UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ÉTUDE D'UNE LOI *A PRIORI* POUR LES ARBRES BINAIRES DE RÉGRESSION

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

ALEXIA JOLICOEUR-MARTINEAU

AVRIL 2016

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.07-2011). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Dans ce mémoire, on introduit les modèles d'arbres de régression fréquentiste et bayésien. On s'intéresse en particulier au modèle d'arbres bayésien proposé par Chipman *et al.* (1998). La loi *a priori* de l'arbre définie par Chipman *et al.* (1998) est spécifiée indirectement par un processus récursif et dépend de deux hyperparamètres. Selon les auteurs, le premier hyperparamètre sert à contrôler la grandeur de l'arbre et le deuxième hyperparamètre sert à contrôler sa forme. On tente de confirmer ces assertions et de mieux comprendre le fonctionnement de la loi *a priori*. Plus précisément, on étudie, dans ce travail, l'impact du choix des hyperparamètres et de la matrice des variables explicatives sur les propriétés de la loi *a priori* de l'arbre de régression. De plus, on étudie l'impact d'un choix approprié d'hyperparamètres sur la loi *a posteriori* de l'arbre lorsque la réelle structure est connue.

On commence par dériver les formules théoriques de l'espérance et de la variance du nombre de feuilles selon la loi *a priori* de l'arbre dans un cas spécial et idéal, c'est-à-dire, avec un hyperparamètre fixé à zéro et une infinité d'observations distinctes. Dans ce cas spécial, la profondeur de l'arbre est directement liée à son nombre de feuilles. Dans le cas général, sans hyperparamètre fixé à zéro, on ne peut dériver analytiquement les formules de l'espérance et de la variance du nombre de feuilles et de la profondeur. On procède donc à l'estimation de ces quantités par simulation des arbres *a priori*. Ensuite, pour étudier l'impact de la loi *a priori* sur les arbres *a posteriori*, on ajuste un arbre de régression avec différents choix d'hyperparamètres, du nombre d'observations et du rapport signal sur bruit.

On constate en premier que le nombre moyen de feuilles des arbres *a priori* est associé positivement avec le premier hyperparamètre et négativement avec le deuxième hyperparamètre. Il est donc possible d'obtenir le même nombre moyen de feuilles avec une infinité de combinaisons différentes des hyperparamètres. En choisissant de grandes valeurs pour les deux hyperparamètres, de façon à obtenir le même nombre de feuilles qu'avec de petites valeurs, on obtient moins de variabilité dans le nombre de feuilles et la profondeur des arbres. On observe de plus que de changer directement un hyperparamètre en fixant l'autre hyperparamètre ne permet pas de réduire la profondeur des arbres sans réduire le nombre de feuilles. Ces résultats révèlent que les deux hyperparamètres contrôlent non seulement la moyenne du nombre de feuilles des arbres, mais aussi la variance de celui-ci.

À cause des restrictions sur les hyperparamètres et des limites imposées par la taille échantillonnale, il est toutefois impossible d'obtenir une loi *a priori* avec une moyenne arbitrairement grande et une variance arbitrairement petite du nombre de feuilles. La loi *a priori* de l'arbre dépend implicitement de la matrice des variables explicatives. On constate que de réduire le nombre d'observations distinctes, lorsqu'on a une seule variable explicative, réduit de façon considérable le nombre de feuilles moyen des arbres *a priori*. Finalement, dans l'exemple synthétique, on constate que le modèle d'arbres bayésien performe mieux lorsque l'on choisit les hyperparamètres de façon à centrer la loi *a priori* de l'arbre sur le véritable nombre de feuilles que lorsqu'on utilise les hyperparamètres par défaut du paquet `tgp` du progiciel `R`.

MOTS-CLÉS : arbres de régression et classification (CART), modèle bayésien, loi *a priori*, hyperparamètres

INTRODUCTION

Le modèle d'arbre de régression et classification (CART) est une méthode statistique moderne qui consiste en l'utilisation d'un arbre binaire pour effectuer des prédictions. Ce modèle comporte plusieurs avantages par rapport aux modèles classiques tels que les régressions linéaire et logistique. Cette méthode a été initialement introduite par Breiman *et al.* (1984) dans un contexte fréquentiste. Un peu plus d'une décennie plus tard, Chipman *et al.* (1998) et Denison *et al.* (1998) ont proposé des analogues bayésiens au modèle CART de Breiman *et al.* (1984). Plus récemment, Wu *et al.* (2007) ont introduit un modèle CART bayésien alternatif dont la particularité se situe dans la spécification de la loi *a priori* de l'arbre (loi "pinball").

