CHAPITRE 7

Régression linéaire bayésienne

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert – Automne 2022

La régression bayésienne peut être vu comme une généralisation du chapitre précédent lorsque la moyenne de la loi normale est une fonction de variables explicatives. La régression linéaire bayésienne est utile pour régler deux problèmes qui peuvent survenir en régression linéaire : la multicolinéarité et la sélection de modèle parmi un vaste ensemble de modèles possibles. Le modèle bayésien de régression linéaire sera d'abord présenté avec une loi *a priori* non informative. Ensuite, un modèle utilisant une loi *a priori* particulière menant au modèle de régression ridge, sera présenté. C'est un modèle très répandu en pratique pour contrôler la multicolinéarité. À la fin du chapitre, vous devriez être en mesure de :

— Estimer les paramètres d'un modèle de régression linéaire avec l'approche bayésienne.
— Implémenter l'échantillonnage de Gibbs pour la régression linéaire bayésienne.
— Limiter les effets de la multicolinéarité en utilisant une loi *a priori* informative.
— Sélectionner le meilleur modèle parmi un ensemble de modèles de régression.

La théorie présentée dans ce chapitre sera illustrée avec la racine cubique des poids des 56 perches pêchées dans le lac Laengelmavesi [1] en Finlande en fonction de la longueur standard (x_1), la longueur non-standard (x_2), la longueur totale (x_3), la hauteur (x_4) et de la largeur (x_5). Une description de ces mesures est disponible dans l'énoncé du TD4. Lors de ce TD, nous avions également vu que la racine cubique permettait de linéariser la relation entre les poids et les caractéristiques.

1. Le fichier `fishweights.csv` est disponible sur le site web du cours.

7.1 Modèle de régression linéaire (rappel)

En supposant les hypothèses 1 à 4 de la régression linéaire (Chapitre 2), la modèle statistique suivant est supposé pour chacune des observations $\{(x_{i1}, \ldots, x_{ip}, y_i) : 1 \leq i \leq n\}$:

$$Y_i \sim \mathcal{N}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2\right), \text{ pour } 1 \leq i \leq n; \tag{7.1}$$

où $\boldsymbol{x}_i$ dénote les variables explicatives associées à Y_i, $\boldsymbol{\beta}$ aux coefficients de régression et σ^2 à la variance de l'erreur. Au chapitre 2, nous avons vu que l'estimation de β par la méthode des moindres carrés correspond à :

$$\hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \boldsymbol{Y}.$$

Remarque. *Dans le cas de la régression linéaire bayésienne, l'hypothèse 4 est nécessaire pour l'estimation des paramètres par le théorème de Bayes. En effet, la vraisemblance est requise pour utiliser le théorème de Bayes.*

À l'aide de la loi normale multidimensionnelle présentée à l'annexe XXX, la vraisemblance de l'échantillon aléatoire pour les n observations peut être réécrit sous la forme suivante :

$$\begin{aligned} f_{(\boldsymbol{Y}|\boldsymbol{\beta},\sigma^2)}(\boldsymbol{y}) &= \prod_{i=1}^{n} \mathcal{N}(y_i \mid \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2) \\ &= \mathcal{N}\left(\boldsymbol{y} \mid X\boldsymbol{\beta}, \sigma^2 I_n\right). \end{aligned}$$

Remarque. *La matrice de covariance de $\boldsymbol{Y}$ est diagonale puisque l'on a supposé que les erreurs étaient indépendantes (hypothèse 3) et les éléments sur la diagonale sont tous égaux à σ^2 en raison de l'hypothèse d'homoscédasticité (hypothèse 2).*

Exemple 1

Partitionnons les données en deux ensembles : un ensemble d'entraînement composé de 44 observations et un ensemble de validation composé de 12 observations. L'ensemble d'entraînement sera utilisé pour estimer les paramètres du modèle et évaluer la qualité d'ajustement tandis que l'ensemble de validation sera utilisé pour évaluer la qualité des prédictions.
Soit le poids des perches [a] de l'ensemble d'entraînement en fonction de la longueur standard x_1. On a donc le modèle de régression linéaire simple suivant :

$$Y_i \sim \mathcal{N}(\beta_0 + x_{i1}\beta_1, \sigma^2), \text{ pour } 1 \leq i \leq n;$$

où β_0, β_1 et σ^2 sont les paramètres inconnus à estimer et où $n = 44$.

a. Je devrais plutôt écrire *la racine cubique du poids* mais je ne souhaite pas alourdir le texte.

7.2 Loi *a priori* non informative

La loi *a priori* sur les paramètres peut être décomposée de la façon suivante :

$$f_{(\boldsymbol{\beta},\sigma^2)}(\boldsymbol{\beta},\sigma^2) = f_{(\boldsymbol{\beta}|\sigma^2)}(\boldsymbol{\beta}) \times f_{(\sigma^2)}(\sigma^2). \tag{7.2}$$

Si aucune information *a priori* n'est disponible sur les paramètres, les lois impropres suivantes peuvent être utilisées :

$$f_{(\boldsymbol{\beta}|\sigma^2)}(\boldsymbol{\beta}) \propto 1;$$
$$f_{(\sigma^2)}(\sigma^2) \propto \frac{1}{\sigma^2},$$

pour former la loi *a priori* non informative suivante :

$$f_{(\boldsymbol{\beta},\sigma^2)}(\boldsymbol{\beta},\sigma^2) \propto \frac{1}{\sigma^2}, \text{ pour } \sigma^2 > 0. \tag{7.3}$$

La forme fonctionnelle de la loi *a posteriori* correspondante s'obtient en multipliant la vraisemblance du modèle de régression linéaire bayésien (éq. 7.1) à la loi *a priori*. Elle possède la forme suivante :

$$\begin{aligned} f_{\{(\boldsymbol{\beta},\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{\beta},\sigma^2) &\propto \left|\sigma^2 I_n\right|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\sigma^2 I_n)^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \times \frac{1}{\sigma^2}; \\ &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta})\right\}. \end{aligned} \tag{7.4}$$

Cette forme fonctionnelle ne correspond à aucune densité connue.

Exercice 1

Pouvez-vous expliquer de façon informelle pourquoi la densité impropre composée des deux lois précédentes correspond bien à une loi *a priori* non informative pour le modèle de régression bayésien de l'équation (7.1) ?

7.2.1 Lois conditionnelles complètes

De la forme fonctionnelle de la loi *a posteriori* exprimée à l'équation (7.4), il est possible d'identifier les lois conditionnelles complètes suivantes :

$$f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\boldsymbol{\beta}) \sim \mathcal{N}\left\{\boldsymbol{\beta}\middle|\hat{\boldsymbol{\beta}}, \sigma^2\left(X^\top X\right)^{-1}\right\}, \tag{7.5}$$

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\beta})}(\sigma^2) \sim \mathcal{I}nverse\mathcal{G}amma\left\{\sigma^2\middle|\frac{n}{2}, \frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta})\right\}; \tag{7.6}$$

où $\hat{\boldsymbol{\beta}}$ correspond à l'estimation par les moindres carrés de $\boldsymbol{\beta}$, *i.e.*

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}.$$

Nous démontrerons ce résultat en classe.

