Stage de Master 2
Biostatistique

Université de Montpellier

Université de Sherbrooke

Sélection d'effets fixes et aléatoires en grande dimension dans les modèles linéaires mixtes

---

Benjamin HEUCLIN

Responsables de stage :
Marie DENIS, CIRAD,
Frédéric MORTIER, CIRAD,

Année universitaire 2016-2017

Remerciments

Je tiens tout d'abord à remercier Marie Denis et Frédéric Mortier pour avoir cru en moi, de m'avoir accepté en stage, pour leur disponibilité et pour la qualité de leurs encadrements.

Je remercie Sébastien Tisné pour le temps et l'aide qu'il m'a consacrés afin que je puise acquérir les bases nécessaires en biologie pour ce stage. Je remercie Cathrine Trottier pour ses nombreuses relectures et ses commentaires constructifs qu'elle a pu m'apporter lors de la rédaction de ce rapport de stage ainsi que mes amies Angèle et Charline pour les corrections.

Je tiens à remercier l'ensemble de l'équipe "Génomique et Sélection" du CIRAD pour leur accueil et plus particulièrement Roseline, Najate, Louise, Virginie, Bénédicte, Gille, Clara, Amandine, Frédéric, Alexandre et Letizia pour leur bonne humeur et les moments partagés.

Je remercie ma famille et tous mes amis qui me soutiennent dans mes choix d'études et qui sont présents lors des moments de doutes.

Je remercie Benoîte de Saporta pour la qualité de l'encadrement du master de Biostatistique de l'Université de Montpellier ainsi que Eric Marchand pour m'avoir permis d'effectuer ce master en échange avec l'Université de Sherbrooke au Canada.

Enfin je tiens à remercier Catherine Trottier, Marie Denis et Frédéric Mortier pour avoir accepté de m'encadrer en thèse à l'issue de ce master ainsi que les membres du jury de l'école doctorale Benoîte de Saporta, Jean-Michel Marin, Ali Gannoun et Jean-Noël Bacro qui ont pu rendre possible cette aventure.

Table des matières

1 Introduction 2
2 Modèle bayésien hiérarchique et méthodes MCMC 3
2.1 Modèle bayésien 4
2.2 Méthodes MCMC 5
2.3 Modèle hiérarchique bayésien pour l'estimation des paramètres d'un modèle linéaire mixte 7
3 Sélection bayésienne d'effets fixes dans un modèle linéaire mixte 8
3.1 Méthode bayésienne de sélection de variables avec *a priori* de type "Spike and Slab" 9
3.2 Simulations 14
3.3 Application sur un jeu de données réelles 19
4 Sélection bayésienne d'effets aléatoires dans un modèle linéaire mixte 22
4.1 Méthode bayésien de sélection d'effets aléatoires 22
4.2 Simulations 24
4.3 Application sur jeu de données réelles 29
5 Prise en compte de la similarité entre les matrices d'apparentement 33
5.1 L'approche Hybrid-Correlation-Based Search 34
5.2 Application sur données simulées 35
5.3 Application sur données réelles 37
6 Conclusion 40

A Lois conditionnelles complètes pour l'estimation des paramètres d'un modèle linéaire mixte **45**

B Sélection d'effets fixes **47**

C Sélection d'effets aléatoires **49**

Résumé

Les programmes d'amélioration génétique du palmier à huile ont pour objectif de sélectionner les meilleurs individus dans des populations fortement apparentées. Depuis quelques années, cette sélection est assistée par une importante information moléculaire acquise par les méthodes de séquençage haut débit. Cette information peut se traduire par des marqueurs moléculaire, considérés comme des facteurs à effet fixe, mais également par des structures d'apparentements, associées à des effets aléatoires. L'objectif est donc de sélectionner des effets fixes ou aléatoires afin d'identifier les positions du génome influençant la variation d'un caractère phénotypique. Ces nouveaux enjeux soulèvent de nouvelles questions méthodologiques.

L'objectif de ce travail a été de mettre en œuvre des méthodes de sélection de variables de type "Spike and Slab" pour les effets fixes ainsi que pour les effets aléatoires dans le cadre des modèles linéaires mixtes. Afin de prendre en compte l'information sur la similarité entre les structures d'apparentement, nous avons adapté un algorithme au cas des effets aléatoires et nous l'avons implémenté.

Les approches mises en œuvre sur des jeux de données simulées et réelles ont mis en évidence l'efficacité des approches de type "Spike and Slab" pour sélectionner les effets fixes et aléatoires. Les résultats obtenus ont également montré que l'intégration de l'information de fortes similarités entre les matrices d'apparentement n'apporte pas de meilleurs résultats.

Mots clés : Sélection de variables, modèles linéaires mixtes, méthodes Bayésiennes, *a priori* "Spike and Slab", génétique quantitative, amélioration génétique.

1 Introduction

Les programmes d'amélioration génétique, que ce soit dans le domaine végétal ou animal, ont pour objectif de sélectionner les meilleurs individus (génotypes) d'une population pour engendrer les générations suivantes. Le succès de ces programmes d'amélioration provient de leur capacité à optimiser un ou plusieurs caractères d'intérêt en se basant sur des dispositifs expérimentaux et l'utilisation de modèles statistiques. Dans le cadre de l'amélioration des espèces pérennes telles que l'eucalyptus ou le palmier à huile, les schémas de sélection récurrente réciproque (SRR) sont largement utilisés. Le principe de la SRR est d'améliorer conjointement deux groupes d'individus, de manière à obtenir des hybrides combinant de façon optimale les caractéristiques des deux groupes parentaux. L'analyse de ces dispositifs a pour objectif principal d'extraire la part génétique de la variation d'un phénotype d'intérêt et de l'isoler de la part des variations environnementales en utilisant les modèles linéaires à effets aléatoires (Linear Mixed models, LMM) (Lynch et al., 1998).

Les LMM sont largement étudiés dans la littérature, on peut citer entre autre Verbeke (1997) et McCulloch and Neuhaus (2001). Un modèle linéaire mixte peut s'écrire sous sa forme générale :

$$Y = \mu\mathbb{1} + X\beta + Zu + \varepsilon \tag{1}$$

Y dénote le vecteur aléatoire réponse, β le vecteur inconnu des paramètres fixes, u le vecteur des réalisations de l'effet aléatoire supposé suivre une loi normale $N(0, D)$, X et Z les matrices de design connues associées aux effets fixes et aléatoires respectivement et ε le vecteur des erreurs résiduelles supposé suivre une loi normale $N(0, R)$. u et ε sont supposés indépendants.

Dans le cadre de la génétique quantitative avec des plans de croisements, deux modèles sont principalement utilisés, le "modèle père-mère" et le "modèle animal" (Sun et al., 2009; Mrode, 2014). Le premier modèle permet d'expliquer la variation d'un caractère d'intérêt avec une décomposition de la variance génétique individuelle comme la somme des variances "mère" et "père". Cette décomposition fait référence à la valeur génétique additive et se traduit par le modèle linéaire mixte suivant :

$$Y = \mu\mathbb{1} + X\beta + Z_m M + Z_p P + \varepsilon \tag{2}$$

M et P sont considérés comme deux vecteurs effets aléatoires car ils peuvent être vus comme le résultat d'un échantillonnage aléatoire dans une population parentale plus large. M est l'effet mère et suit une loi normale $N(0, \sigma_m^2 Id)$. P est l'effet père et suit une loi normale $N(0, \sigma_p^2 Id)$. Z_m et Z_p sont deux matrices d'incidence associées aux effets aléatoires M et P respectivement. Cette modélisation considère que toutes les mères sont indépendantes, tous les pères sont indépendants et que les mères sont indépendantes des pères.

Dans le contexte de plans d'expériences avec une information pedigree sur plusieurs générations, ce type de modèle n'est pas adapté. Prenons l'exemple de deux individus cousins. Avec le modèle 2, la corrélation entre ces deux individus sera nulle car les parents sont supposés tous indépendants entre eux. Or théoriquement cette corrélation n'est pas nulle. Pour pallier ce problème, les "modèles animaux" ont été développés initialement pour des études sur les bovins (Wilson et al., 2010). Ils permettent de tenir compte du pedigree en considérant une structure de dépendance génétique entre les individus, aussi appelée structure d'apparentement. Le "modèle animal" est de la forme :

$$Y = \mu\mathbb{1} + X\beta + Zu + \varepsilon \tag{3}$$

avec u l'effet aléatoire contenant un niveau pour chaque individu, il suit une loi normale $N(0, \sigma_a^2 A)$ où A est la structure d'apparentement connue. Les structures d'apparentement les plus couramment utilisées sont les structures pedigree qui sont calculées à partir de la connaissance du pedigree, Mrode (2014).

Mais, avec les outils de génomique et de génotypage haut débit, les enjeux ont évolué. Il s'agit désormais d'utiliser l'information moléculaire pour assister la sélection et ainsi accélérer les programmes d'amélioration en identifiant les régions du génome impliquées dans la variation phénotypique du caractère d'intérêt. Toutefois, l'utilisation de données génotypiques soulève des questions méthodologiques nouvelles, en particulier, en lien avec la sélection d'effets fixes et plus récemment avec la sélection d'effets aléatoires.

En effet, depuis quelques années, l'acquisition d'un grand nombre de marqueurs moléculaires tels que les SNPs (Single Nucleotide Polymorphisms) a donné accès à une information sur l'ensemble du génome. Les SNPs sont des marqueurs bi-alléliques, qui sont intégrés dans les modèles statistiques de génétique quantitative comme des variables à effet fixe. Dans ce contexte où le nombre de variables est supérieur au nombre d'observations, de nombreux résultats ne sont plus valides en régression linéaire et le nombre de modèles possibles explose. Il est donc nécessaire de mettre en œuvre des méthodes plus automatiques pour sélectionner les variables. Une question se pose donc : sélectionner les marqueurs moléculaires qui influencent la variation d'un caractère morphologique en présence d'information sur le pedigree. Autrement dit comment faire de la sélection d'effets fixes dans le contexte de modèles à effets aléatoires.

Des méthodes développées en génétique humaine ont permis à partir de marqueurs moléculaires et de l'information du pedigree d'obtenir une structure d'apparentement plus fine à l'échelle soit du génome soit à chaque position. Ainsi il est possible de décomposer l'effet individuel global comme la somme d'effets individuels en chaque position. Ces structures associées à différentes positions entraînent un grand nombre d'effets aléatoires qui est souvent supérieur au nombre d'observations. Une nouvelle question se pose alors : comment sélectionner les structures qui influencent la variation d'un caractère morphologique. Autrement dit, comment sélectionner des effets aléatoires.

Dans le cas du palmier à huile où nous avons des plans d'expérimentations complexes, nous disposons du pedigree, nous travaillerons donc sur la sélection de variables dans des "modèles animaux" dans un cadre bayésien. Dans une première partie nous introduirons les concepts de base de la statistique bayésienne et présenterons une application aux modèles linéaires mixtes. Puis dans une deuxième partie nous présenterons les principales méthodes de la sélection bayésienne d'effets fixes et détaillerons différentes distributions *a priori* pour les coefficients de ces effets. Dans une troisième partie, nous travaillerons sur des structures d'apparentement associées à des positions du génome entrainant un grand nombre d'effets aléatoires. Nous étudierons la question de la sélection d'effets aléatoires par une approche bayésienne. Enfin, dans une quatrième partie, nous étudierons la prise en compte de la similarité entre les structures d'apparentement.

2 Modèle bayésien hiérarchique et méthodes MCMC

Dans cette section, nous présentons les concepts de base de l'approche bayésienne puis, des méthodes d'inférence de type Monte Carlo par Chaîne de Markov (MCMC), pour finir par une

présentation des modèles linéaires mixtes dans le cadre bayésien.

2.1 Modèle bayésien

Le cadre bayésien considère les paramètres d'un modèle non plus comme des constantes mais comme des variables aléatoires. Soit Θ l'espace des paramètres et $\mathcal{Y}$ l'espace des observations. On considère le modèle statistique de densité $p(y|\theta)$ dépendant d'un vecteur de paramètres inconnus de dimension $k : \theta \in \Theta$. On dispose de n réalisations $y = (y_1, \ldots, y_n)$ issues de cette distribution.

La connaissance *a priori* sur le paramètre θ est traduite par une loi *a priori* de densité $p(\theta)$. Les paramètres de cette loi sont appelés des hyperparamètres. Le choix de la loi de probabilité et des hyperparamètres peut être influencé par de la connaissance que nous avons acquise avant que l'expérience soit réalisée. L'information sur θ est mise à jour grâce au théorème de Bayes qui permet de prendre en compte l'information apportée par les observations. On obtient ainsi la loi de θ conditionnellement aux observations, appelée loi *a posteriori* de θ de densité $p(\theta|y)$ sur laquelle s'appuie l'inférence statistique :

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{m(y)} \tag{4}$$

où $m(y)$ est la constante d'intégration, appelée également densité marginale de y donnée par :

$$m(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta. \tag{5}$$

Lorsque le dénominateur de (4) ne dépend pas de θ, il est possible de travailler avec les densités *a posteriori* connues à une constante près :

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{6}$$

On distingue deux catégories de loi *a priori*, les lois informatives qui traduisent des connaissances qu'on peut avoir avant l'expérience et les lois non informatives. Pour plus de détails sur les loi *a priori*, nous renvoyons le lecteur au livre de Robert (2006).

Lorsque la loi *a posteriori* est de la même famille que la loi *a priori* on parle alors de loi *a priori* conjuguée. Ce type d'*a priori* permet de simplifier les calculs.

Définition 1. Soit $\mathcal{M}$ une famille de lois de probabilité sur Θ. Elle est dite conjuguée par une fonction de vraisemblance $p(y|\theta)$ si pour toute loi *a priori* de $\mathcal{M}$ la distribution *a posteriori* appartient aussi à $\mathcal{M}$.

Il est aussi possible de décomposer la modélisation en plusieurs niveaux, on parle alors de modèle hiérarchique bayésien. Cela consiste à considèrer les hyperparamètres du premier niveau non plus comme des constantes mais comme des variables aléatoires sur lesquelles on va remettre des lois *a priori* .

Définition 2. Un modèle bayésien hiérarchique est un modèle statistique bayésien $(p(y|\theta), p(\theta))$, dans lequel la loi *a priori* sur θ est décomposée en plusieurs lois conditionnelles

$$p_1(\theta|\theta_1), p_2(\theta_1|\theta_2), \ldots, p_K(\theta_{K-1}|\theta_K)$$

et une loi marginale $p_{K+1}(\theta_K)$ telle que

$$p(\theta) = \int_{\Theta_1 \times \ldots \times \Theta_K} p_1(\theta|\theta_1)p_2(\theta_1|\theta_2)\ldots p_n(\theta_{K-1}|\theta_K)p_{K+1}(\theta_K)d\theta_1 \ldots d\theta_K. \tag{7}$$

Les paramètres associés à la loi *a priori* de θ_i sont appelés hyperparamètres de niveau i $(1 \leq i \leq K)$.

Une difficulté de ces modèles est qu'ils ne permettent pas, en général, un calcul explicite des estimateurs de Bayes ($\mathbb{E}[\theta|y]$) car la densité $p(\theta|y)$ calculée avec $p(\theta)$ à partir de (7) ne donne souvent pas une loi connue. Mais il est toutefois possible d'avoir accès aux lois conditionnelles complètes et donc d'utiliser un échantillonneur de Gibbs et de Metropolis-Hastings within Gibbs dans le but de faire l'inférence statistique.

2.2 Méthodes MCMC

Intéressons nous à l'évaluation de l'intégrale suivante :

$$\mathbb{E}(h(\theta)|Y) = \int h(\theta)p(\theta|Y)d\theta < \infty \tag{8}$$

où h est une fonction mesurable quelconque. Les méthodes de Monte Carlo consistent à réaliser des simulations numériques de variables aléatoires pour obtenir une approximation d'intégrale qui converge avec le nombre de simulations. Ceci est justifié par la loi forte des grands nombres

$$\frac{1}{m}\sum_{i=1}^{m} h(\theta_i) \xrightarrow{p.s} \mathbb{E}[h(\theta)|Y] \tag{9}$$

où les θ_i, $i = 1, \ldots, m$ sont des réalisations issues de la densité *a posteriori* $p(\theta|Y)$.

Toutefois il n'est pas toujours envisageable de simuler suivant la loi *a posteriori* de $\theta|Y$. Les méthodes de Monte Carlo par chaînes de Markov (MCMC) permettent d'obtenir un échantillon $\theta_1, \ldots, \theta_m$ de la loi *a posteriori* de densité $p(\theta|Y)$ sans simuler directement suivant cette loi.

Le principe de ces méthodes repose sur la génération d'une chaîne de Markov ergodique dont la loi stationnaire est la loi *a posteriori*, appelée loi cible ou d'intérêt. Le théorème ergodique garantit la convergence presque sûre de (9). Deux types de techniques sont principalement utilisées pour créer des chaînes de Markov de loi stationnaire donnée, les algorithmes de Metropolis-Hastings et l'échantillonnage de Gibbs.

L'algorithme de Metropolis-Hastings développé initialement par Metropolis et al. (1953) puis généralisé par Hastings (1970), repose sur l'utilisation d'une loi dite instrumentale selon laquelle on peut facilement générer des variables aléatoires. Le principe est de générer des échantillons aléatoires à partir de cette loi instrumentale. Ces échantillons sont ensuite corrigés afin qu'ils se comportent asymptotiquement comme des observations aléatoires de la distribution cible ou stationnaire.

Algorithme de Metropolis-Hastings :

Notons $\theta^{(i)}$ la valeur de la chaîne de Markov à l'itération i de l'algorithme MH et $q(.|.)$ la densité de la loi instrumentale. A chaque itération, l'algorithme effectue les étapes suivantes :

1. Générer $\theta^* \sim q(\theta^*|\theta^{(i)})$,
2. Calculer $\rho(\theta^{(i)}, \theta^*) = min\left\{1, \frac{p(\theta^*|Y).q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)}|Y).q(\theta^*|\theta^{(i)})}\right\}$.

3. Prendre $\theta^{(i+1)} = \begin{cases} \theta^* \text{ avec probabilité } \rho(\theta^{(i)}, \theta^*), \\ \theta^{(i)} \text{ avec probabilité } 1 - \rho(\theta^{(i)}, \theta^*). \end{cases}$

L'intérêt majeur de l'algorithme de Metropolis-Hastings est qu'il permet de simuler suivant une loi d'intérêt de densité $p(\theta|Y)$ connue à une constante près car le calcul de la probabilité d'acceptation fait intervenir cette densité au numérateur et au dénominateur. La convergence de la chaîne de Markov vers la loi cible est théoriquement garantie mais le choix de la loi de proposition est fondamental.

L'autre algorithme MCMC couramment utilisé est l'échantillonneur de Gibbs développé par Geman and Geman (1984). Cet algorithme est très populaire parmi les méthodes MCMC grâce à la simplicité de ses calculs. Considérons le vecteur de paramètres $\theta = (\theta_1, \ldots, \theta_K)$ de dimension K de densité *a posteriori* $p(\theta|Y)$. Supposons qu'il est possible de simuler selon toutes les distributions conditionnelles complètes *a posteriori* $p(\theta_k|Y, \theta_{-k})$, $k = 1, \ldots, K$ et où θ_{-k} désigne le vecteur θ privé de la composante k. Pour simuler suivant la distribution *a posteriori* jointe $p(\theta|Y)$, l'échantillonneur de Gibbs génère les paramètres θ_k , $k = 1, \ldots, K$ itérativement suivant leurs lois conditionnelles complètes.

Algorithme d'échantillonnage de Gibbs :

Soit le vecteur de valeurs initiales $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_K^{(0)})$. $\theta^{(t)}$ sachant $\theta^{(t-1)}$ est obtenu par :

1. Générer $\theta_1^{(t)}$ suivant la distribution conditionnelle complète $p(\theta_1|Y, \theta_2^{(t-1)}, \ldots, \theta_K^{(t-1)})$,
2. Générer $\theta_2^{(t)}$ suivant la distribution conditionnelle complète $p(\theta_2|Y, \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_K^{(t-1)})$,
3. ...
4. Générer $\theta_K^{(t)}$ suivant la distribution conditionnelle complète $p(\theta_1|Y, \theta_1^{(t)}, \ldots, \theta_{K-1}^{(t)})$,
5. $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_K^{(t)})$

L'échantillonneur de Gibbs est un cas particulier de l'algorithme de Metropolis-Hastings dans lequel nous acceptons toujours ($\rho = 1$), il présente donc l'avantage de bouger dans l'espace à chaque itération. Il a également l'avantage de ne pas avoir à choisir une distribution de proposition. L'inconvénient de cette méthode est qu'elle nécessite de savoir générer suivant chacune des distributions conditionnelles complètes. Pour pallier ce problème, lorsque nous avons une distribution conditionnelle complète qui ne découle pas d'une loi connue, il est alors possible d'utiliser une étape de Metropolis-Hastings dans l'échantillonneur de Gibbs, nous parlons alors d'algorithme de Metropolis-Hastings within Gibbs.

Sous les conditions de régularité et après un nombre suffisamment important d'itérations c'est-à-dire $t > t_0$, il a été montré que ces deux méthodes permettent d'obtenir un échantillon $\theta^{(t)}$ qui converge en distribution vers la distribution cible $p(\theta|Y)$ (Robert and Casella, 2004). Ainsi les ensembles $\{\theta^{(t)}, t > t_0\}$ et $\{\theta_i^{(t)}, t > t_0, i = 1, \ldots, K\}$ dans le cas de l'échantillonneur de Gibbs, peuvent êtres considérés comme des échantillons suivant la distribution jointe $p(\theta|Y)$ et suivant les distributions marginales $p(\theta_i|Y)$ respectivement. Les premières t_0 itérations sont appelées la période de "burn-in". Lorsque cela est nécessaire, une quasi-indépendance peut être obtenue par échantillonnage par paquets, c'est-à-dire en ne prenant qu'un point de la chaîne toutes les l itérations, pour un échantillon simulé efficacement, avec par exemple $l = 5$ ou $l = 10$.

2.3 Modèle hiérarchique bayésien pour l'estimation des paramètres d'un modèle linéaire mixte

Pour illustrer les notions qui viennent d'être détaillées, nous présentons maintenant le modèle animal sous sa forme bayésienne ainsi que l'algorithme d'échantillonnage pour inférer les paramètres du modèle. Nous considérons le modèle suivant :

$$Y|\mu,\beta,u,\sigma^2 \sim N(\mu\mathbb{1} + X\beta + Zu, \sigma^2 Id), \tag{10}$$
$$\mu \propto 1, \tag{11}$$
$$\beta \sim N(0, B), \tag{12}$$
$$u|\sigma_u^2 \sim N(0, \sigma_u^2 A), \tag{13}$$
$$\sigma_u^2 \sim IG(a, b), \tag{14}$$
$$\sigma^2 \sim IG(a^*, b^*). \tag{15}$$
$$\tag{16}$$

où μ est l'intercept, $\mathbb{1}$ un vecteur unitaire de dimension n, X la matrice d'incidence associée aux effets fixes de dimension n x p, β le vecteur des coefficients des effets fixes de longueur p, Z la matrice de design associée à l'effet aléatoire de dimension n x n, u le vecteur d'effet aléatoire de longueur n, A une matrice d'apparentement n x n connue, B est une matrice connue, IG désigne la loi Inverse-Gamma "shape and rate". La distribution *a priori* sur μ est impropre et ne favorise aucune valeur. Toutes les autres distributions *a priori* sont conjuguées.

