Les bases de la statistique bayésienne

Jean-Michel Marin
I3M, Université Montpellier 2, Montpellier &
CREST, INSEE, Paris
et
Christian P. Robert
Université Paris Dauphine, Paris &
CREST, INSEE, Paris

Résumé

Dans ce court texte de présentation de la statistique bayésienne, nous nous attachons à démontrer qu'il s'agit d'une approche cohérente et surtout pratique pour résoudre les problèmes d'inférence statistique. Les fondements historiques de cette discipline, pas plus que ses justifications théoriques et philosophiques, ne seront présentés ici, le lecteur étant renvoyé pour cela aux ouvrages de référence que sont Bernardo et Smith (1994); Carlin et Louis (2001); Gelman *et al.* (2001) et Robert (2007) (ou Robert (2006) pour la version française). Notre objet est au contraire de démontrer que cette approche de l'inférence statistique est moderne, adaptée aux outils informatiques de simulation et apte à répondre aux problèmes de modélisation les plus avancés dans toutes les disciplines, plutôt que de l'ancrer sur ses querelles du passé. Dans une première partie, nous présentons les fondements de l'inférence bayésienne, en insistant sur les spécificités de la modélisation a priori et de la construction des tests. Puis, nous mettons en œuvre explicitement les concepts précédemment introduits dans le cas pratique d'un modèle de régression linéaire.

1 Buts et contraintes de l'inférence statistique

1.1 Formalisme

Avant de mettre en place les éléments nécessaires à la construction d'une machine inférentielle bayésienne, nous considérons tout d'abord quels sont les points essentiels définissant la science statistique. Il nous apparaît qu'une définition concise est de mener, grâce à l'observation d'un phénomène aléatoire, une *inférence* soit donc une démarche déductive logique sur la distribution de probabilité à l'origine de ce phénomène, pour ainsi fournir une analyse (ou une description) d'un phénomène passé, ou bien une prévision d'un phénomène à venir (et de même nature). Il va de soi que les étapes nécessaires au recueil des données comme

la construction de plans de sondage ou d'expérience font aussi partie du domaine de la statistique et que l'approche bayésienne peut également apporter un éclairage nouveau sur ces opérations.

L'approche statistique est par essence formelle (ou mathématiquement structurée) parce qu'elle repose sur une formalisation poussée de la réalité objective. En particulier, nous insistons ici sur l'interprétation *décisionnelle* de l'inférence statistique parce que, tout d'abord, les analyses et prédictions mentionnées ci-dessus sont la plupart du temps motivées par un but objectif (comme la construction d'un portefeuille boursier ou la validation d'un contrôle de qualité) ayant des conséquences [quantitativement et souvent monétairement] mesurables (résultats financiers, taux de retour des pièces défectives) et que d'autre part, ce degré supplémentaire de formalisme permet en retour la construction d'une machine automatique d'inférence bayésienne. Notons par ailleurs que la statistique doit être considérée comme l'*interprétation* du phénomène observé, plutôt que comme son explication. En effet, l'inférence statistique est précédé d'une modélisation probabiliste et celle-ci implique nécessairement une étape de formalisation réductrice : sans cette base probabiliste, aucune conclusion utile ne pourrait être obtenue. On pourra peut-être regretter cette apposition d'un modèle probabiliste sur un phénomène inexpliqué, comme il est possible que le phénomène observé soit entièrement déterministe ou tout du moins sans rapport direct avec le modèle pré-supposé. Cependant, cette critique de la modélisation probabiliste n'a guère de consistence si nous considérons la statistique sous l'angle de l'interprétation, évoquée ci-dessus. Ces modèles probabilistes formels permettent en effet d'incorporer simultanément les informations disponibles sur le phénomène (facteurs déterminants, fréquence, amplitude, etc.) et les incertitudes inhérentes à ces informations. Ils autorisent donc un discours qualitatif sur le problème en fournissant, à travers la théorie des probabilités, un véritable *calcul de l'incertain* qui permet de dépasser le stade descriptif des modèles déterministes.

Évidemment la modélisation probabiliste n'a de sens pour l'analyse que si elle fournit une représentation suffisamment proche du phénomène observé. Dans de nombreux cas, la formalisation statistique est bien réductrice au sens où elle n'est qu'une approximation de la réalité, perdant une partie de la richesse de cette réalité mais gagnant en efficacité. Face à ce possible volant de réduction dans la complexité du phénomène observé, deux approches statistiques s'opposent. La première approche suppose que l'inférence statistique doit prendre en compte cette complexité autant que possible et elle cherche donc à estimer la distribution sous-jacente du phénomène sous des hypothèses minimales, en ayant recours en général à l'estimation fonctionnelle (densité, fonction de régression, etc.). Cette approche est dite *non paramétrique*. Par opposition, l'approche *paramétrique* représente la distribution des observations par une fonction de densité $f(x|\theta)$, où seul le paramètre θ (de dimension finie) est inconnu. Les deux approches ont leurs avantages respectifs et, bien que dans cet article, nous ne considérions que l'approche paramétrique pour des raisons pratiques, il existe également des résolutions bayésiennes du problème de l'inférence non-paramétrique (Dey *et al.*, 1997).

1.2 Notations

Le formalisme fondamental d'une approche statistique est de supposer que les observations $x_1, \ldots, x_n$, sur lesquelles l'analyse statistique se fonde, proviennent d'une loi de probabilité paramétrée, ce qui se traduit par les hypothèses selon lesquelles x_1 a une distribution de densité $f_1(x|\theta_1)$ sur un espace mesurable comme $\mathbb{R}^p$ et x_i $(2 \leq i \leq n)$ a, conditionnellement aux observations $x_1, \ldots, x_{i-1}$, une distribution de densité $f_i(x_i|\theta_i, x_1, \ldots, x_{i-1})$ sur le même espace mesurable. Dans ce cas, le paramètre θ_i est inconnu [et constitue un objet d'inférence], mais la fonction générique f_i est connue. Ce modèle peut être réécrit plus succinctement par

$$\mathbf{x} \sim f(\mathbf{x}|\boldsymbol{\theta}) = f_1(x_1|\theta_1) \prod_{i=2}^{n} f_i(x_i|\theta_i, x_1, \ldots, x_{i-1})$$

où $\mathbf{x}$ est le vecteur des observations, $\mathbf{x} = (x_1, \ldots, x_n)$, et $\boldsymbol{\theta}$ l'ensemble des paramètres, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, les composants étant éventuellement tous égaux. Cette représentation est unificatrice dans le sens où elle recouvre les cas d´une observation isolée, d'observations dépendantes, ou d'observations distribuées de façon indépendante et identiquement distribuées (*iid*), les $x_1, \ldots, x_n$ étant tous de même loi, de densité $f(x_1|\theta)$. Dans le dernier cas, $\boldsymbol{\theta} = \theta$ et

$$f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\theta).$$

Les densités $f(x|\theta)$ peuvent ainsi correspondre à des densités binomiales, de Poisson, normales ou gammas, pour citer quelques exemples standard. Une simplification de notation adoptée dans la suite est que les densités des variables aléatoires continues (comme les variables normales) et discrètes (comme les variables de Poisson) sont représentées par les mêmes symboles, la mesure de référence étant fournie naturellement par le contexte. De plus, nous écrirons "*x est distribué selon* f" ou "$x \sim f$" au lieu de "*x est une observation de la distribution de densité* f" par souci de concision.

Cette introduction à la statistique bayésienne est séparée en une première partie (§2), où nous traitons des fondements de l'inférence bayésienne et des outils qui s'y rapportent. Dans une seconde partie (§3), nous appliquons ces techniques dans le cadre du modèle de régression linéaire standard en insistant sur les solutions apportées au problème de choix de variables explicatives.

Notons enfin que la première partie du texte ne comprend que quelques références à des ouvrages que nous considérons comme essentiels pour une compréhension plus poussée des concepts et des méthodes présentés. Ces ouvrages de référence, récents, comprennent eux-mêmes des bibliographies extensives auxquelles le lecteur ou la lectrice peut se rapporter. Nous suggérons en particulier notre livre (Marin et Robert, 2007) pour une introduction pratique plus élaborée aux techniques bayésiennes.

2 Fondements de la statistique bayésienne

Le message que nous voulons communiquer dans ce rapide survol de la statistique bayésienne est on ne peut plus simple : il est possible, sans expertise préalable, de réaliser une analyse bayésienne de tout problème statistique (défini comme ci-dessus par la donnée d'observations provenant d'une distribution de probabilité paramétrée). (Des exemples réalistes et détaillés sont traités dans Gelman *et al.* (2001) et Marin et Robert (2007).) En particulier, nous insistons sur le fait que ce qui est souvent considéré comme les deux difficultés majeures de l'approche bayésienne, à savoir le choix de l'a priori et le calcul des procédures bayésiennes, peuvent être surmontées en suivant des règles simples.

2.1 Le paradigme bayésien

Étant donné un modèle paramétrique d'observation $\mathbf{x} \sim f(\mathbf{x}|\boldsymbol{\theta})$, où $\boldsymbol{\theta} \in \Theta$, un espace de dimension finie, l'analyse statistique bayésienne vise à exploiter le plus efficacement possible l'information apportée par $\mathbf{x}$ sur le paramètre $\boldsymbol{\theta}$, pour ensuite construire des procédures d'*inférence* sur $\boldsymbol{\theta}$. Bien que $\mathbf{x}$ ne soit qu'une réalisation [aléatoire] d'une loi gouvernée par $\boldsymbol{\theta}$, elle apporte une actualisation aux informations préalablement recueillies par l'expérimentateur. Pour des raisons diverses dont certaines apparaitront dans le prochain paragraphe, l'information fournie par l'observation $\mathbf{x}$ est contenue dans la densité $f(\mathbf{x}|\boldsymbol{\theta})$, que l'on représente classiquement sous la forme inversée de *vraisemblance*,

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}), \tag{1}$$

pour traduire qu'il s'agit d'une fonction de $\boldsymbol{\theta}$, qui est *inconnu*, dépendant de la valeur observée $\mathbf{x}$. L'inversion des rôles de $\mathbf{x}$ et de $\boldsymbol{\theta}$ par rapport à la modélisation probabiliste reflète le but premier de la statistique qui est de reconstruire [avec un certain degré de précision] le paramètre $\boldsymbol{\theta}$ au vu de la réalisation aléatoire $\mathbf{x}$. C'est donc pourquoi elle est naturellement liée au *théorème de Bayes* qui formalise l'inversion des conditionnements dans les probabilités : Si A et E sont des événements tels que $\mathbb{P}(E) \neq 0$, $\mathbb{P}(A|E)$ et $\mathbb{P}(E|A)$ sont reliés par

$$\begin{aligned}\mathbb{P}(A|E) &= \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E|A)\mathbb{P}(A) + \mathbb{P}(E|A^c)\mathbb{P}(A^c)} \\ &= \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E)}\,.\end{aligned}$$

Une version continue de ce résultat permet d'inverser les densités conditionnelles, à savoir,

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)\,\mathrm{d}y}\,.$$

Le lien entre ces propriétés probabilistes et l'inférence bayésienne (ainsi appelée en référence au théorème ci-dessus[1]) est que, dans le paradigme bayésien, le paramètre inconnu

[1]Historiquement, le théorème de Bayes apparaît dans un exemple d'inférence bayésienne plutôt que séparément comme une construction probabiliste.

