Sélection bayésienne de variables en régression linéaire

Gilles Celeux[*], Jean-Michel Marin[†] et Christian Robert[‡]

18 mai 2006

Résumé

Nous nous intéressons à la sélection bayésienne de variables en régression linéaire. Nous en abordons tous les aspects afin de fournir au lecteur un guide précis. Nous étudions successivement les cas où les loi a priori sur les paramètres des modèles sont informatives et non informatives. Dans le cas informatif, nous proposons d'utiliser la loi a priori de Zellner pour le modèle contenant toutes les variables et une loi a priori de Zellner compatible avec la précédente pour chaque sous-modèle. Dans le cas non informatif, nous montrons d'abord que l'inférence bayésienne utilisant des loi a priori faiblement informatives construites à partir de la loi de Zellner est très sensible à la valeur prise par un hyperparamètre, ce qui nous amène à déconseiller son utilisation. Nous proposons alors une nouvelle loi a priori hiérarchique basée sur la loi de Zellner. Nous montrons que l'utilisation de cette loi a priori assure d'excellentes performances de sélection, d'un point de vue explicatif, par rapport aux critères fréquentiels classiques. Enfin, lorsque le nombre de variables est important, nous considérons les aspects algorithmiques et, en particulier, nous montrons que l'échantillonneur de Gibbs fonctionne parfaitement bien pour sélectionner les variables pertinentes, contrairement à ce qui est parfois affirmé.

Mots clés : modèle de régression linéaire, sélection bayésienne de variables, loi a priori de Zellner, lois a priori compatibles, modèles hiérarchiques, échantillonneur de Gibbs

Abstract

Bayesian variable selection in linear regression is considered. All its aspects are studied in order to provide a precise and efficient userguide. The informative and non-informative cases are analysed. In the informative case, it is suggested to choose the Zellner G-prior on the full model and to derive compatible prior distributions for each sub-model. In the non-informative case, it is shown that, if a Zellner weakly informative

[*]INRIA FUTURS, Équipe SELECT, gilles.celeux@math.u-psud.fr

[†]Auteur correspondant : INRIA FUTURS, Équipe SELECT et CEREMADE, Université Paris Dauphine, Université Paris-Sud, Laboratoire de Mathématiques, 91425 Orsay, jean-michel.marin@math.u-psud.fr

[‡]CEREMADE, Université Paris Dauphine et CREST, INSEE, xian@ceremade.dauphine.fr

prior is used, the model posterior probabilities are sensitive to the choice of an hyper-parameter. Consequently a new Zellner hierarchical prior is proposed. The use of this prior is shown to outperform penalized likelihood criteria in an explicative point of view. Finally, computational aspects are considered when the number of variables is large and, it is shown that the Gibbs sampling do the job quite well.

Key words : Linear regression model, Bayesian variable selection, Zellner G-prior, compatible priors, hierarchical models, Gibbs sampling

1 Introduction

La sélection bayésienne de variables en régression linéaire a été beaucoup étudiée. On peut citer parmi d'autres les articles suivants : Mitchell et Beauchamp (1988); George et McCulloch (1993); Geweke (1994); Chipman (1996); Smith et Kohn (1996); George et McCulloch (1997); Brown *et al.* (1998); Philips et Guttman (1998); George (2000); Kohn *et al.* (2001); Nott et Green (2004); Schneider et Corcoran (2004); Casella et Moreno (2004). Malgré cela, certaines difficultés n'ont pas été résolues de manière satisfaisante. Dans cet article, nous abordons tous les aspects de ce problème de sélection de variables.

Rappelons que la régression linéaire vise à expliquer ou prédire les valeurs prises par une variable y par une fonction linéaire de paramètres inconnus construite sur un ensemble $\{x_1, \ldots, x_p\}$ de p variables. Les performances d'un critère de choix de modèles dépendent de l'objectif. Le choix de variables en régression peut être vu sous l'angle explicatif ou l'angle prédictif. Dans cet article, le point de vue explicatif est privilégié. C'est alors typiquement un problème de choix de modèles que nous considérons ici dans le paradigme bayésien. Rappelons qu'un modèle bayésien est la conjonction d'un modèle d'occurrence de données (vraisemblance) et d'un modèle d'expertise (loi a priori).

Avant d'entrer dans le vif du sujet, rappelons quelques concepts de choix de modèles bayésiens. Considérons simplement deux modèles bayésiens paramétriques $\mathcal{M}_1$ et $\mathcal{M}_2$ où

$$\mathcal{M}_i : y \sim f_i(y|\theta_i), \theta_i \in \Theta_i, \theta_i \sim \pi_i(\theta_i) \quad i = 1, 2,$$

$f_i(y|\theta_i)$ représentant la vraisemblance du paramètre θ_i et $\pi_i(\theta_i)$ la densité de la loi de probabilité a priori du paramètre θ_i associé au modèle i. Munissons l'espace des modèles d'une loi de probabilité a priori et notons $\mathbb{P}(\mathcal{M}_1)$ et $\mathbb{P}(\mathcal{M}_2)$ les probabilités a priori des deux modèles. Un choix de modèle bayésien est basé sur la loi a posteriori des différents modèles, c'est-à-dire sur les probabilités conditionnelles aux observations $\mathbb{P}(\mathcal{M}_1|y)$ et $\mathbb{P}(\mathcal{M}_2|y)$ où

$$\mathbb{P}(\mathcal{M}_i|y) \propto \mathbb{P}(\mathcal{M}_i) \int_{\Theta_i} f_i(y|\theta_i)\pi_i(\theta_i)d\theta_i \,.$$

Cette distribution a posteriori est très sensible au choix des lois a priori des paramètres des modèles, $\pi_1(\theta_1)$ et $\pi_2(\theta_2)$. Dans le cas où nous disposons d'informations a priori, il est important que les lois a priori entre les différents modèles soient équitables. Par ailleurs,

dans un cas non informatif, il n'est pas possible d'utiliser des lois impropres. En effet, les probabilités a posteriori ne sont alors définies qu'à une constante arbitraire près, ce qui empêche la comparaison de 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, nous proposons d'utiliser des lois a priori de Zellner compatibles et nous en déduisons une procédure de sélection. Dans une deuxième partie, nous traitons du cas non informatif et nous présentons une nouvelle loi a priori hiérarchique basée sur la loi de Zellner. Les performances de sélection sont comparées à celles des critères fréquentiels classiques. Enfin, nous abordons les aspects algorithmiques lorsque le nombre de variables est important et nous traitons des données réelles. Nous résumons les points importants de notre étude dans une courte conclusion.