Pour faire l'inférence de ce modèle, nous devons calculer les lois conditionnelles complètes dans l'objectif d'appliquer la méthode MCMC de l'échantillonneur de Gibbs. Les détails de ces calculs sont donnés en annexe A :

$$\mu|Y,\beta,u,\sigma^2 \sim N\left(\frac{\mathbb{1}'}{n}(Y - X\beta - Zu), \frac{\sigma^2}{n}\right) \tag{17}$$
$$\beta|Y,\mu,u,\sigma^2 \sim N\left(\left(\frac{X'X}{\sigma^2} + B^{-1}\right)^{-1}\frac{X'}{\sigma^2}(Y - \mu - Zu), \left(\frac{X'X}{\sigma^2} + B^{-1}\right)^{-1}\right) \tag{18}$$
$$u|Y,\mu,\beta,\sigma^2 \sim N_n\left(\left(\frac{Z'Z}{\sigma^2} + \frac{A^{-1}}{\sigma_u^2}\right)^{-1}\frac{Z'}{\sigma^2}(Y - \mu\mathbb{1} - X\beta), \left(\frac{Z'Z}{\sigma^2} + \frac{A^{-1}}{\sigma_u^2}\right)^{-1}\right) \tag{19}$$
$$\sigma_u^2|Y,\mu,\beta,u,\sigma^2 \sim IG\left(a + \frac{n}{2}, b + \frac{u'A^{-1}u}{2}\right) \tag{20}$$
$$\sigma^2|Y,\mu,\beta,u \sim IG\left(a^* + \frac{n}{2}, b^* + \frac{1}{2}(Y - \mu\mathbb{1} - X\beta - Zu)'(Y - \mu\mathbb{1} - X\beta - Zu)\right) \tag{21}$$

Toutes ces lois conditionnelles complètes sont des lois connues, nous pouvons donc utiliser l'échantillonneur de Gibbs pour obtenir des échantillons issus des densités marginales de $\mu|Y$, $\beta|Y$, $u|Y$, $\sigma_u^2|Y$ et de $\sigma^2|Y$. Cet échantillonneur est donné par :

Algorithme d'échantillonnage de Gibbs :

Nous commençons l'algorithme à partir de valeurs initiales $\mu^{(0)}, \beta^{(0)}, u^{(0)}, \sigma_u^{2(0)}, \sigma^{2(0)}$. Étapes de l'échantillonneur de Gibbs : A l'itération $t+1$.

1. Générer $\mu^{(t+1)}|Y, \beta^{(t)}, u^{(t)}, \sigma_u^{2(t)}, \sigma^{2(t)}$ suivant 17,

2. Générer $\beta^{(t+1)}|Y, \mu^{(t+1)}, u^{(t)}, \sigma_u^{2(t)}, \sigma^{2(t)}$ suivant 18,

3. Générer $u^{(t+1)}|Y, \mu^{(t+1)}, \beta^{(t+1)}, \sigma_u^{2(t)}, \sigma^{2(t)}$ suivant 19,

4. Générer $\sigma_u^{2(t+1)}|Y, \mu^{(t+1)}, \beta^{(t+1)}, u^{(t+1)}, \sigma^{2(t)}$ suivant 20,

5. Générer $\sigma^{2(t+1)}|Y, \mu^{(t+1)}, \beta^{(t+1)}, u^{(t+1)}, \sigma_u^{2(t+1)}$ suivant 21

---

3 Sélection bayésienne d'effets fixes dans un modèle linéaire mixte

Dans cette partie, nous allons étudier les méthodes de sélection d'effets fixes dans un LMM pour identifier les marqueurs moléculaires influencent la variation d'un caractère phénotypique en présence d'information sur le pedigree. Nous allons travailler avec des marqueurs bi-alléliques de type SNP (Single Nucleotide Polymorphisme). Ce type de marqueur possède 4 états alléliques ; 0|0, 0|1, 1|0 et 1|1, que nous re-coderons en 3 niveaux : 0 pour l'état 0|0, 1 pour les états 0|1 et 1|0 et 2 pour l'état 1|1. Nous considérerons ces variables comme des variables quantitatives.

Dans ce contexte, le nombre de variables est supérieur au nombre d'observations (112 observations pour 2907 marqueurs SNPs dans le cas du palmier à huile), des méthodes de sélection appropriées doivent être mises en œuvre. Nous allons avoir recours aux méthodes de sélection bayésienne de variables. Dans ce domaine, plusieurs distributions *a priori* sur les coefficients de régression ont été développées pour sélectionner un sous-ensemble de variables pertinentes. O'Hara et al. (2009) ont passé en revue les principales méthodes : les méthodes reposant sur des *a priori* de type "Spike and Slab" (George and McCulloch, 1993, 1997), l'adaptative shrinkage (Xu, 2003) ou encore le reversible jump MCMC (Green, 1995).

La première approche est très utilisée et consiste à introduire un vecteur de variables indicatrices latentes pour identifier les variables sélectionnées. Ce type d'*a priori* suppose une loi très resserrée autour de zéro pour les coefficients associés aux variables non sélectionnées (le "Spike") et une loi plus diffuse pour les coefficients associés aux variables sélectionnées (le "Slab"). Cette partie "Slab" permet également de contrôler la dépendance entre les coefficients de régression. Dans la littérature, les lois sur le "Spike" sont soit des masses de Dirac en zéro soit des distributions normales avec de très faibles variances et la loi sur la partie "Slab" est une loi normale (George and McCulloch, 1997; Malsiner-Walli and Wagner, 2016). Sur la partie "Slab", il est aussi possible de prendre en compte une structure de corrélation par exemple le g-prior de Zellner (1986) introduit une structure de corrélation basée sur les données en utilisant la matrice d'information de Fisher.

Nous allons étudier trois variantes de l'*a priori* "Spike and Slab" avec différentes lois puis nous testerons ces modèles sur des simulations.

3.1 Méthode bayésienne de sélection de variables avec *a priori* de type "Spike and Slab"

Dans le cadre du modèle animal (3) l'approche proposée repose sur l'introduction d'un vecteur de variables indicatrices latentes dans le modèle hiérarchique (10) :

$$\gamma_j = \begin{cases} 1 \text{ si variable } j \text{ sélectionnée,} \\ 0 \text{ sinon.} \end{cases}$$

On considère que les coefficients des variables non sélectionnées sont fixés à zéro. Par la suite d_γ désignera le nombre de variables sélectionnées $\sum_{j=1}^{p} \gamma_j$.

Nous ajoutons au modèle hiérarchique (10) la distribution *a priori* suivante sur γ :

$$p(\gamma_j = 1) = \pi. \tag{22}$$

Trois *a priori* sur β seront par la suite étudiés avec différentes lois sur les parties "Spike" et "Slab". Dans un premier temps nous considérons l'*a priori* qui consiste à supposer que tous les β sont indépendants, ceux sélectionnés suivent des lois normales $N(0, \sigma_\beta^2)$ et ceux non sélectionnés suivent des lois de Dirac en zéro. Nous étudierons l'influence du paramètre de variance σ_β^2 des lois normales de la partie "Slab".

Dans un deuxième temps nous utiliserons la distribution *a priori* de Zellner (1986) sur la partie "Slab" qui suppose que les coefficients de régression sélectionnés suivent une loi normale multivariée de matrice de variance-covariance $c\sigma^2(X_\gamma' X_\gamma)^{-1}$ et des masses de Dirac en zéro pour les coefficients de régression non sélectionnés. X_γ étant la matrice X dont les colonnes correspondent aux variables sélectionnées ($\gamma = 1$). Nous étudierons notamment l'influence du paramètre c.

Dans un troisième temps, nous mettons en œuvre l'approche de George and McCulloch (1993). Cette approche suppose *a priori* que tous les β suivent une loi normale multivariée de matrice de variance-covariance $D_\gamma R D_\gamma$, avec D_γ une matrice diagonale telle que $D_\gamma = diag(a_1\tau, \ldots, a_q\tau)$,

$$a_j = \begin{cases} c \text{ si } \gamma_j = 1, \\ 1 \text{ si } \gamma_j = 0, \end{cases}$$

c et τ sont des hyperparamètres. Nous prendrons R égale soit à l'identité, soit à $(X'X)^{-1}$. Cela permet de prendre en compte ou non la structure de dépendance des variables. Nous regarderons l'influence du nombre de variables p sur les estimations des paramètres.

Dans les trois cas, nous nous intéressons aux lois *a posteriori* marginales $\mu|Y$, $\beta|Y$, $\gamma|Y$, $\sigma_u^2|Y$, $\sigma^2|Y$. Nous allons utiliser des méthodes MCMC en utilisant soit un algorithme de Gibbs soit un algorithme de Metropolis-Hastings within Gibbs selon la situation pour échantillonner suivant ces lois.

A priori 1 sur β

Nous nous plaçons dans le cas où la loi *a priori* sur la partie "Spike" est un produit de masse de Dirac en zéro. Sur la partie "Slab" nous supposons que les β sélectionnés sont indépendants. Nous notons β_γ le vecteur de paramètres pour lesquels les γ sont non nuls. L'*a priori* sur β_γ est le suivant :

$$\beta_\gamma|\gamma \sim N_{d_\gamma}(0, \sigma_\beta^2 Id_{d_\gamma}), \tag{23}$$

où σ_β^2 est un facteur d'échelle positif connu que nous fixerons à 1, 4, 10 ou 20. Les lois conditionnelles complètes de μ, u, σ_u^2 et σ^2 restent inchangées par rapport au modèle précédent et suivent respectivement les lois (17), (19), (20) et (21). Il nous faut calculer les lois conditionnelles complètes de β_γ et de γ.

Distribution conditionnelle complète de β_γ :

$$\begin{aligned}
&p(\beta_\gamma|Y,\mu,\gamma,u,\sigma_u^2,\sigma^2) \propto p(Y|\mu,\beta_\gamma,\gamma,u,\sigma_u^2,\sigma^2)p(\beta_\gamma|\gamma,\sigma^2)\\
&\quad\propto exp\Big\{-\frac{1}{2}\Big(\frac{1}{\sigma^2}(Y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)'(Y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)+\frac{1}{\sigma_\beta^2}\beta_\gamma'\beta_\gamma\Big)\Big\}\\
&\quad\propto exp\Big\{-\frac{1}{2}\Big(\beta_\gamma-\Big(\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big)^{-1}\frac{X_\gamma{}'}{\sigma^2}(Y-\mu\mathbb{1}-Zu)\Big)'\Big(\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big)\\
&\qquad\Big(\beta_\gamma-\Big(\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big)^{-1}\frac{X_\gamma{}'}{\sigma^2}(Y-\mu\mathbb{1}-Zu)\Big)\Big\}
\end{aligned}$$

On reconnait une densité gaussienne :

$$\beta_\gamma|Y,\mu,\gamma,u,\sigma_u^2,\sigma^2 \sim N_{d_\gamma}\Big(\Big(\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big)^{-1}\frac{X_\gamma{}'}{\sigma^2}(Y-\mu\mathbb{1}-Zu),\Big(\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big)^{-1}\Big) \tag{24}$$

Distribution conditionnelle complète de γ :

$$\begin{aligned}
&p(\gamma|Y,\mu,\beta_\gamma,u,\sigma_u^2,\sigma^2) \propto p(Y|\mu,\beta_\gamma,\gamma,u,\sigma_u^2,\sigma^2)p(\beta_\gamma|\gamma).p(\gamma)\\
&\quad\propto \prod_{j=1}^{p}\pi_j^{\gamma_j}(1-\pi_j)^{(1-\gamma_j)}.(2\pi)^{-\frac{d_\gamma}{2}}exp\Big\{-\frac{1}{2}\Big(\frac{1}{\sigma^2}(Y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)'(Y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)+\frac{1}{\sigma_\beta^2}\beta_\gamma'\beta_\gamma\Big)\Big\}
\end{aligned} \tag{25}$$

Toutefois nous ne pouvons utiliser la loi (25) pour générer des variables aléatoires γ car elle fait intervenir β. En effet, dans l'algorithme de Metropolis-Hastings, le candidat proposé γ^* dépendra de β^* associé or la valeur de β^* est inconnue (Baragatti, 2011). Ainsi nous allons mettre en œuvre une technique de grouping qui consiste à considérer γ et β ensemble puisqu'il y a une structure de dépendance entre les deux. On s'intéresse donc à la loi conditionnelle complète de (β_γ, γ) :

$$p(\beta_\gamma,\gamma|Y,\mu,u,\sigma_u^2,\sigma^2) \propto p(\gamma|Y,\mu,u,\sigma_u^2,\sigma^2)p(\beta_\gamma|Y,\gamma,\mu,u,\sigma_u^2,\sigma^2)$$

Ainsi, générer (β_γ, γ) suivant $\beta_\gamma, \gamma|Y,\mu,u,\sigma_u^2,\sigma^2$ revient à générer γ suivant $\gamma|Y,\mu,u,\sigma_u^2,\sigma^2$ puis β_γ suivant $\beta_\gamma|Y,\gamma,\mu,u,\sigma_u^2,\sigma^2$. Cet ordre est important pour conserver la structure de dépendance entre β_γ et γ. Nous avons déjà calculé la loi conditionnelle complète de $\beta_\gamma|Y,\gamma,\mu,u,\sigma_u^2,\sigma^2$ (24). Intéressons-nous à la loi de $\gamma|Y,\mu,u,\sigma_u^2,\sigma^2$. Pour ce faire, nous intégrons la loi $\beta_\gamma,\gamma|Y,\mu,u,\sigma_u^2,\sigma^2$ en β_γ :

$$p(\gamma|Y,\mu,u,\sigma_u^2,\sigma^2) \propto p(Y|\mu,\gamma,u,\sigma^2).p(\gamma)$$

Les détails du calcul de $p(Y|\mu,\gamma,u,\sigma^2)$ sont donnés en annexe B

$$p(Y|\mu,\gamma,u,\sigma^2) \propto \frac{\left|\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\right|^{-\frac{1}{2}}}{(\sigma_\beta^2)^{\frac{d_\gamma}{2}}} exp\Big\{-\frac{1}{2}(Y-\mu\mathbb{1}-Zu)'\Big(\frac{Id_n}{\sigma^2}-\frac{1}{(\sigma^2)^2}X_\gamma\Big(\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big)X_\gamma{}'\Big)(Y-\mu\mathbb{1}-Zu)\Big\} \tag{26}$$

La densité conditionnelle complète (26) n'est pas une loi connue, nous allons donc utiliser un algorithme de Metropolis-Hastings pour générer des variables aléatoires suivant cette loi. Pour obtenir un noyau de transition symétrique, nous proposons un γ^* correspondant à $\gamma^{(i)}$ pour lequel r éléments ont été choisis au hasard et modifiés, r reste constant au cours du temps. Nous rappelons également que l'on se place dans le cas où $p(\gamma_j=1)=\pi_j=\pi$ pour tout j, la probabilité d'acceptation se simplifie :

$$\rho(\gamma^{(i)},\gamma^*) = min\Big\{1, \frac{p(\gamma^*|Y,\mu,u,\sigma^2)}{p(\gamma^{(i)}|Y,\mu,u,\sigma^2)}\Big\}$$

avec :

$$\frac{p(\gamma^*|Y,\mu,u,\sigma^2)}{p(\gamma^{(i)}|Y,\mu,u,\sigma^2)} = \Big(\frac{\pi}{1-\pi}\Big)^{d_{\gamma^*}-d_{\gamma^{(i)}}}(\sigma_\beta^2)^{\frac{d_{\gamma^{(i)}}-d_{\gamma^*}}{2}} \frac{\left|\frac{X'_{\gamma^*}X_{\gamma^*}}{\sigma^2}+\frac{Id_{d_{\gamma^*}}}{\sigma_\beta^2}\right|^{-\frac{1}{2}}}{\left|\frac{X'_{\gamma^{(i)}}X_{\gamma^{(i)}}}{\sigma^2}+\frac{Id_{d_{\gamma^{(i)}}}}{\sigma_\beta^2}\right|^{-\frac{1}{2}}}$$
$$exp\Big\{\frac{1}{2\sigma^{2^2}}(Y-\mu\mathbb{1}-Zu)'\Big(X_{\gamma^*}\Big(\frac{X_{\gamma^*}{}'X_{\gamma^*}}{\sigma^2}+\frac{Id_{d_{\gamma^*}}}{\sigma_\beta^2}\Big)X_{\gamma^*}{}'-X_{\gamma^{(i)}}\Big(\frac{X_{\gamma^{(i)}}{}'X_{\gamma^{(i)}}}{\sigma^2}+\frac{Id_{d_{\gamma^{(i)}}}}{\sigma_\beta^2}\Big)X_{\gamma^{(i)}}{}'\Big)(Y-\mu\mathbb{1}-Zu)\Big\}$$

Maintenant que nous savons générer suivant toutes les lois conditionnelles complètes, nous pouvons implémenter un algorithme de Metropolis-Hastings within Gibbs.

A priori 2 sur β

Nous nous plaçons dans le cas où la loi *a priori* sur la partie "Spike" de β est un produit de Dirac en zéro. Sur la partie "Slab" nous prenons la loi *a priori* de Zellner (1986) qui prend en compte une structure de dépendance entre les effets fixes sélectionnés proportionnelle à la matrice d'information de Fisher :

$$\beta_\gamma|\gamma,\sigma^2 \sim N_{d_\gamma}(0, c\sigma^2(X'_\gamma X_\gamma)^{-1}), \tag{27}$$

où c est un facteur d'échelle positif connu.

Les lois conditionnelles complètes de μ, u et σ_u^2 restent inchangées par rapport au modèle précédent et suivent respectivement les lois (17), (19) et (20) mais celle de σ^2 change car σ^2 intervient dans la loi *a priori* de β_γ :

Distribution conditionnelle complète de σ^2 :

$$p(\sigma^2|Y,\mu,\beta_\gamma,\gamma,u,\sigma_u^2) \propto p(Y|\mu,\beta_\gamma,\gamma,u,\sigma_u^2,\sigma^2).p(\beta_\gamma|\gamma,\sigma^2).p(\sigma^2)$$
$$\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}}\frac{1}{(\sigma^2)^{-\frac{d_\gamma}{2}}}exp\Big\{-\frac{1}{2\sigma^2}\Big((Y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)'(Y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)+\beta_\gamma'\frac{X_\gamma{}'X_\gamma}{c}\beta_\gamma\Big)\Big\}\sigma^{2-a^*-1}exp\Big\{-\frac{b^*}{\sigma^2}\Big\}$$

On reconnait ici la densité d'une loi inverse-gamma :

$$\sigma^2|Y,\mu,\beta_\gamma,\gamma,u,\sigma_u^2 \sim IG\Big(a^*+\frac{n+d_\gamma}{2}, b^*+\frac{1}{2}\Big(||y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu||^2+\beta_\gamma'\frac{X_\gamma{}'X_\gamma}{c}\beta_\gamma\Big)\Big) \quad (28)$$

Distribution conditionnelle complète de β_γ :

$$\begin{aligned}
p(\beta_\gamma|Y,\mu,\gamma,u,\sigma_u^2,\sigma^2) &\propto p(Y|\mu,\beta_\gamma,\gamma,u,\sigma_u^2,\sigma^2).p(\beta_\gamma|\gamma,\sigma^2)\\
&\propto exp\Big\{-\frac{1}{2\sigma^2}\Big((Y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)'(Y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)+\beta_\gamma'\frac{X_\gamma{}'X_\gamma}{c}\beta_\gamma\Big)\Big\}\\
&\propto exp\Big\{-\frac{1}{2\sigma^2}\Big(\beta_\gamma-\Big(\frac{1+c}{c}X_\gamma{}'X_\gamma\Big)^{-1}X_\gamma{}'(Y-\mu\mathbb{1}-Zu)\Big)'\Big(\frac{1+c}{c}\Big)X_\gamma{}'X_\gamma\\
&\quad\Big(\beta_\gamma-\Big(\frac{1+c}{c}X_\gamma{}'X_\gamma\Big)^{-1}X_\gamma{}'(Y-\mu\mathbb{1}-Zu)\Big)\Big\}
\end{aligned}$$

On reconnait une densité gaussienne :

$$\beta_\gamma|Y,\mu,\gamma,u,\sigma_u^2,\sigma^2 \sim N_{d_\gamma}\Big(\frac{c}{c+1}\Big(X_\gamma{}'X_\gamma\Big)^{-1}X_\gamma{}'(Y-\mu\mathbb{1}-Zu), \frac{c\sigma^2}{1+c}(X_\gamma{}'X_\gamma)^{-1}\Big) \quad (29)$$

Distribution conditionnelle complète de γ :

Comme précédemment β_γ est un paramètre de nuisance pour la loi conditionnelle complète de γ. Ainsi nous allons mettre en œuvre une technique de grouping qui consiste à considérer γ et β_γ ensemble. On s'intéresse donc à la loi conditionnelle complète de (β_γ,γ) :

$$p(\beta_\gamma,\gamma|Y,\mu,u,\sigma_u^2,\sigma^2) \propto p(\gamma|Y,\mu,u,\sigma_u^2,\sigma^2).p(\beta_\gamma|Y,\gamma,\mu,u,\sigma_u^2,\sigma^2)$$

Le calcul de la densité de $\gamma|Y,\mu,u,\sigma_u^2,\sigma^2$ est donné en annexe B.

$$p(\gamma|Y,\mu,u,\sigma_u^2,\sigma^2) \propto \prod_{j=1}^{p}\pi_j^{\gamma_j}(1-\pi_j)^{(1-\gamma_j)}(2\pi\sigma^2)^{-\frac{n}{2}}.(c+1)^{-\frac{d_\gamma}{2}}exp\Big\{-\frac{1}{2\sigma^2}(Y-\mu\mathbb{1}-Zu)'(Id_n-\frac{c}{c+1}X_\gamma(X_\gamma{}'X_\gamma)^{-1}X_\gamma{}')(Y-\mu\mathbb{1}-Zu)\Big\} \quad (30)$$

Cette densité n'étant pas connue, nous allons utiliser un algorithme de Metropolis-Hastings pour simuler des variables aléatoires selon cette densité dans l'échantillonneur de Gibbs et donc utiliser un algorithme de Metropolis-Hastings within Gibbs.

Pour que la probabilité d'acceptation d'un candidat γ^* à partir de $\gamma^{(i)}$ se simplifie, nous allons prendre un noyau de transition symétrique. Pour obtenir un noyau de transition symétrique, nous proposons un γ^* correspondant à $\gamma^{(i)}$ pour lequel r éléments ont été choisis au hasard et

modifiés, r reste constant au cours du temps. Nous rappelons également que l'on se place dans le cas où $p(\gamma_j = 1) = \pi_j = \pi$ pour tout j.