Dans ce mémoire, on s'intéresse à la méthode de CART introduite par Chipman *et al.* (1998) et plus précisément, à la loi *a priori* de l'arbre proposée par les auteurs. À notre connaissance, cette loi *a priori* n'a jamais été étudiée en détail. Contrairement à Denison *et al.* (1998) et Wu *et al.* (2007), la loi *a priori* de l'arbre de Chipman *et al.* (1998) est spécifiée indirectement par un processus stochastique et dépend de deux hyperparamètres. Selon Chipman *et al.* (1998), le premier hyperparamètre sert à contrôler la grandeur des arbres et le deuxième hyperparamètre sert à contrôler la forme des arbres. Les méthodes proposées par Denison *et al.* (1998) et Wu *et al.* (2007) spécifient plutôt explicitement une loi *a priori* sur la grandeur et la forme des arbres. Nous tentons de mieux comprendre comment les hyperparamètres de la loi *a priori* de Chipman *et al.* (1998) affectent les arbres de régressions, tant *a priori* qu'*a posteriori*.

Dans le premier chapitre, on effectue une mise en contexte et une revue de la littérature ciblée. On introduit la régression linéaire, les arbres de régression fréquentistes et les arbres de régression bayésiens définis par Chipman *et al.* (1998). On compare la régression linéaire avec les arbres de régression de façon à illustrer l'avantage des arbres dans certaines situations de modélisation. Ensuite, on discute des avantages de la méthode bayésienne comparativement à la méthode fréquentiste, pour la création d'arbres de régression.

Dans le deuxième chapitre, on illustre l'impact du choix des hyperparamètres de la loi *a priori* de l'arbre sur sa loi *a posteriori* en introduisant un exemple simple non simulé. Sur la base de l'estimateur du maximum *a posteriori* de l'arbre, on constate que de choisir adéquatement les hyperparamètres de la loi *a priori* de l'arbre permet d'obtenir un arbre ajustant mieux la fonction de régression univariée. La capacité prédictive de l'arbre moyen *a posteriori* est aussi améliorée en choisissant plus adéquatement les hyperparamètres de cette loi.

Dans le troisième chapitre, on étudie en détail le fonctionnement de la loi *a priori* de l'arbre de régression bayésien univarié telle qu'introduite par Chipman *et al.* (1998). On constate que le nombre d'observations distinctes influence les arbres *a priori* car les arbres de régression sont limités dans leur croissance à cause du nombre fini d'observations dans l'échantillon. On s'intéresse donc non seulement à l'effet des hyperparamètres sur les arbres *a priori*, mais aussi à l'impact de la matrice des variables explicatives. On commence par dériver les formules de l'espérance et de la variance du nombre de feuilles des arbres *a priori* pour une spécification particulière de la loi *a priori* de l'arbre qui fixe un hyperparamètre à zéro, tout en présumant que l'on possède une infinité d'observations distinctes. Ensuite, on étudie le cas plus général, où les deux hyperparamètres prennent des valeurs arbitraires et en présumant un nombre fixe d'observations distinctes. Dans ce dernier cas, on estime les caractéristiques de la loi *a priori* du nombre de feuilles

et de la profondeur en simulant des arbres de régression *a priori* à partir d'une version légèrement modifiée du processus stochastique de création d'arbres. Ces résultats nous permettent de bien comprendre comment la loi *a priori* de l'arbre est influencée par ses hyperparamètres et par la matrice des variables explicatives. Par la suite, on propose une modification de la définition de la probabilité de séparation du processus stochastique de création d'arbres, de façon à obtenir un plus grand contrôle sur cette loi *a priori*.

Dans le quatrième chapitre, on explique comment déterminer les caractéristiques de la loi *a priori* de l'arbre qui représentent le mieux notre connaissance *a priori* de la prédiction de la variable expliquée. Ensuite, on discute de comment choisir les hyperparamètres de la loi *a priori* de l'arbre de façon à obtenir la loi désirée.