2 Lois a priori informatives

Nous considérons une variable aléatoire à expliquer y et un ensemble $\{x_1, \ldots, x_p\}$ de p variables explicatives ou 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. Nous notons $X = [1_n, \mathbf{x}_1, \ldots, \mathbf{x}_p]$ la matrice de dimension $n \times (p+1)$ dont les colonnes sont constitués du vecteur 1_n et des p variables explicatives. Le terme constant de la régression fait partie de tous les modèles que nous considèrerons. Ils sont donc au nombre de 2^p. Nous utilisons une représentation hiérarchique et nous désignons chaque modèle à l'aide d'un paramètre binaire $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Gamma = \{0,1\}^{\otimes p}$ 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_γ le nombre de variables entrant dans le modèle γ, $p_\gamma = 1_p'\gamma$ (où A' désigne la transposée de A) ;
- $t_1(\gamma) = \{t_{1,1}(\gamma), \ldots, t_{1,p_\gamma}(\gamma)\}$ l'ensemble des indices de ces variables, ainsi $t_{1,1}(\gamma) = \min\left(\{i \in \{1, \ldots, p\} | \gamma_i = 1\}\right)$;
- $t_0(\gamma) = \{t_{0,1}(\gamma), \ldots, t_{0,(p-p_\gamma)}(\gamma)\}$ l'ensemble des indices des variables n'entrant pas dans le modèle γ, ainsi $t_{0,1}(\gamma) = \min\left(\{i \in \{1, \ldots, p\} | \gamma_i = 0\}\right)$.

Par ailleurs, nous notons $\beta \in \mathbb{R}^{p+1}$ les coefficients de la régression avec $\beta_{t_1(\gamma)} = \left[\beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,p_\gamma}(\gamma)}\right]$, $\beta_{t_0(\gamma)} = \left[\beta_{t_{0,1}(\gamma)}, \ldots, \beta_{t_{0,p-p_\gamma}(\gamma)}\right]$, et enfin $X_{t_1(\gamma)} = \left[1_n | x_{t_{1,1}(\gamma)} | \ldots | x_{t_{1,p_\gamma}(\gamma)}\right]$.

Le modèle de régression linéaire gaussienne γ sur $\mathbf{y} = (y_1, \ldots, y_n)$ est défini ainsi

$$\mathbf{y} | \beta, \gamma, \sigma^2 \sim \mathcal{N}\left(X_{t_1(\gamma)} \beta_{t_1(\gamma)}, \sigma^2 I_n\right)$$

où $\beta_{t_1(\gamma)} \in \mathbb{R}^{p_\gamma + 1}$ et $\sigma^2 \in \mathbb{R}_+^*$ sont les paramètres inconnus.

Nous supposons ainsi que les 2^p modèles ont la même variance σ^2 qui ne dépend donc pas de γ. De nombreux auteurs optent pour ce point de vue (Mitchell et Beauchamp, 1988; George et McCulloch, 1993; Smith et Kohn, 1996; George et McCulloch, 1997; Chipman, 1996; Philips et Guttman, 1998; Schneider et Corcoran, 2004; Nott et Green, 2004) qui revient à interpréter σ^2 comme une variance d'erreur de mesure plutôt que comme une variance résiduelle. Par ailleurs, même si la variance était résiduelle, par exemple par oubli d'une variable explicative, celle-ci serait commune à tous les modèles et justifierait encore le choix d'un σ^2 commun.

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, i.e. contenant toutes les variables explicatives. Pour ce modèle, nous préconisons l'utilisation de la loi a priori informative de Zellner (Zellner, 1986) qui fournit un bon compromis entre une loi informative conjuguée et une loi a priori diffuse. L'idée est de permettre au modélisateur d'introduire des informations sur la valeurs des coefficients de la régression sans être obligé de donner des éléments a priori sur les corrélations entre ces coefficients. Pour $\gamma = (1, \ldots, 1)$, l'approche de Zellner conduit à une loi a priori normale pour β sachant σ^2, $\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X'X)^{-1})$ et à une loi a priori de Jeffreys pour σ^2, $\pi(\sigma^2) \propto \sigma^{-2}$. Finalement, la densité de la loi a priori du modèle complet est telle :

$$\pi(\beta, \sigma^2) \propto \left(\sigma^2\right)^{-(p+1)/2-1} \exp\left(-0.5(c\sigma^2)^{-1}(\beta - \tilde{\beta})'(X'X)(\beta - \tilde{\beta})\right) .$$

Cette loi dépend des données à travers X. Ce n'est pas un problème dans la mesure où nous utilisons la vraisemblance conditionnelle du modèle, i.e. la loi de y sachant X. Si X contenait des variables endogènes, ce serait différent... Le modélisateur choisit l'espérance a priori $\tilde{\beta}$ et c, où c donne la quantité relative d'information a priori par rapport à celle portée par l'échantillon. Par exemple, la valeur $1/c = 0.5$ donne à la loi a priori le même poids que 50% de 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 veiller à ce que les choix de lois a priori soient équitables, c'est-à-dire ne favorisent pas arbitrairement un modèle par rapport aux autres. Certains auteurs (Dawid et Lauritzen, 2000; Leucari et Consonni, 2003; Roverato et Consonni, 2004; Marin, 2006) ont proposé des formalisations pour donner un sens précis à cette notion. Ainsi, Marin (2006) a proposé de mesurer la compatibilité de deux lois a priori, $\pi_1(\theta_1)$ et $\pi_2(\theta_2)$ à l'aide de l'information de Kullback-Leibler entre les deux distributions marginales (ou prédictives) correspondantes, $f_1(y) = \int_{\theta_1} f_1(y|\theta_1)\pi_1(\theta_1)d\theta_1$ et $f_2(y) = \int_{\theta_2} f_2(y|\theta_2)\pi_2(\theta_2)d\theta_2$. 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, de variance commune σ^2 et munis de lois a priori informatives de Zellner

$$\mathcal{M}_1 : \mathbf{y}|X_1, \beta_1, \sigma^2 \sim \mathcal{N}(X_1\beta_1, \sigma^2 I_n), \quad \beta_1|X_1, \sigma^2 \sim \mathcal{N}\left(s_1, \sigma^2 n_1 (X_1'X_1)^{-1}\right), \quad \sigma^2 \sim \pi(\sigma^2)\,,$$

où X_1 est une matrice fixée $(n \times k_1)$ de rang $k_1 \leq n$;

$$\mathcal{M}_2 : \mathbf{y}|X_2, \beta_2, \sigma^2 \sim \mathcal{N}(X_2\beta_2, \sigma^2 I_n), \quad \beta_2|X_2, \sigma^2 \sim \mathcal{N}\left(s_2, \sigma^2 n_2 (X_2'X_2)^{-1}\right), \quad \sigma^2 \sim \pi(\sigma^2)\,,$$

où X_2 est une matrice fixée $(n \times k_2)$ de rang $k_2 \leq n$.
Nous supposons que $\mathcal{M}_2$ est un sous-modèle de $\mathcal{M}_1$ et que les valeurs de (s_1, n_1) sont fixées. Le problème consiste alors à déterminer des valeurs de (s_2, n_2) équitables. Comme σ^2 est un paramètre de nuisance, nous minimisons l'information de Kullback-Leibler entre les deux distributions marginales conditionnellement à σ^2. Elle s'écrit

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

Dans ce cas, Marin (2006) montre que le minimum est atteint pour

$$s_2^* = (X_2'X_2)^{-1}X_2'X_1 s_1 \quad \text{et} \quad n_2^* = n_1\,. \tag{1}$$

Ce résultat est intuitif car $X_2 s_2^*$ est la projection orthogonale de $X_1 s_1$ sur l'espace engendré par les colonnes de X_2. Par exemple, si $X_1 = [1_n, x_1]$, $X_2 = 1_n$, $s_1 = (s_{11}, s_{12})$, $s_2^* = s_{11} + \frac{1}{n}\sum_{i=1}^n x_{1i} s_{12}$. Aussi, si $X_1 = [1_n, x_1]$, $X_2 = x_1$, $s_2^* = \left(\sum_{i=1}^n x_{1i} \Big/ \sum_{i=1}^n x_{1i}^2\right) s_{11} + s_{12}$.

