TD 04 : Régression linéaire

Fondements du Machine Learning, L3 IM2D

Octobre-Novembre 2024 (V2 : 3 novembre 2024)

Dauphine | PSL
UNIVERSITÉ PARIS

Exercice 4.1 : Variation de données

(Adapté d'un exercice d'examen proposé en 2019-2020.)

On considère un jeu de données $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, où $\boldsymbol{x}_i \in \mathbb{R}^n$ et $y_i \in \mathbb{R}$ pour tout $i = 1, \dots, m$. On notera de manière plus compacte

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{m \times n} \quad \text{et} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m.$$

On suppose dans cet exercice que $\operatorname{rang}(\boldsymbol{X}) = n \leq m$, et qu'il existe un vecteur $\boldsymbol{\beta}^* \in \mathbb{R}^n$ tel que:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

où $\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2 I_m\right)$, c'est-à-dire que les composantes de $\boldsymbol{\epsilon}$ sont i.i.d. selon une loi normale de moyenne 0 et de variance $\sigma^2 > 0$.

On considère ensuite le problème aux moindres carrés suivant :

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} f(\boldsymbol{\beta}) := \frac{1}{m} \sum_{i=1}^m (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i)^2 = \frac{1}{m} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2. \tag{1}$$

a) Donner une solution du problème (1) sous les hypothèses considérées. Pourquoi cette solution est-elle unique ?

b) Soit $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$. Montrer que $\boldsymbol{H}$ est symétrique et que $\boldsymbol{H}^2 = \boldsymbol{H}$.

c) Justifier qu'une valeur propre de $\boldsymbol{H}$ est nécessairement égale à 0 ou 1. Sachant que $\operatorname{rang}(\boldsymbol{H}) = n$, en déduire que $\operatorname{trace}(\boldsymbol{H}) = n$.

d) Montrer que $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}$, où $\hat{\boldsymbol{\beta}}$ est la solution de (1) considérée en question a). Montrer alors que

$$f(\hat{\boldsymbol{\beta}}) = \frac{1}{m}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y}).$$

e) Justifier que $\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}\boldsymbol{\beta}^*$, où $\boldsymbol{\beta}^*$ est la vérité terrain du problème. En déduire que :

$$f(\hat{\boldsymbol{\beta}}) = \tfrac{1}{m}(\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\epsilon}).$$

f) On veut maintenant obtenir l'espérance de l'erreur représentée par f lorsque le jeu de données $\mathcal{D}$ (donc le vecteur $\boldsymbol{\epsilon}$) varie. Montrer que l'on a :

$$\mathbb{E}_{\mathcal{D}}\left[f(\hat{\boldsymbol{\beta}})\right] = \sigma^2(1 - \frac{n}{m}).$$

Indication : $\mathbb{E}_{\boldsymbol{u}}\left[\boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{u}\right] = \sigma^2 \operatorname{trace}(\boldsymbol{M})$ *pour toute matrice carrée* $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ *et tout vecteur* $\boldsymbol{u} \sim \mathcal{N}(0, \sigma^2 I_n)$.

g) En déduire que

$$\mathbb{E}_{\mathcal{D}}\left[f(\hat{\boldsymbol{\beta}})\right] \longrightarrow \mathbb{E}_{(\boldsymbol{x},y)}\left[R(\boldsymbol{\beta}^*)\right] \text{ quand } m \to \infty,$$

où $R(\boldsymbol{\beta}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} - y)^2$ et $(\boldsymbol{x}, y)$ suit la même distribution que les éléments de $\mathcal{D}$, c'est-à-dire que l'on a $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Exercice 4.2 : Ajout de données

Soient $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{y} \in \mathbb{R}^m$ avec $\operatorname{rang}(\boldsymbol{X}) = n \leq m$ et $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, où $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_m)$. On rappelle que dans ce contexte, l'estimateur du maximum de vraisemblance (ou l'estimateur des moindres carrés) vaut

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{\beta}^* + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon}.$$

a) Retrouver les formules $\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}^*$ ainsi que
$\operatorname{Cov}\left[\hat{\boldsymbol{\beta}}\right] = \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right])(\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right])^{\mathrm{T}}\right] = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$.