Dans le cinquième chapitre, on présente une étude de simulation visant à vérifier si choisir les hyperparamètres de façon à obtenir une loi *a priori* centrée sur l'arbre de régression à estimer nous permet d'obtenir de meilleures prédictions et des arbres *a posteriori* plus semblables à l'arbre recherché. On examine le comportement du modèle de Chipman *et al.* (1998) en fonction de trois facteurs : le choix des hyperparamètres de la loi *a priori* de l'arbre, la taille échantillonnale et le rapport signal sur bruit. Le modèle bayésien est aussi comparé à son analogue fréquentiste.

Pour conclure, on termine avec une synthèse des résultats, une discussion sur les limitations de ce mémoire et des pistes de recherche futures par rapport à l'analyse de la loi *a priori* de l'arbre de Chipman *et al.* (1998).

CHAPITRE I

MÉTHODES DE RÉGRESSION FLEXIBLES : REVUE DE LA LITTÉRATURE

Dans ce chapitre, le modèle de régression linéaire et le modèle d'arbre de régression sont introduits. Ces deux modèles sont ensuite comparés et analysés. Finalement, l'arbre de régression bayésien est présenté et comparé à son homologue fréquentiste. L'objectif de ce chapitre est d'introduire graduellement l'arbre de régression bayésien et de montrer son potentiel et ses avantages par rapport aux autres modèles.

1.1 Régression linéaire

La régression est une technique fondamentale de la statistique d'aujourd'hui qui sert à modéliser la relation entre plusieurs variables. La régression linéaire est la méthode la plus utilisée due à sa puissance et simplicité. Cette méthode consiste en l'utilisation d'une combinaison linéaire de variables pour expliquer la variable d'intérêt. On représente le modèle de la façon suivante :

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1.1}$$

où y_i est la ième observation de la variable expliquée ($i = 1, \ldots, n$), x_{ij} est la ième observation de la variable explicative x_j ($j = 1, \ldots, p$), avec son coefficient de régression β_j associé, et ε_i est la ième erreur aléatoire non observée. Notons que

dans le modèle de régression linéaire (1.1), on considère les variables explicatives comme fixes. De plus, on présume que $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ et que les erreurs ε_i ne sont pas corrélées entre elles, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$ si $j \neq i$. Les coefficients sont des paramètres inconnus ; on mentionne dans ce qui suit quelques techniques permettant de les estimer.

Il existe plusieurs approches pour ajuster une courbe par régression linéaire, mais la plus utilisée est la méthode des moindres carrés. Ce type d'estimation consiste à déterminer les coefficients $\hat{\beta}_0, ..., \hat{\beta}_p$ qui minimisent la somme des résidus au carré. Dans la régression linéaire, les résidus sont définis comme étant

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \ .$$

La méthode des moindres carrés est donc

$$\min_{\hat{\beta}_0,...,\hat{\beta}_p} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2.$$

En supposant que les erreurs sont distribuées selon une loi normale, l'estimateur des moindres carrés est l'estimateur de maximum de vraisemblance (Montgomery *et al.*, 2012). De plus, si on respecte les conditions énoncées précédemment, il en résulte que l'estimateur des moindres carrés est le meilleur estimateur linéaire non biaisé (Montgomery *et al.*, 2012). L'estimateur est donc optimal dans le cas normal ; un autre avantage est qu'il s'exprime en forme fermée.

La régression linéaire est la méthode appropriée pour ajuster une courbe linéaire. Par contre, dans le cas non linéaire, la régression linéaire n'est pas toujours adéquate. Avec certaines transformations dans la variable expliquée ou dans les variables explicatives, on peut réussir à modéliser certaines relations non linéaires. Il est aussi possible d'approximer des fonctions plus complexes par régression polynômiale. Dans le cas univarié, la régression polynômiale est simplement une régression linéaire avec $x_1 = x, x_2 = x^2, ..., x_p = x^p$. Cette approche se base sur le

théorème de Taylor qui stipule que pour un certain entier $k \geq 1$ et une fonction $f : \Re \to \Re$ dérivable k fois à 0,

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^k(0)}{k!}x^k + h_k(x)x^k$$

où $\lim_{x \to 0} h_k(x) = 0$. Le théorème de Taylor peut aussi être adapté pour les fonctions $f : \Re^k \to \Re$. Il apparaît donc possible d'ajuster la plupart des courbes non linéaires par régression polynômiale. Malheureusement, cette approche peut causer des problèmes de multicolinéarité et en conséquence produire des estimateurs ayant des variances approchant l'infini (Montgomery *et al.*, 2012). De plus, si le degré du polynôme nécessaire pour modéliser la variable expliquée est plus grand que le nombre d'observations, la matrice des variables explicatives

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

devient singulière et l'estimation est donc impossible.