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 les autres propositions de lois a priori équitables. On doit aussi noter qu'indépendamment de leur 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 pénétration.

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

1) définir la loi a priori du modèle complet ;

2) puis, en 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 $U_\gamma = \left(X_{t_1(\gamma)}'X_{t_1(\gamma)}\right)^{-1} X_{t_1(\gamma)}$ et $P_\gamma = X_{t_1(\gamma)}'U_\gamma$. Agissant de la sorte pour un modèle γ, la loi a priori compatible de $\beta_{t_1(\gamma)}$ sachant σ^2 est

$$\mathcal{N}\left(U_\gamma X\tilde{\beta}, c\sigma^2\left(X_{t_1(\gamma)}'X_{t_1(\gamma)}\right)^{-1}\right).$$

La densité de la loi a priori des paramètres du modèle γ est alors telle que

$$\pi(\beta, \sigma^2|\gamma) \propto \left(\sigma^2\right)^{-(p_\gamma+1)/2-1} \mathbb{I}_{0_{p-p_\gamma}}(\beta_{t_0(\gamma)}) \exp\left[\frac{-1}{2c\sigma^2}\left(\beta_{t_1(\gamma)} - U_\gamma X\tilde{\beta}\right)' (X'_{t_1(\gamma)}X_{t_1(\gamma)})\left(\beta_{t_1(\gamma)} - U_\gamma X\tilde{\beta}\right)\right].$$

Le paramètre de sélection γ se situe à un niveau hiérarchique plus élevé, comme l'indique, dans la figure 1, le graphe acyclique orienté (GAO) (Lauritzen, 1996). Ce graphe indique les relations de dépendance conditionnelles entre les variables du modèle. Dans ce graphe, la dépendance au modèle du paramètre β n'est pas explicitement marquée. En effet, dans un but de simplification, tout au long de l'article, nous n'indiquons pas la dépendance évidente au modèle des coefficients β.

Pour ce paramètre, nous utilisons la loi a priori suivante

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

où τ_i correspond à la probabilité a priori que la variable i soit présente dans le modèle. Typiquement, lorsque aucune information a priori n'est présente, on pose $\tau_1 = \ldots = \tau_p = 1/2$. Cela conduit à une loi a priori uniforme $\pi(\gamma) = 2^{-p}$, hypothèse toujours faite dans la suite de cet article.

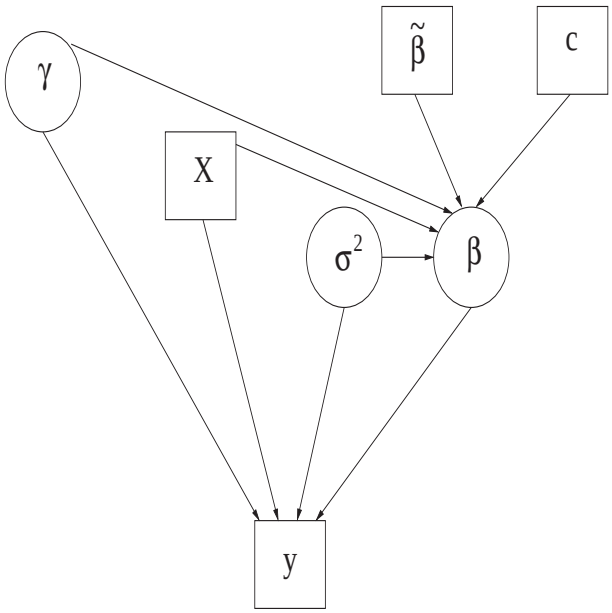

FIG. 1 – Représentation du modèle informatif par son GAO : les quantités aléatoires sont entourées par des ellipses et les quantités fixes ou observées par des rectangles.

2.2 Loi a posteriori de γ

Nous avons

$$\begin{aligned}\pi(\gamma|\mathbf{y}) &\propto f(\mathbf{y}|\gamma)\pi(\gamma)\\ &\propto f(\mathbf{y}|\gamma)\\ &\propto \int\left(\int f(\mathbf{y}|\gamma,\beta,\sigma^2)\pi(\beta|\gamma,\sigma^2)d\beta\right)\pi(\sigma^2)d\sigma^2\,.\end{aligned}$$

Par des calculs élémentaires (George et McCulloch, 1997), il vient

$$\begin{aligned}f(\mathbf{y}|\gamma,\sigma^2) &= \int f(\mathbf{y}|\gamma,\beta,\sigma^2)\pi(\beta|\gamma,\sigma^2)d\beta\\ &= (c+1)^{-(p_\gamma+1)/2}(2\pi)p^{-n/2}\left(\sigma^2\right)^{-n/2}\\ &\quad \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}'\mathbf{y}+\frac{c}{2\sigma^2(c+1)}\mathbf{y}'P_\gamma\mathbf{y}-\frac{1}{2\sigma^2(c+1)}\tilde{\beta}'X'P_\gamma X\tilde{\beta}+\right.\\ &\quad \left.\frac{1}{\sigma^2(c+1)}\mathbf{y}'P_\gamma X\tilde{\beta}\right)\,.\end{aligned}$$

Ainsi,

$$\pi(\gamma|\mathbf{y})\propto\int f(\mathbf{y}|\gamma,\sigma^2)\pi(\sigma^2)d\sigma^2$$

$$\pi(\gamma|\mathbf{y})\propto(c+1)^{-(p_\gamma+1)/2}\left[\mathbf{y}'\mathbf{y}-\frac{c}{c+1}\mathbf{y}'P_\gamma\mathbf{y}+\frac{1}{c+1}\tilde{\beta}'X'P_\gamma X\tilde{\beta}-\frac{2}{c+1}\mathbf{y}'P_\gamma X\tilde{\beta}\right]^{-n/2}. \quad (2)$$