$\boldsymbol{\theta}$ n'est plus considéré comme inconnu et déterministe, mais comme une variable aléatoire. On considère ainsi que l'*incertitude* sur le paramètre $\boldsymbol{\theta}$ d'un modèle peut être décrite par une distribution de *probabilité* π sur Θ, appelée *distribution a priori* [par opposition à la *distribution a posteriori* qui inclut l'information contenue dans l'observation $\mathbf{x}$], ce qui revient à supposer que $\boldsymbol{\theta}$ est distribué suivant $\pi(\boldsymbol{\theta})$, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$, "avant" que $\mathbf{x}$ ne soit généré suivant $f(\mathbf{x}|\boldsymbol{\theta})$, le conditionnement implicite dans cette notation prenant alors tout son sens. Sans vouloir nous engager dans un débat philosophique sur la nature du hasard, notons que le rôle central de la distribution a priori dans l'analyse statistique bayésienne ne réside pas dans le fait que le paramètre d'intérêt $\boldsymbol{\theta}$ puisse (ou ne puisse pas) être perçu comme étant distribué selon π, ou même comme étant une variable aléatoire, mais plutôt dans la démonstration que l'utilisation d'une distribution a priori et de l'appareillage probabiliste qui l'accompagne est la manière la plus efficace [au sens de nombreux critères] de résumer l'information disponible (ou le manque d'information) sur ce paramètre ainsi que l'incertitude résiduelle. Un point plus technique est que le seul moyen de construire une approche mathématiquement justifiée opérant conditionnellement aux observations, tout en restant dans un schéma probabiliste, est d'introduire une distribution correspondante pour les paramètres. Des arguments de nature différente sur l'optimalité de cet outil sont aussi fournis dans Bernardo et Smith (1994) et Robert (2007, chapitre 10).

D'un point de vue pratique, le choix de la loi a priori est souvent perçu comme une difficulté majeure de l'approche bayésienne en ce que l'interprétation de l'information a priori disponible est rarement assez précise pour conduire à la détermination d'une seule et unique loi. D'autre part, il est aisé de constater sur des exemples classiques que des choix contrastés de lois a priori conduisent à des inférences divergentes. Il existe néanmoins des lois calibrées en fonction de la distribution des observations, dites *lois conjuguées*, et des lois à faible contenu informatif, dites *lois non-informatives*, qui permettent d'évaluer l'influence d'une loi a priori donnée. Nous détaillerons cette construction de la loi a priori dans la troisième partie traitant de régression linéaire, mais notons à ce point qu'un choix quasi-automatisé de la loi a priori est réalisable par une modélisation hiérarchique (voir par exemple Marin et Robert, 2007). Signalons également que la notion de loi a priori peut être étendue à des mesures de masse infinie tant que la loi a posteriori donnée par l'équation (2) ci-dessous reste définie.

2.2 Estimation ponctuelle

Par application directe du théorème de Bayes, si $\boldsymbol{\theta}$ est supposé être une variable aléatoire de densité (a priori ou marginale) $\pi(\boldsymbol{\theta})$ et si $f(\mathbf{x}|\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}|\mathbf{x})$ est interprétée comme loi conditionnelle de $\mathbf{x}$ conditionnellement à $\boldsymbol{\theta}$, la loi de $\boldsymbol{\theta}$ conditionnelle à $\mathbf{x}$, $\pi(\boldsymbol{\theta}|\mathbf{x})$, appelée *distribution a posteriori*, est définie par

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\ell(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\int_\Theta \ell(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}\ . \tag{2}$$

Cette densité est centrale pour l'inférence bayésienne en ce qu'elle suffit à déterminer les procédures de décision et, par extension, à conduire toute inférence liée à $\boldsymbol{\theta}$. Celle-ci joue

donc le rôle d'un résumé exhaustif de l'information apportée par les données, ce qui justifie aussi l'utilisation de la vraisemblance (1) comme traduction de l'information contenu dans ces données. Notons que le dénominateur de (2) sert à la fois de constante de normalisation [pour que $\pi(\boldsymbol{\theta}|\mathbf{x})$ soit effectivement une densité de probabilité] et de vraisemblance marginale pour $\mathbf{x}$, utile dans la comparaison de modèles et les tests d'hypothèses. Comme ce dénominateur est uniquement défini en fonction du numérateur, on utilise fréquemment la notation de proportionnalité $\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})$ qui signifie que la densité en $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\mathbf{x})$, est égale au produit $\pi(\boldsymbol{\theta})f(\boldsymbol{\theta}|\mathbf{x})$ à une constante près et que cette constante s'obtient par l'intégration en $\boldsymbol{\theta}$: $1/\int_\Theta \pi(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$. La constante dépend donc de $\mathbf{x}$, vecteur des observations, ce qui n'est pas paradoxal puisque l'analyse bayésienne raisonne intégralement par conditionnement sur $\mathbf{x}$.

Par exemple, si l'on cherche à déterminer un estimateur ponctuel du paramètre $\boldsymbol{\theta}$, on peut comparer les approximations d de $\boldsymbol{\theta}$ au moyen d'une fonction de coût, $L(d, \boldsymbol{\theta})$, qui quantifie les conséquences de l'erreur commise en remplaçant le "vrai" paramètre $\boldsymbol{\theta}$ par son approximation d. Une fois construite la loi a posteriori $\pi(\boldsymbol{\theta}|\mathbf{x})$, les approximations d ont un coût moyen égal à $\mathbb{E}^\pi(L(d, \boldsymbol{\theta})|\mathbf{x})$ où cette notation signifie l'espérance sous $\pi(\boldsymbol{\theta}|\mathbf{x})$ et l'approximation ou estimation optimale est celle qui minimise cette erreur. On définit donc un estimateur bayésien $\delta(\mathbf{x})$ comme la procédure qui à chaque observation associe la solution du problème de minimisation

$$\delta(\mathbf{x}) = \arg\min_{d\in\Theta} \mathbb{E}^\pi[L(d, \boldsymbol{\theta})|\mathbf{x}]\,.$$

Cette solution exprime bien l'importance du choix de π. En pratique, il est souvent difficile de construire la fonction de perte véritablement associée au problème de décision. Une fonction de perte par défaut est alors la fonction de perte quadratique :

$$L(d, \boldsymbol{\theta}) = (d - \boldsymbol{\theta})^2\,.$$

Dans ce cas, l'estimateur bayésien est, si elle existe, l'espérance de la loi a posteriori $\delta(\mathbf{x}) = \mathbb{E}^\pi[\boldsymbol{\theta}|\mathbf{x}]$.

Example 2.1 *Dans le cas d'une observation normale de moyenne inconnue θ, $x \sim \mathcal{N}_1(\theta, 1)$[2], la loi a posteriori associée à la loi a priori $\theta \sim \mathcal{N}_1(0, 10)$ est*

$$\pi(\theta|x) \propto \pi(\theta)f(x|\theta) \propto \exp\frac{-1}{2}\left\{.1\theta^2 + (\theta - x)^2\right\}\,,$$

ce qui équivaut à la loi $\theta|x \sim \mathcal{N}(10x/11, 10/11)$. L'espérance a posteriori de θ est donc $10x/11$. De même l'estimateur par défaut d'une transformation arbitraire $h(\theta)$ de θ sera donné par $\mathbb{E}^\pi[h(\theta)|x]$. ◀

[2]Le vecteur aléatoire X à valeurs dans $\mathbb{R}^p$ distribué suivant une loi normale d'espérance $\boldsymbol{\mu}$ et de structure de covariance $\boldsymbol{\Sigma}$, $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, admet comme densité de probabilité :

$$f_X(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left[-0.5(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]\,,$$

où A^{T} désigne la transposée de la matrice A.

2.3 Estimation par intervalles

La connaissance de la distribution a posteriori permet la détermination des *régions de confiance* sous la forme de régions de plus forte densité a posteriori (*Highest Posterior Density*, HPD), c'est-à-dire des régions de la forme

$$\{\boldsymbol{\theta}; \pi(\boldsymbol{\theta}|\mathbf{x}) \geq k\},$$

dans le cas multidimensionnel comme dans le cas unidimensionnel. La motivation conduisant à cette forme de région de confiance (traditionnellement renommée région de crédibilité dans le cas bayésien pour distinguer la couverture au niveau nominal α en $\boldsymbol{\theta}$ de la couverture au niveau nominal α en $\mathbf{x}$ même si l'utilisation est la même que pour les régions de confiance fréquentielles) est que ces régions sont de volume minimal à un niveau nominal donné.

Example 2.2 *(suite de l'Example 2.1) Dans le cas de la loi a posteriori* $\theta|x \sim \mathcal{N}_1(10x/11, 10/11)$*, la région de crédibilité au niveau nominal* α *est de la forme*

$$C_\alpha = \{\theta\,;\, \pi(\theta|x) \geq k\} = \{\theta\,;\, |\theta - 10x/11| \leq k'\}$$

avec k *et* k' *choisis de manière à ce que* $\pi(C_\alpha|\mathbf{x}) = \alpha$*, et est donc égale à l'intervalle*

$$(10x/11 - q_{1-\alpha/2}\sqrt{10/11}, 10x/11 + q_{1-\alpha/2}\sqrt{10/11})$$

où $q_{1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de la normale centrée réduite.* ◀

2.4 Tests et choix bayésien de modèles

Comme on peut le constater dans les deux exemples précédents, les inférences bayésiennes associées à l'estimation ponctuelle et à l'estimation par intervalle ne constituent pas des mofidications majeures des procédures classiques (même si l'influence de la loi a priori peut être déterminante comme on peut s'en rendre compte en faisant varier la variance a priori quand $\theta \sim \mathcal{N}_1(0, \tau^2)$). L'influence de la loi a priori finit par disparaître sous le poids des observations. A contrario, le domaine des tests d'hypothèses, qui relève intégralement d'une optique de décision de par le choix entre deux hypothèses incompatibles, offre un contraste marquant entre optique classique (selon Neyman-Pearson) et résolution bayésienne, principalement du fait des significations différentes associées aux notions d'hypothèse et de décision.

