CHAPITRE 6

Régression linéaire bayésienne

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

Jonathan Jalbert – Automne 2019

La régression linéaire bayésienne est utile pour régler deux problèmes qui peuvent survenir en régression linéaire, soit la sélection du meilleur modèle parmi un vaste ensemble de modèles et 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.
— Implanter l'échantillonnage de Gibbs pour la régression linéaire bayésienne.
— Corriger les effets de la multicolinéarité en utilisant une loi *a priori* informative.
— Sélectionner le meilleur modèle parmi un ensemble de modèles de régression.

Dans ce chapitre, nous supposerons que les variables explicatives et la variable d'intérêt ont été standardisées au préalable. D'une part, l'ordonnée à l'origine du modèle de régression n'est plus nécessaire. 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.

La standardisation est très utile pour la régression bayésienne avec loi *a priori* informative. Elle est cependant un peu moins un utile dans le cas non informatif.

6.1 Modèle de régression linéaire bayésien

Le problème de la régression linéaire consiste à prédire la variable réponse Y avec un ensemble de variables explicatives $\boldsymbol{x}$. Nous supposons que les variables ont été standardisées

au préalable. Rappelons l'hypothèse fondamentale de linéarité de la régression linéaire :

Hypothèse 1 (Linéarité) :

$$E\left(Y \mid X = \boldsymbol{x}\right) = \boldsymbol{x}\boldsymbol{\beta}$$

où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ sont les coefficients de régression. Cette hypothèse implique le modèle statistique suivant pour un couple d'observations $(\boldsymbol{x}_i, Y_i)$ d'un échantillon aléatoire de taille n, *i.e.* $1 \leq i \leq n$:

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i;$$

où ε_i est un terme d'erreur aléatoire d'espérance nulle. On supposera également les hypothèses 2 et 3 :

Hypothèse 2 (Homoscédasticité de la variance) :

$$\mathbb{V}ar(\varepsilon_i) = \sigma^2 \ ; \ \text{pour } i = 1, \ldots, n.$$

Hypothèse 3 (Indépendance) : Les erreurs doivent être mutuellement indépendantes, c'est-à-dire ε_i indépendante de ε_j pour tout $i \neq j$.

Rappelons qu'en supposant ces trois hypothèses, l'estimateur de $\boldsymbol{\beta}$ par la méthode des moindres carrés correspond à

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

Dans le modèle bayésien, l'hypothèse 4 de normalité est également requise :

Hypothèse 4 (Normalité) : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \ \text{pour } i = 1, \ldots, n.$

La densité de la variable Y_i est alors donnée par

$$f_{(Y_i|\boldsymbol{\beta},\sigma^2)}(y_i) = \mathcal{N}\left(y_i \mid \boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2\right).$$

Puisque les observations sont supposées indépendantes (hypothèse 3), la densité conjointe des Y_i peut s'écrire comme le produit des densités marginales. La densité conjointe des observation peut s'écrire en fonction de la loi normale multidimensionnelle suivante :

$$f_{(\boldsymbol{Y}|\boldsymbol{\beta},\sigma^2)}(\boldsymbol{y}) = \mathcal{N}_n\left(\boldsymbol{y} \mid X\boldsymbol{\beta}, \sigma^2 I_n\right), \tag{6.1}$$

où I_n dénote la matrice identité de taille n. Dans ce modèle, il y a $(p+1)$ paramètres inconnus à estimer : les p coefficients de régression $(\beta_j : 1 \leq j \leq p)$ et la variance de l'erreur σ^2.

6.2 Estimation bayésienne avec une loi *a priori* non informative

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

$$f_{(\boldsymbol{\beta},\sigma^2)}(\boldsymbol{\beta},\sigma^2) = f_{(\boldsymbol{\beta}|\sigma^2)}(\boldsymbol{\beta}) \times f_{(\sigma^2)}(\sigma^2). \tag{6.2}$$

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

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

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

$$\begin{aligned} f_{\{(\boldsymbol{\beta},\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{\beta},\sigma^2) &\propto \left|\sigma^2 I_n\right|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\sigma^2 I_n)^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \times \frac{1}{\sigma^2}; \\ &\propto \frac{1}{(\sigma^2)^{-n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \times \frac{1}{\sigma^2}. \end{aligned} \tag{6.3}$$

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 (6.1) ?

6.2.1 Lois conditionnelles complètes

De la forme fonctionnelle de la loi *a posteriori* exprimée à l'équation (6.3), il est possible d'identifier les lois conditionnelles complètes suivantes :

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

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.