2.3 Lois a posteriori de β et σ^2

Conditionnellement à γ, nous avons

$$\beta_{t_0(\gamma)}|\sigma^2,\mathbf{y},\gamma\sim\delta(0_{p-p_\gamma}),$$

(ce qui signifie simplement que les coefficients des régresseurs qui n'interviennent pas dans le modèle γ sont fixés à zéro),

$$\beta_{t_1(\gamma)}|\sigma^2,\mathbf{y},\gamma\sim\mathcal{N}\left[\frac{c}{c+1}(U_\gamma\mathbf{y}+U_\gamma X\tilde{\beta}/c),\frac{\sigma^2c}{c+1}\left(X'_{t_1(\gamma)}X_{t_1(\gamma)}\right)^{-1}\right],$$

$$\sigma^2|\mathbf{y},\gamma\sim I\mathcal{G}\left[\frac{n}{2},\frac{\mathbf{y}'\mathbf{y}}{2}-\frac{c}{2(c+1)}\mathbf{y}'P_\gamma\mathbf{y}+\frac{\tilde{\beta}'X'P_\gamma X\tilde{\beta}}{2(c+1)}-\frac{1}{c+1}\mathbf{y}'P_\gamma X\tilde{\beta}\right].$$

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

Pour le modèle γ, $U_\gamma \mathbf{y}$ est l'estimateur des moindres carrés ordinaires de $\beta_{t_1(\gamma)}$, noté $\hat{\beta}_{t_1(\gamma)}$, et $U_\gamma X\tilde{\beta}$ est l'espérance a priori compatible de $\beta_{t_1(\gamma)}$, notée $\tilde{\beta}_{t_1(\gamma)}$. En conséquence, l'estimateur bayésien de $\beta_{t_1(\gamma)}$ minimisant la perte quadratique est donné par

$$\mathbb{E}(\beta_{t_1(\gamma)}|\mathbf{y},\gamma) = \mathbb{E}(\mathbb{E}(\beta_{t_1(\gamma)}|\sigma^2,\mathbf{y},\gamma)|\mathbf{y},\gamma) = \frac{c}{c+1}(\hat{\beta}_{t_1(\gamma)} + \tilde{\beta}_{t_1(\gamma)}/c). \tag{3}$$

Si $c = 1$, c'est-à-dire si l'information a priori a le même poids que l'échantillon, l'estimateur bayésien est la moyenne entre l'estimateur des moindres carrés et la moyenne a priori.
De plus, nous pouvons montrer que

$$\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}'P_\gamma\mathbf{y} + \frac{\tilde{\beta}'X'P_\gamma X\tilde{\beta}}{c+1} - \frac{2}{c+1}\mathbf{y}'P_\gamma X\tilde{\beta} =$$

$$(\mathbf{y} - X_{t_1(\gamma)}\hat{\beta}_{t_1(\gamma)})'(\mathbf{y} - X_{t_1(\gamma)}\hat{\beta}_{t_1(\gamma)}) + \frac{1}{c+1}(\tilde{\beta}_{t_1(\gamma)} - \hat{\beta}_{t_1(\gamma)})'X'_{t_1(\gamma)}X_{t_1(\gamma)}(\tilde{\beta}_{t_1(\gamma)} - \hat{\beta}_{t_1(\gamma)}),$$

et alors l'estimateur bayésien minimisant le risque quadratique est donné par $\mathbb{E}\left[\sigma^2|\mathbf{y},\gamma\right]$ soit

$$\frac{(\mathbf{y} - X_{t_1(\gamma)}\hat{\beta}_{t_1(\gamma)})'(\mathbf{y} - X_{t_1(\gamma)}\hat{\beta}_{t_1(\gamma)}) + (\tilde{\beta}_{t_1(\gamma)} - \hat{\beta}_{t_1(\gamma)})'X'_{t_1(\gamma)}X_{t_1(\gamma)}(\tilde{\beta}_{t_1(\gamma)} - \hat{\beta}_{t_1(\gamma)})/(c+1)}{n-2}. \tag{4}$$

Les équations (3) et (4) mettent en lumière que l'influence de la loi a priori disparaît lorsque c tend vers l'infini.

2.4 Influence de c

Dans ce paragraphe, nous étudions l'influence de c sur la loi a posteriori des modèles.

Jeu de données simulées 1. *À l'instar de Casella et Moreno (2004), nous considérons le modèle complet suivant à dix prédicteurs*

$$\mathbf{y}|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{3}\beta_i\mathbf{x}_i + \sum_{i=1}^{3}\beta_{i+3}\mathbf{x}_i^2 + \beta_7\mathbf{x}_1\mathbf{x}_2 + \beta_8\mathbf{x}_1\mathbf{x}_3 + \beta_9\mathbf{x}_2\mathbf{x}_3 + \beta_{10}\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3, \sigma^2 I_n\right)$$

où les composantes des vecteurs $\mathbf{x}_i$, $i = 1,\ldots,10$ *ont été générées indépendamment selon une loi uniforme sur* $]0,10[$. *Le vrai modèle a été bâti tel que l'espérance de* $\mathbf{y}$ *est égale* $\beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2$. *Il est donc caractérisé par* $\gamma^* = (1,1,0,\ldots,0)$ *avec* $n = 50$, $(\beta_0,\beta_1,\beta_2) = (5,1,3)$ *et* $\sigma^2 = 4$.

Le tableau 1 montre que la loi a posteriori des modèles est très sensible au choix de c. Ainsi, lorsque l'on ne dispose pas d'informations a priori, l'idée intuitive consistant à prendre $\tilde{\beta} = 0_{p+1}$ et c grand ne peut pas être utilisée. Cela nous a amené à proposer une loi a priori non informative hiérarchique décrite dans la section suivante.

Modèle	$c = 10$	$c = 100$	$c = 1,000$	$c = 10,000$	$c = 1,000,000$
0,1,2	0.04062	0.35368	0.65858	0.85895	0.98222
0,1,2,7	0.01326	0.06142	0.08395	0.04434	0.00524
0,1,2,4	0.01299	0.05310	0.05805	0.02868	0.00336
0,2,4	0.02927	0.03962	0.00409	0.00246	0.00254
0,1,2,8	0.01240	0.03833	0.01100	0.00126	0.00126

TAB. 1 – **Jeu de données simulées 1** : estimation de la probabilité a posteriori de différents modèles suivant la valeur de l'hyperparamètre c.

3 Lois a priori non informatives

Dans un contexte non informatif, nous proposons d'utiliser la même loi a priori de Zellner avec $\tilde{\beta} = 0_{p+1}$ et d'introduire une loi a priori hiérarchique diffuse sur c, pour de manière classique (Marin et Robert, 2006; Robert, 2006) atténuer le rôle de l'hyperparamètre c

$$\pi(c) \propto c^{-1} \mathbb{I}_{\mathbb{N}^*}(c)\,.$$

La figure 2 donne le GAO représentant le modèle non informatif pour lequel nous avons opté.