Pour décrire l'approche bayésienne, considérons deux modèles dénotés, $\mathfrak{M}_1$ et $\mathfrak{M}_2$, où $(i = 1, 2)$

$$\mathfrak{M}_i : \mathbf{x} \sim f_i(\cdot|\boldsymbol{\theta}_i), \boldsymbol{\theta}_i \in \Theta_i, \boldsymbol{\theta}_i \sim \pi_i(\boldsymbol{\theta}_i)\,,$$

$f_i(\mathbf{x}|\boldsymbol{\theta}_i)$ représentant la vraisemblance du paramètre $\boldsymbol{\theta}_i$ et $\pi_i(\boldsymbol{\theta}_i)$ la densité de la loi a priori du paramètre $\boldsymbol{\theta}_i$ associés au modèle $\mathfrak{M}_i$. Bien que cette présentation puisse sembler trop formalisée en termes de tests d'hypothèses, remarquons que l'un des modèles peut correspondre au cas d'une contrainte sur le paramètre $\boldsymbol{\theta}$ (comme $\theta_1 = 0$) et donc que cette représentation inclut bien les tests d'hypothèses. Alternativement, partant d'un modèle de réference, $\mathfrak{M}_0$,

avec un paramètre générique $\boldsymbol{\theta}$, les tests d'hypothèses correspondent aussi à des lois a priori dégénérées sur certaines composantes de $\boldsymbol{\theta}$.

Pour surmonter la difficulté à manipuler deux modèles bayésiens simultanément, il suffit de construire un méta-modèle englobant les deux cas. Pour ce faire, munissons l'ensemble des modèles d'une loi de probabilité a priori, l'indice du modèle devenant lui aussi une variable aléatoire, et notons $\mathbb{P}(\mathfrak{M}_1)$ et $\mathbb{P}(\mathfrak{M}_2)$ les probabilités a priori des deux modèles avec $\mathbb{P}(\mathfrak{M}_1) + \mathbb{P}(\mathfrak{M}_2) = 1$. Un choix de modèle bayésien est alors basé sur la loi a posteriori des différents modèles, c'est-à-dire sur les probabilités a posteriori, donc conditionnelles aux observations $\mathbf{x}$, $\mathbb{P}(\mathfrak{M}_1|\mathbf{x})$ et $\mathbb{P}(\mathfrak{M}_2|\mathbf{x})$, données par $(i = 1, 2)$

$$\mathbb{P}(\mathfrak{M}_i|\mathbf{x}) \propto \mathbb{P}(\mathfrak{M}_i) \int_{\Theta_i} f_i(\mathbf{x}|\theta_i)\pi_i(\theta_i)\,\mathrm{d}\theta_i\,.$$

On pourra s'en convaincre aisément en calculant la loi a posteriori jointe de $(\mathfrak{M}, \boldsymbol{\theta}_{\mathfrak{M}})$ où $\mathfrak{M}$ est l'indice aléatoire du modèle. En l'absence de fonction de coût particulière qui prendrait en compte les conséquences d'un mauvais choix, on choisira le modèle le plus vraisemblable, c'est à dire celui pour lequel $\mathbb{P}(\mathfrak{M}_i|\mathbf{x}) > 0.5$.

Notons que cette distribution a posteriori est sensible au choix des lois a priori des paramètres des modèles, $\pi_1(\boldsymbol{\theta}_1)$ et $\pi_2(\boldsymbol{\theta}_2)$, et surtout qu'elle dépend directement des a priori sur les différents modèles, $\mathbb{P}(\mathfrak{M}_1)$ et $\mathbb{P}(\mathfrak{M}_2)$. (De plus, crucialement, cette représentation impose l'utilisation de véritables lois de probabilités $\pi_i(\boldsymbol{\theta}_i)$, excluant l'emploi de mesures non-normalisables.) Un outil effaçant l'influence des a priori au niveau des modèles est le facteur de Bayes,

$$B_{12}(\mathbf{x}) = \frac{\mathbb{P}(\mathfrak{M}_1|\mathbf{x})}{\mathbb{P}(\mathfrak{M}_2|\mathbf{x})} \Big/ \frac{\mathbb{P}(\mathfrak{M}_1)}{\mathbb{P}(\mathfrak{M}_2)}\,,$$

qui se comporte comme un rapport de vraisemblance et élimine bien l'influence des poids a priori des deux modèles. Il peut donc être utilisé par rapport à une échelle absolue centrée en 1 pour prendre un décision sur le modèle suggéré par les données : un facteur de Bayes dont le logarithme $\log_{10} B_{12}(\mathbf{x})$ [3] est compris entre 0 et 0.5 donne une évidence (ou une confiance) faible en faveur de $\mathfrak{M}_1$, si $0.5 < \log_{10} B_{12}(\mathbf{x}) < 1$, l'évidence est substantielle, si $1 < \log_{10} B_{12}(\mathbf{x}) < 2$, l'évidence est forte, et si $\log_{10} B_{12}(\mathbf{x}) > 2$, l'évidence est décisive.

Example 2.3 *(suite de l'Example 2.1) Si l'on teste $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$, on compare le modèle $\mathfrak{M}_1$ où $x \sim \mathcal{N}_1(0, 1)$ au modèle $\mathfrak{M}_2$ où $x \sim \mathcal{N}_1(\theta, 1)$ et $\theta \sim \mathcal{N}_1(0, 10)$. Le facteur de Bayes $B_{12}(x)$ est donc le rapport des densités marginales*

$$B_{12}(x) = \frac{(1/\sqrt{2\pi})\exp(-x^2/2)}{(1/\sqrt{2\pi})\displaystyle\int_{\mathbb{R}} \frac{1}{\sqrt{20\pi}} \exp\left(\frac{-\theta^2}{20}\right) \exp\left(\frac{-(x-\theta)^2}{2}\right) \mathrm{d}\theta} = \frac{\exp(-x^2/2)}{\sqrt{1/11}\exp(-x^2/22)}$$

$$= \sqrt{11}\exp(-10x^2/22)\,.$$

[3] Rappelons que $\log_{10}$ correspond au logarithme décimal, il s'agit de la fonction réciproque de la fonction 10^x.

Le maximum de $B_{12}(x)$ *est atteint pour* $x = 0$ *et est favorable [au sense de l'échelle ci-dessus] à* $\mathfrak{M}_1$ *puisqu'il prend la valeur* $\sqrt{11}$*, de logarithme égal à* 1.2*. De plus,* $\log_{10} B_{12}(x)$ *vaut* 0 *pour* $x = 1.62$*, et* -1 *pour* $x = 2.19$*. On peut remarquer la différence avec les bornes classiques, puisque* $x = 1.62$ *correspond presque à un niveau de significativité de* 0.1 *et* $x = 2.19$ *à un niveau de significativité de* $.01$*. On rejettera donc plus difficilement (ou plus conservativement !) l'hypothèse nulle* $H_0 : \theta = 0$ *en utilisant une procédure bayésienne.* ◀

Signalons que la différence notée dans cet exemple tient à une interprétation radicalement différente de l'erreur de mesure : dans l'approche classique des tests, l'erreur traditionnellement de 5% correspond à la probabilité de rejeter à tort l'hypothèse nulle et découle donc d'un choix implicite d'une fonction de coût asymétrique. Dans l'approche bayésienne, le facteur de Bayes compare les deux probabilités, $\mathbb{P}(\mathfrak{M}_1|\mathbf{x})$ et $\mathbb{P}(\mathfrak{M}_2|\mathbf{x})$, et ne conclut que si elles différent suffisamment.

3 Sélection de variables de régression

Dans cette partie, nous nous intéressons à la sélection bayésienne de variables en régression linéaire gausienne. Nous en abordons de nombreux aspects afin de fournir au lecteur ou à la lectrice un guide précis et de lui démontrer ainsi l'applicabilité de l'approche bayésienne dans un contexte concret. Nous étudions successivement deux cas où les loi a priori sur les paramètres des modèles sont respectivement informatives et non informatives.

3.1 Introduction

La sélection bayésienne de variables en régression linéaire a été beaucoup étudiée et fournit un champ d'expérimentation fertile pour la comparaison des lois a priori et des procédures de décision. On peut citer parmi d'autres les articles suivants : Mitchell et Beauchamp (1988); George et McCulloch (1993); Geweke (1994); Chipman (1996); George et McCulloch (1997); Brown *et al.* (1998); Philips et Guttman (1998); George (2000); Fernandez *et al.* (2001); Kohn *et al.* (2001) et plus récemment Casella et Moreno (2006); Celeux *et al.* (2006); Cui et George (2008) et Liang *et al.* (2008).

Rappelons que la régression linéaire vise à expliquer ou à prédire les valeurs prises par une variable y via une combinaison linéaire, reposant sur des paramètres inconnus, d'un ensemble $\{x_1, \ldots, x_p\}$ de p variables dites explicatives.

Les performances d'un critère de choix de modèles dépendent de l'objectif choisi pour la modélisation par régression. Dans le cadre de la régression, les modèles en compétition font intervenir des sous-ensembles distincts de variables $\{x_1, \ldots, x_p\}$ et correspondent à des hypothèses différentes sur les variables véritablement "explicatives". Le choix des variables de régression peut être ainsi vu sous l'angle explicatif (quelle est l'influence de chacune des variables x_i sur y ?) ou sous l'angle prédictif (étant donné une nouvelle valeur du vecteur $\mathbf{x} = (x_1, \ldots, x_p)$ quelle est la valeur associée de y ?). Dans cette partie, le point de vue explicatif est privilégié. C'est alors typiquement un problème de choix de modèles que nous considérons ici (sous l'angle prédictif il est inefficace de rejeter des modèles).

Dans une première partie, nous traitons du cas où nous disposons d'informations a priori sur les paramètres du modèle de régression. Pour tous les modèles en compétition, c'est-à-dire pour tous les sous-ensembles possibles de variables explicatives, nous proposons d'utiliser des lois a priori dites de Zellner compatibles et nous en déduisons une procédure de sélection. Dans une deuxième partie, nous traitons du cas non informatif (où il n'y a plus d'information a priori sur ces paramètres). Puis, nous abordons les aspects algorithmiques associés à cette modélisation lorsque le nombre de variables est important et empêche l'énumération exhaustive de tous les modèles en compétition. Enfin, nous traitons un exemple de données réelles.

Nous considérons donc une variable aléatoire réelle y à expliquer et un ensemble $\{x_1, \ldots, x_p\}$ de p variables explicatives dites régresseurs. Nous faisons l'hypothèse que chaque modèle de régression avec les régresseurs $\{x_{i_1}, \ldots, x_{i_q}\}$, où $\{i_1, \ldots, i_q\} \subseteq \{1, \ldots, p\} \cup \emptyset$, est un modèle plausible pour expliquer la variable y. Nous disposons d'un échantillon de taille n pour estimer le modèle. La variable expliquée $\mathbf{y}$ forme un vecteur de dimension n. Nous notons $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ la matrice de dimension $n \times p$ dont les colonnes sont constituées des n observations des p variables explicatives. Le terme constant de la régression (voir (3) ci-dessous) fait partie de tous les modèles en compétition que nous considérons. Ils sont donc au nombre de 2^p. Nous utilisons une représentation hiérarchique pour les comparer et nous indiqns chaque modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$ à l'aide d'un paramètre binaire $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \Gamma = \{0, 1\}^{\otimes p}$ qui indique quelles sont les variables retenues par le modèle : $\gamma_i = 1$ si la variable x_i est sélectionnée et $\gamma_i = 0$ sinon. Nous notons $p_{\boldsymbol{\gamma}}$ le nombre de variables entrant dans le modèle $\boldsymbol{\gamma}$, soit $p_{\boldsymbol{\gamma}} = \mathbf{1}_p^{\mathrm{T}} \boldsymbol{\gamma}$ (où A^{T} désigne la transposée de la matrice A et $\mathbf{1}_q$ est le vecteur de dimension q dont toutes les composantes sont égales à 1).