b) Pour tout vecteur aléatoire $\boldsymbol{z} \in \mathbb{R}^n$, montrer que

$$\mathbb{E}\left[\|\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right]\|_2^2\right] = \operatorname{trace}\operatorname{Cov}\left[\boldsymbol{z}\right].$$

En déduire que $\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right] = \sigma^2 \operatorname{trace}\left((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right)$.

c) On considère l'ajout d'une nouvelle donnée $(\boldsymbol{v}, w = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon_{m+1}) \in \mathbb{R}^n \times \mathbb{R}$ avec $\epsilon_{m+1} \sim \mathcal{N}(0, \sigma^2)$, et on s'intéresse à l'impact de cet ajout sur l'estimateur du maximum de vraisemblance. Etablir la formule de cet estimateur pour le nouveau jeu de données (on le notera $\tilde{\boldsymbol{\beta}}$) en fonction de $\boldsymbol{X}, \boldsymbol{v}, \boldsymbol{y}, w$ et calculer sa matrice de covariance.

d) Pour toute matrice $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ définie positive et tout vecteur $\boldsymbol{u} \in \mathbb{R}^n$, on a :

$$(\boldsymbol{A}^{-1} + \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})^{-1} = \boldsymbol{A} - \frac{1}{1 + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}}\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}.$$

En utilisant cette relation, montrer que $\operatorname{trace}((\boldsymbol{A}^{-1} + \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})^{-1}) = \operatorname{trace}(\boldsymbol{A}) - \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\boldsymbol{u}}{1+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}}$.

e) Déduire de la question précédente un lien entre $\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right]$ et $\mathbb{E}\left[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right]$. Sous quelle condition est-il avantageux d'introduire une donnée supplémentaire ?

Exercice 4.3 : Maximum de vraisemblance

(Adapté d'un exercice d'examen proposé en 2020-2021.)

On considère un jeu de données sous la forme de deux matrices $\boldsymbol{X} \in \mathbb{R}^{m_1 \times n}$, $\boldsymbol{W} \in \mathbb{R}^{m_2 \times n}$, et de deux vecteurs $\boldsymbol{y} \in \mathbb{R}^{m_1}$, $\boldsymbol{z} \in \mathbb{R}^{m_2}$. On fait l'hypothèse d'un modèle linéaire sous-jacent affecté par un bruit gaussien, c'est-à-dire que l'on suppose qu'il existe deux vecteurs $\boldsymbol{\beta}_1^* \in \mathbb{R}^n$ et $\boldsymbol{\beta}_2^* \in \mathbb{R}^n$ tels que

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_1^* + \boldsymbol{\epsilon}_1 \quad \text{et} \quad \boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}_2^* + \boldsymbol{\epsilon}_2,$$

où $\boldsymbol{\epsilon}_1$ est un vecteur aléatoire de $\mathbb{R}^{m_1}$ distribué selon une loi gaussienne $\mathcal{N}(\mathbf{0}_{\mathbb{R}^{m_1}}, \sigma^2 \boldsymbol{I}_{m_1})$ et $\boldsymbol{\epsilon}_2$ est un vecteur de $\mathbb{R}^{m_2}$ distribué selon une loi gaussienne $\mathcal{N}(\mathbf{0}_{\mathbb{R}^{m_2}}, \zeta^2 \boldsymbol{I}_{m_2})$, avec $\sigma > 0$ et $\zeta > 0$. On supposera également que les matrices $\boldsymbol{X}$ et $\boldsymbol{W}$ sont de même rang n, avec $n \leq \min\{m_1, m_2\}$.

a) Quelles sont alors les distributions des vecteurs $\boldsymbol{y}$ et $\boldsymbol{z}$?