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.

Exercice 2

Avec les lois conditionnelles complètes de $\boldsymbol{\beta}$ et σ^2, pouvez-vous écrire l'échantillonnage de Gibbs permettant de générer un échantillon aléatoire de la loi *a posteriori* des paramètres ?

6.2.2 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 forme fonctionnelle de la loi *a posteriori* exprimée à l'équation (6.3). Nous montrerons en classe que la loi *a posteriori* marginale de $\boldsymbol{\beta}$ s'exprime sous la forme suivante :

$$f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) = t_{n-p}\left\{\boldsymbol{\beta}\middle|\hat{\boldsymbol{\beta}}, s^2\left(X^\top X\right)^{-1}\right\}; \tag{6.4}$$

où

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

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 6.B présente les principales caractéristiques de la loi de Student multidimensionnelle.

Exercice 3

Montrez le résultat suivant :

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

En intégrant les coefficients de régression $\boldsymbol{\beta}$ de la forme fonctionnelle de la *a posteriori* exprimée à l'équation (6.3), la loi *a posteriori* marginale de σ^2 s'exprime sous la forme suivante :

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2) = \mathcal{I}nverse\mathcal{G}amma\left\{\sigma^2\middle|\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right\}. \tag{6.5}$$

Nous montrerons ce résultat en classe.

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 (6.4) et l'estimation de la variance de l'erreur correspond au mode de la loi (6.5).

Rappelons que de façon générale, une loi *a priori* impropre est incompatible avec la sélection de modèle de bayésienne avec le facteur de Bayes. Par conséquent, le facteur de

Bayes ne peut être utilisé pour la sélection de modèle dans cette section puisque la loi *a priori* utilisée est impropre. Le critère BIC peut cependant être utilisé pour la sélection de modèle.

6.3 Estimation bayésienne avec une loi *a priori* partiellement informative

Lorsque la loi *a priori* impropre de la section précédente est utilisée, on peut remarquer que les lois conditionnelles complètes et les lois marginales peuvent être sensibles au problème de la multicolinéarité. En effet, l'inversion de la matrice $(X^\top X)$ 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* partiellement informative. La composante de la loi *a priori* concernant les coefficients de régression sera informative tandis que la composante concernant la variance de l'erreur sera non informative.

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

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

où $\mathbf{0}_p$ dénote le vecteur colonne nul de dimension p. Cette loi suppose *a priori* que tous les effets des variables explicatives sont nuls. La certitude de cet connaissance *a priori* est contrôlée par 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 toutes été standardisées.

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* partiellement informative est donc la suivante :

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

Remarque. *Cette loi* a priori *particulière pour les coefficients de régression conduit vers le modèle de régression Ridge, modèle très populaire en apprentissage machine. Remarquez que d'autres lois* a priori *partiellement informative auraient aussi pu être utilisées.*

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

$$\begin{aligned} f_{\{(\boldsymbol{\beta},\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{\beta},\sigma^2) &\propto \left|\sigma^2 I_n\right|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\sigma^2 I_n)^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \\ &\quad \times \left|\frac{\sigma^2}{\lambda} I_p\right|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0}_p)^\top \left(\frac{\sigma^2}{\lambda} I_p\right)^{-1} (\boldsymbol{\beta}-\mathbf{0}_p)\right) \times \frac{1}{\sigma^2}; \\ &\propto \frac{1}{(\sigma^2)^{\frac{n+p}{2}+1}} \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\}\right]. \end{aligned} \tag{6.6}$$

Cette forme fonctionnelle ne correspond à aucune densité connue.

6.3.1 Lois conditionnelles complètes

De la forme fonctionnelle de la loi *a posteriori* exprimée à l'équation (6.6), il est possible d'identifier les lois conditionnelles complètes suivantes :

$$\begin{aligned} f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\boldsymbol{\beta}) &\sim \mathcal{N}\left\{\boldsymbol{\beta}\,\middle|\,(X^\top X+\lambda I_p)^{-1}X^\top\boldsymbol{y}, \sigma^2\left(X^\top X+\lambda I_p\right)^{-1}\right\}, \\ f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\beta})}(\sigma^2) &\sim \mathcal{I}nverse\mathcal{G}amma\left\{\sigma^2\,\middle|\,\frac{n}{2}, \frac{(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}}{2}\right\}. \end{aligned}$$

Nous montrerons ces résultats en classe.