1.1.1 Exemple

Dans cette section, on présente un exemple d'une courbe non linéaire pour illustrer de façon concrète les difficultés qui peuvent être rencontrées avec la régression linéaire. L'exemple se base sur les données de simulation d'accidents de motos (Silverman, 1985). La variable indépendante x est le temps en millisecondes après l'impact et la variable dépendante y est l'accélération mesurée en g.

La Figure 1.1 montre que l'accélération est stable au début, descend précipitamment, monte précipitamment et s'estompe finalement. L'estimation d'une telle courbe est très difficile avec la régression linéaire. On examine l'ajustement de cette courbe par la régression polynômiale.

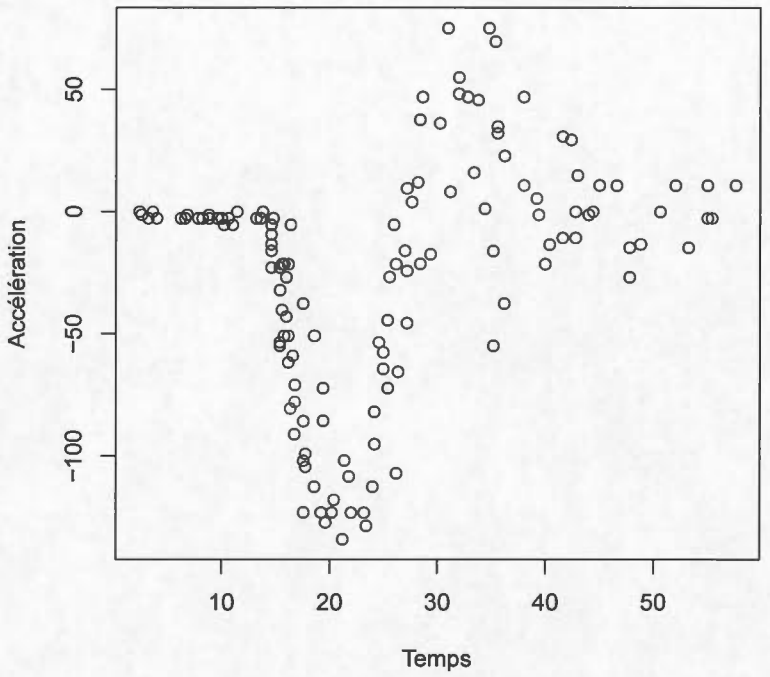


Figure 1.1 : Simulation d'accidents de motos

Dans la Figure 1.2, on voit que, peu importe le polynôme choisi, il n'est pas possible d'ajuster adéquatement la courbe. De plus, il y a un problème de multicolinéarité assez important lorsqu'on augmente le degré du polynôme. Pour le polynôme de degré 4, un des facteurs d'inflation de variance (VIF), une mesure déterminant combien la variance augmente si on enlève une des variables explicatives, est de 528.25. Un VIF de plus de 5 ou 10 est généralement considéré comme étant important (Montgomery *et al.*, 2012). Si on augmente encore plus le degré du polynôme, à un certain point la matrice des variables explicatives devient singulière. Cet exemple illustre qu'il n'est pas possible de bien modéliser le problème des accidents de motos avec la régression linéaire. Il est donc nécessaire de consi-

