]+\lambda(x_1+x_2+x_3-1) \quad (2)$$

Conditions du premier ordre :

$$\nabla_x\mathcal{L}(x,\lambda)=0 \iff \begin{cases}(x_1-2)+\lambda=0\\(x_2-2)+\lambda=0\\(x_3-2)+\lambda=0\end{cases} \quad (3)$$

Cela assure que $x_1=x_2=x_3$, et donc avec la contrainte de satisfiabilité, $x_1+x_2+x_3=1$

Exercice 2.
On prend $y\in\mathbb{R}^n, X\in\mathbb{R}^{n\times p}$ avec $\mathrm{rg}(X)=p$, $R\in\mathbb{R}^{q\times p}$ avec $\mathrm{rg}(R)=q$ et enfin $r\in\mathbb{R}^q$. On définit les moindres carrées sous contraintes de la manière suivante :

$$\widehat{\beta}^c\in\mathop{\arg\min}_{\beta\in\mathbb{R}^p}\tfrac{1}{2}\left\|y-X\beta\right\|^2$$
$$\text{s.c.} \quad R\beta=r$$

1) Montrer qu'avec $\hat{\beta}^{\mathrm{LS}}$ solution des moindres carrés (non contraints), on obtient :

$$\widehat{\beta}^c=\widehat{\beta}^{\mathrm{LS}}+\left(X^\top X\right)^{-1}R^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}(r-R\widehat{\beta}^{\mathrm{LS}})$$

2) Montrer de plus que sous l'hypothèse gaussienne de modèle de régression linéaire classique :

$$\frac{1}{q\widehat{\sigma^2}}(R(\widehat{\beta}^{\mathrm{LS}}-\beta^\star))^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}R(\widehat{\beta}^{\mathrm{LS}}-\beta^\star)\sim\mathcal{F}^q_{n-p}$$

où $\mathcal{F}^q_{n-p}$ est une loi de Fisher.

3) Montrer que $(R\widehat{\beta}^{\mathrm{LS}}-r)^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}(R\widehat{\beta}^{\mathrm{LS}}-r)=\left\|\hat{y}^{\mathrm{LS}}-\hat{y}^c\right\|^2$, et que $\left\|\hat{y}^{\mathrm{LS}}-\hat{y}^c\right\|^2=\|y-\hat{y}^c\|^2-\left\|\hat{y}^{\mathrm{LS}}-y\right\|^2$ où $\hat{y}^{\mathrm{LS}}=X\hat{\beta}^{\mathrm{LS}}$ et $\hat{y}^c=X\hat{\beta}^c$.
4) Prendre $X=[\mathbb{1}_{C_1},\dots,\mathbb{1}_{C_K}]$, et en déduire un test de l'hypothèse "$\mu_1=\cdots=\mu_K$" dans l'ANOVA à un facteur.

Correction:

1)

$$\begin{cases}\frac{\partial\mathcal{L}}{\partial\beta}=-2X^\top Y+2X^\top X\widehat{\beta}^c-R^\top\hat{\lambda}=0\\\frac{\partial\mathcal{L}}{\partial\lambda}=R\widehat{\beta}^c-r=0\end{cases}$$

Ensuite :

$$\begin{aligned}
-2R\left(X^\top X\right)^{-1}X^\top Y + 2R\left(X^\top X\right)^{-1}X^\top X\widehat{\beta}^c - R\left(X^\top X\right)^{-1}R^\top\hat{\lambda} &= 0\\
-2R\left(X^\top X\right)^{-1}X^\top Y + 2R\widehat{\beta}^c - R\left(X^\top X\right)^{-1}R^\top\hat{\lambda} &= 0\\
-2R\left(X^\top X\right)^{-1}X^\top Y + 2r - R\left(X^\top X\right)^{-1}R^\top\hat{\lambda} &= 0\\
-2R\widehat{\beta}^{\mathrm{LS}} + 2r - R\left(X^\top X\right)^{-1}R^\top\hat{\lambda} &= 0
\end{aligned}$$

Puis on en déduit :

$$\hat{\lambda} = 2\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}\left[r - R\widehat{\beta}^{\mathrm{LS}}\right]$$

Ensuite, remplaçons dans la dérivée du lagrangien par rapport à β :

$$-2X^\top Y + 2X^\top X\widehat{\beta}^c - 2R^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}\left[r - R\widehat{\beta}^{\mathrm{LS}}\right] = 0$$