Remarque. *La loi* a posteriori *dépend de* $(X^\top X)^{-1}$. *Il y a alors risque de multicolinéarité en régression bayésienne lorsque la loi* a priori *non informative est utilisée.*

Les lois conditionnelles complètes de $\boldsymbol{\beta}$ et σ^2 sont utiles pour implémenter l'échantillonnage de Gibbs permettant de générer un échantillon de la loi *a posteriori* des paramètres.

Exemple 2: suite de l'exemple 1

La figure 7.1 illustre la chaîne générée par l'échantillonnage de Gibbs avec les lois conditionnelles complètes exprimées aux équations 7.5 et 7.6. L'estimation obtenue par la méthode des moindres carrés est :

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.16 & 0.26 \end{bmatrix}^\top.$$

On constate que la chaîne entre dans la phase d'échantillonnage après seulement quelques itérations. Les itérations pour β_0 et pour β_1 oscillent autour de leurs estimations respectives obtenues par les moindres carrés et les itération pour σ^2 oscillent autour de 0.09.

7.2.2 Lois *a posteriori* marginales

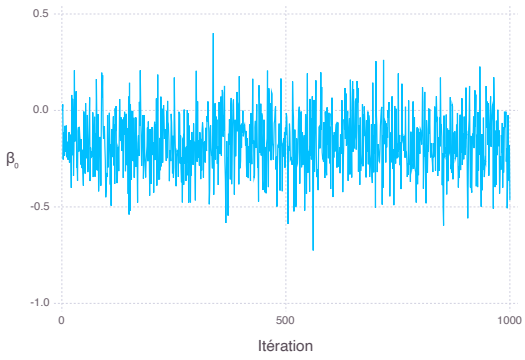
Dans le cas de la régression linéaire bayésienne utilisant la loi *a priori* impropre définie en début de section, les lois *a posteriori* marginales des paramètres s'expriment sous une forme analytique. La loi *a posteriori* marginale de $\boldsymbol{\beta}$ se calcule en intégrant σ^2 de la loi *a posteriori* exprimée à l'équation (7.4). On peut montrer (voir l'annexe 7.C) que la loi *a posteriori* marginale de $\boldsymbol{\beta}$ s'exprime sous la forme suivante :

$$f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) = t_{n-m}\left\{\boldsymbol{\beta}\,\middle|\,\hat{\boldsymbol{\beta}}, \sqrt{s^2\left(X^\top X\right)^{-1}}\right\}; \tag{7.7}$$

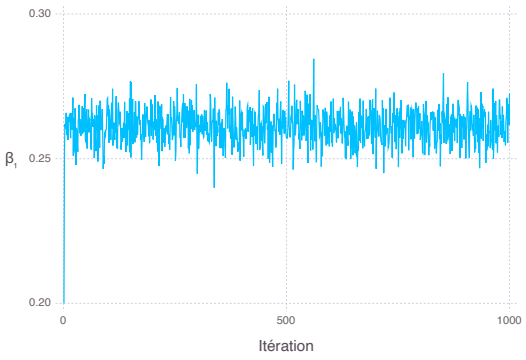
où m est le nombre de colonnes de la matrice X,

$$s^2 = \frac{1}{n-m}\left(\boldsymbol{y} - X\hat{\beta}\right)^\top \left(\boldsymbol{y} - X\hat{\beta}\right);$$

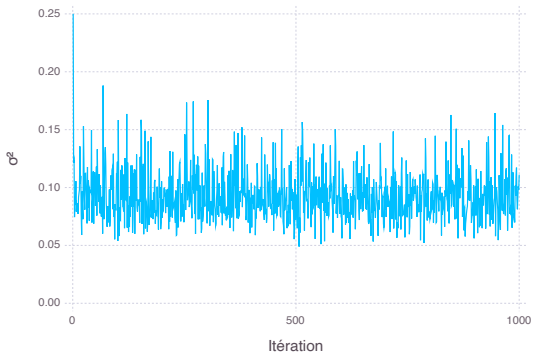
et où $t_\nu(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma)$ dénote la densité de la loi de Student multidimensionnelle à ν degrés de liberté, de paramètre de localisation $\boldsymbol{\mu}$ et de paramètre d'échelle Σ. L'annexe 7.B présente les principales caractéristiques de la loi de Student multidimensionnelle.



(a)



(b)



(c)

FIGURE 7.1 – Chaîne générée par l'échantillonnage de Gibbs pour l'exemple 1.

Rappelons que s^2 est une estimation de σ^2. Dans l'équation (7.7), on constate que lorsque l'on utilise s^2 pour estimer σ^2 inconnue, l'incertitude de cette estimation est prise en compte. On a en effet une loi de Student pour $\boldsymbol{\beta}$ qui a des queues plus lourdes que la loi normale. Les estimations sont toujours centrées à l'estimation obtenue par les moindres carrés mais l'incertitude est proportionnelle est la variance de la loi de Student.

Exemple 3: suite de l'exemple 1

Dans le cas du poids des perches de l'exemple 1, on a que

$$n = 44, \quad m = 2, \quad s^2 = 0.0885$$

et

$$(X^\top X)^{-1} = \begin{bmatrix} 0.2329 & -0.0083 \\ -0.0083 & 0.0003 \end{bmatrix}$$

On a alors que

$$f_{(\beta_0|\boldsymbol{Y}=\boldsymbol{y})}(\beta_0) = t_{42}\left(\beta_0 \mid -0.16,\, 0.14\right)$$

et

$$f_{(\beta_1|\boldsymbol{Y}=\boldsymbol{y})}(\beta_1) = t_{42}\left(\beta_1 \mid 0.26,\, 0.0054\right)$$

La figure 7.2 illustrent ces lois marginales superposées aux échantillons de ces lois marginales générées par l'échantillonnage de Gibbs.