Ainsi nous obtenons :

$$\frac{p(\gamma^*|Y,\mu,u,\sigma_u^2,\sigma^2).q(\gamma^{(i)}|\gamma^*)}{p(\gamma^{(i)}|Y,\mu,u,\sigma_u^2,\sigma^2).q(\gamma^*|\gamma^{(i)})} = \frac{p(\gamma^*|Y,\mu,u,\sigma_u^2,\sigma^2)}{p(\gamma^{(i)}|Y,\mu,u,\sigma_u^2,\sigma^2)}$$

$$= (c+1)^{\frac{\sum_j \gamma_j^{(i)} - \gamma_j^*}{2}} \left(\frac{\pi}{1-\pi}\right)^{\sum_j \gamma_j^* - \gamma_j^{(i)}} \frac{exp\Big\{ -\frac{1}{2\sigma^2}(Y-\mu\mathbb{1}-Zu)'(Id_n - \frac{c}{c+1}X_{\gamma^*}(X_{\gamma^{*\prime}}X_{\gamma^*})^{-1}X_{\gamma^{*\prime}})(Y-\mu\mathbb{1}-Zu)\Big\}}{exp\Big\{ -\frac{1}{2\sigma^2}(Y-\mu\mathbb{1}-Zu)'(Id_n - \frac{c}{c+1}X_{\gamma^{(i)}}(X_{\gamma^{(i)\prime}}X_{\gamma^{(i)}})^{-1}X_{\gamma^{(i)\prime}})(Y-\mu\mathbb{1}-Zu)\Big\}}$$

A priori 3 sur β

Nous allons maintenant mettre en œuvre un *a priori* de type "Stochastic Search Variable Selection" (SSVS) sur β introduit par George and McCulloch (1993) qui consiste à mettre une loi normale multivariée centrée de matrice de variance-covariance de la forme $D_\gamma R D_\gamma$ avec R égale à l'identité ou à $(X'X)^{-1}$ pour prendre en compte la structure de corrélation ou non des données et D_γ une matrice diagonale construite en fonction de γ.

$$\beta|\gamma \sim N_p(0, D_\gamma R D_\gamma), \tag{31}$$

avec :

$$D_\gamma = diag(a_1\tau, \ldots, a_q\tau)$$

$$a_j = \begin{cases} c \text{ si } \gamma_j = 1, \\ 1 \text{ si } \gamma_j = 0, \end{cases}$$

c et τ sont des hyper-paramètres. La variance de β_j sera donc de la forme $a_j^2\tau^2 R_{[j,j]}$ ainsi lorsque le β_j n'est pas sélectionné, seul le paramètre τ a de l'influence, sa valeur doit donc être choisie de telle sorte que $\tau^2 R_{[j,j]}$ soit assez faible pour $j = 1, \ldots, q$. Inversement, lorsque β_j est sélectionné, le produit $c\tau$ a de l'influence, il faut donc choisir c de telle sorte que $c^2\tau^2 R_{[j,j]}$ soit assez élevé pour $j = 1, \ldots, q$.

Les lois conditionnelles complètes de μ, β , u, σ_u^2 et σ^2 restent inchangées par rapport au modèle de la section (2.3) et suivent respectivement les lois (17), (18), (19), (20), (21) avec $B = D_\gamma R D_\gamma$ pour la structure de variance-covariance de β *a priori* .

Distribution conditionnelle complète des γ_j :

$$p(\gamma_j = 1|\beta, \gamma_{-j}) = \frac{p(\beta, \gamma_j = 1|\gamma_{-j})}{p(\beta|\gamma_{-j})} = \frac{p(\beta|\gamma_j = 1, \gamma_{-j})\pi}{p(\beta|\gamma_j = 1, \gamma_{-j})\pi + p(\beta|\gamma_j = 0, \gamma_{-j})(1-\pi)} \tag{32}$$

Ainsi, nous pouvons calculer la probabilité qu'un γ_j soit égal à 1 *a posteriori* et donc le simuler suivant une loi de Bernoulli. Dans le cas où $R = Id_q$ alors les β_j sont indépendants entre eux, la probabilité se simplifie :

$$p(\gamma_j = 1|\beta, \gamma_{-j}) = p(\gamma_j = 1|\beta_j) = \frac{p(\beta_j|\gamma_j = 1)\pi}{p(\beta_j|\gamma_j = 1)\pi + p(\beta_j|\gamma_j = 0)(1-\pi)}$$

Nous pouvons donc implémenter un algorithme de Gibbs pour atteindre les lois *a posteriori* marginales.

3.2 Simulations

Trois jeux de données ont été simulés sur la base du modèle (3) avec 246 observations et un nombre différent de régresseurs ($q = n/2, n, 5n$). Les covariables X sont simulés selon n répétitions indépendantes d'une loi normale $N_q(0, Id_q)$, u suit une loi normale $N_n(0, \sigma_u^2 A)$ avec $\sigma_u^2 = 3$ et A est une matrice d'apparentement de dimension n, ε suit une normale $N(0, \sigma^2 Id_n)$ avec $\sigma^2 = 1.5$. Nous fixons $\mu = -4$ et $\beta = (1, 2, 3, 4, 5, 0, \ldots, 0, 5, 4, 3, 2, 1)$.

Les algorithmes MCMC ont été mis en œuvre sur les trois jeux de données avec 10000 itérations. Nous prendrons un temps de "burn-in" de 2500 itérations.

Nous prenons comme paramètre pour la loi de Bernoulli sur les variables indicatrices $\pi = 0.1$ (10% des variables sont significatives *a priori*), sur la loi Inverse-Gamma de la variance de l'effet aléatoire $a = 2$, $b = 3$ (cela donne un mode de 1 et une espérance de 3) et nous prenons également les mêmes paramètres pour la loi Inverse-Gamma sur la variance résiduelle. Nous prenons pour valeurs initiales $\mu = 0$, $\sigma_u^2 = 1$, $\sigma^2 = 1$, u généré suivant une loi normale $N(0, \sigma_u^2 A)$, $\beta_\gamma = ({X_\gamma}' X_\gamma)^{-1} {X_\gamma}' Y$ et $\beta_{\bar{\gamma}} = 0$. Nous choisissons de mettre $\pi * p$ éléments de γ à un et le reste à zéro.

Pour mesurer la capacité de la méthode à sélectionner les bonnes variables explicatives, nous utiliserons le taux de faux positif FP, c'est-à-dire le nombre de γ_j tel que $\hat{p}(\gamma_j = 1|Y) > 0.5$ parmi les $\gamma_j = 0$ divisé par le nombre de $\gamma_j = 0$, autrement dit :

$$FP = \frac{\#\{j|\hat{p}(\gamma_j = 1|Y) > 0.5, \gamma_j = 0\}}{\#\{j|\gamma_j = 0\}}$$

De même nous définissons le taux de faux négatif FN :

$$FN = \frac{\#\{j|\hat{p}(\gamma_j = 1|Y) < 0.5, \gamma_j = 1\}}{\#\{j|\gamma_j = 1\}}$$

Pour mesurer l'ajustement de l'intercept et des variances, nous donnerons des estimations par la moyenne ainsi qu'un intervalle de crédibilité à 95% des paramètres μ, σ_u^2 et σ^2.

A priori 1 sur β

Nous allons mettre en œuvre le premier *a priori*. Nous allons l'appliquer sur les trois jeux de données simulées. L'algorithme de ce modèle est un Metropolis-Hastings within Gibbs. Nous proposons de changer $r = 2$ valeurs de γ à chaque étape du Metropolis-Hastings et nous fixons donc un nombre d'itérations du Metropolis-Hastings de $k = 30$ (Baragatti et al., 2011). Nous pouvons constater sur le tableau 1 que cet *a priori* donne de bonnes estimations ainsi qu'une bonne sélection des variables quel que soit le nombre de variables. Le paramètre σ_β^2 n'a que peu d'influence sur les estimations.

Nous pouvons voir sur la figure 1 les chaînes de Markov de σ^2, σ_u^2 et μ en fonction du nombre d'itérations ainsi que les estimations par la moyenne des β et des probabilités de sélection des variables pour le jeu de donnée 1 ($p = \frac{n}{2}$) en fixant $\sigma_\beta^2 = 4$ avec 10000 itérations. On constate que les chaînes convergent rapidement.

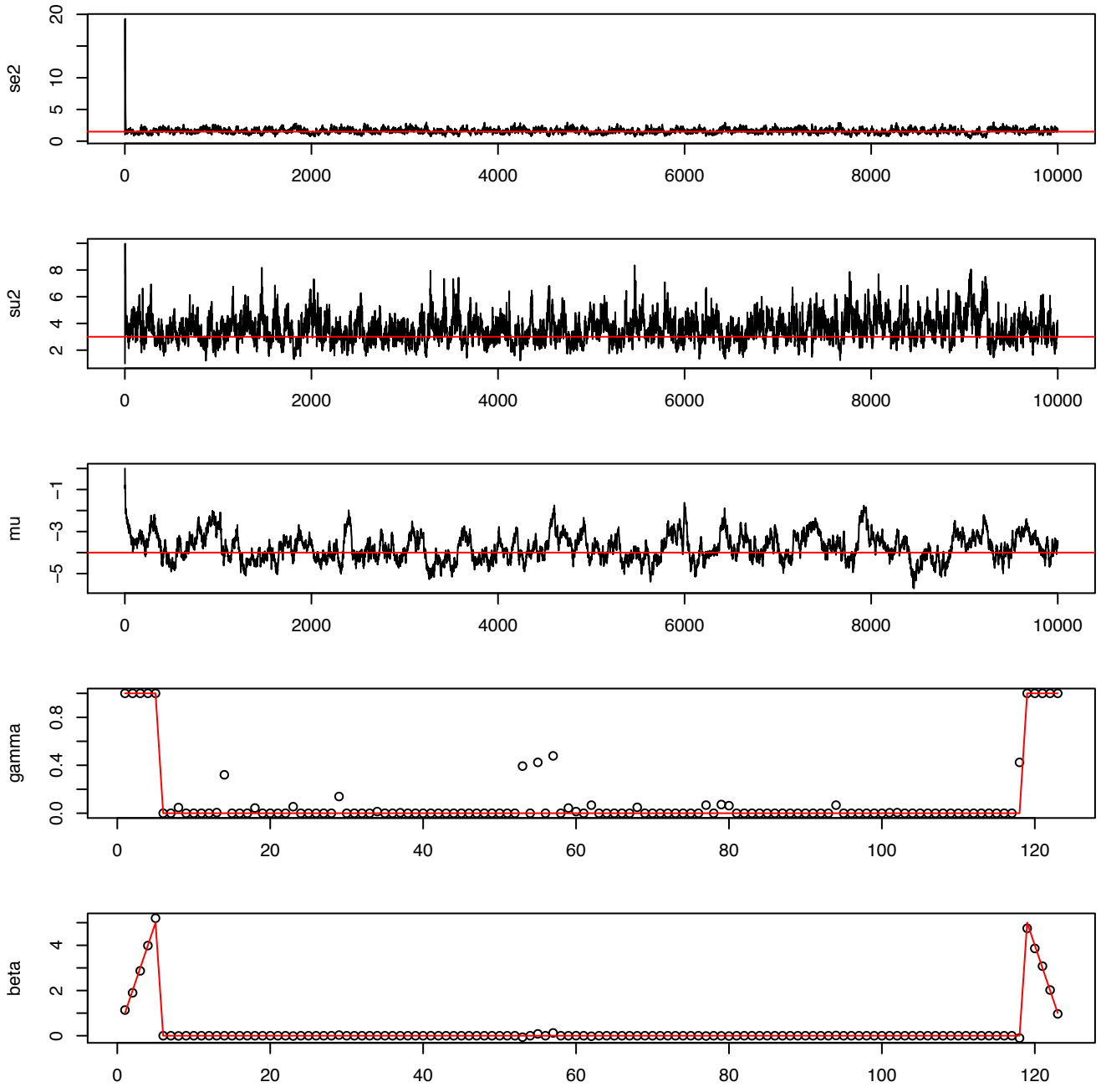

Figure 1 – *Chaîne de Markov des variances σ^2, σ_u^2 et de l'intercept μ générées par un échantillonneur de Metropolis-Hastings within Gibbs sur le modèle utilisant l'a priori 1 appliqué au jeu de données 1 avec $q = n/2$ variables et 100000 itérations ainsi que les estimations par la moyenne des probabilités a posteriori d'inclusion des variables et des coefficients de régression après une période de "burn-in" de 2500 itérations. En rouge les valeurs utilisées pour la simulation des données.*

Jeu 1 : $q = \frac{n}{2}$	$\sigma^2_\beta = 1$	$\sigma^2_\beta = 4$	$\sigma^2_\beta = 10$	$\sigma^2_\beta = 20$
μ (-4)	$-3.89, (-4.36, -3.44)$	$-3.77, (-4.23, -3.33)$	$-3.76, (-4.24, -3.29)$	$-3.5, (-3.93, -3.06)$
σ^2 (1.5)	$1.67, (1.4, 1.93)$	$1.6, (1.35, 1.83)$	$1.6, (1.32, 1.86)$	$1.68, (1.42, 1.95)$
σ^2_u (3)	$3.53, (2.79, 4.18)$	$3.59, (2.87, 4.17)$	$3.57, (2.8, 4.25)$	$3.48, (2.77, 4.09)$
FP	0	0	0.008	0
FN	0	0	0	0
Jeu 2 : $q = n$	$\sigma^2_\beta = 1$	$\sigma^2_\beta = 4$	$\sigma^2_\beta = 10$	$\sigma^2_\beta = 20$
μ (-4)	$-4.56, (-5.04, -4.05)$	$-4.72, (-5.18, -4.24)$	$-4.63, (-5.05, -4.24)$	$-4.66, (-5.12, -4.18)$
σ^2 (1.5)	$1.69, (1.43, 1.93)$	$1.78, (1.52, 2.02)$	$1.78, (1.52, 2.03)$	$1.78, (1.53, 2.02)$
σ^2_u (3)	$3.29, (2.6, 3.88)$	$3.15, (2.46, 3.7)$	$3.19, (2.53, 3.74)$	$3.27, (2.65, 3.84)$
FP	0	0	0	0
FN	0	0	0	0
Jeux 3 : $q = 5n$	$\sigma^2_\beta = 1$	$\sigma^2_\beta = 4$	$\sigma^2_\beta = 10$	$\sigma^2_\beta = 20$
μ (-4)	$-2.65, (-3.08, -2.25)$	$-2.84, (-3.29, -2.44)$	$-2.91, (-3.36, -2.45)$	$-2.9, (-3.3, -2.49)$
σ^2 (1.5)	$1.26, (0.99, 1.51)$	$1.69, (1.35, 2)$	$1.8, (1.46, 2.13)$	$1.8, (1.43, 2.18)$
σ^2_u (3)	$3.05, (2.33, 3.68)$	$3.3, (2.42, 4)$	$3.28, (2.42, 3.98)$	$3.39, (2.41, 4.13)$
FP	0.002	0	0	0
FN	0	0	0	0

Table 1 – *Estimation des paramètres μ, σ^2 et σ^2_u par la moyenne des échantillons issus des lois a posteriori générées par l'échantillonneur de Metropolis-Hastings within Gibbs du modèle utilisant l'a priori 1 ainsi que le taux de faux positifs et de faux négatifs pour différents réglages du paramètre c sur les 3 jeux de données avec une période de "burn-in" de 2500 itérations.*

***A priori* 2 sur β**

Nous allons ajuster le modèle utilisant des masses de Dirac en zéro pour la partie "Spike" et l'*a priori* de Zellner (1986) pour la partie "Slab" sur les trois jeux de données simulées et faire varier le paramètre c. L'algorithme de ce modèle est un Metropolis-Hastings within Gibbs. Nous proposons de changer $r = 2$ valeurs de γ pour chaque étape du Metropolis-Hastings et nous fixons donc un nombre d'itérations du Metropolis-Hastings de $k = 30$.

Sur le tableau 2, nous constatons que le paramètre c a une réelle influence sur l'estimation des lois *a posteriori* marginales. Nous constatons que plus le paramètre c est élevé plus la variance résiduelle est faible et la variance de l'effet aléatoire est élevée. Le choix de ce paramètre est délicat, il est possible de mettre une loi *a priori* sur ce paramètre (Celeux et al., 2006).

Jeu 1 : $q = \frac{n}{2}$	$c = 10$	$c = 100$	$c = 1000$	$c = 10000$
μ (-4)	$-3.94, (-4.33, -3.55)$	$-4.52, (-4.86, -4.18)$	$-4.7, (-5.18, -4.3)$	$-4.48, (-4.99, -3.96)$
σ^2 (1.5)	$15.68, (14.64, 16.67)$	$3.92, (3.64, 4.2)$	$1.83, (1.6, 2.03)$	$1.14, (0.9, 1.35)$
σ_u^2 (3)	$0.98, (0.7, 1.17)$	$1.36, (1.06, 1.62)$	$2.79, (2.26, 3.27)$	$4.35, (3.58, 5.06)$
FP	0	0	0	0.017
FN	0.1	0	0	0
Jeu 2 : $q = n$	$c = 10$	$c = 100$	$c = 1000$	$c = 10000$
μ (-4)	$-3.02, (-3.43, -2.61)$	$-5.32, (-5.62, -5.01)$	$-5.19, (-5.55, -4.82)$	$-5.32, (-5.76, -4.84)$
σ^2 (1.5)	$13.11, (12.25, 13.86)$	$3.56, (3.28, 3.76)$	$1.9, (1.72, 2.08)$	$1.01, (0.18, 1.53)$
σ_u^2 (3)	$1.34, (0.99, 1.64)$	$1.16, (0.89, 1.36)$	$1.87, (1.49, 2.19)$	$4.28, (2.62, 6.37)$
FP	0	0.008	0.008	0.008
FN	0.1	0	0	0
Jeux 3 : $q = 5n$	$c = 10$	$c = 100$	$c = 1000$	$c = 10000$
μ (-4)	$-5.99, (-6.09, -5.89)$	$-5.05, (-5.11, -5)$	$-2.3, (-2.32, -2.27)$	$-2.31, (-2.31, -2.31)$
σ^2 (1.5)	$12.73, (12.48, 12.99)$	$1.31, (1.31, 1.32)$	$0.13, (0.12, 0.13)$	$0.03, (0.03, 0.03)$
σ_u^2 (3)	$1.9, (1.63, 2.17)$	$76.86, (76.71, 77)$	$74.17, (74.13, 74.2)$	$158.92, (154.2, 163.63)$
FP	0.04	0.12	0.12	0.1
FN	0	0.6	0.6	0.8

Table 2 – *Estimation des paramètres μ, σ^2 et σ_u^2 par la moyenne des échantillons issus des lois a posteriori générées par l'échantillonneur de Metropolis-Hastings within Gibbs du modèle utilisant l'a priori 2 ainsi que le taux de faux positifs et de faux négatifs pour différents réglages du paramètre c sur les 3 jeux de données avec une période de "burn-in" de 2500 itérations.*

A priori 3 sur β dans le cas indépendant

Nous allons mettre en œuvre l'*a priori* 3 indépendant où tous les β sont indépendants et suivent une loi normale avec une faible variance pour la partie "Spike" et une variance plus élevée pour la partie "Slab". Dans ce modèle, nous devons choisir les paramètres c et τ de telle sorte que τ soit faible car c'est la variance de la partie "Spike" et que le produit $c\tau$ soit assez élevé car c'est la variance de la partie "Slab". Nous choisissons de prendre $c = 100$ et $\tau = 0.04$ ce qui nous donne une variance de 0.04 pour la partie "Spike" et une variance de 4 pour la partie "Slab".

Sur les jeux de données 1 ($p = n/2$) et 2 ($p = n$), cet *a priori* permet à la chaîne de Markov de converger rapidement vers les lois stationnaires. Sur la figure 2, on peut voir les chaînes de Markov en sorties de l'algorithme de Gibbs ainsi que les estimations par la moyenne des β et des γ sur 10000 itérations après une période de "burn-in" de 2500 itérations. Nous obtenons comme estimation de μ : -4.44, (-4.97, -3.92), pour σ^2 : 1.06, (0.79, 1.29), et pour σ_u^2 : 4.62, (3.82, 5.34).

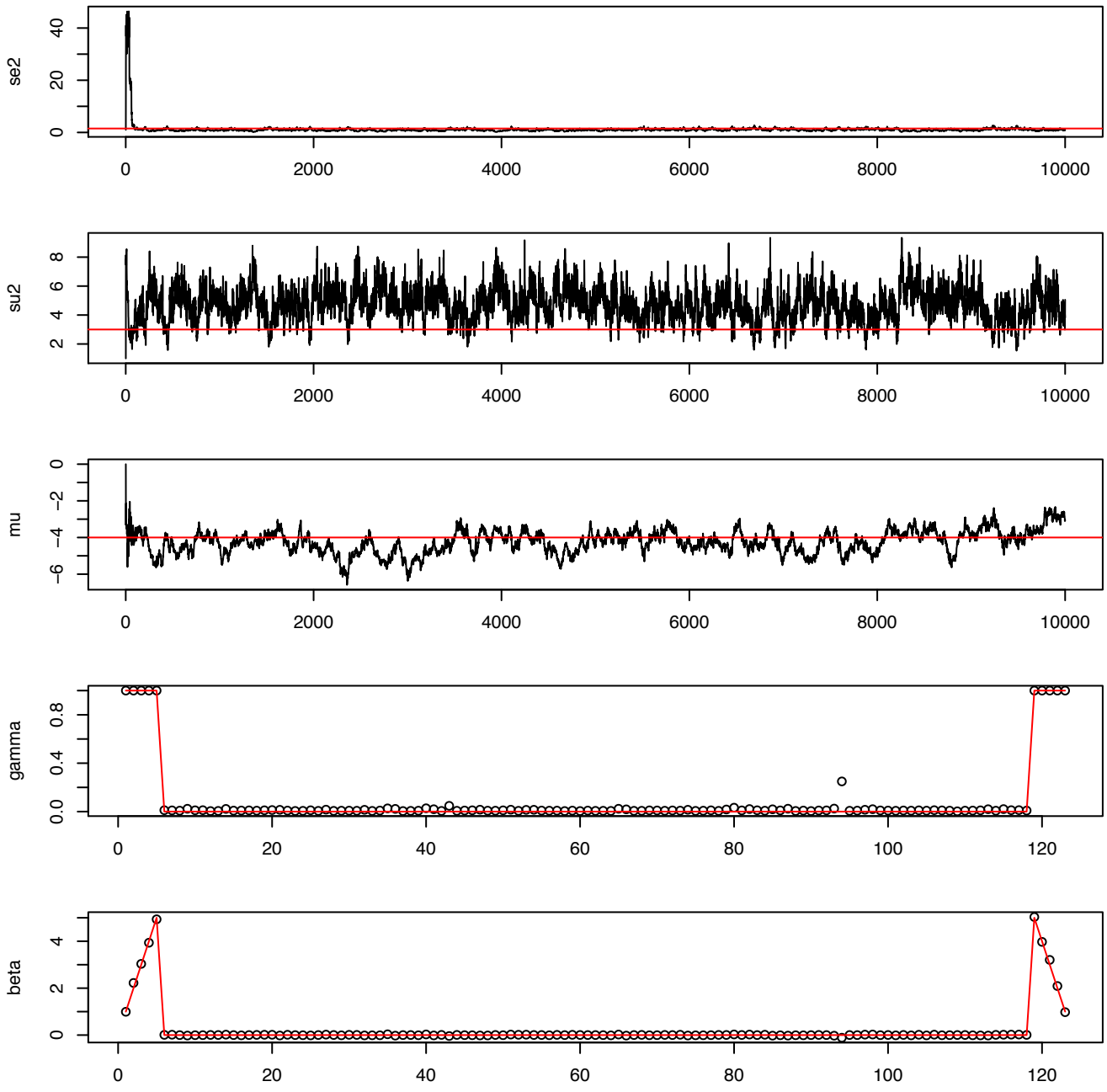

Figure 2 – *Chaînes de Markov des variances et de l'intercept générées par un échantillonneur de Gibbs sur le modèle utilisant l'a priori 3 de type SSVS indépendant appliqué au jeu de données 1 avec $q = n/2$ variables et 100000 itérations ainsi que les estimations des moyennes des probabilités a posteriori d'inclusion des variables et des coefficients de régression après une période de "burn-in" de 2500 itérations et en ne gardant qu'une itération sur trois.*

Sur le jeu de données 3 ($p = 5n$), nous constatons une mauvaise estimation de σ^2 dûe à une sur-estimation causée par un nombre de variables supérieur au nombre d'observations. La variance σ^2 tombe à zéro et n'en bouge quasiment plus.