b) Rappeler la définition d'une fonction de vraisemblance pour m observations réelles $a_1, \ldots, a_m$ dépendant d'un vecteur de paramètres $\boldsymbol{\beta}$, notée $L(a_1, \ldots, a_m; \boldsymbol{\beta})$. En suivant cette définition, établir des formules pour les vraisemblances

$$L(y_1, \ldots, y_{m_1}; \boldsymbol{\beta}_1) \quad \text{et} \quad L(z_1, \ldots, z_{m_2}; \boldsymbol{\beta}_2)$$

c) On s'intéresse à l'estimateur du maximum de vraisemblance pour $\boldsymbol{\beta}_1^*$.

   i) Donner le problème de maximisation de la vraisemblance dont cet estimateur est solution. Pourquoi est-il équivalent de maximiser la vraisemblance et son logarithme ?

   ii) Donner le lien entre l'estimateur du maximum de vraisemblance et l'ensemble des solutions du problème aux moindres carrés

$$\min_{\boldsymbol{\beta}_1 \in \mathbb{R}^n} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{\beta}_1 - \boldsymbol{y}\|^2.$$

   iii) En déduire une formule pour l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}_1^*$, que l'on notera $\hat{\boldsymbol{\beta}}_1$.

d) En utilisant le même raisonnement qu'en question c), donner la formule de l'estimateur du maximum de vraisemblance pour $\boldsymbol{\beta}_2^*$, noté $\hat{\boldsymbol{\beta}}_2$.

e) Justifier que le vecteur $\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}$ est un estimateur non biaisé du vecteur $\begin{bmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \end{bmatrix}$.

Exercice 4.4 : Maximum a posteriori

(Adapté d'un exercice d'examen proposé en 2019-2020.)

Soit un jeu de données $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$. Dans le cadre de la régression linéaire, on recherche un vecteur $\boldsymbol{\beta}$ tel que

$$y_i \approx \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i,$$

où les ϵ_i sont i.i.d. selon une loi normale de moyenne 0 et de variance 1.

On s'intéresse ici à l'estimateur du maximum a posteriori obtenu avec un a priori uniforme sur un intervalle réel $[a, b]$ (avec $a < b$). On rappelle que la densité de probabilité d'une variable z suivant une loi uniforme dans un intervalle $[a, b]$ est la fonction p donnée par :

$$p(z) = \begin{cases} \frac{1}{b-a} & \text{si } z \in [a, b], \\ 0 & \text{sinon.} \end{cases} \tag{2}$$

a) Rappeler la définition de la vraisemblance $\mathcal{L}(y_1, \ldots, y_m; \boldsymbol{\beta})$, et donner sa formule pour le problème considéré.

b) On suppose a priori que les entrées de $\boldsymbol{\beta}$ sont i.i.d. suivant une loi uniforme sur $[a, b]$. Donner alors la formule de la loi a priori de $\boldsymbol{\beta}$, notée $\mathcal{L}(\boldsymbol{\beta})$.

c) Formuler le problème d'estimation du maximum a posteriori en utilisant les logarithmes des lois de probabilité.

d) Montrer que si l'on ne tient pas compte des termes constants, le problème se ramène à celui ci-dessous :

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\{ -\frac{1}{2} \sum_{i=1}^m (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i)^2 + \sum_{j=1}^n \ln \left(\chi_{b-a}([\boldsymbol{\beta}]_j) \right) \right\}, \tag{3}$$

où $\chi_{b-a}(z) = 1$ si $a \leq z \leq b$ et $\chi_{b-a}(z) = 0$ sinon.

e) On peut montrer que l'ensemble des solutions du problème (3) correspond à celui du problème aux moindres carrés avec contraintes suivant :

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^m (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i)^2 \quad \text{sous les contraintes } a \leq \beta_j \leq b \quad \forall j = 1, \ldots, n. \tag{4}$$

En quoi ce nouveau problème modélise-t-il l'a priori que l'on a mis sur le modèle linéaire ?

f) On note $\hat{\boldsymbol{\beta}}$ l'estimateur du maximum de vraisemblance. Si cet estimateur vérifie $a \leq [\hat{\boldsymbol{\beta}}]_j \leq b$ pour tout $j = 1, \ldots, n$, est-il solution du problème (4) ?