Exercice 2

Montrez le résultat suivant :

$$(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = \boldsymbol{y}^\top\boldsymbol{y} - \hat{\boldsymbol{\beta}}^\top X^\top X\hat{\boldsymbol{\beta}}.$$

Il est également possible de trouver la loi *a posteriori* marginale de σ^2 en intégrant les coefficient de régression $\boldsymbol{\beta}$ de la loi *a posteriori*. On peut montrer (voir l'annexe 7.C) que la loi *a posteriori* marginale de σ^2 s'exprime sous la forme suivante :

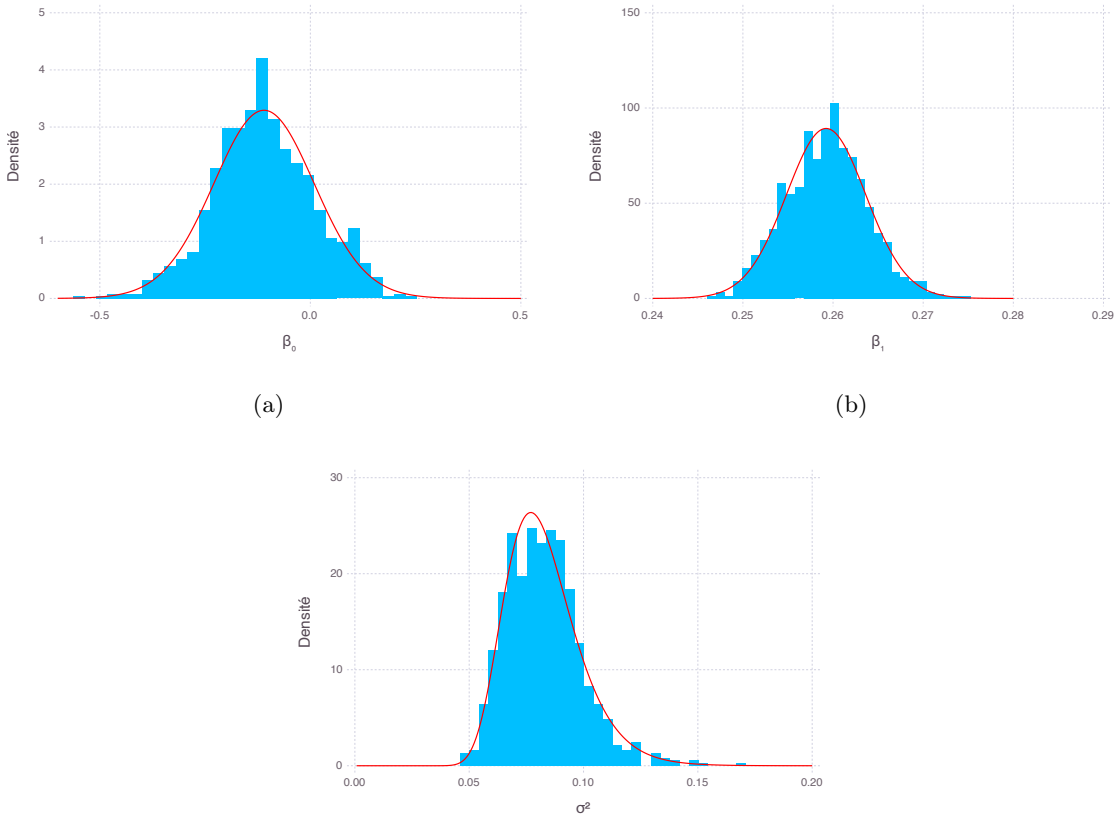
$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2) = \mathcal{I}nverse\mathcal{G}amma\left\{\sigma^2 \,\middle|\, \frac{n-m}{2}, \frac{(n-m)s^2}{2}\right\}. \tag{7.8}$$

Exemple 4: suite de l'exemple 1

Dans le cas du poids des perches de l'exemple 1, on a que

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2) = \mathcal{I}nverse\mathcal{G}amma\left(\sigma^2 \,|21, 1.86\right).$$

La figure 7.2 illustrent ces lois marginales superposées aux échantillons de ces lois marginales générées par l'échantillonnage de Gibbs.



(a) (b)

(c)

FIGURE 7.2 – Lois marginales théoriques (en rouge) superposées aux lois marginales générées par l'échantillonnage de Gibbs (en bleu) pour (a) β_0, (b) pour β_1 et (c) σ^2. Les histogrammes en bleu sont obtenus respectivement avec les itérations de la phase d'échantillonnage des figures 7.1a, 7.1b et 7.1c.

Dans le cas où des estimations ponctuelles des paramètres sont souhaitées, par exemple pour faire de la prédiction très rapidement, les lois marginales peuvent être utilisées. Par exemple, si le mode est utilisé comme estimation ponctuelle, alors l'estimation des coefficients de régression correspond au mode de la loi (7.7) et l'estimation de la variance de l'erreur correspond au mode de la loi (7.8). Avec le mode comme estimation ponctuelle, le BIC pour le modèle de régression peut facilement être calculé.

Remarque. *De façon générale, une loi* a priori *impropre est incompatible avec la sélection de modèle de bayésienne avec le facteur de Bayes. Il faut plutôt utiliser le BIC dans ce cas.*

Exemple 5: suite de l'exemple 1

Les modes des lois *a posteriori* marginales sont

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.11 & 0.26 \end{bmatrix}^\top \quad \text{et} \quad \hat{\sigma}^2 = \frac{(n-m)}{(n-m+2)} s^2 = 0.0845.$$

Le BIC de ce modèle de régression est donné par l'expression suivante :

$$\begin{aligned} BIC &= \sum_{i=1}^{n} \ln f_{(Y_i|\hat{\boldsymbol{\beta}},\hat{\sigma}^2)}(y_i) - \frac{m+1}{2}\ln(n) \\ &= -8.0757 - \frac{3}{2}\ln(44) \\ &= -13.75. \end{aligned}$$

Exemple 6: suite de l'exemple 1

Le modèle de régression peut être ajusté pour chacune des variables explicatives et pour mesurer la qualité d'ajustement, le BIC et le coefficient de détermination sont calculés. Pour évaluer la qualité des prédictions, la racine carré de l'erreur quadratique moyenne est calculé sur les prédictions pour les observations de l'échantillon de validation. Ces mesures sont compilées dans le tableau suivant :

Variable	R^2	BIC	RMSE
Height	0.9987	-5.45	0.3727
TotalLength	0.9982	-12.98	0.21
NonStandardLength	0.9982	-13.27	0.2259
StandardLength	0.9982	-13.75	0.2244
Width	0.9968	-26.26	0.3107

Le modèle qui s'ajuste le mieux aux observations de l'ensemble d'entraînement est celui utilisant la variable *Height* selon le BIC et le R^2. Le modèle qui procure les meilleures prédictions est celui qui utilise la variable *TotalLength* car son RMSE est le plus petit.

7.3 Regression ridge

Lorsque la loi *a priori* impropre de la section précédente est utilisée, on peut remarquer que les lois conditionnelles complètes et les lois marginales demeurent sensibles à la multicolinéarité. En effet, l'inversion de la matrice $(X^\top X)$, présente dans les lois conditionnelles complètes et marginales, peut s'avérer être une opération hasardeuse en présence de multicolinéarité. Une solution élégante pour contrer l'effet de la multicolinéarité consiste à utiliser une loi *a priori* informative. La loi informative présentée dans cette section correspond au modèle de régression ridge, modèle très répandue en pratique.