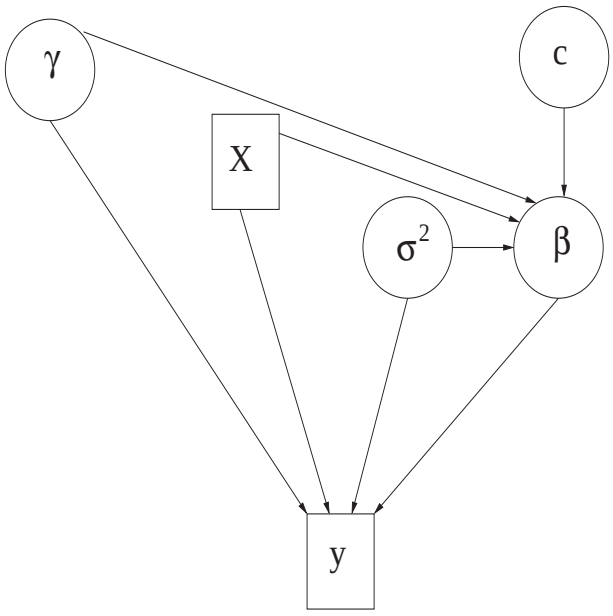

FIG. 2 – Représentation du modèle non informatif par son GAO : les quantités aléatoires sont entourées par des ellipses et les quantités fixes ou observées par des rectangles.

3.1 Loi a posteriori des modèles

Par (2), nous avons

$$\pi(\gamma|\mathbf{y},c) \propto (c+1)^{-(p_\gamma+1)/2}\left[\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}'P_\gamma\mathbf{y}\right]^{-n/2},$$

et

$$\pi(\gamma|\mathbf{y}) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(p_\gamma+1)/2}\left[\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}'P_\gamma\mathbf{y}\right]^{-n/2}. \quad (5)$$

Il est facile de montrer que, pour tout $\mathbf{y}$, la série en jeu dans (5) converge. Pour calculer (5), nous sommes juste amené à tronquer cette somme infinie. Le tableau 2 donne, sur le **jeu de données simulées 1**, les probabilités a posteriori des cinq meilleurs modèles calculées en utilisant deux troncatures différentes. Sur cet exemple, la troncature 10^5 fonctionne bien.

Modèle	$\sum_{i=1}^{10^5} \pi(\gamma|\mathbf{y},c)\pi(c)$	$\sum_{i=1}^{10^6} \pi(\gamma|\mathbf{y},c)\pi(c)$
0,1,2	0.77969	0.78071
0,1,2,7	0.06229	0.06201
0,1,2,4	0.04138	0.04119
0,1,2,8	0.01684	0.01676
0,1,2,5	0.01611	0.01604

TAB. 2 – **Jeu de données 1** : probabilités a posteriori des cinq meilleurs modèles pour $\pi(c) \propto c^{-1}\mathbb{I}_{\mathbb{N}^*}(c)$ et deux valeurs de troncature 10^5 et 10^6.

3.2 Lois a posteriori de β et σ^2

Nous calculons maintenant l'espérance a posteriori de (β,σ^2) conditionnellement à γ.

$$\begin{aligned}
\pi(\beta,\sigma^2|\mathbf{y},\gamma) &= \int \pi(\beta,\sigma^2|\mathbf{y},\gamma,c)f(c|\mathbf{y},\gamma)dc \\
&\propto \int \pi(\beta,\sigma^2|\mathbf{y},\gamma,c)f(c,\mathbf{y},\gamma)dc \\
&\propto \int \pi(\beta,\sigma^2|\mathbf{y},\gamma,c)f(\mathbf{y}|\gamma,c)\pi(c).
\end{aligned}$$

Ainsi,

$$\mathbb{E}(\beta_{t_1(\gamma)}|\mathbf{y},\gamma) = U_\gamma \mathbf{y} \left(\frac{\sum_{c=1}^{\infty} f(\mathbf{y}|\gamma,c)\pi(c)c/(c+1)}{\sum_{c=1}^{\infty} f(\mathbf{y}|\gamma,c)\pi(c)} \right)$$

$$= U_\gamma \mathbf{y} \left(\frac{\sum_{c=1}^{\infty} (c+1)^{-(p_\gamma+1)/2-1} \left[\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}' P_\gamma \mathbf{y} \right]^{-n/2}}{\sum_{c=1}^{\infty} c^{-1}(c+1)^{-(p_\gamma+1)/2} \left[\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}' P_\gamma \mathbf{y} \right]^{-n/2}} \right), \tag{6}$$

et $\mathbb{E}(\sigma^2|\mathbf{y},\gamma)$ s'écrit

$$\frac{\sum_{c=1}^{\infty} \left(\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}' P_\gamma \mathbf{y} \right) c^{-1}(c+1)^{-(p_\gamma+1)/2}(n-2)^{-1} \left[\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}' P_\gamma \mathbf{y} \right]^{-n/2}}{\sum_{c=1}^{\infty} c^{-1}(c+1)^{-(p_\gamma+1)/2} \left[\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}' P_\gamma \mathbf{y} \right]^{-n/2}}. \tag{7}$$

3.3 Performances de sélection

Sur différents jeux de données, nous allons maintenant comparer les performances de sélection d'une procédure bayésienne non informative consistant à sélectionner le modèle ayant la probabilité a posteriori la plus forte à celles des critères fréquentiels classiques : AIC, Cp de Mallows et BIC. Notons que l'objectif des critères AIC et Cp est d'abord prédictif tandis que celui de BIC et de notre critère bayésien est d'abord explicatif. Rappelons que pour le modèle γ :

$$\mathrm{AIC}(\gamma) = -2\log[f(\mathbf{y}|\hat{\beta}_{t_1(\gamma)},\gamma,\hat{\sigma}^2_\gamma)] + 2(p_\gamma+2)\,,$$

où $\hat{\beta}_{t_1(\gamma)}$ et $\hat{\sigma}^2_\gamma$ sont les estimations du maximum de vraisemblance des paramètres $\beta_{t_1(\gamma)}$ et σ^2 pour le modèle γ,

$$\mathrm{BIC}(\gamma) = -2\log[f(\mathbf{y}|\hat{\beta}_{t_1(\gamma)},\gamma,\hat{\sigma}^2_\gamma)] + \log(n)(p_\gamma+2)\,,$$

enfin, si $\tilde{\gamma} = (1,\ldots,1)$,

$$\mathrm{Cp}(\gamma) = (n-p_\gamma-1)\frac{\hat{\sigma}^2_{\tilde{\gamma}}}{\hat{\sigma}^2_\gamma} - n + 2(p_\gamma+2)\,.$$

Jeu de données simulées 2. *Il s'agit d'un modèle avec $p = 10$ variables explicatives potentielles avec deux structures de corrélation entre régresseurs.*

Corrélation forte *: les composantes des vecteurs* $\mathbf{z}_i$ *(*$i = 1, \ldots, 13$*) de dimension* n *sont générées indépendamment suivant une* $\mathcal{N}(0, 1)$*.*

Nous posons

$$\mathbf{x}_i = (\mathbf{z}_i + 6\mathbf{z}_{11})/\sqrt{37} \text{ } (i = 1, 2),$$
$$\mathbf{x}_i = (\mathbf{z}_i + 6\mathbf{z}_{12})/\sqrt{37} \text{ } (i = 3, 4, 5),$$
$$\mathbf{x}_i = (\mathbf{z}_i + 6\mathbf{z}_{13})/\sqrt{37} \text{ } (i = 6, \ldots, 10).$$