A priori 3 sur β avec structure de dépendance

L'*a priori* 3 sur β avec une structure de corrélation basée sur l'information de Fisher ne peut être utilisé dans le cas où $p > n$ car il fait intervenir le calcul de l'inverse de la matrice $X'X$ or cette dernière n'est pas inversible.

Discussion sur le choix des paramètres et *a priori*

L'*a priori* 3 qui consiste à mettre des lois normales sur la partie "Spike" ainsi que sur la partie "Slab" ne permet pas de travailler avec un nombre de variables supérieur au nombre d'observations. Cela est dû au fait que ce modèle associe une loi normale à tous les coefficients de régression.

L'*a priori* 2 qui consiste à mettre des masses des Dirac en zéro sur la partie "Spike" et une loi normale sur la partie "Slab" avec une structure de variance-covariance proportionnelle à l'information de Fisher permet de travailler avec plus de variables que d'observations. Cependant ce *prior* permet de sélectionner au plus un nombre de variables égal au nombre d'observations, ceci est dû à la nécessité d'inverser la matrice $X'_\gamma X_\gamma$ (Baragatti, 2011). De plus, cette modélisation est sensible au choix du coefficient c ce qui confirme les résultats obtenus par Celeux et al. (2006). Enfin l'inversion de la matrice $X'_\gamma X_\gamma$ à chaque itération de l'échantillonneur ralenti l'algorithme.

Finalement l'*a priori* 1 qui consiste à mettre des masses de Dirac en zéro sur la partie "Spike" et des lois normales indépendantes sur la partie "Slab" permet de sélectionner un nombre supérieur de variables que d'observations. Cet *a priori* est peu sensible au choix de l'hyperparamètre de variance σ^2_β. D'un point de vue computationnel, cet algorithme, de par sa simplicité, permet d'avoir un algorithme d'échantillonnage efficace et rapide.

3.3 Application sur un jeu de données réelles

Dans le contexte de l'amélioration génétique du palmier à huile, nous allons appliquer une méthode bayésienne de sélection de variables de type "Spike and Slab" sur des données issues d'un essai génétique mis en place par le CIRAD et PalmElit (filiale du CIRAD). Nous disposons de 2907 marqueurs moléculaires de type SNP sur l'ensemble des 16 chromosomes du palmier à huile pour 112 arbres. L'objectif est de sélectionner des SNPs pertinents pour expliquer les variations du caractère phénotypique, en prenant en compte les apparentements entre les individus. Nous allons nous intéresser au chromosome 12 pour lequel nous disposons de 118 marqueurs moléculaires.

Nous choisissons d'appliquer le modèle utilisant une distribution *a priori* de type "Spike and Slab" sur les coefficients de régression β avec des masses de Dirac en zéro pour la partie "Spike" et des lois normales indépendantes sur la partie "Slab". Cet *a priori* est le moins sensible au choix du paramètre de variance σ^2_β des lois normales sur les coefficients de régression sélectionnés, il est également le plus simple à mettre en œuvre d'un point de vue computationnel et permet ainsi d'obtenir un échantillonneur plus rapide par rapport aux autres distributions *a priori* sur les coefficients de régression β.

Nous faisons 50000 itérations avec un temps de "burn-in" de 10000 itérations et nous prenons comme hyperparamètres $\pi = 0.1$ (pour la loi *a priori* de γ), $\sigma^2_\beta = 1$ (pour la loi *a priori* de β), $(a, b) = (1.5, 1.5)$ (pour la loi *a priori* de σ^2_u) et $(a^*, b^*) = (1.5, 1.5)$ (pour la loi *a priori* de σ^2).

Nous prenons en valeur initiale la moitié des γ à 1, une variance résiduelle à 1, la variance de l'effet aléatoire à 1, un intercept à 0 et pour les coefficients de régression sélectionnés β_γ la solution des moindres carrés.

Les estimations par la moyenne des chaînes de Markov générées par l'échantillonneur de Metropolis-Hastings within Gibbs ainsi que les intervalles de crédibilité à 95% des paramètres de variances (σ^2, σ^2_u) et de l'intercept sont donnés dans le tableau 3. Les probabilités *a posteriori* marginales de sélection des variables ($P(\gamma_i = 1|Y)$), les dix variables ayant les plus fortes probabilités sont données dans le tableau 4 ainsi que les estimations des coefficients de régression associés aux variables. Ghosh and Ghattas (2015) préconisent de s'intéresser également aux probabilités jointes du vecteur γ ($P(\gamma|Y)$), le tableau 5 donne les six modèles les plus probables.

La figure 3 présente les chaînes de Markov générées par l'algorithme de Metropolis-Hastings within Gibbs des variances σ^2 ($se2$), σ^2_u ($su2$), de l'intercept μ (mu) et des quatre coefficients de régression (β_{21} ($beta\_21$), β_{116} ($beta\_116$), β_{45} ($beta\_45$), β_{90} ($beta\_90$)) associés aux quatre variables les plus probables d'être incluses dans le modèle sur la base des probabilités marginales *a posteriori*.

Avec un seuil de sélection à 0.3 sur les probabilités *a posteriori* marginales, nous sélectionnons les variables 21, 45, 90 et 116. On retrouve ces variables dans le deuxième modèle le plus probable (voir tableau 5).

	quantile 0.025	moyenne	quantile 0.975
se2	0.338	0.611	0.999
su2	0.550	1.367	2.331
mu	16.127	17.941	19.616

Table 3 – *Estimation par la moyenne de chaînes de Markov générées par l'échantillonneur de Metropolis-Hasting within Gibbs ainsi que les intervalles de crédibilité des paramètres de variances et de l'intercept.*

Position :	21	116	45	90	27
Marqueur :	S12_3849337	S12_49023523	S12_13480737	S12_39416867	S12_5689236
Probabilité :	0.514	0.371	0.345	0.332	0.226
Estimation :	0.749	-0.569	0.803	0.885	-0.884
Position :	59	103	118	79	86
Marqueur :	S12_19933027	S12_43159345	S12_49951312	S12_29618206	S12_37773038
Probabilité :	0.221	0.216	0.194	0.189	0.178
Estimation :	-0.743	0.566	0.551	-0.807	-0.467

Table 4 – *Estimation des probabilités d'inclusion individuelles (les dix plus élevées) des variables ($P(\gamma_i = 1|Y)$) après un "burn-in" de 10000 itérations.*

Variables sélectionnées :	22 26 30 116	21 22 27 45 55 90 116	21 55 84 86 105 107 116
Probabilité du modèle :	0.00054	0.00042	0.00037
Variables sélectionnées :	21 36 79 103	1 15 45 79 94 103	6 21 50 58 63 86 98 118
Probabilité du modèle :	0.00037	0.00037	0.00034

Table 5 – *Estimation des probabilités jointes d'inclusion (les six élevées) des variables ($P(\gamma|Y)$) après un "burn-in" de 10000 itérations.*

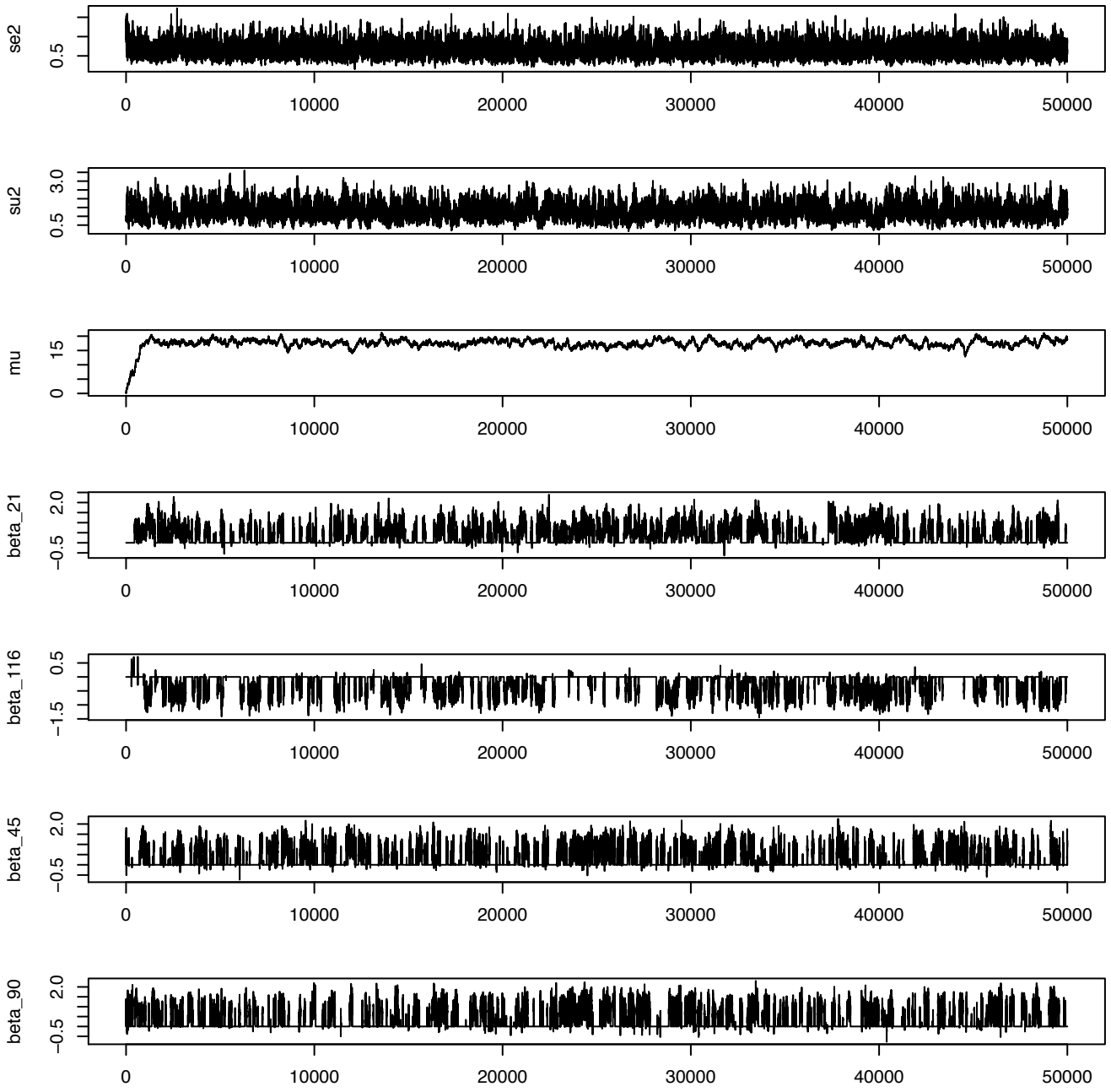

Figure 3 – *Chaînes de Markov générées par un échantillonneur de Metropolis-Hastings within Gibbs des variances, de l'intercept et des coefficients de régression des quatre variables les plus probables avec le modèle utilisant l'a priori de mélange de masses de Dirac et de lois Normale indépendantes sur les β appliqués au jeu de données réelles avec $n = 112$ observations, $q = 118$ variables et 50000 itérations.*

4 Sélection bayésienne d'effets aléatoires dans un modèle linéaire mixte

L'objectif de cette section est de proposer des méthodes de sélection bayésienne d'effets aléatoires dans le cadre du modèle linéaire mixte. Ce développement a été motivé par la décomposition de l'effet individuel global en une somme d'effets individuels sur chaque position du génome. Il en résulte un grand nombre de structures d'apparentement et donc d'effets aléatoires (292 structures) qui peut être supérieur au nombre d'observations (129 observations).

Dans le cadre de la sélection bayésienne de variable, plusieurs approches ont été développées pour l'analyse de données longitudinales par un modèle linéaire mixte. Chen and Dunson (2003), Cai and Dunson (2006), Kinney and Dunson (2007) utilisent une décomposition modifiée de Cholesky de la matrice de variance-covariance de l'effet aléatoire $\Sigma = \Lambda\Gamma\Gamma'\Lambda$ où Λ est une matrice diagonale proportionnelle à l'écart-type de l'effet aléatoire et Γ une matrice triangulaire inférieure liée à la corrélation entre réalisations de l'effet aléatoire. Ils placent sur les éléments de Λ des lois *a priori* de type "Spike and Slab" avec un mélange de masse de Dirac en zéro et d'une loi normale tronquée définie sur $\mathbb{R}^+$.

Dans notre contexte, nous ne disposons pas de données longitudinales, mais de plusieurs effets aléatoires avec une structure d'apparentement associée. En se basant sur le papier de Lu et al. (2015), nous proposons de placer des distributions *a priori* du type "Spike and Slab" sur les écarts-types associés aux effets aléatoires. Nous commençons par détailler le modèle, nous l'utiliserons ensuite pour des jeux de données simulées.

4.1 Méthode bayésien de sélection d'effets aléatoires

Soit le modèle suivant :

$$Y = \mu\mathbb{1} + U\sigma_u + \varepsilon, \qquad \varepsilon \sim N_n(0, \sigma^2 Id_n), \tag{33}$$

où $U = [u_1, \ldots, u_q]$ une matrice d'effets aléatoires pour laquelle chaque u_j suit une loi normale multivariée $N_n(0, IBD_j)$. Les matrices IBD_j sont les structures d'apparentement connues associées à q positions du génome. μ est l'intercept, $\sigma_u = (\sigma_{u_1}, \ldots, \sigma_{u_q})$ est le vecteur d'écarts-types associés à chaque effet aléatoire, chacun ayant n réalisations et une structure d'apparentement de ses réalisations et ε le vecteur des résidus. Avec cette modélisation, la sélection d'effets aléatoires se fait par la sélection des écarts-types correspondants. Sur le vecteur des écarts-types σ_u nous proposons un prior de type "Spike and Slab" défini sur $\mathbb{R}^+$ construit comme un mélange de masse de Dirac en zéro et de loi normale tronquée définie sur $\mathbb{R}^+$. Nous travaillons avec le modèle hiérarchique suivant :

$$\begin{aligned}
\mu &\sim U_{\mathbb{R}} \\
\sigma_{u_j}|\gamma_j &\sim (1-\gamma_j)\mathbb{1}_0 + \gamma_j N^+(0, \omega) \\
\gamma_j &\sim Ber(\pi) \\
\sigma^2 &\sim IG(a, b)
\end{aligned}$$

N^+ désigne une loi normale tronquée de sa partie négative, IG une loi inverse gamma de paramétrisation "shape and rate".

Soit U_γ la matrice contenant les effets aléatoires pour lesquels les γ sont non nuls, de dimension $n|\gamma|$ et σ_{u_γ} le vecteur des écarts-types associé aux γ non nuls, de longueur $|\gamma|$. La loi de Y conditionnelle complète est donc donnée par :

$$Y|\mu, U_\gamma, \sigma_{u_\gamma}, \gamma, \sigma^2 \sim N_n(\mu\mathbb{1} + U_\gamma \sigma_{u_\gamma}, \sigma^2 Id_n)$$

Un échantillonnage de Gibbs est mis en œuvre pour échantillonner selon la loi cible $p(\mu, \sigma_u, \gamma, U, \sigma^2|Y)$. Les calculs des lois *a posteriori* conditionnelles complètes sont donnés en annexe C. Nous posons : $\widetilde{Y}_j = Y - \mu\mathbb{1} - \sum_{i\neq j} u_i \sigma_{ui}$

$$\mu|Y, U_\gamma, \sigma_{u_\gamma}, \gamma, \sigma^2 \sim N\Big(\frac{\mathbb{1}'}{n}(Y - U_\gamma \sigma_{u_\gamma}), \frac{\sigma^2}{n}\Big) \tag{34}$$

$$p(\gamma_j = 1|Y, \mu, U, \gamma_{-j}, \sigma_{u-j}, \sigma^2) = \frac{p(\gamma_j = 1) 2\Big(1 + \frac{\omega}{\sigma^2} u_j' u_j\Big)^{-\frac{1}{2}} exp\Big\{\frac{1}{2\sigma^2}\widetilde{Y}_j'\Big(u_j' u_j + \frac{\sigma^2}{\omega}\Big)^{-1} u_j u_j' \widetilde{Y}_j\Big\} p(Z > 0)}{p(\gamma_j = 1) 2\Big(1 + \frac{\omega}{\sigma^2} u_j' u_j\Big)^{-\frac{1}{2}} exp\Big\{\frac{1}{2\sigma^2}\widetilde{Y}_j'\Big(u_j' u_j + \frac{\sigma^2}{\omega}\Big)^{-1} u_j u_j' \widetilde{Y}_j\Big\} p(Z > 0) + (1 - p(\gamma_j = 1))} \tag{35}$$

$$\sigma_{u\,j}|Y, \mu, U, \gamma, \sigma^2 \sim (1 - \gamma_j)\mathbb{1}_0 + \gamma_j N^+\Big(\Big(u_j' u_j + \frac{\sigma^2}{\omega}\Big)^{-1} u_j' \widetilde{Y}_j, \Big(\frac{u_j' u_j}{\sigma^2} + \frac{1}{\omega}\Big)^{-1}\Big) \tag{36}$$

$$u_j|Y, \mu, u_{-j}, \sigma_u, \sigma^2 \sim N_n\Big(\Big(\frac{{\sigma_{u_j}}^2}{\sigma^2} Id_n + IBD_j^{-1}\Big)^{-1} \frac{\sigma_{u_j}}{\sigma^2}\Big(Y - \mu\mathbb{1} - \sum_{k\neq j} u_k \sigma_{u_k}\Big), \Big(\frac{{\sigma_{u_j}}^2}{\sigma^2} Id_n + IBD_j^{-1}\Big)^{-1}\Big) \tag{37}$$

$$\sigma^2|Y, \mu, U, \sigma_u \sim IG\Big(a + \frac{n}{2}, b + \frac{||Y - \mu\mathbb{1} - U_\gamma \sigma_{u_\gamma}||^2}{2}\Big) \tag{38}$$

Nous pouvons optimiser la simulation des u_j en travaillant sur la matrice de variance-covariance et en utilisant la décomposition SVD des matrices IBD, voir en annexe C.

L'algorithme de Gibbs pour générer des variables suivant la loi cible est donné par :

Algorithme Gibbs :

Nous commençons l'algorithme à partir de valeurs initiales $\mu^{(0)}, \gamma^{(0)}, U^{(0)}, {\sigma_u}^{(0)}, {\sigma^2}^{(0)}$.
A l'itération t

1. Générer $\gamma_j^{(t)}|Y, \mu^{(t-1)}, U^{(t-1)}, {\sigma_u}^{(t-1)}, {\sigma^2}^{(t-1)}$ suivant une loi de Bernoulli de probabilité 35 pour $j = 1, .., q$,

2. Générer ${\sigma_u}^{(t)}|Y, \mu^{(t-1)}, \gamma^{(t)}, U^{(t-1)}, {\sigma^2}^{(t-1)}$ suivant 36,

3. Générer $\mu^{(t)}|Y, U^{(t-1)}, {\sigma_u}^{(t)}, \gamma^{(t)}, {\sigma^2}^{(t-1)}$ suivant 34,

4. Générer $u_j^{(t)}|Y, \mu^{(t)}, \gamma^{(t)}, {\sigma_u}^{(t)}, {\sigma^2}^{(t-1)}$ suivant 37 pour $j = 1, \ldots, q$,

5. Générer ${\sigma^2}^{(t)}|Y, \mu^{(t)}, \gamma^{(t)}, U^{(t)}, {\sigma_u}^{(t)}$ suivant 38

Nous pouvons donc mettre en œuvre ce modèle sur des simulations et tester l'influence de l'hyperparamètre ω.

4.2 Simulations

Les simulations présentées dans cette section se basent sur la méthode (4.1) et sur deux types de matrices d'apparentement (IBD_j). Une première simulation consiste à associer une matrice d'apparentement calculées à partir des SNPs issus de données réelles à chacun des 16 chromosomes. Une mesure de similarité entre ces matrices a été obtenue en calculant le coefficient RV montrant une faible similarité (RV < 0.62) entre ces matrices . Le coefficient RV est une généralisation multivarié du coefficient de corrélation (Abdi, 2007). Une deuxième simulation s'est basée sur des matrices d'apparentement calculées à partir d'information issue du pedigree et de marqueurs moléculaires. Dans ce contexte les matrices associées aux différentes positions du génome sur un même chromosome sont fortement similaires (RV entre 0.625 et 0.998).

Simulation sur des matrices d'apparentement calculées à partir SNPs

Nous allons mettre en œuvre la méthode de sélection d'effets aléatoires (4.1) sur un premier jeu de données simulées à partir de matrices d'apparentement associées à chacun des chromosomes ($q = 16$) et calculées à partir de marqueurs moléculaires de type SNP sur 192 arbres.

Pour cette première simulation, n=192 observations ont été simulées avec un intercept μ égal à 5, une variance résiduelle σ^2 de 1. Parmi les 16 matrices d'apparentement associées aux 16 chromosomes, on associe seulement un effet aux chromosomes 1 et 3, le vecteur des écarts-types associé est donc égal à $\sigma = (\sigma_{u_1}, 0, \sigma_{u_3}, 0, \ldots, 0)$ avec $\sigma_{u_1} = \sigma_{u_3}$. Différentes valeurs de σ_{u_1} et σ_{u_3} ont été testées allant de 0.7 à 2.

Pour chaque simulation 10000 itérations sont réalisées avec une période de "burn-in" de 3000 itérations. Les hyperparamètres ω, a, b et π sont fixés à 10, 2, 3, 0.1 respectivement.

Pour les valeurs initiales, l'intercept μ est fixé à 0, la variance résiduelle σ^2 à 4 et la moitié des écarts-types choisis aléatoirement sont fixés à 2.

La méthode de sélection de variable mise en œuvre permet de sélectionner les variables pertinentes (FP = FN = 0). Les résultats sont représentés sur les figures 4, 5 et 6. On peut voir sur les graphiques de la figure 4 les sorties de l'échantillonneur de Gibbs pour des écarts-types σ_{u1} et σ_{u3} simulés à 0.7.

Sur les graphiques de la figure 5 sont représentées les estimations par la moyenne et les intervalles de crédibilité à 95% des paramètres σ^2, μ et des deux écarts-types σ_{u1} et σ_{u3} en fonction de la valeur des écarts-types σ_{u1} et σ_{u3} utilisés pour simuler les données allant de 0.7 à 2. Les moyennes et les intervalles de crédibilité ont été calculés à partir des échantillons des lois *a posteriori* complètes générées par l'échantillonneur de Gibbs. Sur le graphique du haut de la figure 5, on constate que l'estimation de la variance résiduelle σ^2 à tendance à décroître lorsque les écarts-types utilisés pour la simulation des données augmentent. Sur le dernier graphique de la figure 5 , on constate que les écarts-types estimés sont toujours plus élevés que les écarts-types utilisés pour simuler les données mais que les intervalles de crédibilité contiennent bien la valeur des écarts-types utilisés pour la simulation.