Remarque. *La regression ridge a été développée en Andreï Nikolaïevitch Tikhonov en 1930. Il a proposé d'introduire un terme de pénalité dans la minimisation des moindres carrés pour stabiliser le calculer de l'inverse. Cette technique peut être justifiée rigoureusement dans le cadre de la régression bayésienne avec l'utilisation d'une loi* a priori *informative.*

Exemple 7

Considérons maintenant la racine cubique du poids des perches y en fonction des 5 variables explicatives $(x_1, x_2, x_3, x_4, x_5)$. Nous avons vu au TD4 que les variables explicatives induisent une forte multicolinéarité. La régression ridge permettra de contrôler les effets de la multicolinéarité.

7.3.1 Mise à l'échelle des variables

En général, les échelles des variables explicatives et de la variable d'intérêt peuvent varier de plusieurs ordre de grandeur. La mise à l'échelle des variables (ou *feature scaling* en anglais) est une procédure permettant de ramemer les données à une échelle commune.

Il existe plusieurs méthodes de mise à l'échelle des variables. Dans ce chapitre, nous utiliserons la standardisation :

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}} \qquad \text{et} \qquad y'_i = \frac{y_i - \bar{y}}{s_y};$$

où nous laisserons tomber le $'$ dans la section pour alléger la notation.

Pour la régression, la standardisation des variables entraîne deux conséquences. D'une part, l'ordonnée à l'origine du modèle de régression n'est plus nécessaire. Autrement dit, la matrice de structure X ne commence plus par une colonne de 1 et le vecteur des coefficients de régression n'inclut plus l'ordonnée à l'origine β_0. D'autre part, les effets des variables explicatives sur la variables réponses peuvent être directement comparés puisqu'ils sont tous sur la même échelle. Par exemple, si $\beta_1 > \beta_2$, on pourra conclure que l'effet de la variable x_1 est plus important que celui de la variable x_2.

7.3.2 Loi *a priori*

Si on reprend la décomposition de la loi *a priori* exprimée à l'équation (7.2), la composante de la loi *a priori* concernant les coefficients de régression pour la régression ridge est informative tandis que la composante concernant la variance de l'erreur est non informative.

Soit la loi *a priori* conditionnelle suivante pour les coefficients de régression :

$$f_{(\boldsymbol{\beta}|\sigma^2)}(\boldsymbol{\beta}) = \mathcal{N}\left(\boldsymbol{\beta}\middle|\mathbf{0}_m, \frac{\sigma^2}{\lambda} I_m\right) \text{ avec } \lambda > 0;$$

où $\mathbf{0}_m$ dénote le vecteur colonne nul de dimension m, m étant le nombre de colonnes de la matrice X. Cette loi suppose *a priori* que les effets des variables explicatives sont nuls. La certitude de cette supposition *a priori* est contrôlée avec l'hyperparamètre λ. Remarquez que les variances *a priori* des coefficients de régression sont considérées égales. Cette supposition est raisonnable étant donné que les variables explicatives ont été standardisées au préalable.

Remarque. *Dans le cadre rigoureux de la régression bayésienne, il faudrait que λ soit fixé avant même de voir les données. Nous verrons en pratique que ce n'est pas ce qui est fait dans le cadre de la régression ridge.*

On ajoute ensuite la loi *a priori* marginale impropre pour σ^2 :

$$f_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}.$$

La loi *a priori* est donc la suivante :

$$f_{(\boldsymbol{\beta},\sigma^2)}(\mu, \sigma^2) \propto \mathcal{N}\left(\boldsymbol{\beta}\middle|\mathbf{0}_m, \frac{\sigma^2}{\lambda} I_m\right) \times \frac{1}{\sigma^2}. \tag{7.9}$$

Pour être précis, cette loi *a priori* est partiellement informative : elle est informative pour les coefficients de régression $\boldsymbol{\beta}$ mais non informative pour la variance de l'erreur σ^2.

7.3.3 Lois conditionnelles complètes

La forme fonctionnelle de la loi *a posteriori* s'exprime sous la forme suivante :

$$\begin{aligned} f_{\{(\boldsymbol{\beta},\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{\beta}, \sigma^2) &\propto \left|\sigma^2 I_n\right|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top (\sigma^2 I_n)^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right\} \\ &\quad \times \left|\frac{\sigma^2}{\lambda} I_m\right|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{0}_m)^\top \left(\frac{\sigma^2}{\lambda} I_m\right)^{-1} (\boldsymbol{\beta} - \mathbf{0}_m)\right) \times \frac{1}{\sigma^2}; \\ &\propto \frac{1}{(\sigma^2)^{\frac{n+m}{2}+1}} \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}\right\}\right]. \end{aligned} \tag{7.10}$$

Cette forme fonctionnelle ne correspond à aucune densité connue. Alors les lois conditionnelles complètes peuvent être calculées afin d'implémenter l'échantillonnage de Gibbs permettant de générer un échantillon de la loi *a posteriori*.

La forme fonctionnelle de loi *a posteriori* peut être réécrite sous la forme suivante :

$$f_{\{(\boldsymbol{\beta},\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{\beta},\sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{n+m}{2}+1}} \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)^\top (X^\top X+\lambda I_m)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)+ns_\lambda^2\right\}\right]. \tag{7.11}$$

où

$$\hat{\boldsymbol{\beta}}_\lambda = (X^\top X+\lambda I_m)^{-1}X^\top \boldsymbol{y}$$

et

$$s_\lambda^2 = \frac{1}{n}\left\{\boldsymbol{y}^\top\boldsymbol{y}-\hat{\boldsymbol{\beta}}_\lambda^\top (X^\top X+\lambda I)\hat{\boldsymbol{\beta}}_\lambda\right\}.$$

Nous verrons que $\hat{\boldsymbol{\beta}}_\lambda$ correspond à l'estimation ridge des coefficients de régression et s_λ^2 à l'estimation de l'erreur de la régression ridge.

De la forme fonctionnelle de la loi *a posteriori* exprimée à l'équation (7.11), il est possible d'identifier les lois conditionnelles complètes suivantes :

$$f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\boldsymbol{\beta}) \sim \mathcal{N}\left\{\boldsymbol{\beta}\,\middle|\,\hat{\boldsymbol{\beta}}_\lambda,\ \sigma^2\left(X^\top X+\lambda I_m\right)^{-1}\right\}, \tag{7.12}$$

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\beta})}(\sigma^2) \sim \mathcal{I}nv\mathcal{G}amma\left\{\sigma^2\,\middle|\,\frac{n+m}{2},\ \frac{(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}}{2}\right\}. \tag{7.13}$$

La loi *a priori* ajoute le terme positif λ à la diagonale de la matrice $(X^\top X)$. Cet ajout a pour conséquence de stabiliser le calcul de l'inverse de la matrice $(X^\top X+\lambda I)$, à condition que λ soit suffisamment grand. C'est pourquoi cette méthode est populaire en cas de multicolinéarité.