Cela donne la matrice de corrélation théorique suivante entre les variables explicatives

$$\begin{bmatrix}
1 & 36/37 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
36/37 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 36/37 & 36/37 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 36/37 & 1 & 36/37 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 36/37 & 36/37 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 36/37 & 36/37 & 36/37 & 36/37 \\
0 & 0 & 0 & 0 & 0 & 36/37 & 1 & 36/37 & 36/37 & 36/37 \\
0 & 0 & 0 & 0 & 0 & 36/37 & 36/37 & 1 & 36/37 & 36/37 \\
0 & 0 & 0 & 0 & 0 & 36/37 & 36/37 & 36/37 & 1 & 36/37 \\
0 & 0 & 0 & 0 & 0 & 36/37 & 36/37 & 36/37 & 36/37 & 1
\end{bmatrix}.$$

Le vrai modèle est défini par

$$\mathbf{y} \sim \mathcal{N}(2 + 2\mathbf{x}_2 + 2\mathbf{x}_3 + 2\mathbf{x}_6, I_n).$$

Corrélation nulle *: les composantes des vecteur* $\mathbf{x}_i$ *(*$i = 1, \ldots, 10$*) de dimension* n *sont générées indépendamment suivant une* $\mathcal{N}(0, 1)$*. Le vrai modèle est défini par*

$$\mathbf{y} \sim \mathcal{N}(2 + 2\mathbf{x}_2 + 2\mathbf{x}_3 + 2\mathbf{x}_6, I_n).$$

Sur différents **jeux de données simulées 2**, répliqués 100 fois (toutes les variables étant resimulées), le tableau 3 compare les performances de sélection de la procédure bayésienne non informative par rapport à celle des critères fréquentiels classiques de sélection de variables. Soulignons que, le nombre de variables étant faible, nous effectuons une sélection exhaustive. Ce tableau montre que la procédure bayésienne non informative assure des performances de sélection meilleures que les critères fréquentiels classiques, notamment dans les cas où la taille d'échantillon est faible. De plus la procédure bayésienne résiste assez bien à une forte corrélation entre régresseurs.

Sur ce jeu de données, nous avons fait des expérimentations analogues où aucun modèle proposé n'est le vrai modèle. Pour cela, nous avons enlevé la variable $\mathbf{x}_2$ des choix possibles. Dans le cas indépendant le quasi-vrai modèle, notion clairement définie par Lebarbier et Mary-Huard (2006) ou Burnham et Anderson (2002), s'obtient alors par combinaison des

ρ	36/37		0	
$n = 30$	AIC :	14	AIC :	20
	Cp :	20	Cp :	29
	BIC :	30	BIC :	41
	MAPNI :	38	MAPNI :	78
$n = 50$	AIC :	22	AIC :	24
	Cp :	29	Cp :	33
	BIC :	57	BIC :	60
	MAPNI :	61	MAPNI :	88
$n = 100$	AIC :	26	AIC :	33
	Cp :	30	Cp :	36
	BIC :	79	BIC :	73
	MAPNI :	92	MAPNI :	85
$n = 500$	AIC :	37	AIC :	37
	Cp :	39	Cp :	38
	BIC :	91	BIC :	96
	MAPNI :	94	MAPNI :	97

TAB. 3 – **Jeux de données simulées 2** : comparaison des critères AIC, Cp, BIC et Maximum A Posteriori Non Informatif (MAPNI) : pourcentage de choix du vrai modèle selon différentes tailles d'échantillons, chaque jeu de données étant simulé 100 fois.

variables $\mathbf{x}_3$ et $\mathbf{x}_6$. Dans le cas corrélé, de manière évidente, il s'obtient par combinaison des variables $\mathbf{x}_1$, $\mathbf{x}_3$ et $\mathbf{x}_6$. Le tableau 4 compare les performances de sélection. Il confirme le bon comportement de MAPNI qui fonctionne mieux que BIC pour n assez petit. Notons que les performances de BIC et MAPNI se rejoignent pour n grand. Ceci est assez naturel étant donné la construction de BIC.

ρ	36/37		0	
$n = 30$	AIC :	18	AIC :	20
	Cp :	24	Cp :	32
	BIC :	28	BIC :	55
	MAPNI :	43	MAPNI :	63
$n = 50$	AIC :	24	AIC :	20
	Cp :	30	Cp :	29
	BIC :	54	BIC :	68
	MAPNI :	60	MAPNI :	72
$n = 100$	AIC :	36	AIC :	24
	Cp :	36	Cp :	29
	BIC :	82	BIC :	72
	MAPNI :	86	MAPNI :	73
$n = 500$	AIC :	38	AIC :	33
	Cp :	39	Cp :	33
	BIC :	93	BIC :	90
	MAPNI :	93	MAPNI :	87

TAB. 4 – **Jeux de données simulées 2** : comparaison des critères AIC, Cp, BIC et Maximum A Posteriori Non Informatif (MAPNI) : pourcentage de choix du quasi-vrai modèle selon différentes tailles d'échantillons, chaque jeu de données étant simulé 100 fois.

4 Approximation par échantillonnage de Gibbs

Lorsque le nombre de variables p est grand, typiquement $p > 25$, il est impossible de réaliser une sélection exhaustive. Dans un cadre fréquentiel, des procédures pas à pas avec remise en cause (*stepwise*) ascendante ou descendante sont utilisées (Miller, 1990). Dans un cadre bayésien, il est impossible de calculer explicitement les probabilités a posteriori des 2^p modèles. Un algorithme de simulation permettant d'estimer $\pi(\gamma|\mathbf{y})$ est alors proposé.

Notons γ_{-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

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

et donc

$$\pi(\gamma_i|\mathbf{y}, \gamma_{-i}) \propto \pi(\gamma|\mathbf{y}). \tag{8}$$

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

Échantillonnage de Gibbs

- Itération 0 : tirage de γ^0 selon la distribution uniforme sur Γ.
- Itération t, pour $i = 1, \ldots, p$, tirage de γ_i^t selon $\pi(\gamma_i|\mathbf{y}, \gamma_1^t, \ldots, \gamma_{i-1}^t, \ldots, \gamma_{i+1}^{t-1}, \ldots, \gamma_p^{t-1})$.

L'échantillonnage de Gibbs est décomposé en deux phases. La première phase est une période d'échauffement T_0 à la fin de laquelle, on considère que la chaîne de Markov générée a atteint son régime stationnaire. La deuxième phase vise à approximer sa mesure stationnaire $\pi(\gamma|\mathbf{y})$. Les γ_i générés durant cette phase sont conservés pour mener à bien l'inférence. Pour un modèle γ donné, l'estimateur de $\pi(\gamma|\mathbf{y})$ déduit de l'échantillonnage de Gibbs est

$$\widehat{\pi(\gamma|\mathbf{y})}^{GIBBS} = \left(\frac{1}{T - T_0}\right) \sum_{t=T_0+1}^{T} \mathbb{I}_\gamma(\gamma^t), \tag{9}$$

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

$$\widehat{\mathbb{P}(\gamma_i = 1|\mathbf{y})}^{GIBBS} = \left(\frac{1}{T - T_0}\right) \sum_{t=T_0+1}^{T} \mathbb{I}_{\gamma_i}(\gamma_i^t). \tag{10}$$

Pratiquement, l'échantillonnage de Gibbs conduit à ne pas considérer tous les modèles. Les modèles visités par Gibbs sont ceux de plus forte probabilité. Ainsi, le critère de visite semble plus raisonnable que les critères utilisés par les techniques pas à pas.