Enfin,

$$\begin{aligned}
\widehat{\beta}^c &= \left(X^\top X\right)^{-1}X^\top Y + \left(X^\top X\right)^{-1}R^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}(r - R\widehat{\beta}^{\mathrm{LS}})\\
&= \widehat{\beta}^{\mathrm{LS}} + \left(X^\top X\right)^{-1}R^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}(r - R\widehat{\beta}^{\mathrm{LS}})
\end{aligned}$$

2) $R\hat{\beta} \sim \mathcal{N}(R\beta^\star, \sigma^2 R(X^\top X)^{-1}R^\top)$ avec $R(X^\top X)^{-1}R^\top \in \mathbb{R}^{q\times q}$ inversible (car X et R sont de plein rang). On rappelle que pour l'estimateur des moindres carrés $\hat{\beta}$ et $\widehat{\sigma^2}$ sont indépendants :

$$\frac{1}{\sigma^2}(R(\widehat{\beta}^{\mathrm{LS}} - \beta^\star))^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}R(\widehat{\beta}^{\mathrm{LS}} - \beta^\star) \sim \chi^2_q \ .$$

On termine en rappelant que $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ où $\widehat{\sigma^2} = \frac{1}{n-p}\left\|y - X\widehat{\beta}^{\mathrm{LS}}\right\|^2$, et que les deux χ^2 sont indépendants :

$$\widehat{\beta}^{\mathrm{LS}} = \left(X^\top X\right)^{-1}X^\top Y = \left(X^\top X\right)^{-1}X^\top\left(X\left(X^\top X\right)^{-1}X^\top\right)Y = \left(X^\top X\right)^{-1}X^\top\Pi_X Y \ ,$$

et en rappelant que par le théorème de Cochran $\Pi_X(y)$ et $(\mathrm{Id} - \Pi_X)(y)$ sont indépendants et donc il en est de même pour les deux χ^2.

3) On repart de la première question pour obtenir :

$$\left(X^\top X\right)(\widehat{\beta}^c - \widehat{\beta}^{\mathrm{LS}}) = R^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}(r - R\widehat{\beta}^{\mathrm{LS}})$$

En multipliant $(\widehat{\beta}^c - \widehat{\beta}^{\mathrm{LS}})^\top$ on obtient

$$\begin{aligned}
\underbrace{(\widehat{\beta}^c - \widehat{\beta}^{\mathrm{LS}})^\top\left(X^\top X\right)(\widehat{\beta}^c - \widehat{\beta}^{\mathrm{LS}})}_{=\|\hat{y}^c - \hat{y}^{\mathrm{LS}}\|^2} &= (\widehat{\beta}^c - \widehat{\beta}^{\mathrm{LS}})^\top R^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}(r - R\widehat{\beta})\\
&= (r - R\widehat{\beta})^\top\left[R\left(X^\top X\right)^{-1}R^\top\right]^{-1}(r - R\widehat{\beta}) \ .
\end{aligned}$$

Enfin on montre que

$$\begin{aligned}
\|\hat{y}^c - y\|^2 &= \left\|\hat{y}^c - \hat{y}^{\mathrm{LS}}\right\|^2 + \left\|\hat{y}^{\mathrm{LS}} - y\right\|^2 + 2\langle\hat{y}^c - \hat{y}^{\mathrm{LS}}, \hat{y}^{\mathrm{LS}} - y\rangle\\
&= \left\|\hat{y}^c - \hat{y}^{\mathrm{LS}}\right\|^2 + \left\|\hat{y}^{\mathrm{LS}} - y\right\|^2 + 2\langle X\widehat{\beta}^c - X\hat{\beta}, X\hat{\beta} - y\rangle
\end{aligned}$$

Le dernier produit scalaire est nul car le vecteur des résidus dans les moindres carrés est orthogonal à toutes les variables explicatives.

4) Ici on a $\hat{\beta} = (\bar{y}_1, \ldots, \bar{y}_K)^\top$, donc $(X^\top X)^{-1} = \operatorname{diag}(\frac{1}{n_1}, \ldots, \frac{1}{n_K})$ et donc $\hat{y} = \sum_{k=1}^K \mathbb{1}_{C_k} \bar{y}_{C_k}$. Dans ce contexte $R = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(K-1)\times K}$ et $r = 0$ et donc $q = K - 1$. On peut aussi voir que la solution du problème

$$\begin{aligned} \hat{\beta}^c \in \underset{\beta \in \mathbb{R}^p}{\arg\min}\, & \tfrac{1}{2}\left\|y - X\beta\right\|^2 \\ \text{s.c.} \quad & \beta_1 = \cdots = \beta_K \end{aligned}$$

donne simplement $\hat{\beta}_1 = \cdots = \hat{\beta}_K = \bar{y}_n$ et donc $\hat{y}^c = \bar{y}_n \mathbb{1}_n$. En rappelant que $R\beta^\star = 0$ et que $\left\|\hat{y} - \hat{y}^c\right\|^2 = \left\|y - \hat{y}^c\right\|^2 - \left\|\hat{y} - y\right\|^2$, on obtient :

$$\frac{\frac{1}{K-1} \sum_{k=1}^K (\bar{y}_{C_k} - \bar{y}_n)^2}{\frac{1}{n-K} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \sim \mathcal{F}_{n-K}^{K-1} .$$