Le modèle de régression linéaire gaussien correspondant à $\mathfrak{M}_{\boldsymbol{\gamma}}$ est tel que

$$\mathbf{y} | \boldsymbol{\gamma}, \alpha, \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2 \sim \mathcal{N}_n \left(\alpha \mathbf{1}_n + \mathbf{X}^{\boldsymbol{\gamma}} \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2 \, \mathbf{I}_n \right) , \tag{3}$$

où

- $\mathbf{I}_n$ désigne la matrice identité d'ordre n ;
- $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ désigne la loi gausienne n-dimensionnelle d'espérance $\boldsymbol{\mu}$ (vecteur de $\mathbb{R}^n$) et de structure de covariance $\boldsymbol{\Sigma}$ (matrice définie positive de dimension $n \times n$) ;
- $\mathbf{X}^{\boldsymbol{\gamma}}$ est la sous-matrice de $\mathbf{X}$ contenant seulement les colonnes pour lesquelles $\gamma_i = 1$;
- $\alpha \in \mathbb{R}$, $\boldsymbol{\beta}^{\boldsymbol{\gamma}} \in \mathbb{R}^{p_{\boldsymbol{\gamma}}}$ et $\sigma^2 > 0$ sont des paramètres inconnus.

Insistons sur le fait que les paramètres σ^2 et α sont considérés comme étant commun à tous les modèles : ils ne sont pas indexés par γ. Concernant σ^2, de très nombreux auteurs optent pour ce point de vue, c'est le cas par exemple des articles récents de Casella et Moreno (2006), Celeux *et al.* (2006), Cui et George (2008) et Liang *et al.* (2008). Cela revient à interpréter σ^2 comme la variance d'une erreur de mesure plutôt que comme une variance résiduelle (une fois éliminé l'effet des variables explicatives). Par ailleurs, notons que, même si la variance était résiduelle, par exemple par oubli d'une variable explicative, cette variance serait commune à tous les modèles et justifierait encore le choix d'un σ^2 commun. Par ailleurs, le fait que le paramètre α soit commune à tous les modèles revient à supposer que les variables explicatives sont centrées. Il s'agit d'une pratique courante en régression qui

n'a aucun effet sur les résultats obtenus et qui facilite l'interprétation des valeurs estimées des coefficients. Aussi, pour ces mêmes raisons liées à l'interprétation, nous supposerons que les variables explicatives sont centrées réduites : pour tout $i \in \{1, \ldots, p\}$, nous avons $\mathbf{x}_i^{\mathrm{T}} \mathbf{1}_n = 0$ et $\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i = n$.

3.2 Lois a priori informatives

3.2.1 Distributions a priori informatives compatibles

Tout d'abord, on peut s'interroger sur ce que peut être une loi informative réaliste pour le modèle de régression complet, soit celui contenant toutes les variables explicatives. Pour ce modèle, nous préconisons l'utilisation de la loi a priori informative avancée par Zellner (Zellner, 1986) qui fournit un bon compromis entre une loi informative conjuguée et une loi a priori diffuse donc peu informative. L'idée est de permettre au modélisateur ou à la modélisatrice d'introduire des informations sur la valeur des coefficients de la régression sans être obligé(e) de fournir des éléments a priori sur les corrélations entre ces coefficients, plus difficiles à spécifier. Lorsque $\boldsymbol{\gamma} = \boldsymbol{\gamma}^c = (1, \ldots, 1)$ correspondant au modèle complet, l'approche de Zellner pour le modèle (3) conduit à une loi a priori normale pour $\boldsymbol{\beta}^{\boldsymbol{\gamma}^c}$ conditionnelle au couple (α, σ^2),

$$\boldsymbol{\beta}^{\boldsymbol{\gamma}^c} | \boldsymbol{\gamma}^c, \alpha, \sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}^c}} \left(\tilde{\boldsymbol{\beta}}, g\sigma^2 (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \right) , \tag{4}$$

(notons que $\mathbf{X}^{\boldsymbol{\gamma}^c} = \mathbf{X}$) et à une loi a priori non informative pour le couple (α, σ^2),

$$\pi \left(\alpha, \sigma^2 | \boldsymbol{\gamma}^c \right) \propto \sigma^{-2} . \tag{5}$$

Notons que dans la proposition initiale de Zellner (Zellner, 1986) le paramètre α n'était pas commun à tous les modèles et n'avait pas une loi a priori plate. Il était considéré comme une composante du vecteur $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$. C'est l'approche adoptée par Celeux *et al.* (2006) et Marin et Robert (2007). Le fait que α soit commun à tous les modèles et muni d'une loi a priori plate assure une propriété importante d'invariance par changement d'échelle et par translation sur la variable explicative y de la procédure de choix de modèle. Par contre, dans ce cas, il n'est plus possible dans un contexte non informatif d'utiliser une loi impropre sur l'hyper-paramètre g. Dans cet article, nous ne considérons pas que l'hyper-paramètre g est une variable aléatoire mais lui attribuons une valeur. Ainsi, nous nous plaçons dans le cas assurant l'invariance, c'est-à-dire le cas où α est commun à tous les modèles et est muni d'une loi a priori plate. Ce débat passionnant dépasse largement l'objectif de cet article.

Par ailleurs, comme les paramètres α et σ^2 sont communs à tous les modèles, le fait que (5) ait une masse infinie ne pose pas de problème. Rappelons que, de manière générale, on ne peut pas utiliser des lois a priori impropres (de masse infinie) lorsque l'on met en oeuvre une procédure de choix de modèle. En effet, dans ce cas, les probabilités a posteriori des modèles dépendent de constantes arbitraires. Ce problème disparaît lorsque les lois impropres sont utilisées sur des paramètres communs à tous les modèles.

Cette loi (4)-(5) dépend des données à travers $\mathbf{X}$. Ce n'est pas un problème non plus dans la mesure où nous utilisons la vraisemblance conditionnelle du modèle, i.e. la loi de $\mathbf{y}$ sachant $\mathbf{X}$. (Si $\mathbf{X}$ contenait des variables endogènes, ce serait différent...) Le modélisateur choisit l'espérance a priori $\tilde{\boldsymbol{\beta}}$ et le facteur d'échelle g, où, comme nous le verrons dans les sections suivantes, g donne la quantité relative d'information a priori par rapport à celle portée par l'échantillon.

Dans un contexte de sélection bayésienne de variables, le danger est de favoriser de manière involontaire un ou des modèles à cause de lois a priori mal calibrées les unes par rapport aux autres. Il faut donc veiller à ce que les choix de lois a priori soient équitables. Certains auteurs (notamment Dawid et Lauritzen (2000)) ont proposé des formalisations pour donner un sens précis à cette notion d'équité. A leur suite, nous proposons de mesurer la compatibilité de deux lois a priori, $\pi_1(\theta_1)$ et $\pi_2(\theta_2)$, évoluant dans deux espaces différents, à l'aide de l'information de Kullback-Leibler entre les deux distributions marginales (ou prédictives) correspondantes, $f_1(y) = \int_{\Theta_1} f_1(y|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1$ et $f_2(y) = \int_{\Theta_2} f_2(y|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2$. Plus cette information est faible plus les lois a priori sont jugées équitables. Étudions cette approche dans le cas qui nous intéresse : soient $\mathcal{M}_1$ et $\mathcal{M}_2$ deux modèles bayésiens de régression linéaire gaussienne, de variance σ^2 et d'ordonnée à l'origine α communes comme dans (3). Munissons ces deux modèles des lois a priori informatives de Zellner définies ci-dessus.

$$\mathfrak{M}_1 : \mathbf{y}|\alpha, \boldsymbol{\beta}^1, \sigma^2 \sim \mathcal{N}_n(\alpha\mathbf{1}_n + \mathbf{X}^1\boldsymbol{\beta}^1, \sigma^2\mathbf{I}_n),\ \boldsymbol{\beta}^1|\alpha, \sigma^2 \sim \mathcal{N}_{p_1}\left(s_1, \sigma^2 g_1 \left\{\left(\mathbf{X}^1\right)^{\mathrm{T}}\mathbf{X}^1\right\}^{-1}\right),$$

$$\left(\alpha, \sigma^2\right) \sim \pi_1(\alpha, \sigma^2),$$

où $\mathbf{X}^1$ est une matrice fixée $(n \times p_1)$ de rang $p_1 \leq n$;

$$\mathfrak{M}_2 : \mathbf{y}|\alpha, \boldsymbol{\beta}^2, \sigma^2 \sim \mathcal{N}_n(\alpha\mathbf{1}_n + \mathbf{X}^2\boldsymbol{\beta}^2, \sigma^2\mathbf{I}_n),\ \boldsymbol{\beta}^2|\alpha, \sigma^2 \sim \mathcal{N}_{p_2}\left(s_2, \sigma^2 g_2 \left\{\left(\mathbf{X}^2\right)^{\mathrm{T}}\mathbf{X}_2\right\}^{-1}\right),$$

$$\left(\alpha, \sigma^2\right) \sim \pi_2(\alpha, \sigma^2),$$

où $\mathbf{X}^2$ est une matrice fixée $(n \times p_2)$ de rang $p_2 \leq n$.
Nous supposons sans perte de généralité que $\mathfrak{M}_2$ est un sous-modèle de $\mathfrak{M}_1$, c'est à dire que le sous-espace vectoriel engendré par les colonnes de $\mathbf{X}^2$ est inclus dans le sous-espace vectoriel engendré par les colonnes de $\mathbf{X}^1$. Nous pouvons alors chercher à déterminer les paramètres (s_2, g_2), en fonction des paramètres (s_1, g_1) fixés, équitables. Comme σ^2 et α sont des paramètres communs aux deux modèles, nous proposons de minimiser l'information de Kullback-Leibler entre les deux distributions marginales conditionnellement au couple (σ^2, α). Elle s'écrit

$$\int \log\left\{\frac{f_1(\mathbf{y}|\sigma^2, \alpha)}{f_2(\mathbf{y}|\sigma^2, \alpha)}\right\} f_1(\mathbf{y}|\sigma^2, \alpha)\mathrm{d}\mathbf{y}\,,$$

en prenant $\mathfrak{M}_1$ comme modèle de référence. Dans ce cas, on peut montrer que le minimum est atteint pour

$$s_2^* = \left\{\left(\mathbf{X}^2\right)^{\mathrm{T}}\mathbf{X}^2\right\}^{-1}\left(\mathbf{X}^2\right)^{\mathrm{T}}\mathbf{X}^1 s_1 \quad \text{et} \quad g_2^* = g_1\,. \tag{6}$$

Ce résultat est intuitif car $\mathbf{X}^2 s_2^*$ est la projection orthogonale de $\mathbf{X}^1 s_1$ sur l'espace engendré par les colonnes de $\mathbf{X}^2$.