Tout d'abord, nous illustrons les performances de l'échantillonnage de Gibbs sur le **jeu de simulées données 1**. Le tableau 5 résume les résultats obtenus avec une loi a priori faiblement informative caractérisée par $\tilde{\beta} = 0_{11}$ et $c = 100$. Le tableau 6 donne les résultats pour une loi a priori non informative. Il s'avère que les résultats déduits de l'échantillonnage de Gibbs sont excellents.

$\beta_i \neq 0$	$\pi(\gamma\|\mathbf{y})$	$\widehat{\pi(\gamma\|\mathbf{y})}^{GIBBS}$
0,1,2	0.3537	0.3510
0,1,2,7	0.0614	0.0604
0,1,2,4	0.0531	0.0537
0,2,4	0.0396	0.0399
0,1,2,8	0.0383	0.0388
0,1,2,5	0.0377	0.0376
0,1,2,9	0.0355	0.0347
0,1,2,3	0.0355	0.0351
0,1,2,6	0.0353	0.0352
0,1,2,10	0.0352	0.0345
0,2,3,8	0.0127	0.0133

TAB. 5 – **Jeu de données simulées 1** : résultats de l'échantillonnage Gibbs ($T = 100000$ et $T_0 = 10000$) pour $\tilde{\beta} = 0_{11}$ et $c = 100$. La première colonne indique les indices des variables sélectionnées, la seconde les probabilités a posteriori associées, et la troisième les estimations de ces probabilités tirées de l'échantillonneur de Gibbs.

$\beta_i \neq 0$	$\pi(\gamma\|\mathbf{y})$	$\widehat{\pi(\gamma\|\mathbf{y})}^{GIBBS}$
0,1,2	0.7797	0.7804
0,1,2,7	0.0623	0.0617
0,1,2,4	0.0414	0.0410
0,1,2,8	0.0168	0.0167
0,1,2,5	0.0161	0.0160
0,1,2,9	0.0136	0.0143
0,1,2,3	0.0136	0.0143
0,1,2,6	0.0134	0.0126
0,1,2,10	0.0134	0.0134
0,2,4	0.0029	0.0029

TAB. 6 – **Jeu de données simulées 1** : résultats de l'échantillonnage de Gibbs ($T = 100000$ et $T_0 = 10000$) pour $\pi(c) \propto c^{-1}\mathbb{I}_{\mathbb{N}^*}(c)$. La première colonne indique les indices des variables sélectionnées, la seconde les probabilités a posteriori associées, et la troisième les estimations de ces probabilités tirées de l'échantillonneur de Gibbs.

Jeu de données simulées 3. *Nous considérons un jeu de données où les vingt régresseurs potentiels sont très corrélés. Nous avons procédé de la manière suivante. Nous avons posé* $x_i = z_i + 3z$*, les* z_i *et* z *étant simulé indépendamment selon une loi* $\mathcal{N}_n(0_n, I_n)$*. La corrélation entre les régresseurs* $x_1, \ldots, x_{20}$ *est de l'ordre de* 0.9*. Le vrai modèle est construit sur les sept variables suivantes* $x_1, x_3, x_5, x_6, x_{12}, x_{18}, x_{20}$*, avec* $(\beta_0, \beta_1, \beta_3, \beta_5, \beta_6, \beta_{12}, \beta_{18}, \beta_{20}) = (3, 4, 1, -3, 12, -1, 5, -6)$*,* $\sigma^2 = 4$ *et* $n = 180$*.*

Nous illustrons maintenant les performances de l'échantillonnage de Gibbs sur le **jeu de données simulées 3**. Du fait de la forte corrélation entre les régresseurs, ce jeu de données est susceptible de mettre l'échantillonneur de Gibbs en difficulté. Le tableau 7 donne les résultats pour une loi a priori faiblement informative, caractérisée par $\tilde{\beta} = 0_{21}$ et $c = 100$. Manifestement, la vitesse de l'échantillonnage de Gibbs n'a pas eu à souffrir de la forte corrélation entre les variables explicatives.

$\beta_i \neq 0$	$\pi(\gamma\|\mathbf{y})$	$\widehat{\pi(\gamma\|\mathbf{y})}^{GIBBS}$
0,1,3,5,6,12,18,20	0.1893	0.1822
0,1,3,5,6,18,20	0.0588	0.0598
0,1,3,5,6,9,12,18,20	0.0223	0.0236
0,1,3,5,6,12,14,18,20	0.0220	0.0193
0,1,2,3,5,6,12,18,20	0.0216	0.0222
0,1,3,5,6,7,12,18,20	0.0212	0.0233
0,1,3,5,6,10,12,18,20	0.0199	0.0222
0,1,3,4,5,6,12,18,20	0.0197	0.0182
0,1,3,5,6,12,15,18,20	0.0196	0.0196
0,1,3,5,6,8,12,18,20	0.0193	0.0197

TAB. 7 – **Jeu de données simulées 3** : résultats de l'échantillonnage de Gibbs ($T = 100000$ et $T_0 = 10000$) pour $\tilde{\beta} = 0_{21}$ et $c = 100$. La première colonne indique les indices des variables sélectionnées, la seconde les probabilités a posteriori associées, et la troisième les estimations de ces probabilités tirées de l'échantillonneur de Gibbs.

5 Données réelles

5.1 Chenilles processionnaires

Jeu de données 1. *Ces données sont issues d'un étude datant de 1973. Elles sont notamment analysées dans Tomassone* et al. *(1992, 1993). 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* $n = 33$ *observations et des dix 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 *: le diamètre du pin se trouvant le plus au centre de la parcelle,*
x_6 *: un indice de densité de population,*
x_7 *: l'orientation de la parcelle (de* 1 *si orienté sud à* 2 *sinon),*
x_8 *: la taille (en mètres) du plus grand pin de la parcelle,*
x_9 *: le nombre de strates de végétation,*
x_{10} *: un indice de mélange de végétation (de* 1 *si pas mélangée à* 2 *si mélangée).*

Sur ces données, nous obtenons les résultats suivants

	Moy. Post.	Var. Post.
(Constante)	9.2714	9.1164
X1	-0.0037	2e-06
X2	-0.0454	0.0004
X3	0.0573	0.0086
X4	-1.0905	0.2901
X5	0.1953	0.0099
X6	-0.3008	2.1372
X7	-0.2002	0.8815
X8	0.1526	0.0490
X9	-1.0835	0.6643
X10	-0.3651	0.4716

Par ailleurs, la moyenne a posteriori de σ^2 est égale à 0.7732. Enfin, malgré la faible taille d'échantillon, tous les critères expérimentés sélectionnent les mêmes variables, à savoir les variables x_1, x_2, x_4, et x_5. Tomassone *et al.* (1993) obtiennent le même résultat.