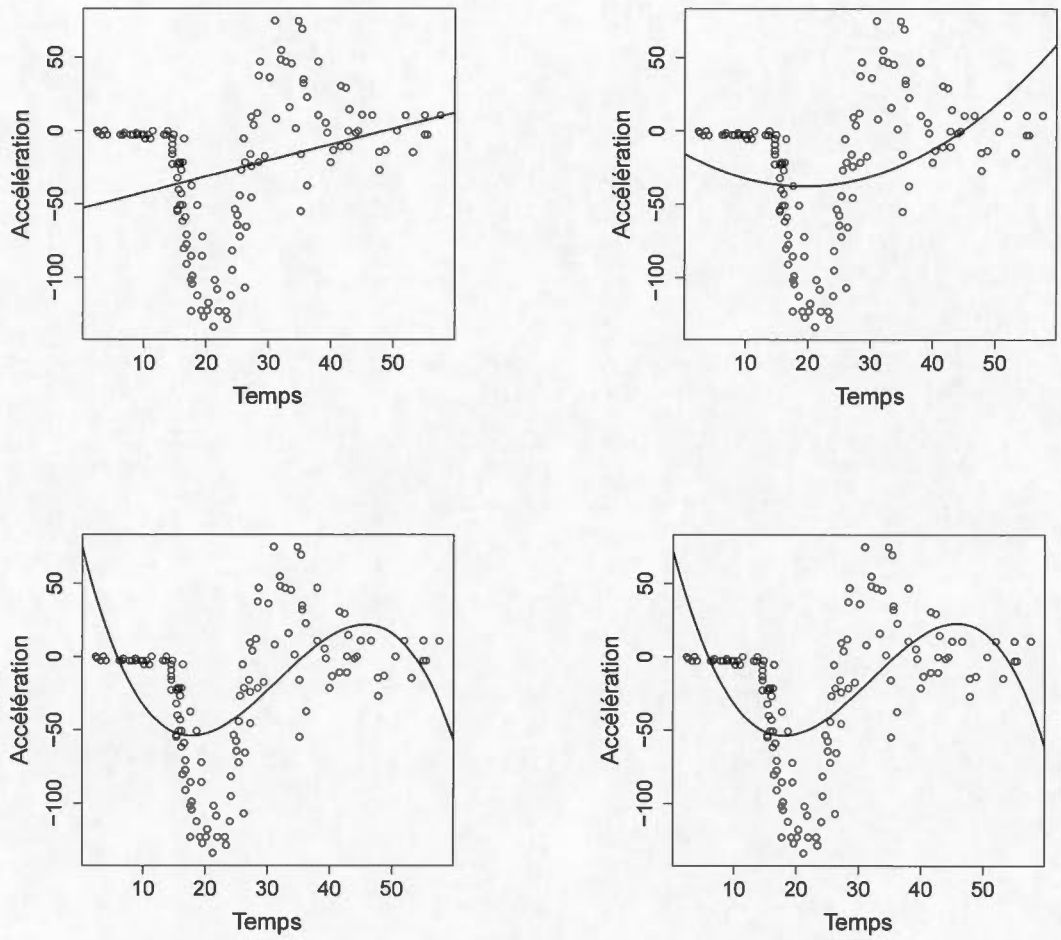


Figure 1.2 : Ajustement de courbe par régression polynômiale de degré 1, 2, 3 et 4 (simulation d'accidents de motos)

dérer l'utilisation des modèles plus sophistiqués étant capables de s'adapter aux courbes non linéaires. Dans le reste de ce mémoire, on discute exclusivement du modèle d'arbre de régression, mais il est à noter que les splines, séparateurs à vaste marge et réseaux de neurones artificiels sont d'autres exemples de modèles non paramétriques permettant d'approximer des courbes non linéaires (Hastie *et al.*, 2009).

1.2 CART

La méthode de Classification And Regression Trees (CART) (Breiman *et al.*, 1984) est une méthode non paramétrique qui consiste en la construction d'un arbre de décision binaire pour modéliser la relation entre la variable expliquée et les variables explicatives. Un arbre de décision est un graphe orienté, acyclique et connexe. Dans un arbre binaire, chaque noeud peut donner naissance à deux noeuds enfants. Les noeuds qui n'ont pas d'enfant sont appelés des feuilles, tandis que les noeuds qui ont des enfants sont appelés des parents. La racine est le seul noeud ne possédant pas de parent. Alternativement, il est aussi possible d'interpréter l'arbre de décision comme étant un cas spécial d'une fonction constante par morceaux (Thathachar et Sastry, 2004).