Remarquons que cette proposition de lois a priori équitables correspond à celle donnée par Ibrahim et Laud (1994) et Ibrahim (1997), et qu'elle aurait également été obtenue avec d'autres propositions de lois a priori équitables. On doit aussi noter qu'indépendamment de leurs qualités ou de leurs limites, toutes les propositions faites pour définir des lois a priori compatibles induisent pour bien des modèles des difficultés de calcul qui limitent leur dissémination (le cas linéaire gaussien étant d'une certaine manière l'exception).

Finalement, nous suggérons de procéder de la manière suivante pour obtenir des modèles bayésiens de régression équitables :

1) définir une loi a priori sur le modèle complet ;
2) puis déduire les lois a priori des $2^p - 1$ modèles restants en prenant pour chaque modèle la loi a priori équitable par rapport au modèle complet précédemment introduit.

Notons $\mathbf{U}^{\boldsymbol{\gamma}} = \left\{(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}})\right\}^{-1}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}$ et $\mathbf{P}^{\boldsymbol{\gamma}} = \mathbf{X}^{\boldsymbol{\gamma}}\mathbf{U}^{\boldsymbol{\gamma}}$. Agissant suivant le principe énoncé ci-dessus, nous déduisons la loi a priori équitable du modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$ comme donnée par

$$\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \alpha, \sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}}}\left(\mathbf{U}^{\boldsymbol{\gamma}}\mathbf{X}\tilde{\boldsymbol{\beta}}, g\sigma^2\left\{(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}\right\}^{-1}\right),$$

si la loi du modèle complet est

$$\boldsymbol{\beta}^{\boldsymbol{\gamma}^c}|\boldsymbol{\gamma}^c, \alpha, \sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}^c}}\left(\tilde{\boldsymbol{\beta}}, g\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right).$$

La modélisation du paramètre de sélection $\boldsymbol{\gamma}$ se situe à un niveau hiérarchique plus élevé. Nous utilisons la loi a priori suivante

$$\pi(\boldsymbol{\gamma}) = \prod_{i=1}^{p}\left(\tau_i^{\gamma_i}(1-\tau_i)^{1-\gamma_i}\right),$$

où $0 \leq \tau_i \leq 1$ correspond à la probabilité a priori que la variable i soit présente dans le modèle. Typiquement, lorsque aucune information a priori n'est disponible, on pose $\tau_1 = \ldots = \tau_p = 1/2$. Cela conduit à une loi a priori uniforme $\pi(\boldsymbol{\gamma}) = 2^{-p}$ sur les modèles, hypothèse toujours faite dans la suite. Une alternative consiste à supposer que $\tau_1 = \ldots = \tau_p = \tau$ et à inclure un niveau hiérarchique supplémentaire en supposant que τ est une variable aléatoire distribuée suivant une loi uniforme sur l'intervalle $[0, 1]$.

3.2.2 Lois a posteriori de α, $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ et σ^2

Dans ce paragraphe, nous montrons que l'on peut expliciter les lois a posteriori des paramètres α, $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ et σ^2. Malheureusement, la mise en évidence de ces lois nécessite quelques développements matriciels (sans difficultés). Nous avons comme vraisemblance du modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$ tel que $p_{\boldsymbol{\gamma}} \neq 0$

$$f(\mathbf{y}|\alpha, \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}\right)^{\mathrm{T}}\left(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}\right)\right).$$

Notons $\bar{\mathbf{y}} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ la moyenne empirique des composantes du vecteur $\mathbf{y}$, il vient alors que la norme se décompose par le théorème de Pythagore :

$$\begin{aligned}
& (\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}) \\
= \; & (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}} + (\bar{\mathbf{y}} - \alpha)\mathbf{1}_n)^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}} + (\bar{\mathbf{y}} - \alpha)\mathbf{1}_n) \\
= \; & (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}) + 2(\bar{\mathbf{y}} - \alpha)\mathbf{1}_n^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}) + n(\bar{\mathbf{y}} - \alpha)^2 \\
= \; & (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}) + n(\bar{\mathbf{y}} - \alpha)^2 .
\end{aligned}$$

En effet, $\mathbf{1}_n^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}) = (n\bar{\mathbf{y}} - n\bar{\mathbf{y}}) = 0$. Ainsi,

$$\begin{aligned}
f(\mathbf{y}|\alpha, \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}) \;=\; & \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}) \right) \times \\
& \exp\left(-\frac{n}{2\sigma^2} (\bar{\mathbf{y}} - \alpha)^2 \right) .
\end{aligned}$$

Par ailleurs, la loi a priori s'écrit sous la forme

$$\begin{aligned}
\pi\left(\alpha, \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2 | \boldsymbol{\gamma}\right) \;\propto\; & (\sigma^2)^{-p_{\boldsymbol{\gamma}}/2} \exp\left(-\frac{1}{2g\sigma^2} \left\{ (\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} \boldsymbol{\beta}^{\boldsymbol{\gamma}} - 2 (\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} \mathbf{U}^{\boldsymbol{\gamma}} \mathbf{X} \tilde{\boldsymbol{\beta}} \right\} \right) \times \\
& \exp\left(-\frac{1}{2g\sigma^2} \tilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{X} \mathbf{P}^{\boldsymbol{\gamma}} \mathbf{X} \tilde{\boldsymbol{\beta}} \right) \times \sigma^{-2} .
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\pi\left(\alpha, \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2 | \boldsymbol{\gamma}, \mathbf{y}\right) \;\propto\; & (\sigma^2)^{-n/2 - p_{\boldsymbol{\gamma}}/2 - 1} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}) \right) \times \\
& \exp\left(-\frac{n}{2\sigma^2} (\bar{\mathbf{y}} - \alpha)^2 \right) \times \exp\left(-\frac{1}{2g\sigma^2} \tilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{X} \mathbf{P}^{\boldsymbol{\gamma}} \mathbf{X} \tilde{\boldsymbol{\beta}} \right) \times \\
& \exp\left(-\frac{1}{2g\sigma^2} \left\{ (\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} \boldsymbol{\beta}^{\boldsymbol{\gamma}} - 2 (\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} \mathbf{U}^{\boldsymbol{\gamma}} \mathbf{X} \tilde{\boldsymbol{\beta}} \right\} \right) .
\end{aligned}$$

Et donc, comme $\mathbf{1}_n^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} = \mathbf{0}_p$,

$$\begin{aligned}
\pi\left(\alpha, \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2 | \boldsymbol{\gamma}, \mathbf{y}\right) \;\propto\; & (\sigma^2)^{-n/2 - p_{\boldsymbol{\gamma}}/2 - 1} \exp\left(-\frac{1}{2\sigma^2} \left\{ (\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} \boldsymbol{\beta}^{\boldsymbol{\gamma}} - 2 \mathbf{y}^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} \boldsymbol{\beta}^{\boldsymbol{\gamma}} \right\} \right) \times \\
& \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n) \right) \times \\
& \exp\left(-\frac{n}{2\sigma^2} (\bar{\mathbf{y}} - \alpha)^2 \right) \times \exp\left(-\frac{1}{2g\sigma^2} \tilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{X} \mathbf{P}^{\boldsymbol{\gamma}} \mathbf{X} \tilde{\boldsymbol{\beta}} \right) \times \\
& \exp\left(-\frac{1}{2g\sigma^2} \left\{ (\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} \boldsymbol{\beta}^{\boldsymbol{\gamma}} - 2 (\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}} (\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}} \mathbf{X}^{\boldsymbol{\gamma}} \mathbf{U}^{\boldsymbol{\gamma}} \mathbf{X} \tilde{\boldsymbol{\beta}} \right\} \right) .
\end{aligned}$$

Nous en déduisons aisément que conditionnellement à $\boldsymbol{\gamma}$, $\mathbf{y}$ et σ^2 les variables α et $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ sont indépendantes et telles que

$$\alpha|\boldsymbol{\gamma},\mathbf{y},\sigma^2 \sim \mathcal{N}_1\left(\bar{\mathbf{y}},\sigma^2/n\right)\,,$$

$$\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y},\sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}}}\left(\frac{g}{g+1}\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}+\mathbf{U}^{\boldsymbol{\gamma}}\mathbf{X}\tilde{\beta}/g\right),\frac{\sigma^2 g}{g+1}\left\{(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}\right\}^{-1}\right)$$

où $\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}=\left\{(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}\right\}^{-1}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{y}$ correspond à l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$. L'indépendance entre α et $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ découle du fait que $\mathbf{X}$ (et donc $\mathbf{X}^{\boldsymbol{\gamma}}$) est centrée et que la modélisation a priori suppose α et $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ indépendants.

Par ailleurs, pour le modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$ tel que $p_{\boldsymbol{\gamma}}\neq 0$ la loi a posteriori de σ^2 est

$$\sigma^2|\boldsymbol{\gamma},\mathbf{y}\sim I\mathcal{G}\left[(n-1)/2,\left\{(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)-g/(g+1)\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{y}+\right.\right.$$

$$\left.\left.\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{X}\tilde{\boldsymbol{\beta}}/(g+1)-2\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{X}\tilde{\boldsymbol{\beta}}/(g+1)\right\}\Big/2\right]\,,$$

où $I\mathcal{G}(a,b)$ est une loi inverse gamma[4] de moyenne $b/(a-1)$.

On remarque que

$$\kappa_{\boldsymbol{\gamma}}=(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)-g/(g+1)\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{y}+\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{X}\tilde{\boldsymbol{\beta}}/(g+1)-2\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{X}\tilde{\boldsymbol{\beta}}/(g+1)=$$

$$s^2_{\boldsymbol{\gamma}}+(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})/(g+1)\,,$$

où $s^2_{\boldsymbol{\gamma}}=(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}^{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}^{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})$ correspond à la somme des carrés résiduels et $\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}=\mathbf{U}^{\boldsymbol{\gamma}}\mathbf{X}\tilde{\boldsymbol{\beta}}$ est l'espérance a priori compatible de $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$.