Exercice 3. Soit $X = \left[\frac{\mathbb{1}_{C_1}}{\sqrt{n_1}}, \ldots, \frac{\mathbb{1}_{C_K}}{\sqrt{n_K}}\right]$ (avec $\mathbf{x}_k = \frac{\mathbb{1}_{C_k}}{\sqrt{n_k}}$), avec $\hat{\mu}_k = \dfrac{1}{n_k} \displaystyle\sum_{i \in C_k} y_i$ et $n_k = \#\{i \in C_k, i \in [\![1, n]\!]\}$ avec $n_1 + \cdots + n_K = n$, et on suppose que les C_k forment une partition de $[\![1, n]\!]$.

On propose l'estimateur suivant :

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K)^\top = \underset{\beta_0, \ldots, \beta_K}{\arg\min} \left\| \mathbf{y} - \beta_0 \mathbb{1}_n - \sum_{j=1}^K \beta_j \mathbf{x}_j \right\|^2 + \lambda \sum_{j=1}^K \beta_j^2$$

1) Donner la valeur de $X^\top X$ et de $X^\top y$.
2) Donner une formule explicite pour l'estimateur Ridge, en fonction de y, $\hat{\mu}_1, \ldots, \hat{\mu}_K$, et des $n_1, \ldots, n_k$.
3) Comparer avec l'estimateur Ridge classique obtenu en forçant la contrainte $\beta_0 = 0$.

__Correction:__

1) $X^\top X = \mathrm{Id}_K$ et $X^\top \mathbf{y} = (\sqrt{n_1}\hat{\mu}_1, \ldots, \sqrt{n_K}\hat{\mu}_K)^\top$.
2) Notons $e = \left(\sqrt{n_1}, \ldots, \sqrt{n_K}\right)^\top$ et vérifions que $\mathbb{1}_n^\top X = e^\top$ Partons des conditions nécessaires du premier ordre :

$$\begin{cases} \langle \mathbb{1}_n, X\tilde{\boldsymbol{\beta}} + \mathbb{1}_n \hat{\beta}_0 - y \rangle & = 0 \\ X^\top \left(X\tilde{\boldsymbol{\beta}} + \mathbb{1}_n \hat{\beta}_0 - y \right) + \lambda \tilde{\boldsymbol{\beta}} & = 0 \end{cases} \iff \begin{cases} e^\top \tilde{\boldsymbol{\beta}} & = n\bar{y}_n - n\hat{\beta}_0 \\ (1+\lambda)\tilde{\boldsymbol{\beta}} + e\hat{\beta}_0 - (\sqrt{n_1}\hat{\mu}_1, \ldots, \sqrt{n_K}\hat{\mu}_K)^\top & = 0 \end{cases}$$

En multipliant la dernière équation à gauche et à droite par e on obtient :

$$(1+\lambda)e^\top \tilde{\boldsymbol{\beta}} + n\hat{\beta}_0 - \sum_{k=1}^K n_k \mu_k = (1+\lambda)e^\top \tilde{\boldsymbol{\beta}} + n\hat{\beta}_0 - n\bar{y}_n = 0$$

En y substituant l'expression de $e^\top \tilde{\boldsymbol{\beta}}$ obtenue dans le système précédent, on obtient :

$$(1+\lambda)(n\bar{y}_n - n\hat{\beta}_0) + n\hat{\beta}_0 - n\bar{y}_n = 0$$

et donc comme $\lambda > 0$,

$$\boxed{\hat{\beta}_0 = \bar{y}_n} .$$

Enfin la dernière équation du système précédent donne alors :

$$\boxed{\tilde{\boldsymbol{\beta}} = \frac{1}{1+\lambda}\left(\sqrt{n_1}(\hat{\mu}_1 - \bar{y}_n), \dots, \sqrt{n_K}(\hat{\mu}_K - \bar{y}_n)\right)} \ .$$