7.3.4 Lois *a posteriori* marginales

À l'instar de la section précédente, il est possible d'obtenir une expression analytique pour les lois *a posteriori* marginales des coefficients de régression et de la variance de l'erreur. Nous montrerons en classe que la loi *a posteriori* marginale de $\boldsymbol{\beta}$ est la suivante :

$$f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) = t_n\left\{\boldsymbol{\beta}\,\middle|\,\hat{\boldsymbol{\beta}}_\lambda,\ \sqrt{s_\lambda^2\left(X^\top X+\lambda I_m\right)^{-1}}\right\};$$

En classe, nous montrerons également que la loi *a posteriori* marginale de σ^2 s'exprime sous la forme suivante :

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2) = \mathcal{I}nverse\mathcal{G}amma\left\{\sigma^2\,\middle|\,\frac{n}{2},\ \frac{ns_\lambda^2}{2}\right\}$$

L'annexe 7.D contient tous les détails.

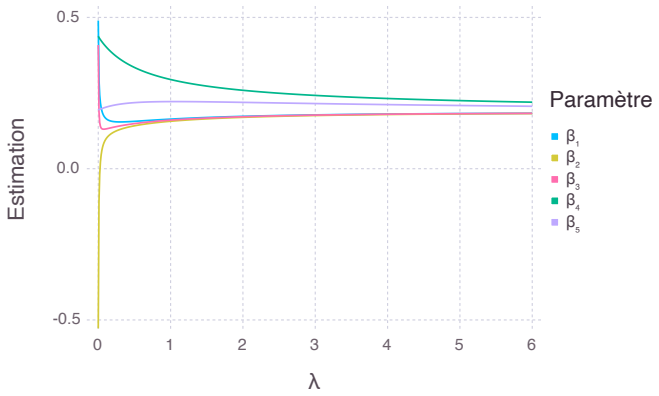


FIGURE 7.3 – Estimation des coefficients de régression ridge en fonction de λ.

Remarque. *Lorsque $\lambda = 0$, les coefficients de régression ridge correspondent aux coefficients de régression par les moindres carrés ordinaires.*

Exemple 8

La figure 7.3 illustre l'estimation des coefficients de la régression ridge en fonction de λ pour le problème du poids des perches décrit à l'exemple 7. L'estimation des coefficients se stabilisent à mesure que λ augmente. Lorsque $\lambda = 0$, les estimations des coefficients sont très instables en raison de la multicolinéarité.

Remarque. *Les coefficients de la régression ridge $\hat{\boldsymbol{\beta}}_\lambda$ sont biaisés, c'est à dire que*

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_\lambda) - \boldsymbol{\beta} \neq \mathbf{0}.$$

contrairement aux estimations obtenues par les moindres carrés. En cas de multicolinéarité, la variance des coefficients de régression ridge est cependant susceptible d'être plus petite que celle des estimations obtenues par les moindres carrés.

7.3.5 Spécification de λ

Dans une approche bayésienne rigoureuse, l'hyperparamètre λ devrait être fixé avant même d'avoir vu les données. Cependant en pratique lorsque la régression ridge est utilisée en apprentissage machine, λ est la plupart du temps estimé avec données. En particulier, la valeur de λ est déterminée par celle qui minimise l'erreur de prédiction sur l'ensemble de validation.

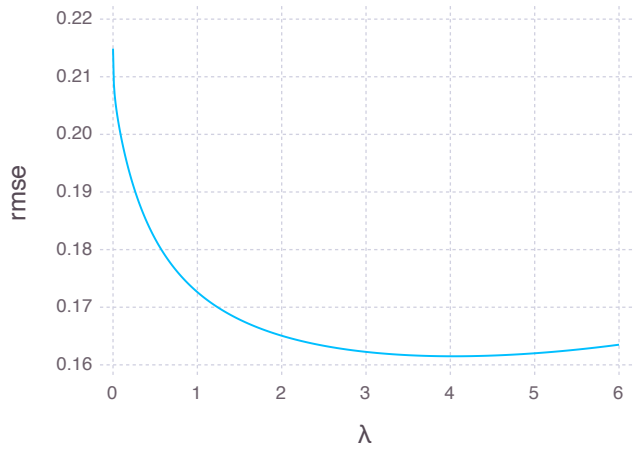


FIGURE 7.4 – La racine carré de l'erreur quadratique moyenne (rmse) des prédictions en fonction de λ. L'erreur est calculée sur les prédictions de la racine cubique du poids des perches de l'ensemble de validation.

Exemple 9

La figure 7.4 illustre la racine carrée de l'erreur quadratique moyenne en fonction de λ pour le problème du poids des perches décrit à l'exemple 7. La valeur de l'hyperparamètre λ qui minimise le rmse est $\hat{\lambda} = 4.04$ avec un $rmse = 0.1615$.

En utilisant la régression ridge, on a pu utiliser les 5 variables explicatives même si elles étaient multicorrélées. Ce modèle a été avantageux parce que l'erreur de prédiction est plus petite que celle du meilleur modèle unidimensionnel. Si vous avez le temps, vous pourriez comparer le rmse de la régression ridge au rmse de la régression en utilisant les composantes principales, l'autre méthode que nous avons vu pour contrôler la multicolinéarité.

7.3.6 Lien avec l'optimisation

On remarque que le poids demeure toujours la variable la plus importante relative la valeur des diamants. Les coefficients correspondants aux variables de dimensions (x, y, z) des diamants changent le plus. Certains coefficients changent même de signe. Cela est dû par le fait que la relation qui unit les variables explicatives à la valeur du diamant n'est pas unique. En ajoutant une loi *a priori* centrée en 0 pour les coefficient de régression, cela a pour effet d'ajouter une contrainte à la solution : on cherche la relation avec les près de 0

possible. Le poids de cette contrainte est contrôlé par λ.

7.4 Sélection de modèle (OPTIONEL)

Tel que mentionné au chapitre sur la régression linéaire, le modèle de régression optimal conserve le nombre minimal de variables explicatives tout en maximisant le pouvoir prédictif sur la variable réponse. Dans le cas où de nombreuses variables explicatives sont considérées, un grand nombre de modèles de régression sont possibles. Par exemple, si l'on possède p variables explicatives, il y a plus de 2^p modèles de régression possibles. La régression bayésienne facilite le choix du modèle lorsqu'il est impossible de tous les dénombrer. Il suffit de parcourir l'espace des modèles avec un algorithme MCMC.