Pour le modèle $\mathfrak{M}_{\boldsymbol{\gamma}^0}$ où $\boldsymbol{\gamma}^0=(0,\ldots,0)$, par des calculs similaires, nous obtenons que

$$\alpha|\boldsymbol{\gamma}^0,\mathbf{y},\sigma^2\sim\mathcal{N}_1\left(\bar{\mathbf{y}},\sigma^2/n\right)\,,$$

$$\sigma^2|\boldsymbol{\gamma}^0,\mathbf{y}\sim I\mathcal{G}\left[(n-1)/2,(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)/2\right]\,.$$

Quelle que soit la fonction de perte considérée, l'estimateur bayésien de α sera le même pour tous les modèles. En effet la loi a posteriori de α sachant $\boldsymbol{\gamma}$, σ^2 et $\mathbf{y}$ ne dépend pas de $\boldsymbol{\gamma}$. Ceci est parfaitement cohérent avec l'hypothèse selon laquelle le paramètre α est commun à tous les modèles. Aussi, l'estimateur de α correspondant à la fonction de perte quadratique est égal à la moyenne empirique des composantes du vecteur $\mathbf{y}$. Ce résultat extrêmement intuitif vient du fait que[5]

$$\mathbb{E}^{\pi}\left[\alpha|\boldsymbol{\gamma},\mathbf{y}\right]=\mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\left(\alpha|\boldsymbol{\gamma},\mathbf{y},\sigma^2\right)|\boldsymbol{\gamma},\mathbf{y}\right]=\mathbb{E}^{\pi}\left[\bar{\mathbf{y}}|\boldsymbol{\gamma},\mathbf{y}\right]=\bar{\mathbf{y}}\,.$$

[4]La variable aléatoire X à valeurs réelles distribuée suivant une loi inverse gamma de paramètres $\alpha>0$ et $\beta>0$ admet comme densité de probabilité :

$$f_X(x|\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{-\alpha-1}\exp(-\beta/x)\mathbb{I}_{]0,+\infty]}(x)\,.$$

[5]$\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right]=\mathbb{E}\left[X\right]$.

On peut aussi calculer la variance a posteriori de α, en effet[6]

$$\mathbb{V}^{\pi}(\alpha|\boldsymbol{\gamma},\mathbf{y}) = \mathbb{V}(\bar{\mathbf{y}}|\boldsymbol{\gamma},\mathbf{y}) + \mathbb{E}\left[\frac{\sigma^2}{n}\middle|\boldsymbol{\gamma},\mathbf{y}\right] = \kappa_{\boldsymbol{\gamma}}/(n(n-3))\,.$$

Par ailleurs, nous remarquons que

$$\pi(\alpha,\sigma^2|\boldsymbol{\gamma},\mathbf{y}) \propto (\sigma^2)^{-1/2}\exp\left(-\frac{n}{2\sigma^2}(\alpha-\bar{\mathbf{y}})^2\right)(\sigma^2)^{-(n-1)/2-1}\exp\left(-\frac{1}{2\sigma^2}\kappa_{\boldsymbol{\gamma}}\right)\,,$$

$$\Longrightarrow \pi(\alpha|\boldsymbol{\gamma},\mathbf{y}) \propto \left[1+\frac{n(\alpha-\bar{\mathbf{y}})^2}{\kappa_{\boldsymbol{\gamma}}}\right]^{-n/2}\,.$$

Ainsi, la loi marginale a posteriori de α, à savoir la loi de α sachant $\boldsymbol{\gamma}$ et $\mathbf{y}$, est une loi de Student à $n-1$ degrés de libertés de paramètres de décentrage $\bar{\mathbf{y}}$ et d'échelle $\kappa_{\boldsymbol{\gamma}}/(n(n-1))$[7]. Cette loi marginale est notamment utilisée pour le calcul de la région de crédibilité de plus forte densité a posteriori.

Le paramètre $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ n'a de sens que pour le modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$ tel que $p_{\boldsymbol{\gamma}} \neq 0$. Aussi contrairement à α, l'estimateur de $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ va dépendre de $\boldsymbol{\gamma}$. Celui qui minimise la perte quadratique est donné par

$$\mathbb{E}^{\pi}\left[\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y}\right] = \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\left(\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y},\sigma^2\right)\middle|\boldsymbol{\gamma},\mathbf{y}\right] = \mathbb{E}^{\pi}\left[\frac{g}{g+1}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}+\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}/g)\middle|\boldsymbol{\gamma},\mathbf{y}\right] =$$

$$\frac{g}{g+1}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}+\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}/g)\,.$$

Si $g=1$, c'est-à-dire si l'information a priori a le même poids que l'échantillon, l'estimation bayésienne est la moyenne entre l'estimation des moindres carrés et l'espérance a priori. Plus g est grand, plus l'information a priori est faible et plus l'estimation bayésienne se rapproche de l'estimation des moindres carrés. Lorsque g tend vers l'infini, la moyenne a posteriori converge vers $\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}$. On peut aussi calculer la variance a posteriori de $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$, en effet

$$\mathbb{V}(\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y}) = \mathbb{V}\left[\frac{g}{g+1}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}+\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}/g)|\boldsymbol{\gamma},\mathbf{y}\right] + \mathbb{E}\left[\frac{g\sigma^2}{g+1}((\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}})^{-1}\right]$$

$$= \frac{\kappa_{\boldsymbol{\gamma}} g}{(g+1)(n-3)}((\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}})^{-1}\,.$$

[6] $\mathbb{V}(X|Y) = \mathbb{E}[(X-\mathbb{E}(X|Y))^2\big|Y]$ et $\mathbb{V}(X) = \mathbb{V}[\mathbb{E}(X|Y)] + \mathbb{E}[\mathbb{V}(X|Y)]$.

[7] Le vecteur aléatoire X à valeurs dans $\mathbb{R}^p$ distribué suivant une loi Student à d degrés de liberté de paramètre de décentrage $\boldsymbol{\theta}$ et d'échelle Σ admet comme densité de probabilité :

$$f_X(\mathbf{x}|\boldsymbol{\theta},\Sigma) \propto \left[1+\frac{(\mathbf{x}-\boldsymbol{\theta})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\theta})}{d}\right]^{-(d+p)/2}\,.$$

Par ailleurs, nous remarquons que

$$\begin{aligned}\pi(\boldsymbol{\beta}^{\boldsymbol{\gamma}},\sigma^2|\boldsymbol{\gamma},\mathbf{y}) &\propto (\sigma^2)^{-p_{\boldsymbol{\gamma}}/2}\exp\left(-\frac{g+1}{2g\sigma^2}\big\{\boldsymbol{\beta}^{\boldsymbol{\gamma}}-\mathbb{E}^{\pi}\left[\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y}\right]\big\}^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}\big\{\boldsymbol{\beta}^{\boldsymbol{\gamma}}-\mathbb{E}^{\pi}\left[\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y}\right]\big\}\right)\\ &\quad\times(\sigma^2)^{-(n-1)/2-1}\exp\left(-\frac{1}{2\sigma^2}\kappa_{\boldsymbol{\gamma}}\right),\end{aligned}$$

$$\Longrightarrow \pi(\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y})\propto\left[1+\frac{g+1}{g\kappa_{\boldsymbol{\gamma}}}\big\{\boldsymbol{\beta}^{\boldsymbol{\gamma}}-\mathbb{E}^{\pi}\left[\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y}\right]\big\}^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}\big\{\boldsymbol{\beta}^{\boldsymbol{\gamma}}-\mathbb{E}^{\pi}\left[\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbf{y}\right]\big\}\right].$$

Ainsi, pour tout $\boldsymbol{\gamma}$ tel que $p_{\boldsymbol{\gamma}}\neq 0$, la loi marginale a posteriori de $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$, à savoir la loi de $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ sachant $\boldsymbol{\gamma}$ et $\mathbf{y}$, est une loi de Student à $n-1$ degrés de libertés de paramètres de décentrage $\frac{g}{g+1}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}+\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}/g)$ et d'échelle $\frac{g\kappa_{\boldsymbol{\gamma}}}{(g+1)(n-1)}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}})^{-1}$. Comme pour α, cette loi marginale peut être utilisée pour le calcul de régions de crédibilité de plus forte densité a posteriori.

Enfin, pour le modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$ tel que $p_{\boldsymbol{\gamma}}\neq 0$, l'estimateur de σ^2 minimisant le risque quadratique est donné par

$$\mathbb{E}\left[\sigma^2|\boldsymbol{\gamma},\mathbf{y}\right]=\frac{\kappa_{\boldsymbol{\gamma}}}{n-3}=\frac{s^2_{\boldsymbol{\gamma}}+(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})/(g+1)}{n-3}.$$

Aussi, pour le modèle $\mathfrak{M}_{\boldsymbol{\gamma}^0}$ ($\boldsymbol{\gamma}^0=(0,\ldots,0)$), nous avons

$$\mathbb{E}\left[\sigma^2|\boldsymbol{\gamma}^0,\mathbf{y}\right]=\frac{(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)}{n-3}.$$

3.2.3 Loi a posteriori de $\boldsymbol{\gamma}$

Dans ce paragraphe, nous mettons en lumière le fait que la loi a posteriori de $\boldsymbol{\gamma}$ est disponible de manière explicite ce qui autorise directement la comparaison de modèles, c'est-à-dire la sélection de variables. Pour tout $\boldsymbol{\gamma}$ tel que $p_{\boldsymbol{\gamma}}\neq 0$, nous avons

$$f(\mathbf{y}|\boldsymbol{\gamma})=\int\left(\int\int f(\mathbf{y}|\boldsymbol{\gamma},\alpha,\boldsymbol{\beta}^{\boldsymbol{\gamma}},\sigma^2)\pi(\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\alpha,\sigma^2)\pi(\sigma^2,\alpha|\boldsymbol{\gamma})\mathrm{d}\alpha\mathrm{d}\boldsymbol{\beta}^{\boldsymbol{\gamma}}\right)\mathrm{d}\sigma^2.$$

Aussi,

$$\begin{aligned}f(\mathbf{y}|\boldsymbol{\gamma},\alpha,\boldsymbol{\beta}^{\boldsymbol{\gamma}},\sigma^2)\pi(\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\alpha,\sigma^2) &= \frac{\left|(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}\right|^{1/2}}{(2\pi\sigma^2)^{(n+p_{\boldsymbol{\gamma}})/2}g^{p_{\boldsymbol{\gamma}}/2}}\exp\left(-\frac{n}{2\sigma^2}(\alpha-\bar{\mathbf{y}})^2\right)\\ &\quad\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}})\right)\\ &\quad\exp\left(-\frac{1}{2g\sigma^2}(\boldsymbol{\beta}^{\boldsymbol{\gamma}}-\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}(\boldsymbol{\beta}^{\boldsymbol{\gamma}}-\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})\right).\end{aligned}$$

Dans la mesure où $\pi(\alpha,\sigma^2|\boldsymbol{\gamma}) = \delta\sigma^{-2}$ (où δ est une constante arbitraire), par des calculs élémentaires mais somme toute assez lourds, il vient que pour tout $\boldsymbol{\gamma}$ tel que $p_{\boldsymbol{\gamma}} \neq 0$