À chaque noeud intérieur de l'arbre, on sépare les observations $(y_i, \mathbf{x}_i)$, où $i = 1, \ldots, n$ et $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, selon une règle déterminée par une certaine variable de séparation x_j. Si la variable de séparation est continue, on envoie les observations qui répondent au critère $\{x_j \leq s\}$, où s est une constante, dans l'enfant à gauche et les autres observations sont envoyées dans l'enfant à droite. Si la variable de séparation est catégorique, on envoie les observations qui répondent au critère $\{x_j \in C\}$, où C est une catégorie de la variable x_j, dans l'enfant à gauche et les autres observations dans l'enfant à droite. On obtient donc une partition des

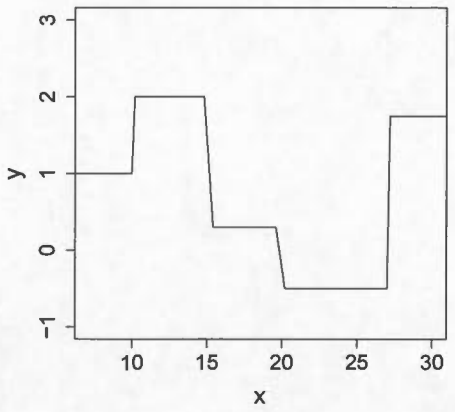


Figure 1.3 : Exemple de séparation par arbre avec une seule variable explicative (x) et 5 feuilles

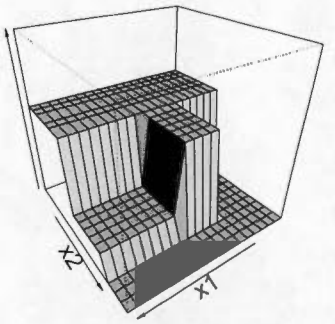


Figure 1.4 : Exemple de séparation par arbre avec deux variables explicatives (x_1 et x_2) et 5 feuilles

observations en régions $R_1, R_2, ..., R_b$, où b est le nombre de feuilles. On définit la profondeur d'un noeud comme étant le plus petit nombre d'arêtes reliant la racine et le noeud ; la profondeur de la racine est donc nulle. La profondeur d'un arbre est définie comme étant le maximum des profondeurs des noeuds de l'arbre (Knuth, 1998).

Dans la Figure 1.3, on a un exemple de séparation produite par un arbre avec une seule variable explicative et 5 feuilles. La variable explicative est désignée par x et la variable expliquée est désignée par y. Dans la Figure 1.4, on peut voir un exemple de séparation produite par un arbre avec deux variables explicatives (x_1 et x_2). La 3-ième dimension représente la valeur de la variable expliquée.

On caractérise également les arbres selon leur degré d'équilibre. Un arbre est équilibré lorsque la profondeur du sous-arbre gauche de chaque noeud ne diffère que de ± 1 par rapport à la profondeur du sous-arbre droit (Knuth, 1998). Par exemple, la Figure 1.5 représente un arbre équilibré. Le sous-arbre gauche de la

racine a une profondeur de 2 et le sous-arbre droit a une profondeur de 1. De plus, le sous-arbre gauche et le sous-arbre droit de l'enfant à gauche de la racine ont la même profondeur. Dans la Figure 1.6, on peut voir un arbre déséquilibré. Le sous-arbre gauche de la racine a une profondeur de 3 tandis que le sous-arbre droit a une profondeur de 1.

1.2.1 Arbre de régression

Pour modéliser la relation entre des variables explicatives et une variable expliquée continue, on assigne un paramètre c_k $(k = 1, \ldots, b)$ à chaque feuille de l'arbre. Ces paramètres représentent les prédictions de y dans la région R_k, où $R_1, \ldots, R_b$, est la partition induite par l'arbre. Pour une certaine observation $(y, \mathbf{x})$, le modèle d'arbre de régression est défini de la façon suivante :

$$f(\mathbf{x}) = \sum_{k=1}^{b} c_k \mathbb{1}(\mathbf{x} \in R_k).$$

Pour estimer les coefficients $c_1, ..., c_b$, on peut utiliser la méthode des moindres carrés

$$\min_{\hat{c}_1,\ldots,\hat{c}_b} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{b} \hat{c}_k \mathbb{1}(\mathbf{x}_i \in R_k) \right)^2.$$

Les coefficients optimaux sont

$$\hat{c}_k = \frac{\sum_{i=1}^{n} y_i \mathbb{1}(\mathbf{x}_i \in R_k)}{\sum_{i=1}^{n} \mathbb{1}(\mathbf{x}_i \in R_k)}, \; k = 1, \ldots, b.$$

Ils représentent la moyenne de la variable expliquée dont les valeurs des variables explicatives appartiennent à la région R_k.