Solutions

Solution de l'exercice 4.1: Variation de données

a) Une solution du problème est donnée par $\boldsymbol{X}^\dagger \boldsymbol{y}$, qui est ici égale à $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ car on $\mathrm{rang}(\boldsymbol{X}) = n$. On sait dans ce cas que la solution est unique.

b) Un calcul explicite donne

$$\boldsymbol{H}^{\mathrm{T}} = \left(\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}} = (\boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}}\left((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right)^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{X}\left((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}},$$

ce qui prouve que $\boldsymbol{H}$ est symétrique. On vérifie également que

$$\boldsymbol{H}^2 = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{H}.$$

c) Soit $\lambda \in \mathbb{R}$ une valeur propre de $\boldsymbol{H}$, et $\boldsymbol{v}$ un vecteur propre associé. Alors $\boldsymbol{H}\boldsymbol{x} = \lambda\boldsymbol{x}$ et $\boldsymbol{H}^2\boldsymbol{x} = \lambda^2\boldsymbol{x}$. Comme $\boldsymbol{H}^2 = \boldsymbol{H}$ d'après la question précédente, on en déduit que $\lambda^2 = \lambda$, d'où $\lambda \in \{0, 1\}$. Comme le rang de $\boldsymbol{H}$ est égal à n, cette matrice possède exactement n valeurs propres non nulles. On en déduit que la trace de $\boldsymbol{H}$, qui est égale à la somme des valeurs propres, vaut $n \times 1 + (m - n) \times 0 = n$.

d) En utilisant l'expression $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, il vient

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

En utilisant cela dans l'expression de $f(\hat{\boldsymbol{\beta}})$, il vient alors

$$\begin{aligned} f(\hat{\boldsymbol{\beta}}) &= \frac{1}{m}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}\|^2 \\ &= \frac{1}{m}\|\boldsymbol{H}\boldsymbol{y} - \boldsymbol{y}\|^2 \\ &= \frac{1}{m}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}^2\boldsymbol{y} - 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\right), \end{aligned}$$

où l'on a utilisé la symétrie de $\boldsymbol{H}$ pour obtenir le premier et le second termes. En utilisant ensuite que $\boldsymbol{H}^2 = \boldsymbol{H}$, on obtient que

$$\begin{aligned} f(\hat{\boldsymbol{\beta}}) &= \frac{1}{m}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y} - 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\right) \\ &= \frac{1}{m}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y}\right). \end{aligned}$$

e) On tire de l'expression de $\boldsymbol{H}$ que

$$\begin{aligned} \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}^* &= \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* \\ &= \boldsymbol{X}\boldsymbol{\beta}^*. \end{aligned}$$

Par conséquent, on a $\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}^*$. On a ainsi

$$\begin{aligned} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} &= (\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon})^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}) \\ &= (\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* + 2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}. \end{aligned}$$

et

$$\begin{aligned} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y} &= (\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}) \\ &= (\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}^* + 2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\epsilon} \\ &= (\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* + 2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\epsilon} \end{aligned}$$

où l'on a utilisé la symétrie de $\boldsymbol{H}$ puis $\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}\boldsymbol{\beta}^*$.

En appliquant cela au résultat de la question précédente, on obtient finalement

$$\begin{aligned} f(\hat{\boldsymbol{\beta}}) &= \frac{1}{m}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y}\right) \\ &= \frac{1}{m}\big((\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* + 2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} \\ &\quad -(\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* - 2\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\epsilon}\big) \\ &= \frac{1}{m}\left(\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\epsilon}\right). \end{aligned}$$

f) On note que la source d'aléatoire provient du vecteur $\boldsymbol{\epsilon}$, et donc que $\mathbb{E}_{\mathcal{D}}\left[f(\hat{\boldsymbol{\beta}})\right] = \mathbb{E}_{\boldsymbol{\epsilon}}\left[f(\hat{\boldsymbol{\beta}})\right]$.