Afin d'étudier l'influence du paramètre de variance ω de la loi *a priori* associée aux écarts-types sur la sélection des effets significatifs, nous avons mis en œuvre plusieurs échantillonneurs de Gibbs avec des valeurs de ω allant de 1 à 10 dans le contexte de la simulation avec des écarts-types égaux à 0.7. La figure 6 des résultats obtenus montrent que ce paramètre n'a pas

d'influence significative sur les estimations.

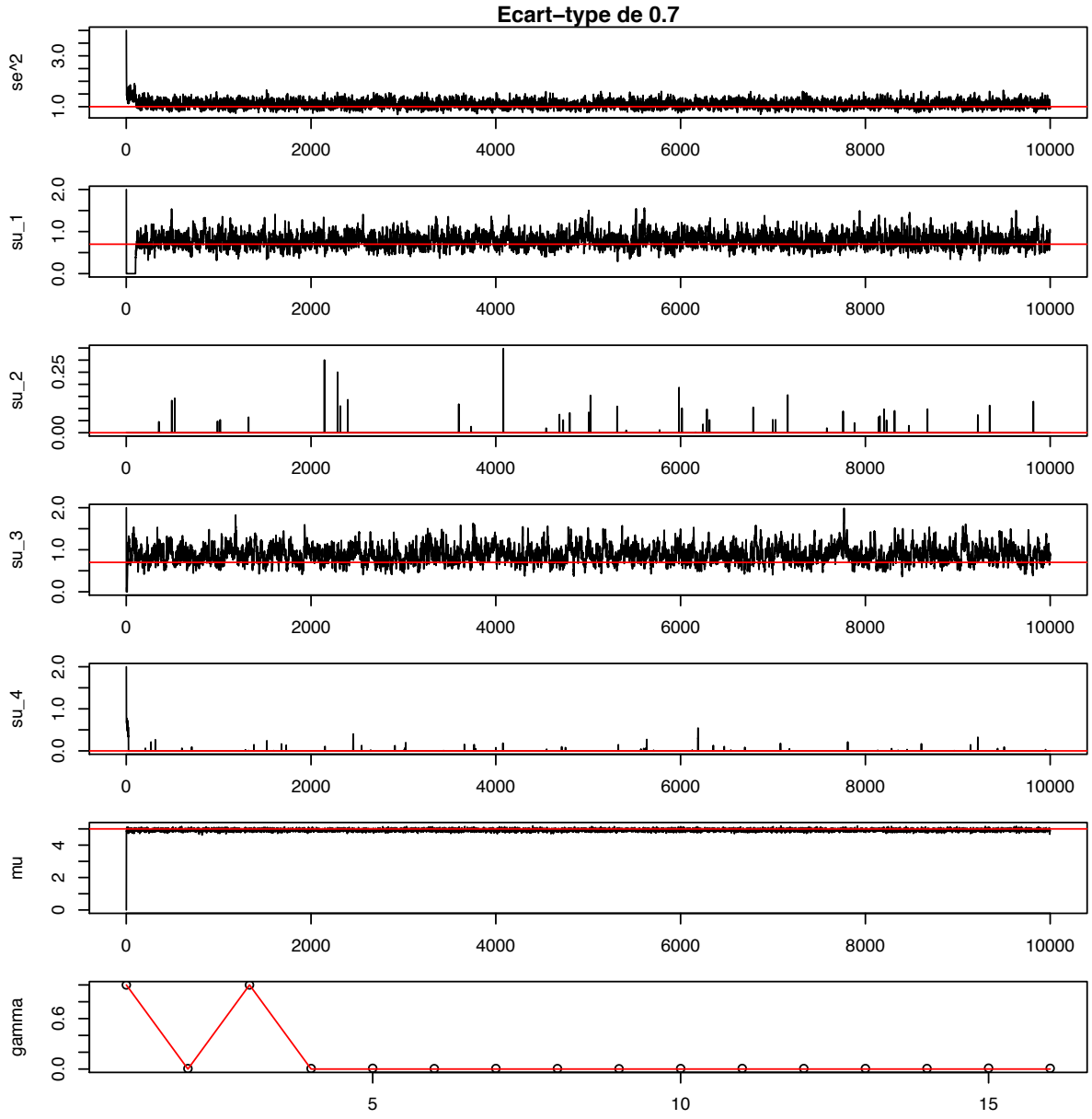

Figure 4 – *Chaînes de Markov de la variance résiduelle, des quatre premiers écarts-types et de l'intercept générées par un échantillonneurs de Gibbs de la méthode de sélection d'effets aléatoires ainsi que l'estimation de la probabilité a posteriori marginale d'inclusion des écarts-types après un "burn-in" de 3000 itérations sur des structures d'apparentement non corrélées.*

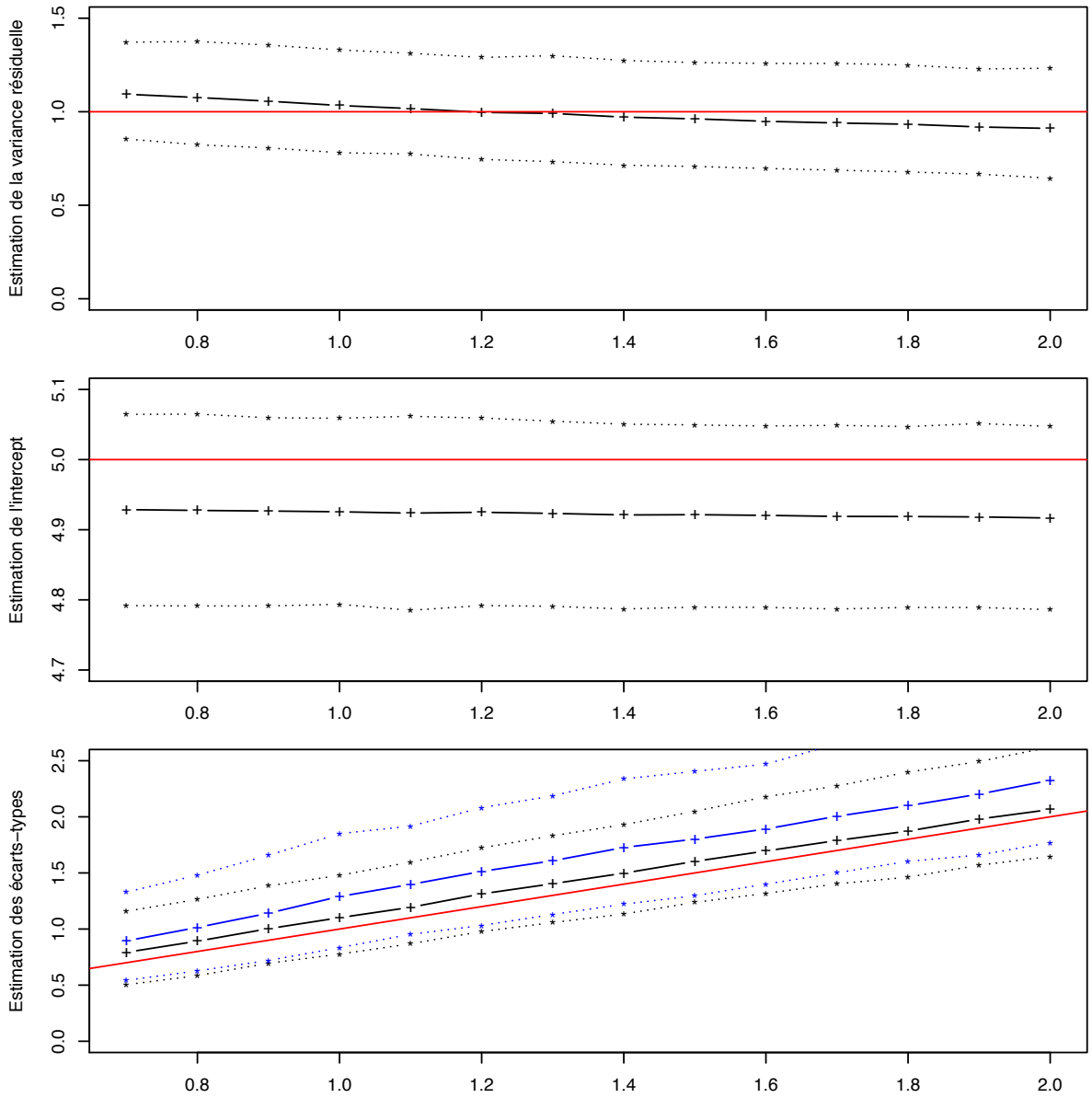

Figure 5 – *Graphiques des estimations par la moyenne avec intervalles de crédibilité à 95% de la variance résiduelle, de l'intercept et des écarts-types non nuls en fonction des écarts-types utilisés pour simuler les données* (σ_{u1}, σ_{u3}).

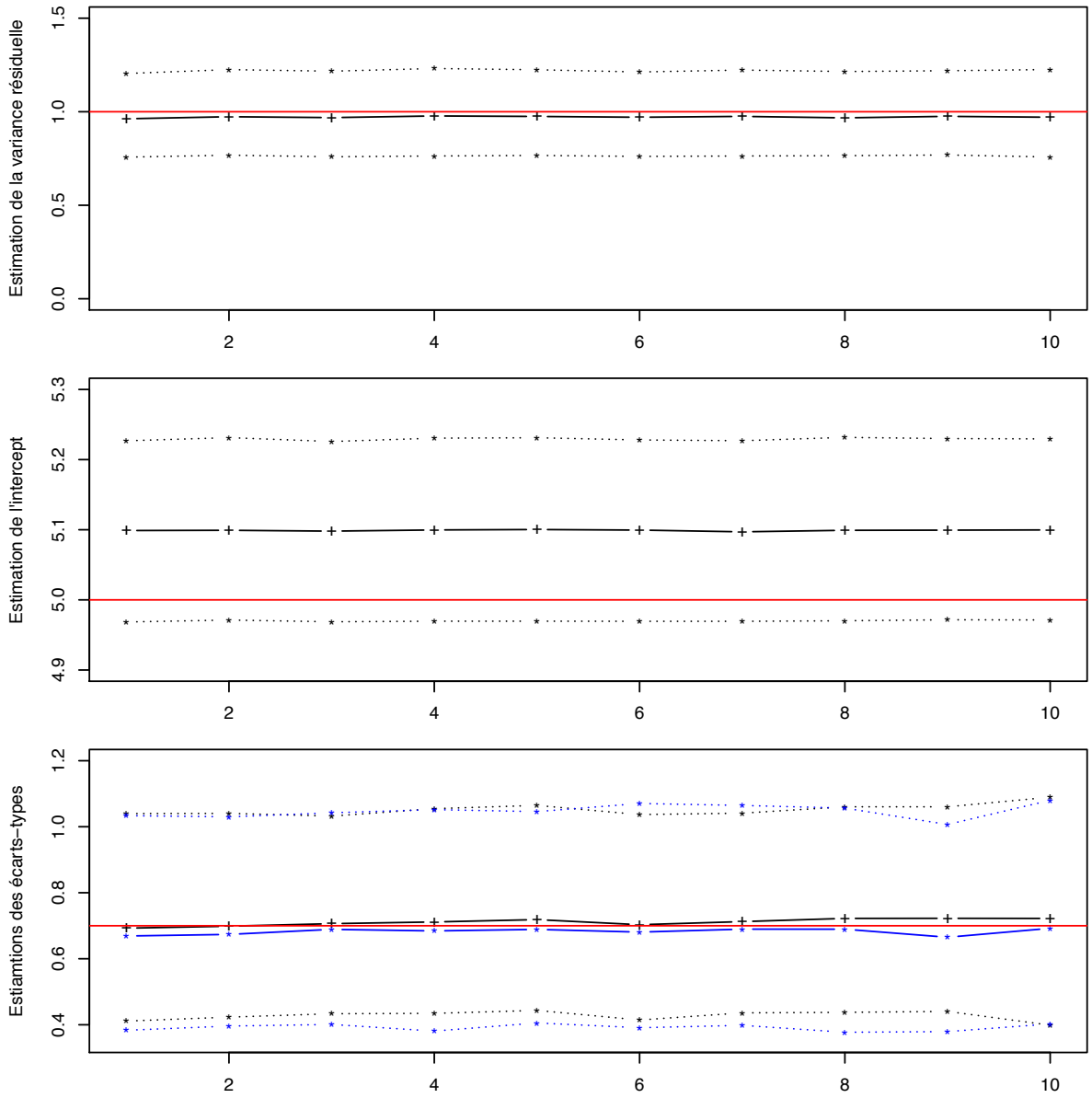

Figure 6 – *Graphiques des estimations par la moyenne avec intervalle de crédibilité à 95% de la variance résiduelle, de l'intercept et des écarts-types non nuls en faisant varier l'hyperparamètre* ω.

Simulation avec des matrices d'apparentement calculées à partir d'information pedigree et d'information moléculaire

Les simulations précédentes se basent sur 16 structures d'apparentement différentes associées aux seize chromosomes avec de faible similarité entre elles (coefficient RV < 0.62). Dans ce contexte la méthode de sélection de variables mise en œuvre permet de sélectionner les bons effets significatifs (FP = FN = 0). L'étape suivante consiste à sélectionner des effets aléatoires auxquels sont associés des matrices d'apparentement présentant de fortes similarités entre elles (coefficient RV élevé). Nous allons donc étudier le comportement de la méthode de sélection d'effets aléatoires (4.1) précédemment développée sur ce type de données.

Nous allons mettre en œuvre la méthode (4.1) sur un jeu de données simulé à partir de matrices d'apparentement associées à des positions successives du génome calculées à partir d'information pedigree et d'information moléculaire sur $n = 146$ arbres. Le coefficient RV de ces structures est compris entre 0.625 et 0.998. Nous nous concentrons uniquement sur les chromosomes 1, 2, 3, 5, 10, 12 comptabilisant $q = 103$ structures d'apparentement.

Pour cette simulation $n = 146$ observations ont été simulées avec un intercept μ égal à 18, une variance résiduelle σ^2 de 1. Parmi les 103 structures d'apparentement, on associe seulement un effet à la $19^{\text{ème}}$ structure du chromosome 1 (noté 1-19) et à la $10^{\text{ème}}$ structure du chromosome 12 (noté 12-10), le vecteur des écarts-types associé est donc égal à $\sigma = (0, \ldots, 0, \sigma_{u_{1-19}}, 0, \ldots, 0, \sigma_{u_{2-10}}, 0, \ldots, 0)$ avec $\sigma_{u_{1-19}} = 1.5$ et $\sigma_{u_{12-10}} = 1.2$.

Pour les valeurs initiales, l'intercept μ est fixé à 0, la variance résiduelle σ^2 à 1 et 30% des écarts-types choisis aléatoirement sont fixés à 1.

Les hyperparamètres ω, a, b et π sont fixés à 4, 1.5, 1.5, 0.1 respectivement. 50000 itérations ont été réalisées avec une période de "burn-in" de 10000 itérations.

Chromosome :	12-9	12-10	1-16	1-15	1-17	1-13	1-14	1-19	1-18	12-13
Probabilité :	0.45	0.35	0.28	0.21	0.21	0.20	0.18	0.14	0.13	0.10
Estimation :	0.88	0.89	0.95	0.84	0.87	0.80	0.76	0.81	0.76	0.59

Table 6 – *Estimation des probabilités a posteriori marginales d'inclusion (les dix plus élevées) des structures d'apparentement ($P(\gamma_i = 1|Y)$) après un "burn-in" de 10000 itérations.*

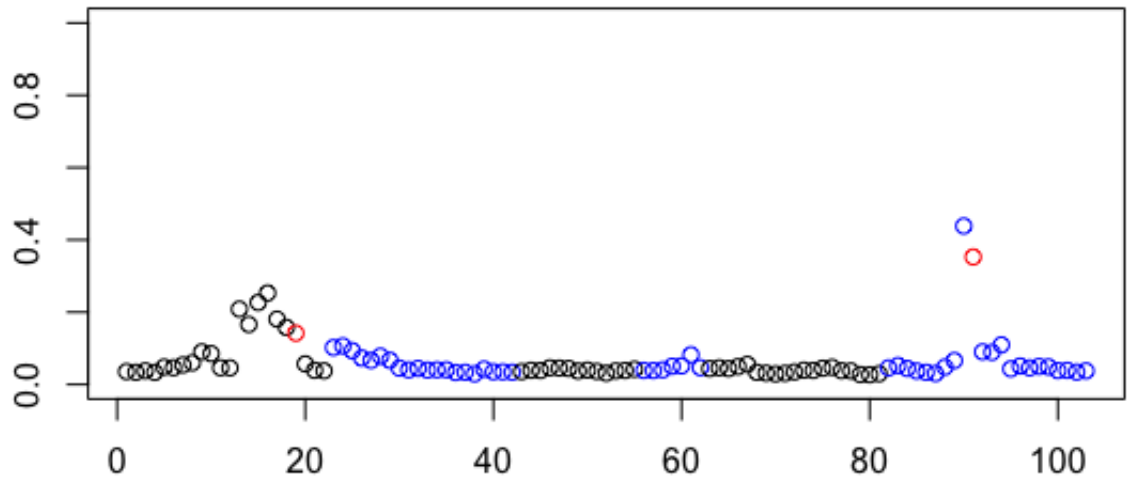

Figure 7 – *Estimation des probabilités a posteriori marginales d'inclusion des effets aléatoires avec des structures d'apparentement présentant de fortes ressemblances. En rouge les structures utilisées pour la simulation de données.*

Le tableau 6 et le graphique 7 donnent les estimations des probabilités *a posteriori* marginales d'inclusion des effets aléatoires. On constate que la méthode de sélection d'effets aléatoires détecte deux régions qui contiennent bien les variables significatives utilisées pour la simulation

(1-19, 12-10). Le tableau 7 donne les probabilités *a posteriori* jointes d'inclusion des variables. On constate que les modèles les plus probables contiennent uniquement une variable de la première région et une de la deuxième région. Les coefficients RV entre les structures d'apparentement de la première région sont donnés dans le tableau 8 et les structures de la deuxième région (12-9 et 12-10) ont un coefficient RV de 0.995. Ces coefficients sont très élevés et cela met en évidence le fait que la méthode de sélection d'effets aléatoires a des difficultés à identifier les structures significatives étant donné qu'il existe d'autres structures très similaires qui apportent quasiment la même information. L'algorithme a tendance à "switcher" entre les variables fortement similaires mais arrive à identifier les régions qui contiennent les variables significatives.

Modèle :	1-16 12-10	1-16 12-9	1-15 12-9	1-13 12-9	1-17 12-9
Probabilité :	0.00107	0.00095	0.00055	0.00055	0.00045
Modèle :	1-15 12-10	1-14 1-16 12-9	1-13 12-10	1-19 12-9	1-14 12-9
Probabilité :	0.00037	0.00035	0.00035	0.00027	0.00027

Table 7 – *Estimation des probabilités jointes d'inclusion (les dix élevées) des structures d'apparentement ($P(\gamma|Y)$) après un "burn-in" de 10.000 itérations.*

	1-13	1-14	1-15	1-16	1-17	1-18	1-19
1-13	1.00	0.99	0.96	0.94	0.92	0.91	0.82
1-14	0.99	1.00	0.98	0.97	0.95	0.94	0.85
1-15	0.96	0.98	1.00	0.99	0.98	0.97	0.89
1-16	0.94	0.97	0.99	1.00	0.99	0.98	0.92
1-17	0.92	0.95	0.98	0.99	1.00	0.99	0.95
1-18	0.91	0.94	0.97	0.98	0.99	1.00	0.97
1-19	0.82	0.85	0.89	0.92	0.95	0.97	1.00

Table 8 – *Coefficient RV entre les structures 13 à 19 du chromosome 1.*

4.3 Application sur jeu de données réelles

Nous allons appliquer la méthode de sélection d'effets aléatoires sur deux jeux de données réelles issus des essais génétiques mis en place par le CIRAD et PalmElit (filiale du CIRAD). Le premier jeu de données comporte 16 structures d'apparentement (une par chromosome) calculées à partir de marqueurs moléculaires de type SNPs, conduisant à des structures peu similaires. Nous travaillerons ensuite sur un jeu de données comportant des structures d'apparentement sur l'ensemble du génome avec des similarités plus fortes entre les structures.

Application sur des structures d'apparentement peu similaires

Nous allons appliquer la méthode (4.1) sur un premier jeu de données réelles, nous disposons de 112 individus pour lesquels nous avons accès au nombre moyen de régimes produits annuellement ainsi que qu'une structure d'apparentement par chromosomes calculée à partir de marqueurs moléculaires de type SNP ($q = 16$).

Nous effectuons 50000 itérations d'échantillonneur de Gibbs avec les hyperparamètres $\pi = 0.1$, $\omega = 4$ (variance de la loi normale tronquée sur les écarts-types sélectionnés) et $(a, b) = (1.5, 1.5)$ (paramètre de la loi inverse-gamma sur la variance résiduelle).

Nous prenons en valeurs initiales la moitié des γ à 1, une variance résiduelle à 1, les écarts-types sélectionnés à 0.7 et un intercept de 10.

La moyenne de la loi *a posteriori* ainsi que les intervalles de crédibilité à 95% de la variance résiduelle σ^2 et de l'intercept μ sont donnés dans le tableau 9. Ces estimations sont calculées sur les chaînes de Markov qui convergent en loi vers les lois *a posteriori* marginales générées par l'échantillonneur de Gibbs. Le tableau 10 donne les probabilités *a posteriori* marginales de sélection des écarts-types $P(\gamma_i = 1|Y)$ ainsi que les estimations par la moyenne des écarts-types associés. Si on fixe un seuil de sélection à 0.5 alors seul le chromosome 12 est sélectionné. Le tableau 11 donne les probabilités *a posteriori* jointes des six modèles les plus probables. On constate que le modèle le plus probable contient uniquement le chromosome 12 qui est également le seul chromosome qui ait une probabilité *a posteriori* marginale supérieure à 0.5, on en déduit que le chromosome 12 influence le plus la variation du caractère phénotypique. Les chromosomes 4, 1, 10, 8 ont également des probabilités *a posteriori* marginales non négligeables (supérieur à 0.2) et interviennent dans les six modèles les plus probables.

	quantile 0.025	moyenne	quantile 0.975
σ^2	0.392	0.733	1.155
μ	17.266	17.429	17.590

Table 9 – *Estimation des paramètres de la variance résiduelle et de l'intercept.*

Chromosome :	12	4	1	10	8	13	2	15	5	7
Probabilité :	0.599	0.465	0.450	0.318	0.270	0.179	0.073	0.055	0.044	0.035
Estimation :	1.067	0.884	0.695	0.688	0.882	0.604	0.509	0.6	0.568	0.444

Table 10 – *Estimation des probabilités a posteriori marginales d'inclusion (les dix plus élevées) des structures d'apparentement ($P(\gamma_i = 1|Y)$) après un "burn-in" de 10000 itérations.*

Chromosomes sélectionnées :	12	4, 10	8, 12	1, 12	1, 8, 13	1, 4
Probabilité du modèle :	0.0930	0.0588	0.0556	0.0555	0.0523	0.0486

Table 11 – *Estimation des probabilités jointes d'inclusion (les six élevées) des structures d'apparentement ($P(\gamma|Y)$) après un "burn-in" de 10000 itérations.*

Application sur des structures d'apparentement fortement similaires

Nous allons travailler sur les structures d'apparentement utilisées dans l'article de Tisné et al. (2015). Ces structures d'apparentement sont calculées à partir d'information moléculaire et d'information pedrigree. Elles sont associées à différentes positions tout le long du génome. Ces structures présentent de fortes similarités entre elles (coefficient RV allant jusqu'à 0.999). Nous prenons les structures d'apparentement sur l'ensemble du génome soit $q = 296$ structures pour $n = 144$ observations.