Enfin on obtient donc comme prédicteur associé $\hat{y}$ donné par :

$$\begin{aligned}
\hat{y}^{\mathrm{LS}} &= X\tilde{\boldsymbol{\beta}} + \hat{\beta}_0 \mathbb{1}_n \\
\hat{y}^{\mathrm{LS}} &= \frac{1}{1+\lambda}\sum_{k=1}^{K}(\hat{\mu}_k - \bar{y}_n)\mathbb{1}_{C_k} + \bar{y}_n \mathbb{1}_n \\
\hat{y}^{\mathrm{LS}} &= \frac{1}{1+\lambda}\sum_{k=1}^{K}\left(\hat{\mu}_k \mathbb{1}_{C_k} + \lambda \bar{y}_n \mathbb{1}_n\right)^\top \ .
\end{aligned}$$

Ainsi cela signifie que pour un élément $i \in C_k$, le prédicteur associé est donné par :

$$\boxed{\hat{y}_k = \frac{1}{1+\lambda}(\lambda \bar{y}_n + \hat{\mu}_k)}$$

Interprétation : λ permet d'osciller entre le prédicteur globale ($\bar{y}_n$, si $\lambda = +\infty$) et le prédicteur par modalité ($\hat{\mu}_k$, si $\lambda = 0$)

3) Dans ce cas l'estimateur est défini par :

$$\underset{\beta_1,\dots,\beta_K}{\arg\min} \left\| \mathbf{y} - \sum_{j=1}^{K} \beta_j \mathbf{x}_j \right\|^2 + \lambda \sum_{j=1}^{K} \beta_j^2$$

et l'on trouve :

$$\boxed{\hat{y}_k = \frac{1}{1+\lambda}\hat{\mu}_k}$$

Exercice 4. On observe K classes $C_1, \dots, C_K$ et n observations d'un phénomène (*e.g.*, le rendement de la variété) sont consignées. On fait l'hypothèse que les classes C_k sont disjointes et forment une partition des observations : $\cup_{k=1}^{K} C_k = [\![1, n]\!]$ et $\forall (k, k') \in [\![1, K]\!], C_k \cap C_{k'} = \varnothing$. Enfin on suppose que la cardinalité de chaque classe C_k est n_k, et donc que $n = \sum_{k=1}^{K} n_k$.

Le modèle linéaire associé peut s'écrire de la façon suivante : on définit les α_k, coefficients qui correspondent au niveau d'influence de la k^{e} classe

$$y = \underbrace{\begin{bmatrix} \mathbb{1}_n & \mathbb{1}_{C_1} & \dots & \mathbb{1}_{C_K} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}}_{\beta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_K \end{bmatrix} \ .$$

Donner alors l'estimateur des moindres carrés sous contraintes associé :

$$\begin{aligned}
&\min_{\beta \in \mathbb{R}^{K+1}} \frac{1}{2}\|y - X\beta\|^2 \ , \\
&\qquad \text{t.q. } \beta = (\mu, \alpha_1, \dots, \alpha_K)^\top \text{ et } \sum_{k=1}^{K} c_k \alpha_k = 0 \ ,
\end{aligned}$$

où le vecteur $c = (c_1, \dots, c_K)^\top \in \mathbb{R}^K$ est un vecteur encodant les contraintes choisies tel que $\sum_{k=1}^{K} c_k \neq 0$.

Exercice 5. Soit $X = \left[\frac{\mathbb{1}_{C_1}}{\sqrt{n_1}}, \dots, \frac{\mathbb{1}_{C_K}}{\sqrt{n_K}}\right]$ (avec $\mathbf{x}_k = \frac{\mathbb{1}_{C_j}}{\sqrt{n_k}}$), avec $\hat{\mu}_k = \dfrac{1}{n_k}\sum_{i \in C_k} y_i$ et $n_k = \#\{i \in C_k, i \in [\![1, n]\!]\}$ avec $n_1 + \dots + n_K = n$.

L'estimateur Lasso (sans pénalité sur la constante) est solution de

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K)^\top = \underset{\beta_0, \ldots, \beta_K}{\arg\min} \frac{1}{2} \left\| \mathbf{y} - \beta_0 \mathbb{1}_n - \sum_{j=1}^{K} \beta_j \mathbf{x}_j \right\|_2^2 + \lambda \sum_{j=1}^{K} |\beta_j|$$

1) Donner la valeur de $X^\top X$ et de $X^\top y$

2) Donner une formule explicite pour l'estimateur Lasso, en fonction de y, $\hat{\mu}_1, \ldots, \hat{\mu}_K$, et des $n_1, \ldots, n_k$.