En utilisant l'indication et la linéarité de l'espérance, on a

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\epsilon}}\left[f(\hat{\boldsymbol{\beta}})\right] &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}\right] - \frac{1}{m}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\epsilon}\right] \\ &= \frac{1}{m}\operatorname{trace}(\sigma^2\boldsymbol{I}_m) - \frac{1}{m}\operatorname{trace}(\sigma^2\boldsymbol{H}) \\ &= \frac{1}{m}m\sigma^2 - \frac{1}{m}n\sigma^2 = \sigma^2\left(1 - \frac{n}{m}\right). \end{aligned}$$

g) On tire de la question précédente que $\mathbb{E}_{\mathcal{D}}\left[f(\hat{\boldsymbol{\beta}})\right] \to \sigma^2$ lorsque $m \to \infty$ (mais que n reste fixe !). Par ailleurs, on a

$$\begin{aligned} \mathbb{E}_{(\boldsymbol{x},y)}\left[(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - y)^2\right] &= \mathbb{E}_{(\boldsymbol{x},y)}\left[(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - \epsilon)^2\right] \\ &= \mathbb{E}_{\epsilon}\left[(\epsilon^2)\right] = \sigma^2, \end{aligned}$$

où la dernière égalité provient de la définition même de la variance.

Solution de l'exercice 4.2: Ajout de données

a) On note en préambule que pour tout vecteur aléatoire $\boldsymbol{z} \in \mathbb{R}^n$ et toute matrice $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ qui ne dépend pas de $\boldsymbol{z}$, on a :

$$\mathbb{E}\left[\boldsymbol{A}\boldsymbol{z}\right] = \boldsymbol{A}\,\mathbb{E}\left[\boldsymbol{z}\right] \quad \text{et} \quad \mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}\right] = \mathbb{E}\left[\boldsymbol{z}^{\mathrm{T}}\right]\boldsymbol{A}.$$

(Dans le cas d'un vecteur aléatoire, l'espérance est linéaire à gauche et à droite.)

On étudie d'abord l'espérance de $\hat{\boldsymbol{\beta}}$. On a :

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= \mathbb{E}\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\right] \\
&= \mathbb{E}\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon})\right] \\
&= \mathbb{E}\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^* + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon}\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^* + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon}\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^*\right] + \mathbb{E}\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon}\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^*\right] + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{\epsilon}\right] \\
&= \boldsymbol{\beta}^* + 0 = \boldsymbol{\beta}^*.
\end{aligned}$$

Pour la matrice de covariance, on introduit à nouveau la définition de $\boldsymbol{y}$ impliquant $\boldsymbol{\beta}^*$. Cela donne :

$$\begin{aligned}
\mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right])(\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right])^{\mathrm{T}}\right] &= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[(\boldsymbol{\beta}^* + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon} - \boldsymbol{\beta}^*)(\boldsymbol{\beta}^* + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon} - \boldsymbol{\beta}^*)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon})((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon})^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right] \\
&= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right]\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1},
\end{aligned}$$

où l'on a utilisé le fait que $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}$, et donc l'inverse de la transposée de $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}$ est égale à l'inverse de $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$.

Par définition, $\mathbb{E}\left[\boldsymbol{\epsilon}\right] = 0$ et $\mathrm{Cov}\left[\boldsymbol{\epsilon}\right] = I_m$. Comme $\mathrm{Cov}\left[\boldsymbol{\epsilon}\right] = \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right]$, on obtient :

$$\begin{aligned}
\mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right])(\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right])^{\mathrm{T}}\right] &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right]\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1},
\end{aligned}$$

d'où le résultat.

b) On utilise le fait que pour toutes matrices $\boldsymbol{A} \in \mathbb{R}^{n\times 1}$ et $\boldsymbol{B} \in \mathbb{R}^{1\times n}$, $\mathrm{trace}\,\boldsymbol{B}\boldsymbol{A} = \boldsymbol{B}\boldsymbol{A} = \mathrm{trace}\,\boldsymbol{A}\boldsymbol{B}$. En effet,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right]\|_2^2\right] &= \mathbb{E}\left[(\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right])^{\mathrm{T}}(\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right])\right] \\
&= \mathbb{E}\left[\mathrm{trace}\,(\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right])^{\mathrm{T}}(\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right])\right] \\
&= \mathbb{E}\left[\mathrm{trace}\,(\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right])(\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right])^{\mathrm{T}}\right] \\
&= \mathrm{trace}\,\mathbb{E}\left[(\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right])(\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right])^{\mathrm{T}}\right] = \mathrm{trace}\,\mathrm{Cov}\left[\boldsymbol{z}\right],
\end{aligned}$$

où l'on peut passer de l'avant-dernière à la dernière ligne parce que la trace est une application linéaire (on exploite encore la linéarité de l'espérance).