7.4.1 Une variable indicatrice représentant les variables incluses dans le modèle

Une notation pratique pour indiquer quelles variables sont incluses dans un modèle particulier consiste à utiliser un vecteur ligne $\boldsymbol{\gamma}$ de dimension p composé de 0 et de 1 où l'élément j, dénoté γ_j correspond à

$$\gamma_j = \begin{cases} 0 \text{ si la } j^e \text{ variable n'est pas incluse dans le modèle,} \\ 1 \text{ si la } j^e \text{ variable est incluse dans le modèle.} \end{cases}$$

Le modèle de régression correspondant à une valeurs particulière du vecteur $\boldsymbol{\gamma}$ peut être dénoté par $\mathcal{M}_\gamma$. Par exemple, la notation $\mathcal{M}_{[1100]}$ correspondrait au modèle de régression composée des deux premières variables explicatives.

Posons la variable scalaire $q_\gamma = \boldsymbol{\gamma}\mathbf{1}_p$ où $\mathbf{1}_p$ dénote le vecteur colonne composé de uns de dimension p. La variable q_γ indique le nombre de variables explicatives considérées dans le modèle de régression $\mathcal{M}_{\boldsymbol{\gamma}}$.

7.4.2 Sélection de variables si le nombre de modèles n'est pas trop grand

Si le nombre de modèles possibles de régression n'est pas trop grand, il est alors possible de calculer le critère BIC pour chacun d'eux. Le meilleur modèle sera sélectionné comme celui ayant le BIC le plus élevé. Pour un modèle particulier, il suffira de calculer :

— le mode $\hat{\boldsymbol{\beta}}$ de la loi *a posteriori* des coefficients de régression ;
— le mode $\hat{\sigma}^2$ de la loi *a posteriori* de la variance de l'erreur ;
— le nombre de paramètres du modèle $(q_\gamma + 1)$;

pour calculer le BIC de ce modèle :

$$BIC = \ln f_{(\boldsymbol{Y}|\hat{\boldsymbol{\beta}},\hat{\sigma}^2)}(\boldsymbol{y}) - \frac{q_{\boldsymbol{\gamma}} + 1}{2} \ln n.$$

S'il y a présence de multicolinéarité, la loi *a priori* partiellement informative présentée à la section précédente s'avère très utile. Dans le cas contraire, la loi *a priori* impropre peut être utilisée.

7.4.3 Recherche stochastique du meilleur modèle

Dans le cas où le nombre de variables explicatives est très grand, il sera pratiquement impossible de calculer le critère BIC pour chacun des sous-modèles. Par exemple, s'il y a 30 variables explicatives, il y aura plus de $2^{30} \approx 1 \times 10^9$ sous-modèles possibles. L'alternative consiste à implémenter l'échantillonnage de Gibbs pour la sélection de modèle.

Supposons tous les modèles équiprobables *a priori*, on a alors que

$$f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \begin{cases} \frac{1}{2^p} & \text{si } \boldsymbol{\gamma} \in \{0,1\}^p, \\ 0 & \text{sinon.} \end{cases}$$

La loi conditionnelle complète de cette variable est une loi de Bernoulli étant donné que γ_j ne peut prendre que la valeur 0 ou 1. L'idée consiste à initialiser les variables explicatives incluses dans le modèle de régression à une valeur arbitraire, disons $\boldsymbol{\gamma}^{(1)}$. Ensuite, l'état suivant de la première composante, *i.e.* $\gamma_1^{(2)}$ est une réalisation de la loi suivante :

$$\gamma_1^{(2)} \sim \mathcal{B}ernoulli(\theta_1);$$

avec

$$\theta_1 = \frac{\exp\left[\text{BIC}\left\{\mathcal{M}_{(1,\gamma_2^{(1)},\ldots,\gamma_p^{(1)})}\right\}\right]}{\exp\left[\text{BIC}\left\{\mathcal{M}_{(0,\gamma_2^{(1)},\ldots,\gamma_p^{(1)})}\right\}\right] + \exp\left[\text{BIC}\left\{\mathcal{M}_{(1,\gamma_2^{(1)},\ldots,\gamma_p^{(1)})}\right\}\right]}$$

En répétant cette procédure pour toutes les autres composantes du vecteur $\boldsymbol{\gamma}$ et un très grand nombre de fois, on obtient l'échantillonnage de Gibbs résumé à l'algorithme 1. Le modèle qui sera le plus souvent sélectionné correspond au modèle le plus probable.

Algorithm 1 Échantillonnage de Gibbs pour la sélection de variables

Initialiser l'état des variables indicatrices $\boldsymbol{\gamma}^{(0)} = \left(\gamma_1^{(0)}, \ldots, \gamma_p^{(0)}\right)$.

for t = 1 à N **do**

1. Tirer $\gamma_1^{(t)}$ de la loi $\mathcal{B}ernoulli(\theta_1)$, où

$$\theta_1 = \frac{\exp\left[\text{BIC}\left\{\mathcal{M}_{(1,\gamma_2^{(t-1)},\ldots,\gamma_p^{(t-1)})}\right\}\right]}{\exp\left[\text{BIC}\left\{\mathcal{M}_{(0,\gamma_2^{(t-1)},\ldots,\gamma_p^{(t-1)})}\right\}\right] + \exp\left[\text{BIC}\left\{\mathcal{M}_{(1,\gamma_2^{(t-1)},\ldots,\gamma_p^{(t-1)})}\right\}\right]}.$$

$\vdots$

p. Tirer $\gamma_p^{(t)}$ de la loi $\mathcal{B}ernoulli(\theta_p)$, où

$$\theta_1 = \frac{\exp\left[\text{BIC}\left\{\mathcal{M}_{(\gamma_1^{(t)},\ldots,\gamma_{p-1}^{(t)},1)}\right\}\right]}{\exp\left[\text{BIC}\left\{\mathcal{M}_{(\gamma_1^{(t)},\ldots,\gamma_{p-1}^{(t)},0)}\right\}\right] + \exp\left[\text{BIC}\left\{\mathcal{M}_{(\gamma_1^{(t)},\ldots,\gamma_{p-1}^{(t)},1)}\right\}\right]}.$$

end for

7.5 Exercices

1. Évaluez les intégrales suivantes.
   (a) $\displaystyle\int_{-\infty}^{\infty} \exp\left(-2y^2 + 4y - 4\right) dy$
   (b) $\displaystyle\int_{0}^{\infty} y^{-5} \exp\left(-\frac{1}{y}\right) dy$
   Vous pouvez utilisez le fait que pour n entier positif, $\Gamma(n) = (n-1)!$.

2. Supposez que l'on utilise la loi *a priori* partiellement informative suivante :
$$f_{(\boldsymbol{\beta},\sigma^2)}(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(\sigma^2)^2} \exp\left(-\frac{1}{2\sigma^2}\right)$$
   pour les paramètres $\boldsymbol{\beta}$ et σ^2 du modèle de régression linéaire bayésienne.
   (a) Pour quel(s) paramètre(s) la loi *a priori* est informative ? Justifiez.
   (b) Quelle est la forme fonctionnelle de la loi *a posteriori* des paramètres ?
   (c) Quelle est la loi conditionnelle complète de $\boldsymbol{\beta}$? Autrement dit, identifiez la densité $f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\boldsymbol{\beta})$.