On remarque que l'utilisation de la loi partiellement informative introduite dans cette section a pour effet d'ajouter le terme 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.

6.3.2 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 :

$$f_{(\boldsymbol{\beta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\beta}) = t_{n-p}\left\{\boldsymbol{\beta}\,\middle|\,\hat{\boldsymbol{\beta}}_{Ridge},\, s^2\left(X^\top X+\lambda I\right)^{-1}\right\};$$

où

$$\hat{\boldsymbol{\beta}}_{Ridge} = \left(X^\top X+\lambda I\right)^{-1}X^\top\boldsymbol{y}$$

et

$$s^2 = \frac{1}{n-p}\left(\boldsymbol{y}-X\hat{\boldsymbol{\beta}}_{Ridge}\right)^\top\left(\boldsymbol{y}-X\hat{\boldsymbol{\beta}}_{Ridge}\right).$$

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

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

6.3.3 Spécification de λ

En pratique, le paramètre λ est souvent estimé avec une procédure de validation croisée. La valeur de λ qui minimise l'erreur sur l'ensemble de validation est choisie. C'est une approche pragmatique en contradiction avec la philosophie bayésienne.

6.4 Sélection de variables

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.

6.4.1 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 lignt $\boldsymbol{\gamma}$ 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 $\mathcal{M}_{\boldsymbol{\gamma}}$.

6.4.2 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.

6.4.3 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

6.5 Exercices

1. 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*.

2. 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.

3. 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.

4. (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.

6.A La loi normale multidimensionnelle

Avant de présenter la régression linéaire bayésienne, révisons d'abord la loi normale multidimensionnelle qui est essentielle pour la régression linéaire bayésienne. 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.

Les trois propositions suivantes montrent qu'une combinaison linéaire de lois normales est normale et que toutes les lois conditionnelles sont aussi normales.

Proposition 1. *Soit le vecteur aléatoire $\boldsymbol{Y}$ de dimension n distribué selon la loi normale multidimensionnelle de paramètres $\boldsymbol{\mu}$ et Σ. Soit la matrice A de dimension $(m \times n)$ et le vecteur colonne $\boldsymbol{b}$ de dimension m. Si $\boldsymbol{Z} = A\boldsymbol{Y} + \boldsymbol{b}$, alors*

$$\boldsymbol{Z} = \mathcal{N}\left(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^\top\right).$$

Montrer que la moyenne de $\boldsymbol{Z}$ est $A\boldsymbol{\mu} + \boldsymbol{b}$ et que sa variance est $A\Sigma A^\top$ ne devrait pas être trop difficile. Montrer que le vecteur $\boldsymbol{Z}$ est distribué selon la loi normale multidimensionnelle est cependant difficile.

Proposition 2. *Soit $\boldsymbol{Z}$, un vecteur aléatoire distribué selon la loi normale multidimensionnelle de dimension m de moyenne $\boldsymbol{\mu}_z$ et de matrice de covariance Σ_z :*

$$\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \Sigma_z).$$

Soit A une matrice de dimension $(n \times m)$ et $\boldsymbol{b}$ un vecteur de dimension $(n \times 1)$ telle que la combinaison linéaire $A\boldsymbol{z} + \boldsymbol{b}$ corresponde à la moyenne de la variable aléatoire $(\boldsymbol{Y}|\boldsymbol{Z} = \boldsymbol{z})$:

$$(\boldsymbol{Y}|\boldsymbol{Z} = \boldsymbol{z}) \sim \mathcal{N}(A\boldsymbol{z} + \boldsymbol{b}, \Sigma_y).$$

avec la matrice de covariance Σ_y de dimensions $(n \times n)$. Alors

$$\boldsymbol{Y} \sim \mathcal{N}(A\boldsymbol{\mu}_z + \boldsymbol{b}\,,\, \Sigma_y + A\Sigma_z A^\top).$$

et

$$(\boldsymbol{Z}|\boldsymbol{Y} = \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}_{z|y}\,,\, \Sigma_{z|y}),$$

où

$$\begin{aligned}
\Sigma_{z|y}^{-1} &= \Sigma_z^{-1} + A^\top \Sigma_y^{-1} A; \\
\boldsymbol{\mu}_{z|y} &= \Sigma_{z|y} \left\{ A^\top \Sigma_y^{-1} (\boldsymbol{y} - \boldsymbol{b}) + \Sigma_z^{-1} \boldsymbol{\mu}_z \right\}.
\end{aligned}$$

La démonstration de cette proposition n'est pas facile.

6.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

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

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 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.