En appliquant la relation ci-dessus avec $\boldsymbol{z} = \hat{\boldsymbol{\beta}}$, on obtient :

$$\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right] = \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right]\|_2^2\right] = \mathrm{trace}\,(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}.$$

c) La matrice $\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} \\ \boldsymbol{v}^{\mathrm{T}} \end{bmatrix}$ est de rang n, comme $\boldsymbol{X}$. L'estimateur du maximum de vraisemblance est donc donné par $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{y}}$, avec $\tilde{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \\ w \end{bmatrix}$.

d) On utilise la linéarité de la trace ainsi que la propriété de ce même opérateur déjà évoquée en question b). On a ainsi :

$$\begin{aligned} \operatorname{trace}(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})^{-1} &= \operatorname{trace}\boldsymbol{A} - \frac{1}{1+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}}\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A} \\ &= \operatorname{trace}\boldsymbol{A} - \frac{1}{1+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}}\operatorname{trace}\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A} \\ &= \operatorname{trace}\boldsymbol{A} - \frac{1}{1+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}}\operatorname{trace}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\boldsymbol{u} \\ &= \operatorname{trace}\boldsymbol{A} - \frac{1}{1+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\boldsymbol{u}. \end{aligned}$$

e) D'après les questions c) et d), on a :

$$\begin{aligned} \mathbb{E}\left[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right] &= \operatorname{trace}\operatorname{Cov}\left[\tilde{\boldsymbol{\beta}}\right] \\ &= \operatorname{trace}(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1} \\ &= \operatorname{trace}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}})^{-1} \\ &= \operatorname{trace}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} - \frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-2}\boldsymbol{v}}{1+\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{v}} \\ &= \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right] - \frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-2}\boldsymbol{v}}{1+\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{v}} \end{aligned}$$

Par conséquent, $\tilde{\boldsymbol{\beta}}$ est un meilleur estimateur (en moyenne quadratique) que $\hat{\boldsymbol{\beta}}$ lorsque $\boldsymbol{v}$ est choisi tel que $\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-2}\boldsymbol{v}$, c'est-à-dire $\|\boldsymbol{v}\| \neq 0$.

Solution de l'exercice 4.3: Maximum de vraisemblance

Le but de l'exercice est de réviser les concepts fondamentaux du maximum de vraisemblance à travers l'utilisation d'un jeu de données double. On pourra rappeler les formules concernant la loi d'une variable gaussienne.

a) $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}_1^*, \sigma^2\boldsymbol{I}_{m_1})$ et $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{\beta}_2^*, \zeta^2\boldsymbol{I}_{m_2})$.

b) On a

$$\mathcal{L}(y_1, \ldots, y_{m_1}; \boldsymbol{\beta}_1) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{m_1} \exp\left[-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{\beta}_1 - \boldsymbol{y}\|^2\right]$$

et

$$\mathcal{L}(z_1, \dots, z_{m_2}; \boldsymbol{\beta}_2) = \left(\frac{1}{\sqrt{2\pi}\zeta}\right)^{m_1} \exp\left[-\frac{1}{2\zeta^2}\|\boldsymbol{W}\boldsymbol{\beta}_2 - \boldsymbol{z}\|^2\right]$$

c)

i) Le problème de maximisation de la vraisemblance peut être écrit comme

$$\max_{\boldsymbol{\beta}_1 \in \mathbb{R}^n} L(y_1, \dots, y_{m_1}; \boldsymbol{\beta}_1)$$

soit, de manière équivalente par croissance du logarithme, comme

$$\max_{\boldsymbol{\beta}_1 \in \mathbb{R}^n} \ln\left(L(y_1, \dots, y_{m_1}; \boldsymbol{\beta}_1)\right).$$

ii) L'estimateur du maximum de vraisemblance est solution du problème aux moindres carrés, car les deux problèmes ont le même ensemble de solutions.