(d) Quelle est la loi conditionnelle complète de σ^2 ? Autrement dit, identifiez la densité $f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\beta})}(\sigma^2)$.

3. Reprenez le jeu de données `bodyfat.csv` contenant le taux de gras de 20 femmes en santé en fonction de
   — x_1 : l'épaisseur des plis de la peau des triceps (en mm) ;
   — x_2 : le tour de cuisse (en mm) ;
   — x_3 : la circonférence du bras en (mm).
   Avec la loi *a priori* non informative, générez un échantillon aléatoire de la loi *a posteriori* à l'aide de l'échantillonnage de Gibbs. Obtenez les estimations bayésiennes ponctuelles définies comme la moyenne de la loi *a posteriori*.

4. Toujours avec le jeu de données `bodyfat.csv`, implémentez la régression linéaire bayésienne avec la loi *a priori* informative introduite à la section 6.4 pour contrer l'effet de la multicolinéarité.
   a) Calculez l'estimation de λ avec une méthode bayésienne empirique.
   b) Est-ce que les estimations bayésiennes ponctuelles sont différentes de celles calculées au numéro précédent ? Si oui, expliquez pourquoi. Sinon, expliquez aussi pourquoi.

5. Toujours avec le jeu de données `bodyfat.csv`, effectuez la sélection de modèle en calculant le BIC pour chacun des modèles possibles. Puisqu'il y a 3 variables explicatives, il y a 8 modèles de régression possibles.

6. (Optionel) Avec les BIC calculés au numéro précédent, implémentez l'échantillonnage de Gibbs permettant d'effectuer une recherche stochastique du meilleur modèle. Vous constaterez que le modèle choisi le plus souvent au fil des itérations correspond au meilleur modèle identifié au numéro précédent.

7.A La loi normale multidimensionnelle

La loi normale multidimensionnelle est la densité de probabilité qui généralise la loi normale unidimensionnelle en plusieurs dimensions. Soit le vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de dimension n. On dit que $\boldsymbol{Y}$ est distribuée selon la loi normale multidimensionnelle de dimension n si la densité conjointe du vecteur s'exprime sous la forme suivante :

$$f_{(\boldsymbol{Y}|\boldsymbol{\mu},\Sigma)}(\boldsymbol{y}) = \frac{(2\pi)^{-n/2}}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\},$$

où $\boldsymbol{\mu}$ est un vecteur colonne de taille n et Σ est une matrice carrée de taille n semi-définie positive. On dénote la phrase *le vecteur aléatoire de dimension n est distribuée selon la loi normale multidimensionnelle de paramètres μ et Σ* par

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma).$$

Voici quelques propriétés de la loi normale multidimensionnelle :

1. $Y_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$, où μ_i est l'élément i du vecteur $\boldsymbol{\mu}$ et Σ_{ii} est l'élément (i,i) de la matrice Σ.
2. $\mathbb{C}ov(Y_i, Y_j) = \Sigma_{ij}$.
3. Si Y_i et Y_j sont indépendantes, alors $\Sigma_{ij} = \Sigma_{ji} = 0$.

La propriété 1 stipule que la loi marginale de la composante Y_i du vecteur aléatoire est distribuée selon la loi normale de moyenne μ_i et de variance Σ_{ii}. Le paramètre $\boldsymbol{\mu}$ de la loi normale multidimensionnelle correspond donc aux vecteur des moyennes marginales. La propriété 2 indique que le paramètre Σ correspond à la matrice de covariance des Y_i, où les variances marginales se retrouvent sur la diagonale.

7.B La loi de Student multidimensionnelle

La loi t de Student multidimensionnelle est la densité de probabilité qui généralise la loi t unidimensionnelle en plusieurs dimensions. Soit le vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de dimension n. On dit que $\boldsymbol{Y}$ est distribuée selon la loi t de Student multidimensionnelle de dimension n si la densité conjointe du vecteur s'exprime sous la forme suivante :

$$f_{(\boldsymbol{Y}|\boldsymbol{\mu},\Sigma)}(\boldsymbol{y}) = \frac{\Gamma\left[(\nu+p)/2\right]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+p)/2}$$

où $\nu > 0$ correspond au nombre de degrés de liberté, $\boldsymbol{\mu} \in \mathbb{R}^n$ au paramètre de localisation et Σ, qui est une matrice définie positive de taille n, au paramètre d'échelle. On dénote la phrase *le vecteur aléatoire de dimension n est distribuée selon la loi de Student multidimensionnelle à nu degrés de liberté de localisation μ et d'échelle Σ* par

$$\boldsymbol{Y} \sim t_\nu(\boldsymbol{\mu}, \Sigma).$$

Voici quelques propriétés de la loi de Student multidimensionnelle :

1. La moyenne, la médiane et le mode de la loi de Student multidimensionnelle sont donnés par $\boldsymbol{\mu}$ si $\nu > 1$.
2. La matrice de variance de la loi de Student multidimensionnelle est donnée par
$$\frac{\nu}{\nu - 2}\Sigma \text{ si } \nu > 2.$$
3. $\mathbb{C}ov(Y_i, Y_j) = \frac{\nu}{\nu-2}\Sigma_{ij}$ si $\nu > 2$.
4. $Y_i \sim t_\nu(\mu_i, \Sigma_{ii})$, où μ_i est l'élément i du vecteur $\boldsymbol{\mu}$ et Σ_{ii} est l'élément (i, i) de la matrice Σ.

La propriété 4 stipule que toutes les lois marginales de la loi de Student multidimensionnelle sont des lois de Student unidimensionnelle. Le paramètre $\boldsymbol{\mu}$ correspond aux vecteur des moyennes marginales tandis que la matrice Σ correspond aux covariance entre les composantes.