Nous effectuons 50000 itérations d'échantillonneur de Gibbs avec les hyperparamètres $\pi = 10/q$, $\omega = 4$ (variance de la loi normale tronquée sur les écarts-types sélectionnés) et $(a, b) = (1.5, 1.5)$ (paramètre de la loi inverse-gamma sur la variance résiduelle).

Nous prenons en valeur initiale 10% des γ à 1, une variance résiduelle à 2, les écarts-types sélectionnés à 1 et un intercept égal à la moyenne des observations.

Le tableau 12 donne les estimations par la moyenne ainsi que les intervalles de crédibilité à 95% de la variance résiduelle σ^2 et de l'intercept μ. Ces estimations sont calculées sur les chaînes générées par l'échantillonneur de Gibbs. Sur le tableau 13 sont données les probabilités *a posteriori* marginales de sélection des écarts-types $P(\gamma_i = 1|Y)$ ainsi que les estimations par la moyenne des écarts-types associés. Le tableau 14 donne les probabilités *a posteriori* jointes des six modèles les plus probables.

	quantile 0.025	moyenne	quantile 0.975
σ^2	0.323	0.675	1.149
μ	15.101	16.347	17.708

Table 12 – *Estimation des paramètres de la variance résiduelle et de l'intercept.*

Structure :	1-19	11-16	15-24	15-25	4-5	15-23	4-29	15-22	4-8	11-17
Probabilité :	0.413	0.341	0.283	0.236	0.223	0.175	0.174	0.145	0.145	0.141
Estimation :	1.148	0.955	1.434	1.304	0.863	1.290	1.213	1.220	0.874	0.982

Table 13 – *Estimation des probabilités a posteriori marginales d'inclusion (les dix plus élevées) des structures d'apparentement ($P(\gamma_i = 1|Y)$) après un "burn-in" de 10000 itérations.*

Structures sélectionnées :	Probabilité du modèle :
1-19, 4-29, 8-31, 11-16, 15-5, 15-25	0.00177
1-16, 5-1, 8-14, 8-31, 11-16, 12-10, 15-23	0.00160
1-18, 4-29, 8-30, 11-16, 15-10, 15-22	0.00113
1-19, 4-5, 4-30, 8-27, 15-25	0.00110
1-19, 4-8, 8-29, 11-16, 15-20	0.00097
1-19, 4-2, 4-29, 8-23, 11-16, 15-10, 15-24	0.00093

Table 14 – *Estimation des probabilités a posteriori jointes d'inclusion (les six élevées) des structures d'apparentement ($P(\gamma|Y)$) après un "burn-in" de 10000 itérations.*

Nous pouvons constater sur le tableau 13 que quatre structures d'apparentement successives sur le chromosome 15 font parties des dix variables les plus probables d'être incluses dans le modèle (15-22, 15-23, 15-24, 15-25). Nous observons sur le tableau des probabilités jointes que ces structures ne sont pas sélectionnées simultanément. Sur la figure 8, nous pouvons visualiser que l'algorithme "switch" entre ces structures d'apparentement. La méthode arrive à identifier une région du génome impliquée dans la variation du caractère phénotypique mais elle a des difficultés à identifier la structure la plus pertinente. Effectivement, sur le tableau 15 sont donnés les coefficients RV des structures d'apparentement 22, 23, 24 et 25 du chromosome 15 et nous constatons que ces coefficients sont élevés. Ce qui signifie que ces structures sont fortement similaires et qu'elles apportent quasiment la même information dans le modèle. Ainsi on retrouve le problème "switch" mis en évidence sur les simulations.

Ghosh and Ghattas (2015) préconisent de s'intéresser aux probabilités *a posteriori* marginales et jointes en présence de forte colinéarité dans le cas de la sélection d'effets fixes. Si on

applique cette recommandation alors six structures sont sélectionnées : 1-19, 4-29, 8-31, 11-16, 15-5 et 15-25. Sur le même jeu de données, Tisné et al. (2015) identifie les structures d'apparentement 1-19, 4-33, 8-25, 9-12, 11-17 et 15-25 par une méthode de sélection de variables basée sur des tests de rapport de vraisemblance. Sur la base des probabilités *a posteriori* marginales et jointes la méthode mise en place est cohérente avec les résultats obtenus dans l'article de Tisné et al. (2015). Effectivement nous sélectionnons 5 structures identiques ou très proches des six sélectionnées par Tisné et al. (2015).

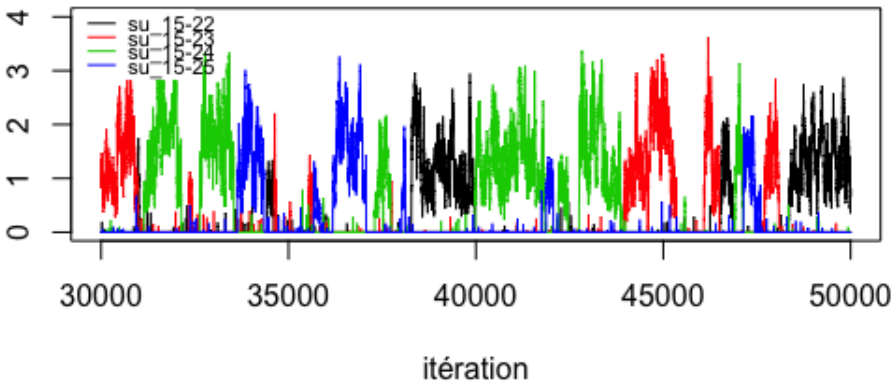

Figure 8 – *Superposition des chaînes de Markov des écarts-types associés aux structures d'apparentement 22, 23, 24 et 25 du chromosome 15.*

	15-22	15-23	15-24	15-25
15-22	1.000	0.982	0.981	0.978
15-23	0.982	1.000	0.999	0.997
15-24	0.981	0.999	1.000	0.997
15-25	0.978	0.997	0.997	1.000

Table 15 – *Coefficient RV des structures d'apparentement 22, 23, 24, 25 du chromosome 15.*

Nous pouvons constater que la méthode développée pour la sélection d'effets aléatoires avec des structures de variance-covariance connues à une constante près donne des résultats satisfaisants lorsque la similarité entre les structures d'apparentement n'est pas élevée. Dans ce contexte, les simulations montrent que la méthode permet de sélectionner et d'estimer des écarts-types avec de faibles valeurs, nous avons pu constater que cette modélisation est peu sensible au choix de l'hyperparamètre ω de la variance sur la loi des écarts-types sélectionnés.

Les simulations sur des structures d'apparentement fortement similaires montrent les limites de cette modélisation. Certaines structures apportent quasiment la même information. Ainsi cette modélisation a du mal à identifier les structures pertinentes, l'algorithme a tendance à "switcher" entre les variables ce qui a un impact sur les probabilités *a posteriori* marginales de sélection des variables. L'application sur données réelles dans le cas des structures d'apparentement fortement similaires a confirmé ce phénomène de "switch". Comme préconisé par Ghosh and Ghattas (2015), la considération des probabilités *a posteriori* jointes de sélection des va-

riables en cas de forte similarité apporte une information complémentaire et permet d'identifier un modèle.

5 Prise en compte de la similarité entre les matrices d'apparentement

Les résultats obtenus tant sur les données simulées que réelles ont mis en évidence la difficulté de sélectionner les effets aléatoires lorsque les matrices d'apparentement présentent de fortes similarités (RV $\in [0.662, 0.998]$). Cela n'est bien évidemment pas surprenant. Deux matrices d'apparentement similaires portent intrinsèquement la même information. L'objectif de cette section est de discuter d'outils qui pourraient être utilisés, adaptés ou encore développés pour pallier cette difficulté. Ce problème dans le cadre de la sélection des effets aléatoires peut s'apparenter au problème bien connu de la sélection des effets fixes en présence de fortes colinéarités entre les covariables. Cela soulève aussi une nouvelle question biologique : l'objectif est-il de sélectionner les effets (ponctuels) ou cherche-t-on à sélectionner la ou les régions (groupes) du génome impliqués dans la variation du caractère phénotypique. Ce nouvel enjeu peut orienter les choix de modélisation.

Dans le contexte de la sélection des effets fixes, au regard de ces deux nouveaux enjeux, différentes stratégies ont d'ores et déjà été proposées dans le cadre fréquentiste. Parmi ces approches, "elastic net" (Zou and Hastie, 2005) ou "group lasso" (Kyung et al., 2010). Ces deux méthodes sont une généralisation de la méthode Lasso introduite par (Tibshirani, 1996) qui, on peut le noter, bien que n'ayant pas été développées pour la gestion de la colinéarité, sont couramment utilisées et permet de sélectionner quelques représentants du groupe de variables corrélées (Kyung et al., 2010). L'eslatic net a été spécifiquement développé pour gérer la multicolinéarité tandis que le "group lasso" permet la sélection de groupes importants ainsi que les variables pertinentes dans ces groupes. Ces méthodes, Lasso, elastic net ou encore group lasso, ont été adaptées dans une approche bayésienne. Cependant, dans le cadre bayésien, elles présentent l'inconvénient de ne pas mettre à zéro les coefficients associés aux variables non sélectionnées. L'alternative que nous avons utilisée dans ce travail se fonde sur les méthodes de type "Spike and Slab". Parmi les approches de type "Spike and Slab" traitant de la colinéarité, Kwon et al. (2011) ont développé un algorithme permettant la prise en compte de corrélations élevées entre les effets fixes lors de la proposition du vecteur γ. Nous présenterons cette approche plus en détail dans la section suivante. Récemment, Ghosh and Ghattas (2015) mènent une étude qui compare différentes distributions *a priori* pour la partie "Slab". Ils concluent qu'en présence de forte colinéarité, l'utilisation de lois normales indépendantes, comme nous le proposons, se comportent particulièrement bien. Dans l'objectif de la construction de groupes, Xu et al. (2015) proposent aussi de combiner les méthodes de type "Spike and Slab" et la méthode "Bayesian group Lasso".

Finalement, dans le contexte de la sélection de régions du génome impliquées dans la variation du caractère, l'utilisation de loi *a priori* d'Ising (Li and Zhang, 2010) pourrait être envisagée. Cette loi *a priori* favorise l'inclusion de variables sur la base des relations identifiées au travers d'un graphe connu. Néanmoins, dans notre situation, les méthodes se basant sur des groupes de variables connus *a priori* est compliqué. En effet, la construction des matrices d'apparentement à partir de marqueurs moléculaires et de l'information de pedigree conduit à une forte similarité entre les matrices le long des chromosomes à des positions proches ou éloignées. Une alternative consisterait à inférer les groupes comme proposé par Peterson et al. (2016) dans le cadre du Ising prior et de la sélection d'effets fixes.

Dans cette partie, nous proposons de mettre en œuvre dans le contexte des effets aléatoires l'approche développée par Kwon et al. (2011) et de la comparer avec la méthode développée dans la partie 4.1.

5.1 L'approche Hybrid-Correlation-Based Search

La méthode Hybrid-Correlation-Based Search (H-CBS) a été développée dans un cadre de sélection d'effets fixes présentant de fortes corrélations. Cette méthode permet de tenir compte d'une structure de corrélation entre les effets fixes au travers de la loi de proposition d'un échantillonneur de Metropolis-Hastings pour générer la variable aléatoire γ. Trois mouvements sont permis : l'ajout d'une variable faiblement corrélée aux variables sélectionnées, la suppression d'une variable fortement corrélée aux autres variables sélectionnées et le troisième mouvement qui combine l'ajout d'une variable faiblement corrélée et la suppression d'une variable fortement corrélée.

Cette méthode modifie uniquement la façon de générer le vecteur de variables indicatrices latentes γ dans l'algorithme d'échantillonnage. Soient Υ la matrice de similarité des structures d'apparentement, $\rho_{i,j}$ les éléments de cette matrice, $d_\gamma = |\gamma|$ le nombre de variables sélectionnées, $\iota_\gamma = \{i : \gamma_i = 1, i = 1, \ldots, q\}$ l'ensemble des indices des variables inclues dans le modèle et $\xi_\gamma = \{i : \gamma_i = 0, i = 1, \ldots, q\}$ l'ensemble des indices des variables non sélectionnées. Si l'ajout d'une variable est choisi, l'algorithme va choisir au hasard une variable parmi celles sélectionnées $i' \in \iota_\gamma$ et inclure la variable $j' \in \xi_\gamma$ telle que la similarité entre les structures i' et j' soit minimale : $j' = \underset{j \in \xi_\gamma}{argmin} |\rho_{i',j}|$. Si le mouvement de suppression est choisi alors l'algorithme va choisir au hasard une variable parmi celles sélectionnées $i' \in \iota_\gamma$ et exclure la variable $j' \in \iota_\gamma$ telle que la similarité entre les structures i' et j' soit maximale : $j' = \underset{j \in \iota_\gamma}{argmax} |\rho_{i',j}|$. Le mouvement d'addition et de suppression va combiner les deux mouvements précédents. La loi de proposition se simplifie et est égale à :

$$q(\gamma^*|\gamma) = \begin{cases} \frac{\phi}{2d_\gamma} & \text{if } |d_\gamma - d_{\gamma^*}| = 1, \\ \frac{1-\phi}{d_\gamma} & \text{if } |d_\gamma - d_{\gamma^*}| = 0. \end{cases}$$

Notons que les auteurs proposent de reprendre la loi *a priori* utilisée par (Chipman et al., 2001) $p(\gamma) \propto \binom{p}{p_\gamma}^{-1} \frac{1}{p_\gamma}$. On peut voir cette distribution *a priori* comme un produit de loi, une loi *a priori* sur le nombre de variables sélectionnées qui favorise un nombre faible de variables ($\frac{1}{p_\gamma}$) et la probabilité de choisir un modèle contenant p_γ variables parmi les p variables ($\binom{p}{p_\gamma}^{-1}$).

Les détails de l'étape de Metropolis-Hastings pour générer γ dans l'echantillonneur de Metropolis-Hastings within Gibbs pour la sélection d'effets aléatoires sont donnés en annexe (C).

Pour assurer la convergence de la chaîne, l'algorithme combine des échantillonnages de γ sans prise en compte de la similarité avec probabilité ϕ et avec prise en compte de la similarité par la méthode "Correlation-Based Search" avec probabilité $1-\phi$. Le choix de ϕ étant arbitraire, Kwon et al. (2011) trouvent que les résultats ne sont pas fortement influencés par le choix de cette proportion et la fixe à 0.1.

5.2 Application sur données simulées

Nous reprenons les simulations de la section 4.2 avec des structures d'apparentement fortement similaires. Nous allons appliquer la méthode H-CBS et la comparer à la méthode de sélection d'effets aléatoires ne tenant pas compte de la similarité que nous proposons dans la section 4.1 pour ϕ égale à 0.5, 0.3 et 0.2.

Nous effectuons 50000 itérations avec un "burn-in" de 10000 itérations. Nous prenons comme hyperparamètres $\pi = 10/q$ (proportion de variable sélectionnée *a priori*), $(a, b) = (1.5, 1.5)$ (paramètre de l'Inverse-Gamma sur σ^2), $\omega = 4$ (variance de la loi de β_γ) et $k = 5$ (nombre d'itérations pour l'étape de Metropolis-Hastings pour générer γ).

Nous prenons en valeur initiale 10% des γ à 1, une variance résiduelle σ^2 de 2, un écart-type de 1 pour les variables sélectionnées ($\sigma_{u\gamma}$) et un intercept μ de 0.

Le graphique 9 et le tableau 16 donnent l'estimation des probabilités *a posteriori* marginales de sélection des structures d'apparentement. Nous constatons que le paramètre de proportion ϕ a une influence sur les probabilités *a posteriori* marginales de sélection. La première structure d'apparentement utilisée pour la simulation (1-19) est fortement similaire avec les structures l'entourant, on constate que la méthode 4.1 ($\phi = 1$) n'arrive pas à la détecter avec un seuil de 0.2 alors que les méthodes H-CBS arrivent à la sélectionner. La deuxième structure d'apparentement utilisée pour la simulation (12-10) et fortement similaire à la structure 12-9 (coefficient RV de 0.977), on constate alors que la méthode 4.1 ($\phi = 1$) "switch" entre les deux structures tout comme l'approche H-CBS. Le nouvel algorithme ne semble dans ce contexte pas faire mieux que le précédent.

Nous constatons que la méthode H-CBS réduit le bruit dans la sélection de variables lorsque la structure d'apparentement significative appartient à un groupe de structures fortement similaires comme pour la structure 1-19. Lorsque la structure significative est fortement similaire à une seule autre, alors le modèle H-CBS peut "switcher" entre les deux structures comme c'est le cas pour les structures 12-9 et 12-10. Toutefois, cela permet d'identifier une zone assez précise sur la position de la structure d'apparentement significative. Le tableau 17 donne le nombre de faux positifs (NFP) et de faux négatifs (NFN) pour la méthode 4.1 ($\phi = 1$) et la méthode H-CBS pour $\phi = 0.5$, 0.3 et 0.2. On constate que la méthode H-CBS permetrait de réduire le nombre de faux positifs et qu'elle ne génère pas de faux négatifs.

Le tableau 18 donne les cinq modèles ayant les probabilités *a posteriori* jointes les plus élevées suite à la mise en œuvre de la modélisation 4.1 ($\phi = 1$) et de la modélisation H-CBS pour $\phi = 0.5$, 0.3 et 0.2. On constate que seule la modélisation H-CBS arrive à retrouver le modèle contenant uniquement les structures d'apparentement utilisées pour simuler la variable réponse. Nous constatons également que la modélisation H-CBS permet d'avoir des probabilités *a posteriori* jointes plus élevées. Ceci est dû au fait que la méthode H-CBS cible l'exploration de l'espace.

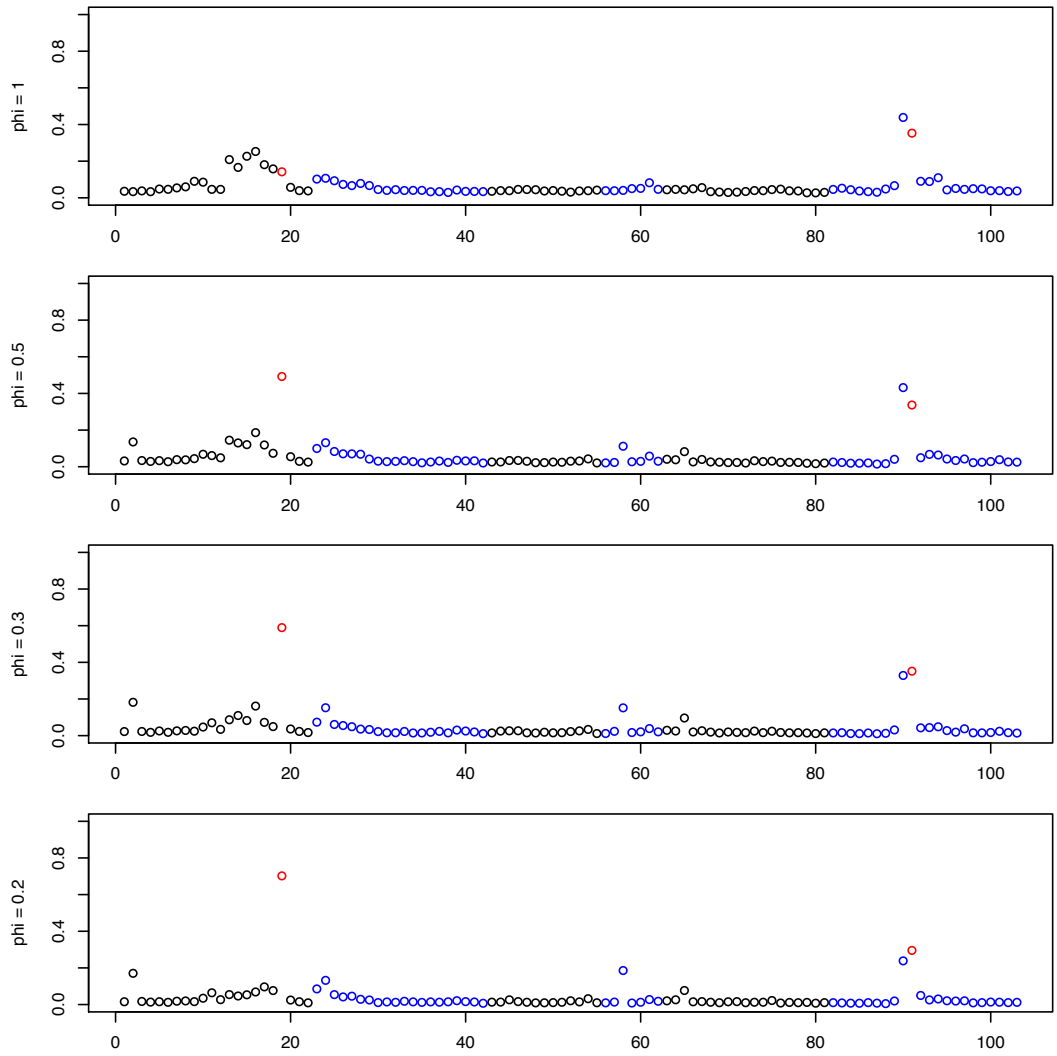

Figure 9 – *Comparaison des probabilités a posteriori marginales d'inclusion des structures d'apparentement pour différentes valeurs du paramètre de réglage ϕ (proportion de mouvement classique pour l'étape de mise à jour de γ dans l'algorithme d'échantillonnage). En rouge les positions utilisées pour simuler la variable réponse.*

$\phi = 1$	Structures :	12-9	12-10	1-16	1-15	1-13	1-17
	Probabilité	0.465	0.339	0.267	0.231	0.219	0.192
H-CBS	Structures	1-19	12-9	12-10	1-16	1-13	1-15
$\phi = 0.5$	Probabilité	0.483	0.399	0.357	0.169	0.144	0.140
H-CBS	Structures	1-19	12-10	12-9	1-2	1-16	5-3
$\phi = 0.3$	Probabilité	0.591	0.401	0.302	0.179	0.164	0.156
H-CBS	Structures	1-19	12-10	12-9	5-3	1-2	2-2
$\phi = 0.2$	Probabilité	0.667	0.278	0.243	0.183	0.176	0.128

Table 16 – *Tableau des six structures d'apparentement ayant des probabilités a posteriori marginales les plus élevées pour la méthode 4.1 ($\phi = 1$) et la modélisation H-CBS pour différentes valeurs de ϕ (0.5, 0.3 et 0.2). En rouge les structures utilisées pour la simulation de la variable réponse et en bleu les structures ayant une probabilité a posteriori marginale supérieure à 0.2.*

	$\phi = 1$	$\phi = 0.5$	$\phi = 0.3$	$\phi = 0.2$
NFP	3	1	1	1
NFN	1	0	0	0

Table 17 – *Nombre de faux positifs (NFP) et de faux négatifs (NFN) sur la sélection de variable basée sur les probabilités a posteriori marginales d'inclusion pour la méthode 4.1($\phi = 1$) et la méthode H-CBS pour $\phi = 0.5$, 0.3 et 0.2.*

$\phi = 1$	Modèle :	1-16, 12-10	1-16, 12-9	1-15, 12-9	1-13, 12-9	1-17, 12-9
	Probabilité :	0.00103	0.00087	0.00063	0.00063	0.00057
H-CBS	Modèle :	1-19, 12-9	1-19, 12-10	1-19	1-13, 12-9	1-16, 12-9
$\phi = 0.5$	Probabilité :	0.01070	0.00640	0.00560	0.00257	0.00250
H-CBS	Modèle :	1-19, 12-10	1-19	1-19, 12-9	1-16, 12-10	1-16
$\phi = 0.3$	Probabilité :	0.03880	0.03293	0.02360	0.00587	0.00513
H-CBS	Modèle :	1-19	1-19, 12-10	1-19, 12-9	1-17	1-16
$\phi = 0.2$	Probabilité :	0.10393	0.05410	0.03433	0.01293	0.00950

Table 18 – *Tableau des cinq modèles ayant les probabilités a posteriori jointes les plus élevées pour le modèle 4.1 ($\phi = 1$) et le modèle H-CBS pour différentes valeurs de ϕ (0.5, 0.3 et 0.2). En rouge le modèle utilisé pour la simulation de la variable réponse.*

5.3 Application sur données réelles

Nous reprenons le jeu de données réelles avec les structures d'apparentement fortement similaires de la section 4.3. Ce jeu de données contient $q = 296$ structures d'apparentement pour $n = 144$ observations. Nous appliquons la méthode H-CBS avec différentes valeurs de proportion ϕ égale à 0.5, 0.3 et 0.2 que nous comparons à la méthode 4.1 qui ne tient pas compte de la similarité entre les structures d'apparentement ($\phi = 1$).