iii) Dans le cadre de l'exercice, comme $\operatorname{rang}(\boldsymbol{X}) = n$, on sait que la solution du problème est unique et égale à $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$.

d) Par analogie avec la question précédente, on a $\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{z}$.

e) Il suffit de remarquer que $\mathbb{E}_{\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_2}\left[\hat{\boldsymbol{\beta}}_1\right] = \boldsymbol{\beta}_1^*$ et $\mathbb{E}_{\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_2}\left[\hat{\boldsymbol{\beta}}_2\right] = \boldsymbol{\beta}_2^*$, puis d'utiliser la formule sur l'espérance d'un vecteur.

Solution de l'exercice 4.4: Maximum a posteriori

a) La vraisemblance est donnée par la loi jointe des y_i étant donné le vecteur $\boldsymbol{\beta}$. Par hypothèse, on considère que $y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ suit une loi normale de moyenne 0 et de variance 1, et que les y_i sont indépendants. Cela conduit à la vraisemblance suivante :

$$\mathcal{L}(y_1, \dots, y_{m_1}; \boldsymbol{\beta}_1) = \left(\frac{1}{\sqrt{2\pi}}\right)^{m} \exp\left[-\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2\right].$$

b) Sous l'hypothèse de la question, la loi a priori de $\boldsymbol{\beta}$ est donnée par

$$\mathcal{L}(\boldsymbol{\beta}) = \begin{cases} \frac{1}{(b-a)^n} & \text{si } \beta_j \in [a,b]\ \forall j = 1, \dots, n \\ 0 & \text{sinon.} \end{cases}$$

Par anticipation de la question d), on écrira

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{(b-a)^n}\prod_{j=1}^{n}\chi_{b-a}([\boldsymbol{\beta}]_j)$$

c) Le problème d'estimation du maximum a posteriori s'écrit

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^n} \ln\left(\mathcal{L}(y_1, \dots, y_m; \boldsymbol{\beta})\mathcal{L}(\boldsymbol{\beta})\right). \tag{5}$$

d) En utilisant les formules obtenues en questions a) et b), on obtient

$$\begin{aligned}\ln\left(\mathcal{L}(y_1,\ldots,y_m;\boldsymbol{\beta})\mathcal{L}(\boldsymbol{\beta})\right) &= \ln(\mathcal{L}(y_1,\ldots,y_m;\boldsymbol{\beta})) + \ln(\mathcal{L}(\boldsymbol{\beta})) \\ &= m\,\ln(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2 + n\ln(\frac{1}{b-a}) + \sum_{j=1}^{n}\ln\left(\chi_{b-a}([\boldsymbol{\beta}]_j)\right).\end{aligned}$$

Les termes ne dépendant pas de $\boldsymbol{\beta}$ n'ont pas d'influence sur la solution du problème d'optimisation (5), on peut les omettre dans la formulation du problème, qui devient donc

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^n}\left\{-\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2 + \sum_{j=1}^{n}\ln\left(\chi_{b-a}([\boldsymbol{\beta}]_j)\right)\right\},$$

soit le résultat attendu.

e) La formulation (4) contraint les valeurs des coefficients de $\boldsymbol{\beta}$ dans l'intervalle $[a,b]$. En supposant un a priori uniformément distribué dans $[a,b]^n$, c'est bien ce que l'on cherche à faire.

f) On sait que $\hat{\boldsymbol{\beta}}$ est la solution du problème

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n}\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2$$

sans contraintes. Si cet estimateur vérifie de plus $a \leq [\hat{\boldsymbol{\beta}}]_j \leq b$, alors il s'agit aussi de la solution du problème (4). En ce sens, le maximum de vraisemblance est compatible avec l'a priori.