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\gamma}) &= \delta n^{-1/2}(g+1)^{-p_\gamma/2}(2\pi)^{-(n-1)/2}\int(\sigma^2)^{-(n-1)/2-1}\exp\left(-\frac{1}{2\sigma^2}\kappa_{\boldsymbol{\gamma}}\right)\mathrm{d}\sigma^2 \\ &= \frac{\delta\Gamma((n-1)/2)}{\pi^{(n-1)/2}n^{1/2}}(g+1)^{-p_\gamma/2}\left[s^2_{\boldsymbol{\gamma}}+(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})/(g+1)\right]^{-(n-1)/2}. \end{aligned}$$

Par ailleurs, pour $\boldsymbol{\gamma}^0 = (0,\ldots,0)$, nous avons

$$f(\mathbf{y}|\boldsymbol{\gamma}^0) = \int\left(\int f(\mathbf{y}|\boldsymbol{\gamma}^0,\alpha,\sigma^2)\pi(\alpha,\sigma^2|\boldsymbol{\gamma}^0)\mathrm{d}\alpha\right)\mathrm{d}\sigma^2\,.$$

Aussi,

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\gamma}^0,\alpha,\sigma^2)\pi(\alpha,\sigma^2|\boldsymbol{\gamma}^0) &= \frac{(\sigma^2)^{-1}}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)\right) \\ &= \frac{(\sigma^2)^{-n/2-1}}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)\right)\times \\ &\quad \exp\left(-\frac{n}{2\sigma^2}(\alpha-\bar{\mathbf{y}})^2\right). \end{aligned}$$

Par des calculs élémentaires, il vient que

$$f(\mathbf{y}|\boldsymbol{\gamma}^0) = \frac{\delta\Gamma((n-1)/2)}{\pi^{(n-1)/2}n^{1/2}}\left[(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)\right]^{-(n-1)/2}.$$

Ainsi, pour tout $\boldsymbol{\gamma}$ tel que $p_{\boldsymbol{\gamma}} \neq 0$,

$$\pi(\boldsymbol{\gamma}|\mathbf{y}) \propto (g+1)^{-p_\gamma/2}\left[s^2_{\boldsymbol{\gamma}}+(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})/(g+1)\right]^{-(n-1)/2},$$

et

$$\pi(\boldsymbol{\gamma}^0|\mathbf{y}) \propto \left[(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n)\right]^{-(n-1)/2}.$$

D'après les calculs précédents, le facteur de Bayes du modèle $\mathfrak{M}_{\boldsymbol{\gamma}^1}$ (tel que $p_{\boldsymbol{\gamma}^1} \neq 0$) contre le modèle $\mathfrak{M}_{\boldsymbol{\gamma}^2}$ (tel que $p_{\boldsymbol{\gamma}^2} \neq 0$) est donné par

$$B_{12}(\mathbf{y}) = \frac{(g+1)^{-p_{\gamma^1}/2}\left[s^2_{\boldsymbol{\gamma}^1}+(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}^1}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}^1})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}^1})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}^1}(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}^1}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}^1})/(g+1)\right]^{-(n-1)/2}}{(g+1)^{-p_{\gamma^2}/2}\left[s^2_{\boldsymbol{\gamma}^2}+(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}^2}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}^2})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}^2})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}^2}(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}^2}-\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}^2})/(g+1)\right]^{-(n-1)/2}}.$$

3.3 Lois a priori non informatives

Dans un contexte non informatif, nous proposons d'utiliser les mêmes lois a priori que précédemment avec $\tilde{\boldsymbol{\beta}} = \mathbf{0}_p$ et $g = n$ (où $\mathbf{0}_p$ désigne le vecteur de dimension p dont toutes les composantes sont nulles). Dans ce cas, l'hyper-paramètre g joue le rôle d'un paramètre de rétrécissement. Divers choix sont proposés pour fixer le niveau rétrécissement et ainsi la valeur de g (Fernandez *et al.*, 2001). Récemment, des auteurs ont proposé des approches hiérarchiques en considérant g comme une variable aléatoire (Celeux *et al.*, 2006; Marin et Robert, 2007; Cui et George, 2008; Liang *et al.*, 2008). Diverses lois a priori sur g sont alors possibles. Ici, nous privilégions la valeur $g = n$. Ce choix, en relation étroite avec le critère *Bayesian Information Criterion* (BIC) (Kass et Raftery, 1995), correspond à donner globalement à la loi a priori le poids d'une observation.

Pour $g = n$, l'estimateur de $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ est

$$\frac{n}{n+1}\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}\,.$$

Ainsi, le coefficient de rétrécissement est égal à $\dfrac{n}{n+1}$.

Par ailleurs, pour le modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$ tel que $p_{\boldsymbol{\gamma}} \neq 0$, nous avons

$$s^2_{\boldsymbol{\gamma}} + \frac{(\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}}{n+1} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - n\bar{\mathbf{y}} - \mathbf{y}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{y} + \frac{\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{y}}{n+1} = \mathrm{STC} - \mathrm{SCEM}_{\boldsymbol{\gamma}} + \frac{\mathrm{SCEM}_{\boldsymbol{\gamma}}}{n+1}\,,$$

où $\mathrm{STC} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - n\bar{\mathbf{y}} = (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)$ est la somme totale des carrés et $\mathrm{SCEM}_{\boldsymbol{\gamma}} = \mathbf{y}^{\mathrm{T}}\mathbf{P}^{\boldsymbol{\gamma}}\mathbf{y}$ la somme des carrés expliqués par le modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$. Ainsi, pour le modèle $\mathfrak{M}_{\boldsymbol{\gamma}}$ tel que $p_{\boldsymbol{\gamma}} \neq 0$,

$$\begin{aligned}
\pi(\boldsymbol{\gamma}|\mathbf{y}) \;\propto\; & (n+1)^{-p_\gamma/2}\left[s^2_{\boldsymbol{\gamma}} + (\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{X}^{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}^{\boldsymbol{\gamma}}(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}})/(n+1)\right]^{-(n-1)/2}\,, \\
& (n+1)^{(n-1-p_\gamma)/2}\left[(n+1)(\mathrm{STC} - \mathrm{SCEM}_{\boldsymbol{\gamma}}) + \mathrm{SCEM}_{\boldsymbol{\gamma}}\right]^{-(n-1)/2}\,, \\
& (n+1)^{(n-1-p_\gamma)/2}\mathrm{STC}^{-(n-1)/2}\left[(n+1)(1 - R^2_{\boldsymbol{\gamma}}) + R^2_{\boldsymbol{\gamma}}\right]^{-(n-1)/2}\,, \\
& (n+1)^{(n-1-p_\gamma)/2}\mathrm{STC}^{-(n-1)/2}\left[1 + n(1 - R^2_{\boldsymbol{\gamma}})\right]^{-(n-1)/2}\,,
\end{aligned}$$

où $R^2_{\boldsymbol{\gamma}} = \dfrac{\mathrm{SCEM}_{\boldsymbol{\gamma}}}{\mathrm{STC}}$ est le coefficient de détermination du modèle $\boldsymbol{\gamma}$.
(Rappelons que $\pi(\boldsymbol{\gamma}^0|\mathbf{y}) \propto \mathrm{STC}^{-(n-1)/2}$.)

3.4 Approximation par échantillonnage de Gibbs

Lorsque le nombre de variables p est grand, typiquement pour $p > 25$, il est impossible de réaliser une sélection exhaustive des variables explicatives puisque elle impliquerait 2^p termes. Dans un cadre fréquentiel, des procédures pas à pas avec remise en cause (*step-wise*) ascendante ou descendante sont utilisées (Miller, 1990). Dans un cadre bayésien, il

est impossible de calculer l'intégralité des probabilités a posteriori des 2^p modèles. Un algorithme de simulation de Monte-Carlo permettant d'estimer $\pi(\boldsymbol{\gamma}|\mathbf{y})$ est cependant disponible pour approcher la sélection bayésienne des jeux de variables explicatives les plus probables. Nous ne détaillerons pas ici les principes ni les méthodes de Monte-Carlo en statistique, renvoyant la lectrice ou le lecteur à Robert et Casella (2004) ou à Marin et Robert (2007). Rappelons cependant que l'utilisation de techniques fondées sur des chaînes de Markov, techniques dites MCMC, permet la simulation suivant des lois a posteriori complexes, en évitant la simulation directe. Deux méthodes particulièrement efficaces sont les algorithmes de Metropolis-Hastings et les échantillonneurs de Gibbs. La méthode de l'échantillonneur de Gibbs repose sur la simulation successive suivant les diverses lois conditionnelles associées à la loi a posteriori cible. Cette décomposition en lois simples, bien que lente, permet de faire face à des problèmes de simulation dans des espaces dont les dimensions sont arbitrairement grandes en les ramenant à de nombreuses problèmes de petite dimension.

Notons $\boldsymbol{\gamma}_{-i}$ le vecteur $(\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p)$. D'après la règle de Bayes, la distribution de γ_i sachant $\mathbf{y}$ et γ_{-i} est telle que

$$\pi(\gamma_i|\mathbf{y}, \boldsymbol{\gamma}_{-i}) = \frac{\pi(\gamma|\mathbf{y})}{\pi(\boldsymbol{\gamma}_{-i}|\mathbf{y})}\,,$$

et donc

$$\pi(\gamma_i|\mathbf{y}, \boldsymbol{\gamma}_{-i}) \propto \pi(\boldsymbol{\gamma}|\mathbf{y})\,. \tag{7}$$

Ainsi, comme γ_i est binaire, la distribution conditionnelle $\pi(\gamma_i|\mathbf{y}, \boldsymbol{\gamma}_{-i})$ est obtenue par le calcul normalisé de $\pi(\boldsymbol{\gamma}|\mathbf{y})$ pour $\gamma_i = 0$ et $\gamma_i = 1$. Il est donc possible d'utiliser ici l'échantillonneur de Gibbs qui se base sur la simulation successive des γ_i suivant leur loi conditionnelle sachant $\boldsymbol{\gamma}_{-i}$ et $\mathbf{y}$, puisque $\pi(\boldsymbol{\gamma}|\mathbf{y})$ est disponible sous forme explicite.