Nous effectuons 50000 itérations avec un "burn-in" de 10000 itérations. Nous prenons comme hyperparamètres $\pi = 10/q$ (proportion de variable sélectionnées *a priori*), $(a, b) = (1.5, 1.5)$ (paramètre de l'Inverse-Gamma sur σ^2), $\omega = 4$ (variance de la loi de β_γ) et $k = 5$ (nombre

d'itérations pour l'étape de Metropolis-Hastings pour générer γ).

Nous prenons en valeur initiale 10% des γ à 1, une variance résiduelle σ^2 de 2, un écart-type de 1 pour les variables sélectionnées ($\sigma_{u\gamma}$) et un intercept μ égale à la moyenne de la variable réponse Y.

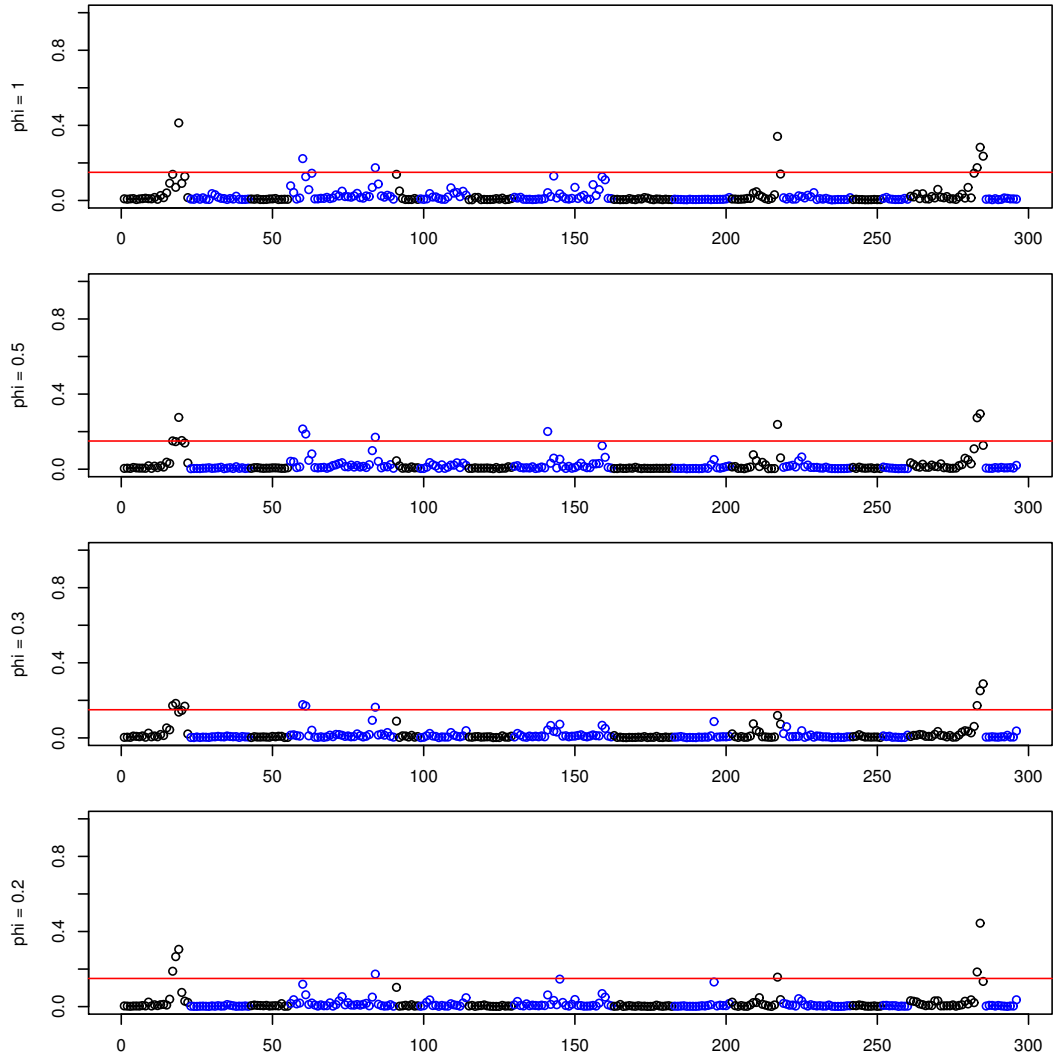

Figure 10 – *Comparaison des probilités a posteriori marginales d'inclusion des structures d'apparentement pour différentes valeurs du paramètre de proportion ϕ. En rouge le seuil de sélection de 0.15.*

Le graphique 10 et le tableau 19 donnent l'estimation des probabilités *a posteriori* marginales de sélection des structures d'apparentement. L'apport de la méthode H-CBS n'est pas évident sur les données réelles. Le "switch" entre les structures d'apparentement 22, 23, 24 et 25 du chromosome 15 mis en évidence par la méthode 4.1 qui ne prend pas en compte la similarité est également observé avec la méthode H-CBS. De plus, les probabilités *a posteriori* marginales font apparaître dans les méthodes H-CBS un "switch" entre les structures 17, 18 et 19 du chromosome 1 qui n'est pas présent avec la méthode 4.1.

$\phi = 1$	Structures :	1-19	11-16	15-24	15-25	4-5	15-23	4-29	15-22
	Probabilité :	0.413	0.341	0.283	0.236	0.223	0.175	0.174	0.145
H-CBS	Structures :	1-19	15-24	11-16	15-25	15-23	4-28	4-5	4-6
$\phi = 0.5$	Probabilité :	0.300	0.247	0.232	0.214	0.209	0.198	0.196	0.164
H-CBS	Structures :	15-25	15-24	1-17	15-23	4-5	1-18	1-19	4-6
$\phi = 0.3$	Probabilité :	0.310	0.232	0.222	0.200	0.187	0.179	0.178	0.143
H-CBS	Structures :	15-24	1-19	1-18	15-23	1-17	4-29	10-14	4-5
$\phi = 0.2$	Probabilité :	0.381	0.289	0.255	0.188	0.186	0.170	0.141	0.140

Table 19 – *Tableau des six structures d'apparentement ayant les probabilités a posteriori marginales les plus élevées pour la méthode 4.1 ($\phi = 1$) et la modélisation H-CBS pour différentes valeurs de ϕ (0.5, 0.3 et 0.2).*

	Structures sélectionnées :	Probabilité du modèle :
$\phi = 1$	1-19, 4-29, 8-31, 11-16, 15-5, 15-25	0.00177
	1-16, 5-1, 8-14, 8-31, 11-16, 12-10, 15-23	0.00160
	1-18, 4-29, 8-30, 11-16, 15-10, 15-22	0.00113
	1-19, 4-5, 4-30, 8-27, 15-25	0.00110
	1-19, 4-8, 8-29, 11-16, 15-20	0.00097
	1-19, 4-2, 4-29, 8-23, 11-16, 15-10, 15-24	0.00093
H-CBS	Structures sélectionnées :	Probabilité du modèle :
$\phi = 0.5$	1-20, 4-6, 11-9, 15-23	0.00265
	1-17, 8-12, 15-24	0.00210
	1-19, 4-6, 8-12, 15-24	0.00200
	1-20, 4-1, 4-29, 8-13, 15-23	0.00195
	1-21, 4-1, 4-21, 8-14, 11-16, 12-7, 13-1, 15-23	0.00185
	1-19, 4-29, 8-30, 11-16, 15-11, 15-17	0.00185
H-CBS	Structures sélectionnées :	Probabilité du modèle :
$\phi = 0.3$	1-17, 15-25	0.01015
	15-25	0.00880
	1-21, 4-6, 15-25	0.00455
	1-21, 4-6, 8-13, 15-24	0.00455
	1-17, 4-6, 6-4, 8-13, 11-16, 15-24	0.00450
	1-15, 15-25	0.00395
H-CBS	Structures sélectionnées :	Probabilité du modèle :
$\phi = 0.2$	1-17, 15-24	0.01280
	15-23	0.00905
	1-19, 15-23	0.00860
	15-24	0.00810
	1-18, 15-24	0.00665
	1-19, 8-16	0.00655

Table 20 – *Estimation des probabilités a posteriori jointes d'inclusion (les six élevées) des structures d'apparentement ($P(\gamma|Y)$) après un "burn-in" de 10000 itérations.*

Le tableau des probabilités *a posteriori* jointes 20 montre que le paramètre ϕ permet d'avoir un modèle plus parcimonieux. Les probabilités *a posteriori* jointes obtenues avec la méthode H-CBS sont également plus élevées comme on avait pu le constater sur les simulations. Ceci est

dû à une exploration de l'espace plus ciblée.

Pour le jeu de données simulées, la méthode H-CBS permet d'identifier les effets aléatoires significatifs avec une forte probabilité jointe *a posteriori*. On observe également que les probabilités *a posteriori* marginales de sélection augmentent lorsque la proportion ϕ diminue et le phénomène de "switch" est atténué. Les résultats concernant l'application sur données réelles sont moins concluants, au regard des probabilités *a posteriori* marginales. En effet, le phénomène de "switch" persiste avec la méthode H-CBS.

Une stratégie pourrait être envisagée en effectuant une sélection en deux étapes. Dans un premier temps, considérer les structures d'apparentement associées à chacun des chromosomes calculées à partir de marqueurs SNPs. Puis dans un deuxième temps, effectuer une sélection plus fine en utilisant les structures d'apparentement calculées à partir d'information moléculaire et d'information pedigree, appliquée uniquement sur les chromosomes sélectionnés dans la première étape. Cela permettrait de réduire le nombre de structures d'apparentement et ainsi de faciliter la sélection que ce soit pour la méthode H-CBS ou la méthode 4.1.

6 Conclusion

Ce travail a permis de mettre en œuvre des méthodes bayésiennes de sélection d'effets fixes de type "Spike and Slab" tenant compte d'un effet aléatoire. Nous avons comparé plusieurs distributions *a priori* et étudié l'influence des différents paramètres. Les modèles ont été implémentés et mis en œuvre avec le logiciel R. Des scripts R sont disponibles pour permettre aux généticiens la mise en œuvre de ces approches dans le cadre de la génétique d'association.

Lorsque l'information pedigree est couplée à de l'information moléculaire, on obtient des structures d'apparentement associées à différentes positions du génome. Afin de sélectionner ce type de variable, nous avons étendu les méthodes bayésiennes de sélection d'effets fixes de type "Spike and Slab" à la sélection d'effets aléatoires. Nous avons travaillé sur l'approche de Lu et al. (2015) en modifiant la loi *a priori* utilisée sur les coefficients d'écarts-types associés aux effets aléatoires. L'implémentation avec le logiciel R de cette méthode donne aux généticiens un outil de sélection de structures d'apparentement satisfaisant dans le cas où les structures sont peu similaires.

Dans le contexte de structures d'apparentement associées à différentes positions sur le génome, une forte similarité est observée entraînant des difficultés pour les méthodes de sélection de variable dans l'identification des variables pertinentes. Ce problème est bien connu dans le cadre de la sélection d'effets fixes, plusieurs stratégies ont été développées. Nous avons étendu la méthode H-CBS prometteuse de Kwon et al. (2011) pour la sélection d'effets aléatoires permettant de tenir compte de la similarité entre les structures d'apparentement. Bien que cette approche ait donné des résultats encourageants sur les simulations, les résultats sur les données réelles ne sont pas concluants. L'approche classique ne tenant pas compte de la similarité permet d'obtenir des résultats similaires. Ce résultat va dans le même sens que les résultats de Ghosh and Ghattas (2015) obtenus pour la sélection des effets fixes en présence de fortes colinéarités.

Pour traiter de la similarité, une perspective intéressante consisterait à utiliser des méthodes de pénalisation de type Lasso ou Elastic net pour la partie "Slab" de la loi *a priori* dans le cadre de la sélection d'effets aléatoires, comme proposé par Tibshirani (1996), Kyung et al. (2010),

Zou and Hastie (2005) pour la sélection d'effets fixes.

Toujours dans le cadre des méthodes de type "Spike and Slab", il existe des distributions *a priori* de type "Ising prior" sur le vecteur de variables indicatrices latentes γ tenant compte de groupes de variables qui sont reliées entre elles au travers d'un graphe. Cet *a priori* favorise l'inclusion de variables reliée à celles déjà sélectionnées.

Une modification de la méthode "Ising prior" pourrait être apportée pour favoriser l'inclusion de variables faiblement reliées avec celles déjà sélectionnées. Toutefois l'utilisation de distribution *a priori* de type "Ising prior" implique d'avoir accès à une matrice associée à un graphe de liaison entre les variables. Nos structures d'apparentement présentent un continuum de similarités entre elles ainsi il nous est difficile de définir des groupes *a priori* . Une perspective serait de mettre en œuvre une méthode utilisant un "Ising prior" tout en estimant le graphe de liaison (Peterson et al., 2016).

La considération de groupes de structure au niveau des chromosomes est une possibilité qui a du sens d'un point de vue biologique. Effectivement, l'application sur données réelles montre que certains chromosomes n'influencent pas la variation du caractère phénotypique. Il serait faisable de mettre en place une approche bayésienne "bi-level" pour sélectionner les chromosomes et les positions à l'interieur des chromosomes avec une modélisation de type "Spike and Slab".

Bibliographie

Abdi, H. (2007). Rv coefficient and congruence coefficient. *Encyclopedia of measurement and statistics*, pages 849–853.

Baragatti, M. (2011). *Sélection bayésienne de variables et méthodes de type Parallel Tempering avec et sans vraisemblance*. PhD thesis, Aix Marseille 2.

Baragatti, M. et al. (2011). Bayesian variable selection for probit mixed models applied to gene selection. *Bayesian Analysis*, 6(2) :209–229.

Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics*, 62(2) :446–457.

Celeux, G., Marin, J.-M., and Robert, C. (2006). Sélection bayésienne de variables en régression linéaire. *Journal de la société française de statistique*, 147(1) :59–79.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4) :762–769.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., and Stine, R. A. (2001). The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6) :721–741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423) :881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Ghosh, J. and Ghattas, A. E. (2015). Bayesian variable selection under collinearity. *The American Statistician*, 69(3) :165–173.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, pages 711–732.

Gupta, M. and Ibrahim, J. G. (2009). An information matrix prior for bayesian analysis in generalized linear models with high dimensional data. *Statistica Sinica*, 19(4) :1641.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1) :97–109.

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, 63(3) :690–698.

Kwon, D., Landi, M. T., Vannucci, M., Issaq, H. J., Prieto, D., and Pfeiffer, R. M. (2011). An efficient stochastic search for bayesian variable selection with high-dimensional correlated predictors. *Computational statistics & data analysis*, 55(10) :2807–2818.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2) :369–411.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American statistical association*, 105(491) :1202–1214.

Lu, Z.-H., Zhu, H., Knickmeyer, R. C., Sullivan, P. F., Williams, S. N., and Zou, F. (2015). Multiple snp set analysis for genome-wide association studies through bayesian latent variable selection. *Genetic epidemiology*, 39(8) :664–677.

Lynch, M., Walsh, B., et al. (1998). *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Malsiner-Walli, G. and Wagner, H. (2016). Comparing spike and slab priors for bayesian variable selection. *Austrian Journal of Statistics*, 40(4) :241–264.

McCulloch, C. E. and Neuhaus, J. M. (2001). *Generalized linear mixed models*. Wiley Online Library.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6) :1087–1092.

Mrode, R. A. (2014). *Linear models for the prediction of animal breeding values*. Cabi.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods : what, how and which. *Bayesian analysis*, 4(1) :85–117.

Peterson, C. B., Stingo, F. C., and Vannucci, M. (2016). Joint bayesian variable and graph selection for regression models with network-structured predictors. *Statistics in medicine*, 35(7) :1017–1031.

Robert, C. (2006). *Le choix bayésien : Principes et pratique*. Springer Science & Business Media.

Robert, C. and Casella, G. (2004). Monte carlo statistical methods springer. *New York*.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological information into linear models : A bayesian approach to the selection of pathways and genes. *The annals of applied statistics*, 5(3).

Stingo, F. C. and Vannucci, M. (2010). Variable selection for discriminant analysis with markov random field priors for the analysis of microarray data. *Bioinformatics*, 27(4) :495–501.

Sun, C., Madsen, P., Nielsen, U., Zhang, Y., Lund, M., and Su, G. (2009). Comparison between a sire model and an animal model for genetic evaluation of fertility traits in danish holstein population. *Journal of dairy science*, 92(8) :4063–4071.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tisné, S., Denis, M., Cros, D., Pomiès, V., Riou, V., Syahputra, I., Omoré, A., Durand-Gasselin, T., Bouvet, J.-M., and Cochard, B. (2015). Mixed model approach for ibd-based qtl mapping in a complex oil palm pedigree. *BMC genomics*, 16(1) :798.

Verbeke, G. (1997). Linear mixed models for longitudinal data. In *Linear mixed models in practice*, pages 63–153. Springer.

Wilson, A. J., Reale, D., Clements, M. N., Morrissey, M. M., Postma, E., Walling, C. A., Kruuk, L. E., and Nussey, D. H. (2010). An ecologist's guide to the animal model. *Journal of Animal Ecology*, 79(1) :13–26.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2) :789–801.

Xu, X., Ghosh, M., et al. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4) :909–936.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques : Essays in Honor of Bruno De Finetti*, 6 :233–243.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320.

Annexe A

Lois conditionnelles complètes pour l'estimation des paramètres d'un modèle linéaire mixte

Distribution conditionnelle complète de μ :

$$\begin{aligned}
p(\mu|Y,\beta,u,\sigma_u^2,\sigma^2) &\propto p(Y|\mu,\beta,u,\sigma_u^2,\sigma^2)\ p(\mu)\\
&\propto exp\Big\{ -\frac{1}{2\sigma^2}(Y-\mu\mathbb{1}-X\beta-Zu)'(Y-\mu\mathbb{1}-X\beta-Zu)\Big\}\\
&\propto exp\Big\{ -\frac{1}{2(\frac{\sigma^2}{n})}\Big(\mu^2-2\mu\frac{\mathbb{1}'}{n}(Y-X\beta-Zu)\Big)\Big\}
\end{aligned}$$

On reconnait une loi normale :

$$\mu|Y,\beta,u,\sigma_u^2,\sigma^2 \sim N\Big(\frac{\mathbb{1}'}{n}(Y-X\beta-Zu),\frac{\sigma^2}{n}\Big) \tag{A.1}$$

Distribution conditionnelle complète de u :

$$\begin{aligned}
&p(u|Y,\mu,\beta,\sigma_u^2,\sigma^2) \propto p(Y|\mu,\beta,u,\sigma^2)p(u|\sigma_u^2)\\
&\propto exp\Big\{ -\frac{1}{2\sigma^2}(Y-\mu\mathbb{1}-X\beta-Zu)'(Y-\mu\mathbb{1}-X\beta-Zu)-\frac{1}{2\sigma_u^2}u'A^{-1}u\Big\}\\
&\propto exp\Big\{ -\frac{1}{2}\Big(u'\Big(\frac{Z'Z}{\sigma^2}+\frac{A^{-1}}{\sigma_u^2}\Big)u-\frac{1}{\sigma^2}(Y-\mu\mathbb{1}-X\beta)'Zu-\frac{1}{\sigma^2}u'Z'(Y-\mu\mathbb{1}-X\beta)\Big)\Big\}\\
&\propto exp\Big\{ -\frac{1}{2}\Big(u-\Big(\frac{Z'Z}{\sigma^2}+\frac{A^{-1}}{\sigma_u^2}\Big)^{-1}\frac{Z'}{\sigma^2}(Y-\mu\mathbb{1}-X\beta)\Big)'\Big(\frac{Z'Z}{\sigma^2}+\frac{A^{-1}}{\sigma_u^2}\Big)\\
&\qquad\Big(u-\Big(\frac{Z'Z}{\sigma^2}+\frac{A^{-1}}{\sigma_u^2}\Big)^{-1}\frac{Z'}{\sigma^2}(Y-\mu\mathbb{1}-X\beta)\Big)\Big\}
\end{aligned}$$

On reconnait ici la densité d'une loi normale :

$$u|Y,\mu,\beta,\sigma_u^2,\sigma^2 \sim N_n\Big(\Big(\frac{Z'Z}{\sigma^2}+\frac{A^{-1}}{\sigma_u^2}\Big)^{-1}\frac{Z'}{\sigma^2}(Y-\mu\mathbb{1}-X\beta),\Big(\frac{Z'Z}{\sigma^2}+\frac{A^{-1}}{\sigma_u^2}\Big)^{-1}\Big) \tag{A.2}$$

Distribution conditionnelle complète de σ_u^2 :

$$\begin{aligned} p(\sigma_u^2|Y,\mu,\beta,u,\sigma^2) &\propto p(u|Y,\mu,\beta,\sigma_u^2,\sigma^2)p(\sigma_u^2) \\ &\propto p(u|\sigma_u^2)p(\sigma_u^2) \\ &\propto \frac{1}{{\sigma_u^2}^{\frac{n}{2}}} exp\Big\{ -\frac{1}{2\sigma_u^2} u'A^{-1}u \Big\} {\sigma_u^2}^{-a-1} exp\Big\{ -\frac{b}{\sigma_u^2} \Big\} \\ &\propto {\sigma_u^2}^{-(a+\frac{n}{2})-1} exp\Big\{ -\frac{1}{\sigma_u^2}\Big(b + \frac{u'A^{-1}u}{2}\Big) \Big\} \end{aligned}$$

On reconnait ici la densité d'une loi inverse-gamma :

$$\sigma_u^2|Y,\mu,\beta,u,\sigma^2 \sim IG\bigg(a + \frac{n}{2}, b + \frac{u'A^{-1}u}{2}\bigg) \tag{A.3}$$