5.2 Ozone

Jeu de données 2. *Il s'agit de données d'ozone analysées notamment dans Breiman et Friedman (1985). Ce sont des mesures journalières de la pollution à l'ozone dans la région de Los Angeles en 1976. Il y a* 366 *observations pour* 10 *variables. Du fait de données manquantes, nous utilisons seulement* 330 *observations. La variable à expliquer est l'enregistrement maximum d'ozone, moyenné sur une heure à Upland, Californie.*

Les variables explicatives sont

- x_1 *: le mois : 1 = janvier, ..., 12 = décembre,*
- x_2 *: le jour du mois,*
- x_3 *: le jour de la semaine : 1 = lundi,..., 7 = dimanche,*
- x_4 *: la pression,*
- x_5 *: la vitesse du vent,*
- x_6 *: le taux humidité,*
- x_7 *: la température,*
- x_8 *: l'altitude,*
- x_9 *: un gradient de pression,*
- x_{10} *: la visibilité.*

Les trois premières variables sont considérées artificiellement comme quantitatives. Sur ces données, nous obtenons les résultats suivants

	Moy. Post.	Var. Post.
(Constante)	-7.8825	1004.3520
X1	-0.2423	0.0100
X2	-0.0097	0.0012
X3	-0.0237	0.0215
X4	-0.0002	3e-05
X5	-0.0213	0.0259
X6	0.0774	0.0005
X7	0.3286	0.0020
X8	-0.0007	5e-08
X9	-0.0120	0.0002
X10	-0.0078	2e-05

Par ailleurs, la moyenne a posteriori de σ^2 est égale à 30.4804. Bien que la taille d'échantillon soit relativement importante, tous les critères ne sélectionnent pas les mêmes variables. L'approche bayésienne non informative fournit la sélection la plus parcimonieuse : AIC et Cp sélectionnent les variables x_1, x_6, x_7, x_8, et x_{10} ; BIC sélectionne les variables x_1, x_6, x_7, et x_8 et MAPNI les variables x_6, x_7 et x_8. Il est intéressant de remarquer que MAPNI ne sélectionne pas la variable x_1 (le mois) dont on peut douter de l'effet linéaire sur y.

6 Conclusion

Dans cet article, nous avons développé une procédure bayésienne de sélection de variables en régression linéaire gaussienne. À cette occasion, dans le cas informatif, nous avons proposé une procédure de construction de lois a priori compatibles. Elle consiste à définir la loi a

priori du modèle complet et à en déduire de manière cohérente les lois a priori de tous les sous-modèles. Par ailleurs, dans le cas non informatif, nous avons proposé une loi a priori hiérarchique qui évite l'arbitraire d'un hyperparamètre.

Des expérimentations illustrent la capacité de cette procédure bayésienne à sélectionner un ensemble parcimonieux de variables. D'un point explicatif, elle produit de meilleurs résultats que les critères classiques, notamment pour de faibles tailles d'échantillon. Par ailleurs, lorsque le nombre de variables est important, nous avons montré que l'échantillonneur de Gibbs fournit rapidement[1] une approximation précise de la loi a posteriori des modèles. Et contrairement à ce que l'on pouvait craindre, l'échantillonneur de Gibbs n'est pas perturbé par de fortes corrélations entre régresseurs.

D'un point de vue prédictif, l'approche bayésienne consiste à pondérer les résultats obtenus pour chaque modèle par leur probabilité a posteriori (Madigan *et al.*, 1997). À titre illustratif, le tableau 8 compare les performances prédictives de l'approche bayésienne non informative et celle obtenue par différents critères pour le **jeu de données simulées 2**. Les 10000 valeurs tests ont été obtenues par la simulation de 100 réplications de Monte-Carlo sur lesquelles 100 échantillons tests sont considérés. Ces résultats confirment le bon comportement de notre approche bayésienne. Il conviendrait de les développer. Par ailleurs, lorsque le nombre de variables est important, il serait intéressant de comparer les performances de l'approche bayésienne utilisant l'échantillonneur de Gibbs avec les procédures classiques de sélection (régression ascendante, descendante, stepwise...).

ρ	36/37		0	
	AIC :	5.3 (0.6)	AIC	: 5.2 (0.5)
	Cp :	5.3 (0.6)	Cp	: 5.2 (0.5)
	BIC :	5.2 (0.5)	BIC	: 5.1 (0.5)
	BMA :	1.0 (0.2)	BMA	: 1.0 (0.1)

TAB. 8 – **Jeux de données simulées 2** et $n = 100$: carrés des écarts moyens (écart-types) entre la valeur observée et la valeur prédite pour 10000 points tests en utilisant les critères AIC, Cp et BIC et la combinaison bayésienne de tous les modèles (BMA).

Nous remercions chaleureusement l'éditeur et un rapporteur pour leurs commentaires judicieux.

[1]Dans cet article, nous ne mentionnons pas les temps de calcul car ils ne sont pas handicapants.

Références

Breiman, L. et Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. American Statist. Assoc.*, pages 580–598.

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

Burnham, K. et Anderson, D. (2002). *Model selection and multi-model inference*. Springer-Verlag.

Casella, G. et Moreno, E. (2004). Objective Bayesian Variable Selection. Rapport technique, University of Florida.

Chipman, H. (1996). Bayesian Variable Selection with Related Predictors. *Canadian Journal of Statistics*, 1:17–36.

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.

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.

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

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Lebarbier, E. et Mary-Huard, T. (2006). Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la Société Française de Statistique*, 121.

Leucari, V. et Consonni, G. (2003). Compatible priors for causal Bayesian networks. *In Bayesian Statistics 7*, pages 597–606. Oxford University Press, Oxford.

Madigan, D., Raftery, A. et Hoeting, J. (1997). Bayesian model averaging for linear regression models. *J. American Statist. Assoc.*, 92:179–191.

Marin, J.-M. (2006). Conjugate compatible prior distributions. *Soumis*.

Marin, J.-M. et Robert, C. (2006). *The Bayesian Core : A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag. à paraître.

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.

Nott, D. J. et Green, P. J. (2004). Bayesian Variable selection and the Swendsen-Wang Algorithm. *J. Comput. Graph. Statist.*, 13:1–17.

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 Implémentation*. Springer-Verlag.

Roverato, A. et Consonni, G. (2004). Compatible Prior Distributions for DAG models. *J. Royal Statist. Soc. Series B*, 66:47–61.

Schneider, U. et Corcoran, J. (2004). Perfect sampling for Bayesian variable selection in a linear regression model. *J. Statist. Plann. Inference*, 126:153–171.

Smith, M. et Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–343.

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

Tomassone, R., Dervin, C. et Masson, J.-P. (1993). *Biométrie : modélisation de phénomènes biologiques*. Masson.

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.