Échantillonnage de Gibbs

- Itération 0 : tirage de $\boldsymbol{\gamma}^{\text{init}}$ selon la distribution uniforme sur Γ ou sélection d'un $\boldsymbol{\gamma}^{\text{init}}$ arbitraire, ou encore sélection du $\boldsymbol{\gamma}^{\text{init}}$ maximisant $\pi(\boldsymbol{\gamma}|\mathbf{y})$ (dont on connaît l'expression à une constante près) par procédures pas à pas[8] ;
- Itération $t = 1, \ldots, T$: pour $i = 1, \ldots, p$, tirage de γ_i^t selon

$$\pi(\gamma_i|\mathbf{y}, \gamma_1^t, \ldots, \gamma_{i-1}^t, \gamma_{i+1}^{t-1}, \ldots, \gamma_p^{t-1})\,.$$

L'échantillonnage de Gibbs présenté ci-dessus est décomposé en deux phases. La première phase correspond à une période d'échauffement de longueur T_0 à la fin de laquelle, on considère que la chaîne de Markov générée a atteint son régime stationnaire, c'est-à-dire produit marginalement des simulations suivant $\pi(\boldsymbol{\gamma}|\mathbf{y})$. La deuxième phase, démarrant au temps T_0, vise à approximer cette mesure stationnaire $\pi(\boldsymbol{\gamma}|\mathbf{y})$. Les γ_i générés durant cette phase sont conservés pour mener à bien l'approximation. Pour un modèle $\boldsymbol{\gamma}$ donné, l'estimateur de $\pi(\boldsymbol{\gamma}|\mathbf{y})$ déduit de l'échantillonnage de Gibbs est la moyenne empirique

$$\hat{\pi}(\boldsymbol{\gamma}|\mathbf{y})^{GIBBS} = \sum_{t=T_0+1}^{T} \mathbb{I}_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^t} \Big/ (T - T_0)\,, \tag{8}$$

[8]Cette procédure d'un coût très modéré est sans aucun doute la plus efficace.

et celui de $\mathbb{P}(\gamma_i = 1|\mathbf{y})$ s'écrit

$$\hat{\mathbb{P}}(\gamma_i = 1|\mathbf{y})^{GIBBS} = \sum_{t=T_0+1}^{T} \mathbb{I}_{\gamma_i=\gamma_i^t} \Big/ (T - T_0)\,. \tag{9}$$

Cette seconde approximation de la présence ou non de la i-éme variable explicative est utile pour déterminer les variables à sélectionner dans le modèle car, si p est grand, le nombre de modèles ignorés par l'algorithme est trop grand pour garantir que le modèle le plus probable est visité. Notons de plus que, pour chaque modèle visité, il est possible de calculer la valeur de $\pi(\boldsymbol{\gamma}|\mathbf{y})$ (à une constante près), donc de garder la trace du modèle le plus probable rencontré lors des simulation, et de comparer le modèle retenu à celui incluant les variables les plus probables.

Pratiquement, l'échantillonnage de Gibbs permet de ne pas considérer tous les modèles en ce que les modèles visités par Gibbs sont de forte probabilité. Les modèles de faible probabilité ne sont que rarement visités mais il peut demeurer des modèles de forte probabilité qui restent ignorés par l'échantillonneur de Gibbs parce qu'isolés des autres modèles de fortes probabilités. On peut se prémunir contre cette difficulté en répétant l'échantillonnage plusieurs fois et en comparant les valeurs de $\pi(\boldsymbol{\gamma}|\mathbf{y})$ obtenues. Ainsi, le critère de visite semble plus exhaustif que les critères utilisés par les techniques pas à pas.

Nous illustrons les performances de l'échantillonnage de Gibbs sur les données réelles présentées dans le paragraphe suivant.

3.5 Chenilles processionnaires

Ces données sont issues d'une étude datant de 1973. Elles sont notamment analysées dans Tomassone *et al.* (1992). L'objectif est d'étudier l'influence de caractéristiques végétales sur le développement des chenilles processionnaires. La variable à expliquer est le logarithme du nombre moyen de nids de chenilles par arbre sur des parcelles de 500 mètres carrés. Nous disposons de $n = 33$ parcelles et des huit variables explicatives suivantes :

x_1 : l'altitude (en mètres),
x_2 : la pente (en degrés),
x_3 : le nombre de pins dans la parcelle,
x_4 : la taille (en mètres) du pin se trouvant le plus au centre de la parcelle,
x_5 : l'orientation de la parcelle,
x_6 : la taille (en mètres) du plus grand pin de la parcelle,
x_7 : le nombre de strates de végétation,
x_8 : un indice de mélange de végétation (de 1 si pas mélangée à 2 si mélangée).

Deux variables issues des données initiales ont été retirées de l'analyse. Il s'agit : du diamètre du pin se trouvant le plus au centre de la parcelle et d'un indice de densité de population. Ces deux variables étant, respectivement, très corrélées avec les variables x_4 et x_3, elles apportent de la confusion quant à l'interprétation. Nous obtenons les résultats suivants :

	Moy. Post.	Ecart-type Post.	Facteur de Bayes
(Constante)	-0.8133	0.1406	
X1	-0.5039	0.1883	0.7224 (**)
X2	-0.3755	0.1508	0.5392 (**)
X3	0.6225	0.3436	-0.0443
X4	-0.2776	0.2804	-0.5422
X5	-0.2069	0.1499	-0.3378
X6	0.2806	0.4759	-0.6857
X7	-1.0420	0.4178	0.5435 (**)
X8	-0.0221	0.1531	-0.7609

Confiance contre l'hypothèse H0 : (****) décisive, (***) forte,
(**) substantielle, (*) faible

Dans la dernière colonne du tableau précédent, on trouve le logarithme décimal du facteur de Bayes du modèle complet contre le modèle contenant toutes les variables exceptée celle correspondant à la ligne concernée. Par exemple, le logarithme décimal du facteur de Bayes du modèle complet contre le modèle contenant toutes les variables exceptée x_1 est égal 0.7224. Ce tableau est l'analogue bayésien de la sortie standard d'une régression fréquentielle, le facteur de Bayes jouant le rôle du test de Student. Si l'on se contente de ce dernier, on retient le modèle contenant les variables x_1, x_2 et x_7 : l'altitude, la pente, le nombre de strate de végétation. D'après leurs coefficients estimés, ces trois variables jouent un rôle négatif sur le nombre de nids. Par ailleurs, la moyenne a posteriori de σ^2 est égale à 0.6528.

Si l'on utilise une approche globale, on trouve que le modèle composé des variables x_1, x_2 et x_7 est celui qui a la probabilité a posteriori la plus forte, cette dernière est égale environ à 0.0767. Cette probabilité est très faible : attention quant aux conclusions trop tranchées.

Le nombre de variables étant limité, il est possible de calculer explicitement la probabilité a posteriori des 256 modèles en compétition. Le tableau suivant montre que, sur cet exemple très simple, l'échantillonneur de Gibbs se comporte très bien en terme d'estimation des probabilités a posteriori des modèles. On y retrouve les estimations des probabilités a posteriori des huit meilleurs modèles pour 6000 itérations de l'échantillonneur de gibbs incluant 1000 itérations de temps de chauffe.

Variables	Probabilité a posteriori	Estimation par échantillonnage de Gibbs
1,2,7	0.0767	0.0744
1,7	0.0689	0.0692
1,2,3,7	0.0685	0.0650
1,3,7	0.0376	0.0388
1,2,6	0.0369	0.0362
1,2,3,5,7	0.0326	0.0348
1,2,5,7	0.0294	0.0322
1,6	0.0205	0.0200

4 Conclusion

Nous renvoyons à Marin et Robert (2007) pour des exemples convaincants dans le cadre des modèles linéaires généralisés, des modèles de capture-recapture, des modèles de mélange, des séries temporelles... Il existe dans chaque cas une modélisation a priori par défaut et une résolution algorithme qui permettent de fournir une solution bayésienne de référence pour le problème considéré. Bien entendu d'autres lois a priori peuvent être considérées, le modèle de référence servant alors à évaluer l'impact de ce choix a priori. Nous voulions simplement communiquer ici l'idée selon laquelle il est possible de mener une inférence bayésienne sur un problème réaliste sans disposer d'une expertise particulière pour la construction de lois a priori.

Références

Bernardo, J. et Smith, A. (1994). *Bayesian Theory*. John Wiley, New York.

Brown, P., Vannucci, M. et Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *J. Royal Statist. Society Series B*, pages 627–641.

Carlin, B. et Louis, T. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York, second édition.

Casella, G. et Moreno, E. (2006). Objective Bayesian variable selection. *J. American Statist. Assoc.*, 101(473):157–167.

Celeux, G., Marin, J.-M. et Robert, C. (2006). Sélection bayésienne de variables en régression linéaire. *Journal de la Société Française de Statistique*, 147(1):59–79.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 1:17–36.

Cui, W. et George, E. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138:888–900.

Dawid, A. et Lauritzen, S. (2000). Compatible prior distribution. *In Bayesian Methods with Application to Science Policy and Official Statistics. The sixth world meeting of the ISBA*, pages 109–118.

Dey, D., Müller, P. et Sinha, D. (1997). *Practical Nonparametric and Semiparametric Bayesian Analysis*, volume 133 de *Lecture Notes in Statistics*. Springer-Verlag, New York.

Fernandez, C., Ley, E. et Steel, M. (2001). Benchmark priors for bayesian model averaging. *J. Econometrics*, 100:381–427.

Gelman, A., Carlin, J., Stern, H. et Rubin, D. (2001). *Bayesian Data Analysis*. Chapman and Hall, New York, New York, second édition.

George, E. (2000). The variable selection problem. *J. American Statist. Assoc.*, 95:1304–1308.

George, E. et McCulloch, R. (1993). Variable selection via Gibbbs sampling. *J. American Statist. Assoc.*, 88:881–889.

George, E. et McCulloch, R. (1997). Approaches to Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Geweke, J. (1994). Variable selection and model comparison in regression. Rapport technique, University of Minnesota.

Ibrahim, G. (1997). On properties of predictive priors in linears models. *The American Statistician*, 51(4):333–337.

Ibrahim, G. et Laud, P. (1994). A predictive approach to the analysis of designed experiments. *J. American Statist. Assoc.*, 89(425):309–319.

Kass, R. et Raftery, A. (1995). Bayes factor and model uncertainty. *J. American Statist. Assoc.*, 90:773–795.

Kohn, R., Smith, M. et Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11:313–322.

Liang, F., Paulo, R., Molina, G., Clyde, M. et Berger, J. (2008). Mixtures of g-priors for Bayesian variable selection. *J. American Statist. Assoc.*, 103(481):410–423.

Marin, J.-M. et Robert, C. (2007). *Bayesian Core : A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag, New York.

Miller, A. (1990). *Subset Selection in Regression*. Chapman and Hall.

Mitchell, T. et Beauchamp, J. (1988). Bayesian variable selection in linear regression. *J. American Statist. Assoc.*, 83:1023–1032.

Philips, R. et Guttman, I. (1998). A new criterion for variable selection. *Statist. Prob. Letters*, 38:11–19.

Robert, C. (2006). *Le choix bayésien : Principes et pratique*. Springer-Verlag France, Paris.

Robert, C. (2007). *The Bayesian Choice, From Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag, New York, 2 édition.

Robert, C. et Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second édition.

Tomassone, R., Audrain, S., Lesquoy, E. et Millier, C. (1992). *La Régression : nouveaux regards sur une ancienne méthode statistique*. Masson, 2 édition.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distribution regression using bayesian variable selection. *In Bayesian inference and decision techniques : Essays in Honor of Bruno De Finetti*, pages 233–243. North-Holland / Elsevier.