Distribution conditionnelle complète de σ^2 :

$$\begin{aligned} &p(\sigma^2|Y,\mu,\beta,u,\sigma_u^2) \propto p(Y|\mu,\beta,u,\sigma_u^2,\sigma^2).p(\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} exp\bigg\{ -\frac{1}{2\sigma^2}(Y - \mu\mathbb{1} - X\beta - Zu)'(Y - \mu\mathbb{1} - X\beta - Zu) \bigg\} (\sigma^2)^{-a^*-1} exp\Big\{ -\frac{b^*}{\sigma^2} \Big\} \end{aligned}$$

On reconnait ici la densité d'une loi inverse-gamma :

$$\sigma^2|Y,\mu,\beta,u,\sigma_u^2 \sim IG\bigg(a^* + \frac{n}{2}, b^* + \frac{1}{2}||Y - \mu\mathbb{1} - X\beta - Zu||^2\bigg) \tag{A.4}$$

Annexe B

Sélection d'effets fixes

Calcul de la loi de $\gamma|Y,\mu,u,\sigma_u^2,\sigma^2$ pour l'*a priori* 1 pour la sélection d'effets fixes :

$$
\begin{aligned}
p(Y|\mu,\gamma,u,\sigma^2) &= \int p(Y,\beta_\gamma|\mu,\gamma,u,\sigma^2)\partial\beta_\gamma \\
&= \int p(Y|\beta_\gamma,\mu,\gamma,u,\sigma^2)p(\beta_\gamma|\gamma)\partial\beta_\gamma \\
&\propto \int \frac{1}{(\sigma_\beta^2)^{\frac{d_\gamma}{2}}} exp\Big\{ -\frac{1}{2}\Big(\frac{1}{\sigma^2}(y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)'(y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)+\frac{1}{\sigma_\beta^2}\beta_\gamma'\beta_\gamma\Big)\Big\}\partial\beta_\gamma \\
&\propto \frac{1}{(\sigma_\beta^2)^{\frac{d_\gamma}{2}}} exp\Big\{ -\frac{1}{2\sigma^2}(y-\mu\mathbb{1}-Zu)'(y-\mu\mathbb{1}-Zu)\Big\} \\
&\int exp\Big\{ -\frac{1}{2}\Big(\beta_\gamma'\Big(\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big)\beta_\gamma-\beta_\gamma'\frac{X_\gamma{}'}{\sigma^2}(Y-\mu\mathbb{1}-Zu)-(Y-\mu\mathbb{1}-Zu)'\frac{X_\gamma}{\sigma^2}\beta_\gamma\Big)\Big\}\partial\beta_\gamma \\
&\propto \frac{\Big|\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big|^{-\frac{1}{2}}}{(\sigma_\beta^2)^{\frac{d_\gamma}{2}}} exp\Big\{ -\frac{1}{2}(Y-\mu\mathbb{1}-Zu)'\Big(\frac{Id_n}{\sigma^2}-\frac{1}{(\sigma^2)^2}X_\gamma\Big(\frac{X_\gamma{}'X_\gamma}{\sigma^2}+\frac{Id_{d_\gamma}}{\sigma_\beta^2}\Big)X_\gamma{}'\Big)(Y-\mu\mathbb{1}-Zu)\Big\}
\end{aligned}
\tag{B.1}
$$

Calcul de la loi de $\gamma|Y,\mu,u,\sigma_u^2,\sigma^2$ pour l'*a priori* 2 pour la sélection d'effets fixes :

$$
\begin{aligned}
p(\gamma|Y,\mu,u,\sigma_u^2,\sigma^2) &\propto \int_{\mathbb{R}} p(\beta_\gamma,\gamma|Y,\mu,u,\sigma_u^2,\sigma^2)d\beta_\gamma \\
&\propto \int_{\mathbb{R}} p(Y|\mu,\beta_\gamma,\gamma,u,\sigma_u^2,\sigma^2)p(\beta_\gamma|\gamma,\sigma^2)p(\gamma)d\beta_\gamma \\
&\propto \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{(1-\gamma_j)}(2\pi\sigma^2)^{-\frac{n}{2}}(2\pi)^{-\frac{d_\gamma}{2}}|c\sigma^2({X_\gamma}'X_\gamma)^{-1}|^{-\frac{1}{2}} \\
&\quad \int_{\mathbb{R}} exp\Big\{ -\frac{1}{2\sigma^2}\Big((y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)'(y-\mu\mathbb{1}-X_\gamma\beta_\gamma-Zu)+\beta_\gamma'\frac{{X_\gamma}'X_\gamma}{c}\beta_\gamma\Big)\Big\} \\
&\propto \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{(1-\gamma_j)}(2\pi\sigma^2)^{-\frac{n}{2}}(2\pi)^{-\frac{d_\gamma}{2}}|c\sigma^2({X_\gamma}'X_\gamma)^{-1}|^{-\frac{1}{2}} \\
&\quad exp\Big\{ -\frac{1}{2\sigma^2}\Big((y-\mu\mathbb{1}-Zu)'(Id_n-\frac{c}{c+1}X_\gamma({X_\gamma}'X_\gamma)^{-1}{X_\gamma}')(y-\mu\mathbb{1}-Zu)\Big\} \\
&\quad \int_{\mathbb{R}} exp\Big\{ -\frac{1}{2}\Big(\beta_\gamma-\frac{c}{c+1}({X_\gamma}'X_\gamma)^{-1}{X_\gamma}'(Y-\mu\mathbb{1}-Zu)\Big)'\frac{c+1}{c\sigma^2}{X_\gamma}'X_\gamma \\
&\quad \Big(\beta_\gamma-\frac{c}{c+1}({X_\gamma}'X_\gamma)^{-1}{X_\gamma}'(Y-\mu\mathbb{1}-Zu)\Big)d\beta_\gamma \\
&\propto \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{(1-\gamma_j)}(2\pi\sigma^2)^{-\frac{n}{2}}(c+1)^{-\frac{d_\gamma}{2}} \\
&\quad exp\Big\{ -\frac{1}{2\sigma^2}\Big((y-\mu\mathbb{1}-Zu)'(Id_n-\frac{c}{c+1}X_\gamma({X_\gamma}'X_\gamma)^{-1}{X_\gamma}')(y-\mu\mathbb{1}-Zu)\Big\}
\end{aligned}
$$

Annexe C

Sélection d'effets aléatoires

Loi conditionnelle complète de γ

σ_{uj} est un paramètre de nuisance pour mettre à jour γ_j, pour $j = 1, \ldots, q$. Ainsi nous utilisons la technique de grouping et nous intégrons la loi conditionnelle complète de γ_j en σ_{uj}. Nous posons : $\widetilde{Y}_j = Y - \mu\mathbb{1} - \sum_{i \neq j} u_i' \sigma_{ui}$

$$\begin{aligned} p(\gamma_j = 1|Y, \mu, U, \gamma_{-j}, \sigma_{u-j}, \sigma^2) &= \frac{p(\gamma_j = 1, Y|\mu, U, \gamma_{-j}, \sigma_{u-j}, \sigma^2)}{p(\gamma_j = 1, Y|\mu, U, \gamma_{-j}, \sigma_{u-j}, \sigma^2) + p(\gamma_j = 0, Y|\mu, U, \gamma_{-j}, \sigma_{u-j}, \sigma^2)} \\ &= \frac{p(Y|\mu, U, \gamma_j = 1, \gamma_{-j}, \sigma_{u-j}, \sigma^2) p(\gamma_j = 1)}{p(Y|\mu, U, \gamma_j = 1, \gamma_{-j}, \sigma_{u-j}, \sigma^2) p(\gamma_j = 1) + p(Y|\mu, U, \gamma_j = 0, \gamma_{-j}, \sigma_{u-j}, \sigma^2) p(\gamma_j = 0)} \end{aligned} \tag{C.1}$$

Pour ce faire, nous devons calculer $p(Y|\mu, U, \gamma_j = 1, \gamma_{-j}, \sigma_{u-j}, \sigma^2)$ et $p(Y|\mu, U, \gamma_j = 0, \gamma_{-j}, \sigma_{u-j}, \sigma^2)$:

$$\begin{aligned} p(Y|\mu, U, \gamma_j = 1, \gamma_{-j}, \sigma_{u-j}, \sigma^2) &= \int p(Y|\mu, U, \gamma_j = 1, \gamma_{-j}, \sigma_u, \sigma^2) p(\sigma_{uj}|\gamma_j = 1) d\sigma_{uj} \\ &= \int \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\Big\{ -\frac{1}{2\sigma^2} (\widetilde{Y}_j - u_j\sigma_{uj})'(\widetilde{Y}_j - u_j\sigma_{uj}) \Big\} \frac{2}{\sqrt{2\pi\omega}} exp\Big\{ -\frac{1}{2\omega} \sigma_{uj}^2 \Big\} \mathbb{1}_{\sigma_{uj} > 0} d\sigma_{uj} \\ &= \int exp\Big\{ -\frac{1}{2} \Big[\sigma_{uj}^2 \Big(\frac{u_j'u_j}{\sigma^2} + \frac{1}{\omega} \Big) - 2\sigma_{uj} \frac{u_j'}{\sigma^2} \widetilde{Y}_j \Big] \Big\} d\sigma_{uj} \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \frac{2}{\sqrt{2\pi\omega}} exp\Big\{ -\frac{1}{2\sigma^2} \widetilde{Y}_j' \widetilde{Y}_j \Big\} \\ &= \frac{2}{(2\pi\sigma^2)^{\frac{n}{2}}} \Big(1 + \frac{\omega}{\sigma^2} u_j'u_j \Big)^{-\frac{1}{2}} exp\Big\{ -\frac{1}{2\sigma^2} \widetilde{Y}_j' \Big(1 - \Big(u_j'u_j + \frac{\sigma^2}{\omega} \Big)^{-1} u_j u_j' \Big) \widetilde{Y}_j \Big\} p(Z > 0) \end{aligned}$$

Avec $Z \sim N\Big(\Big(u_j'u_j + \frac{\sigma^2}{\omega} \Big)^{-1} u_j' \widetilde{Y}_j, \Big(u_j'u_j + \frac{\sigma^2}{\omega} \Big)^{-1} \sigma^2 \Big)$

$$\begin{aligned} p(Y|\mu, U, \gamma_j = 0, \gamma_{-j}, \sigma_{u-j}, \sigma^2) &= \int p(Y|\mu, U, \gamma_j = 0, \gamma_{-j}, \sigma_u, \sigma^2) p(\sigma_{uj}|\gamma_j = 0) d\sigma_{uj} \\ &= \int \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\Big\{ -\frac{1}{2\sigma^2} (\widetilde{Y}_j - u_j\sigma_{uj})'(\widetilde{Y}_j - u_j\sigma_{uj}) \Big\} \frac{2}{\sqrt{2\pi\omega}} \mathbb{1}_{\sigma_{uj} = 0} d\sigma_{uj} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\Big\{ -\frac{1}{2\sigma^2} \widetilde{Y}_j' \widetilde{Y}_j \Big\} \end{aligned}$$

Nous obtenons ainsi la probabilité conditionnelle complète que $\gamma_j = 1$ et nous pouvons donc générer les γ_j selon des lois de Bernoulli

$$p(\gamma_j = 1|Y, \mu, U, \gamma_{-j}, \sigma_{u-j}, \sigma^2) = \frac{p(\gamma_j = 1) 2 \Big(1 + \frac{\omega}{\sigma^2} u_j'u_j \Big)^{-\frac{1}{2}} exp\Big\{ \frac{1}{2\sigma^2} \widetilde{Y}_j' \Big(u_j'u_j + \frac{\sigma^2}{\omega} \Big)^{-1} u_j u_j' \widetilde{Y}_j \Big\} p(Z > 0)}{p(\gamma_j = 1) 2 \Big(1 + \frac{\omega}{\sigma^2} u_j'u_j \Big)^{-\frac{1}{2}} exp\Big\{ \frac{1}{2\sigma^2} \widetilde{Y}_j' \Big(u_j'u_j + \frac{\sigma^2}{\omega} \Big)^{-1} u_j u_j' \widetilde{Y}_j \Big\} p(Z > 0) + (1 - p(\gamma_j = 1))}$$

Loi conditionnelle complète de σ_u

Nous allons mettre à jours les $\sigma_{uj}, j = 1, \ldots, q$ individuellement selon leur lois conditionnelles complètes dans le but de travailler avec les lois normales univariées tronquées plutôt qu'une seule loi normale multivariée tronquée car cela est plus rapide à générer dans le logiciel R.

$$\begin{aligned} p(\sigma_{uj}|Y, \mu, U, \gamma_j = 1, \gamma_{-j}, \sigma^2) &\propto p(Y|\mu, U, \sigma_u, \sigma^2) p(\sigma_{uj}|\gamma_j = 1) \\ &\propto exp\Big\{ -\frac{1}{2\sigma^2}(Y - \mu\mathbb{1} - U\sigma_u)'(Y - \mu\mathbb{1} - U\sigma_u)\Big\} exp\Big\{ -\frac{1}{2\omega}\sigma_{uj}^2\Big\}\mathbb{1}_{\sigma_{uj}>0} \\ &\propto exp\Big\{ -\frac{1}{2}\Big[\sigma_{uj}^2\Big(\frac{u_j'u_j}{\sigma^2} + \frac{1}{\omega}\Big) - 2\sigma_{uj}\frac{u_j'}{\sigma^2}\widetilde{Y}_j\Big]\Big\}\mathbb{1}_{\sigma_{uj}>0} \end{aligned}$$

Nous en déduisons que la loi conditionnelle complète de σ_{u_γ} est une loi normale tronquée en zéro.

$$\sigma_{uj}|Y, \mu, U, \gamma_j = 1, \gamma_{-j}, \sigma^2 \sim N^+\Bigg(\Big(u_j'u_j + \frac{\sigma^2}{\omega}\Big)^{-1} u_j'\widetilde{Y}_j, \Big(\frac{u_j'u_j}{\sigma^2} + \frac{1}{\omega}\Big)^{-1}\Bigg) \tag{C.2}$$

Si $\gamma_j = 0$ alors la loi *a posteriori* de σ_{uj} reste une masse de Dirac en zéro.

Loi conditionnelle complète de μ

$$\begin{aligned} p(\mu|Y, U_\gamma, \sigma_{u_\gamma}, \gamma, \sigma^2) &\propto p(Y|\mu, U_\gamma, \sigma_{u_\gamma}, \gamma, \sigma^2) p(\mu) \\ &\propto exp\Big\{ -\frac{1}{2\frac{\sigma^2}{n}}\Big(\mu^2 - 2\mu\frac{\mathbb{1}'}{n}(Y - U_\gamma\sigma_{u_\gamma})\Big)\Big\} \end{aligned}$$

Ainsi nous en déduisons que μ suit une loi conditionnelle complète normale :

$$\mu|Y, U_\gamma, \sigma_{u_\gamma}, \gamma, \sigma^2 \sim N\Big(\frac{\mathbb{1}'}{n}(Y - U_\gamma\sigma_{u_\gamma}), \frac{\sigma^2}{n}\Big) \tag{C.3}$$

Loi conditionnelle complète de u_j

$$\begin{aligned} p(u_j|Y, \mu, u_{-j}, \sigma_u, \sigma^2) &\propto p(Y\mu, U, \sigma_u, \sigma^2) p(u_j) \\ &\propto exp\Big\{ -\frac{1}{2\sigma^2}\Big(Y - \mu\mathbb{1} - \sum_{j'=1}^{q} u_{j'}\sigma_{u_{j'}}\Big)'\Big(Y - \mu\mathbb{1} - \sum_{j'=1}^{q} u_{j'}\sigma_{u_{j'}}\Big) - \frac{1}{2}u_j' IBD_j^{-1} u_j\Big\} \\ &\propto exp\Big\{ -\frac{1}{2}\Big[u_j'\Big(\frac{\sigma_{u_j}{}^2}{\sigma^2} + IBD^{-1}\Big)u_j - u_j'\Big(Y - \mu\mathbb{1} - \sum_{k\neq j} u_k\sigma_{u_k}\Big)\frac{\sigma_{u_j}}{\sigma^2} - \Big(Y - \mu\mathbb{1} - \sum_{k\neq j} u_k\sigma_{u_k}\Big)' u_j\frac{\sigma_{u_j}}{\sigma^2}\Big]\Big\} \end{aligned}$$

Nous en déduisons que u_j suit une loi conditionnelle complète :

$$u_j|Y, \mu, u_{-j}, \sigma_u, \sigma^2 \sim N_n\Bigg(\Sigma_j^{-1}\frac{\sigma_{u_j}}{\sigma^2}\Big(Y - \mu\mathbb{1} - \sum_{k\neq j} u_k\sigma_{u_k}\Big), \Sigma_j^{-1}\Bigg) \tag{C.4}$$

avec :

$$\Sigma_j^{-1} = \left(\frac{{\sigma_{u_j}}^2}{\sigma^2} Id_n + IBD_j^{-1} \right)^{-1}$$

Nous pouvons optimiser la simulation des u_j en travaillant la matrice de variance-covariance et en utilisant la décomposition SVD des matrices IBD :

$$IBD_j = Q_j V_j Q_j'$$

avec V_j une matrice diagonale de rang r :

$$V_j = \begin{pmatrix} v_{j_1} & & \\ & \ddots & \\ & & v_{j_r} \end{pmatrix}$$

et Q_j une matrice orthonormale ($Q_j Q_j' = Id$). On a alors :

$$\begin{aligned}
IBD_j^{-1} &= Q_j V_j^{-1} Q_j' \\
\frac{{\sigma_{u_j}}^2}{\sigma^2} Id_n + IBD_j^{-1} &= Q_j \begin{pmatrix} \frac{{\sigma_{u_j}}^2}{\sigma^2} + \frac{1}{v_{j_1}} & & \\ & \ddots & \\ & & \frac{{\sigma_{u_j}}^2}{\sigma^2} + \frac{1}{v_{j_r}} \end{pmatrix} Q_j' \\
\Sigma_j^{-1} &= Q_j \underbrace{\begin{pmatrix} \frac{1}{\frac{{\sigma_{u_j}}^2}{\sigma^2} + \frac{1}{v_{j_1}}} & & \\ & \ddots & \\ & & \frac{1}{\frac{{\sigma_{u_j}}^2}{\sigma^2} + \frac{1}{v_{j_r}}} \end{pmatrix}}_{\tilde{V}^{-1}} Q_j'
\end{aligned}$$

On obtient ainsi $\Sigma_j^{-1} = Q_j \tilde{V}^{-1} Q_j'$. Une fois les décompositions SVD calculées, le calcul de Σ_j^{-1} s'apparente à un produit matriciel. Pour générer u_j on peut générer :

$$c_j \sim N\left(\frac{\sigma_{u_j}}{\sigma^2} \tilde{V}^{-1} Q_j' (Y - \mu \mathbb{1} - U_{-j} \sigma_{u-j}), \tilde{V}^{-1} \right)$$

alors $Q_j c_j$ suit la loi C.4.

Loi conditionnelle complète de σ^2

$$\begin{aligned}
p(\sigma^2 | Y, \mu, U, \sigma_u) &\propto p(Y | \mu, U, \sigma_u, \sigma^2) p(\sigma^2) \\
&\propto \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left\{ -\frac{1}{2\sigma^2} (Y - \mu\mathbb{1} - U_\gamma \sigma_{u_\gamma})' (Y - \mu\mathbb{1} - U_\gamma \sigma_{u_\gamma}) \right\} \sigma^{2-a-1} esp\left\{ -\frac{b}{\sigma^2} \right\} \\
&\propto \sigma^{2(-a-\frac{n}{2}-1)} exp\left\{ -\frac{1}{\sigma^2} \left(b + \frac{||Y - \mu\mathbb{1} - U_\gamma \sigma_{u_\gamma}||}{2} \right) \right\}
\end{aligned}$$

Nous en déduisons que la loi conditionnelle complète de σ^2 est une loi inverse-gamma :

$$\sigma^2|Y,\mu,U,\sigma_u \sim IG\left(a+\frac{n}{2}, b+\frac{||Y-\mu\mathbb{1}-U_\gamma\sigma_{u_\gamma}||^2}{2}\right) \tag{C.5}$$

Echantillonnage de γ suivant sa loi conditionnelle complète pour la méthode H-CBS pour la prise en compte de la similarité entre les structures d'apparentement :

Etape de Metropolis-Hastings pour générer γ suivant la méthode "Hybrid - Correlation Based Search".

Algorithme de Metropolis-Hastings :

Notons $\theta^{(i)}$ la valeur de la chaîne de Markov à l'itération i de l'algorithme MH et $q(.|.)$ la densité d'un noyau de transition. A chaque itération, l'algorithme effectue les étapes suivantes :

1. Générer γ^* suivant les mouvements d'addition, suppression ou les deux combinés,
2. Calculer $\rho(\gamma^{(i)},\gamma^*) = min\left\{1, \frac{p(\gamma^*)p(Y|\mu,\gamma^*,U,\sigma^2)q(\gamma^i|\gamma^*)}{p(\gamma^i)p(Y|\mu,\gamma^i,U,\sigma^2)q(\gamma^*|\gamma^i)}\right\}$
3. Prendre $\gamma^{(i+1)} = \begin{cases} \gamma^* \text{ avec probabilité } \rho(\gamma^{(i)},\gamma^*), \\ \gamma^{(i)} \text{ avec probabilité } 1-\rho(\gamma^{(i)},\gamma^*). \end{cases}$

Avec

$$p(Y|\mu,\gamma,U,\sigma^2) \propto \frac{{\sigma^2}^{\frac{d_\gamma}{2}}}{\omega^{\frac{d_\gamma}{2}}|\Sigma_\gamma|^{\frac{1}{2}}} exp\left\{\frac{1}{2\sigma^2}\widetilde{Y}_j' U_\gamma \Sigma_\gamma^{-1} U_\gamma' \widetilde{Y}_j\right\} p(Z>0)$$

Avec $Z \sim N\left(\Sigma_\gamma^{-1}\frac{U_\gamma'}{\sigma^2}\widetilde{Y}_j,\ \ \Sigma_\gamma^{-1}\right)$ et $\Sigma_\gamma = U_\gamma' U_\gamma + \frac{\sigma^2}{\omega} Id_{d_\gamma}$.

Si le mouvement d'addition est choisi alors le mouvement inverse est le mouvement de suppression et vice versa ainsi le rapport se simplifie :

$$\frac{q(\gamma^i|\gamma^*)}{q(\gamma^*|\gamma^i)} = \begin{cases} 1 \quad \text{si mouvement d'addition et suppression,} \\ \frac{\phi}{2d_{\gamma^*}}\frac{2d_{\gamma^i}}{\phi} = \frac{d_{\gamma^i}}{d_{\gamma^*}} \quad \text{si mouvement d'addition ou de suppression,} \end{cases}$$

$$\frac{p(\gamma^*)}{p(\gamma^i)} = \frac{\binom{q}{d_{\gamma^*}}^{-1}\frac{1}{d_{\gamma^*}}}{\binom{q}{d_{\gamma^i}}^{-1}\frac{1}{d_{\gamma^i}}} = \frac{(q-d_{\gamma^*})!(d_{\gamma^*}-1)!}{(q-d_{\gamma^i})!(d_{\gamma^i}-1)!} = \begin{cases} 1 \quad \text{si mouvement d'addition et suppression,} \\ \frac{d_{\gamma^i}}{q-d_{\gamma^i}} \quad \text{si mouvement d'addition,} \\ \frac{q-d_{\gamma^*}}{d_{\gamma^*}} \quad \text{si mouvement de suppression.} \end{cases}$$