7.C Lois marginales dans le cas non informatif

En utilisant la loi *a priori* non informative exprimée à l'équation (7.3), la loi *a posteriori* marginale de $\boldsymbol{\beta}$ peut être trouver en intégrant σ^2 de la loi *a posteriori* :

$$\begin{aligned}
f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) &\propto \int_0^\infty \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top(\boldsymbol{y} - X\boldsymbol{\beta})\right\} d\sigma^2 \\
&\propto \frac{\Gamma\left(\frac{n}{2}\right)}{\left\{\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top(\boldsymbol{y} - X\boldsymbol{\beta})\right\}^{\frac{n}{2}}} \\
&\propto \left\{(\boldsymbol{y} - X\boldsymbol{\beta})^\top(\boldsymbol{y} - X\boldsymbol{\beta})\right\}^{-\frac{n}{2}} \\
&\propto \left\{\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{\beta}^\top X^\top\boldsymbol{y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}\right\}^{-\frac{n}{2}} \\
&\propto \left\{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top(X^\top X)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{y}^\top\boldsymbol{y} - \hat{\boldsymbol{\beta}}X^\top X\hat{\boldsymbol{\beta}}\right\}^{-\frac{n}{2}}
\end{aligned}$$

On peut montrer (exercice XXX) que

$$(\boldsymbol{y} - X\boldsymbol{\beta})^\top(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^\top\boldsymbol{y} - \hat{\boldsymbol{\beta}}X^\top X\hat{\boldsymbol{\beta}}.$$

On alors que

$$\begin{aligned}
f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) &\propto \left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top(X^\top X)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta})\right\}^{-\frac{n}{2}} \\
&\propto \left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top(X^\top X)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (n-m)s^2\right\}^{-\frac{n}{2}} \\
&\propto \left\{1 + \frac{1}{(n-m)}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\frac{(X^\top X)}{s^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}^{-\frac{n}{2}} \\
&\propto \left\{1 + \frac{1}{(n-m)}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\frac{(X^\top X)}{s^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}^{-\frac{(n-m)+m}{2}}.
\end{aligned}$$

On reconnaît la forme fonctionnelle de loi de Student multidimensionnelle :

$$f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) = t_{n-m}\left(\boldsymbol{\beta}\left|\hat{\boldsymbol{\beta}}, \sqrt{s^2(X^\top X)^{-1}}\right.\right).$$

De façon analogue, on peut trouver la loi *a posteriori* marginale de σ^2 en intégrant le vecteur $\boldsymbol{\beta}$ de la loi *a posteriori* :

$$\begin{aligned}
f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2) &\propto \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{1}{(\sigma^2)^{\frac{n}{2}+1}}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta})\right\}\,d\boldsymbol{\beta} \\
&\propto \frac{1}{(\sigma^2)^{\frac{n}{2}+1}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top(X^\top X)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (n-m)s^2\right\}\right]\,d\boldsymbol{\beta} \\
&\propto \frac{\exp\left\{-\frac{(n-m)s^2}{2\sigma^2}\right\}}{(\sigma^2)^{\frac{n}{2}+1}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\frac{(X^\top X)}{\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}\right]\,d\boldsymbol{\beta} \\
&\propto \frac{\exp\left\{-\frac{(n-m)s^2}{2\sigma^2}\right\}}{(\sigma^2)^{\frac{n}{2}+1}}\left|\sigma^2(X^\top X)\right|^{\frac{1}{2}} \\
&\propto \frac{(\sigma^2)^{\frac{m}{2}}}{(\sigma^2)^{\frac{n}{2}+1}}\exp\left\{-\frac{(n-m)s^2}{2\sigma^2}\right\} \\
&\propto \frac{1}{(\sigma^2)^{\frac{n-m}{2}+1}}\exp\left\{-\frac{(n-m)s^2}{2\sigma^2}\right\}
\end{aligned}$$

On reconnaît la forme fonctionnelle de la loi inverse gamma. On a alors que

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2) = \mathcal{I}nv\mathcal{G}amma\left\{\sigma^2\left|\frac{(n-m)}{2}, \frac{(n-m)s^2}{2}\right.\right\}.$$

7.D Lois marginales dans le cas de la régression ridge

En utilisant la loi *a priori* informative exprimée à l'équation (7.9), la loi *a posteriori* marginale de $\boldsymbol{\beta}$ peut être trouver en intégrant σ^2 de la loi *a posteriori* :

$$
\begin{aligned}
&f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) \\
&\quad\propto \int_0^\infty \frac{1}{(\sigma^2)^{\frac{n+m}{2}+1}} \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)^\top (X^\top X+\lambda I_m)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)+ns_\lambda^2\right\}\right]\, d\sigma^2 \\
&\quad\propto \frac{\Gamma(\frac{n+m}{2})}{\left[\frac{1}{2}\left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)^\top (X^\top X+\lambda I_m)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)+ns_\lambda^2\right\}\right]^{\frac{(n+m)}{2}}} \\
&\quad\propto \left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)^\top (X^\top X+\lambda I_m)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)+ns_\lambda^2\right\}^{-\frac{(n+m)}{2}} \\
&\quad\propto \left\{1+\frac{1}{n}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)^\top \frac{(X^\top X+\lambda I_m)}{s_\lambda^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)\right\}^{-\frac{(n+m)}{2}}
\end{aligned}
$$

On reconnaît la forme fonctionnelle de la loi de Student multidimensionnelle à n degrés de liberté, vecteur de localisation $\hat{\boldsymbol{\beta}}$ et de matrice d'échelle $\sqrt{s_\lambda^2(X^\top X+\lambda I)^{-1}}$. Alors on a que

$$
f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) = t_n\left\{\boldsymbol{\beta}\,\middle|\,\hat{\boldsymbol{\beta}}, \sqrt{s_\lambda^2(X^\top X+\lambda I)^{-1}}\right\}
$$

De façon analogue, on peut trouver la loi *a posteriori* marginale de σ^2 en intégrant le vecteur $\boldsymbol{\beta}$ de la loi *a posteriori* :

$$
\begin{aligned}
&f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2) \\
&\quad\propto \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \frac{1}{(\sigma^2)^{\frac{n+m}{2}+1}} \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)^\top (X^\top X+\lambda I_m)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)+ns_\lambda^2\right\}\right]\, d\boldsymbol{\beta} \\
&\quad\propto \frac{\exp\left(-\frac{ns_\lambda^2}{2\sigma^2}\right)}{(\sigma^2)^{\frac{n+m}{2}+1}} \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)^\top (X^\top X+\lambda I_m)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\lambda)\right\}\right]\, d\boldsymbol{\beta} \\
&\quad\propto \frac{1}{(\sigma^2)^{\frac{n+m}{2}+1}} \exp\left(-\frac{ns_\lambda^2}{2\sigma^2}\right) \left|\sigma^2(X^\top X+\lambda I_m)\right|^{\frac{1}{2}} \\
&\quad\propto \frac{1}{(\sigma^2)^{\frac{n+m}{2}+1}} \exp\left(-\frac{ns_\lambda^2}{2\sigma^2}\right) (\sigma^2)^{\frac{m}{2}}
\end{aligned}
$$

$$\propto \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} \exp\left(-\frac{ns_\lambda^2}{2\sigma^2}\right).$$

On reconnaît la forme fonctionnelle de la loi inverse gamma. On a alors que

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2) = \mathcal{I}nv\mathcal{G}amma\left\{\sigma^2 \middle| \frac{n}{2}, \frac{ns_\lambda^2}{2}\right\}.$$