Contrairement à l'estimation des coefficients, déterminer quelles règles de séparation produisent la meilleure partition des données est une tâche beaucoup plus difficile. Il n'est généralement pas possible d'utiliser la méthode des moindres

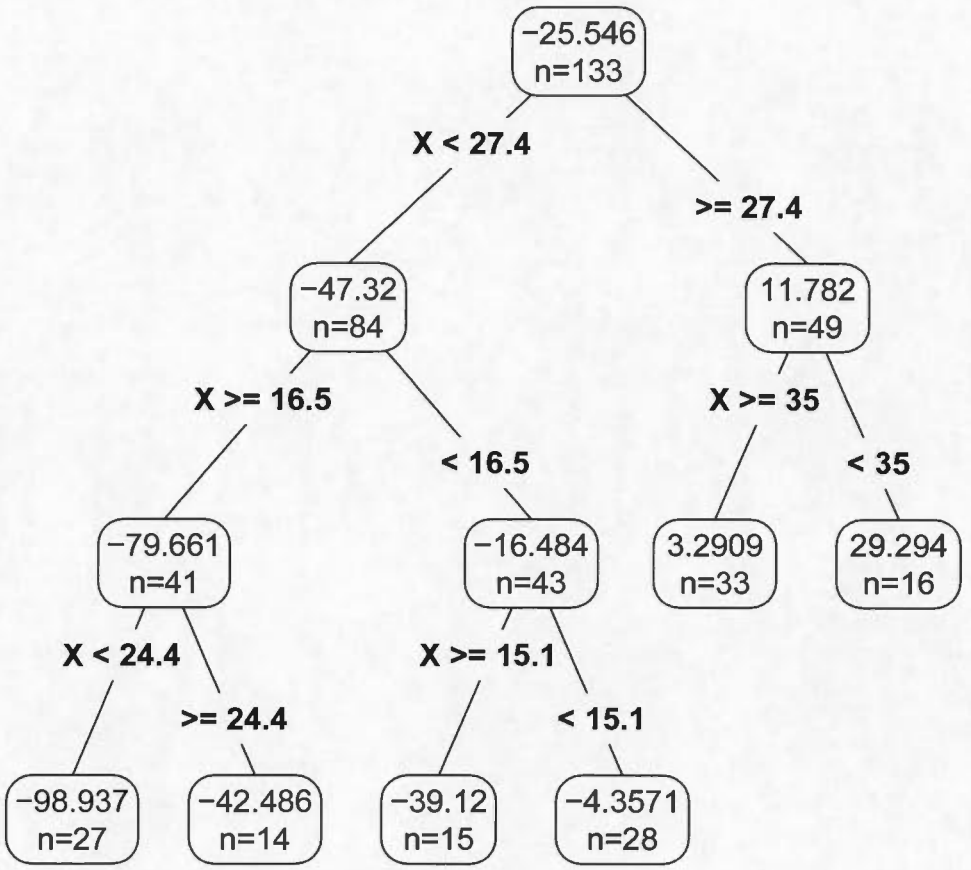


Figure 1.5 : Exemple d'arbre équilibré

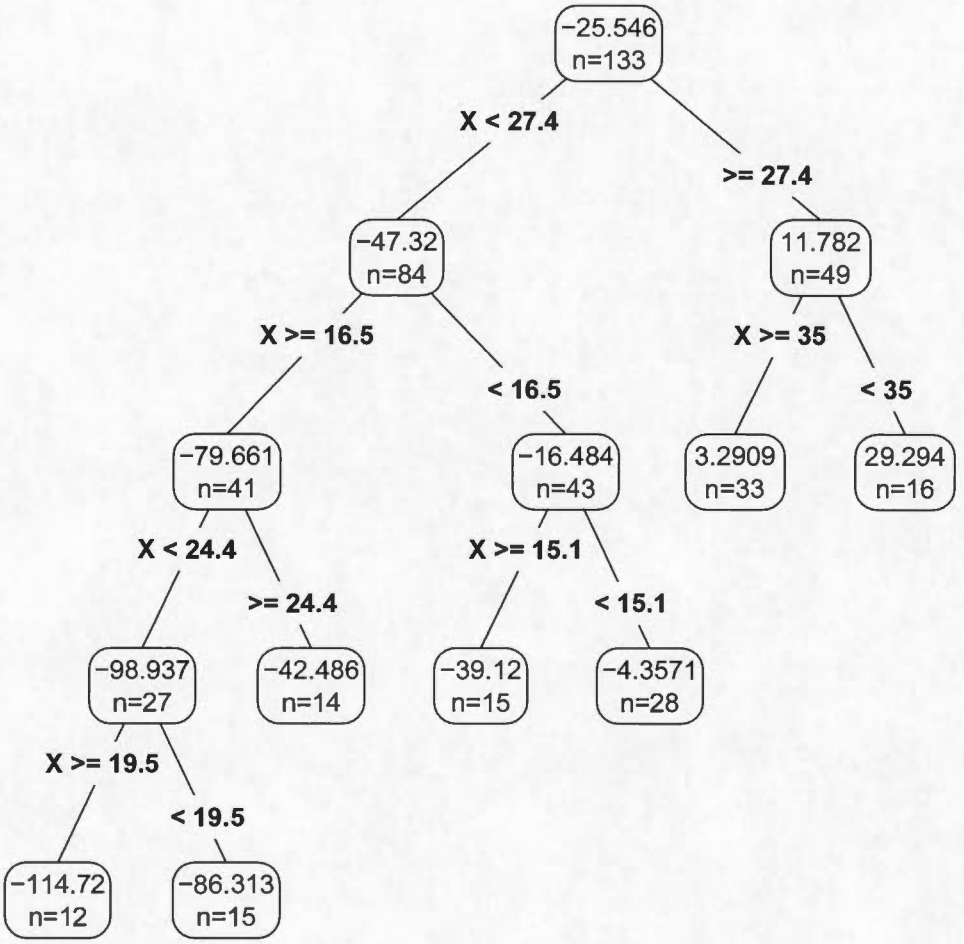


Figure 1.6 : Exemple d'arbre déséquilibré

carrés car cela demande une trop grande puissance computationnelle. On utilise plutôt un algorithme de type glouton qui détermine, une étape à la fois, quelle est la prochaine règle de séparation idéale. À chaque étape on agrandit donc l'arbre à partir d'une feuille de région R en résolvant l'équation suivante (Hastie *et al.*, 2009) :

$$\min_{j,s}\left[\min_{\hat{c}_1}\sum_{\mathbf{x}_i\in R_1(j,s)}(y_i-\hat{c}_1)^2+\min_{\hat{c}_2}\sum_{\mathbf{x}_i\in R_2(j,s)}(y_i-\hat{c}_2)^2\right]$$

où $R_1(j,s)=\{\mathbf{x}|x_j\leq s,\mathbf{x}\in R\}$ et $R_2(j,s)=\{\mathbf{x}|x_j>s,\mathbf{x}\in R\}$.

Ce processus est répété jusqu'à l'obtention d'un arbre de la taille désirée. Pour déterminer quand arrêter l'algorithme, on peut utiliser la validation croisée ou un critère d'information tel que le AIC ou le BIC (Hastie *et al.*, 2009).

Il est à noter qu'il existe d'autres méthodes permettant de déterminer des règles de séparation quasi-optimales. Les algorithmes gloutons basés sur le concept d'entropie tels que le C4.5 et le C5.0 (Quinlan, 1993) sont fréquemment utilisés. Ces algorithmes font croître les arbres de façon à maximiser le gain d'information.

1.2.2 Exemple d'arbre de régression

On va maintenant ajuster la courbe de l'exemple de simulation d'accidents de motos pour illustrer les capacités de CART. On utilise le paquet `rpart` (Therneau *et al.*, 2014) du langage de programmation `R` pour ajuster un tel modèle. Les paramètres de la fonction sont ceux par défaut sauf pour *minsplit* qui définit le nombre d'observations minimal requis dans une feuille pour que l'algorithme tente de séparer celle-ci. On choisit $minsplit = 15$ plutôt que $minsplit = 20$ (défaut) car nous n'avons que 133 observations.

La Figure 1.7 montre l'arbre de régression résultant de l'algorithme de création

d'arbres de régression du paquet `rpart`. Dans la Figure 1.8, on peut voir que l'arbre de régression produit une courbe assez bien ajustée aux données. Le résultat est bien meilleur que les résultats qui avaient été obtenus par la régression polynômiale.

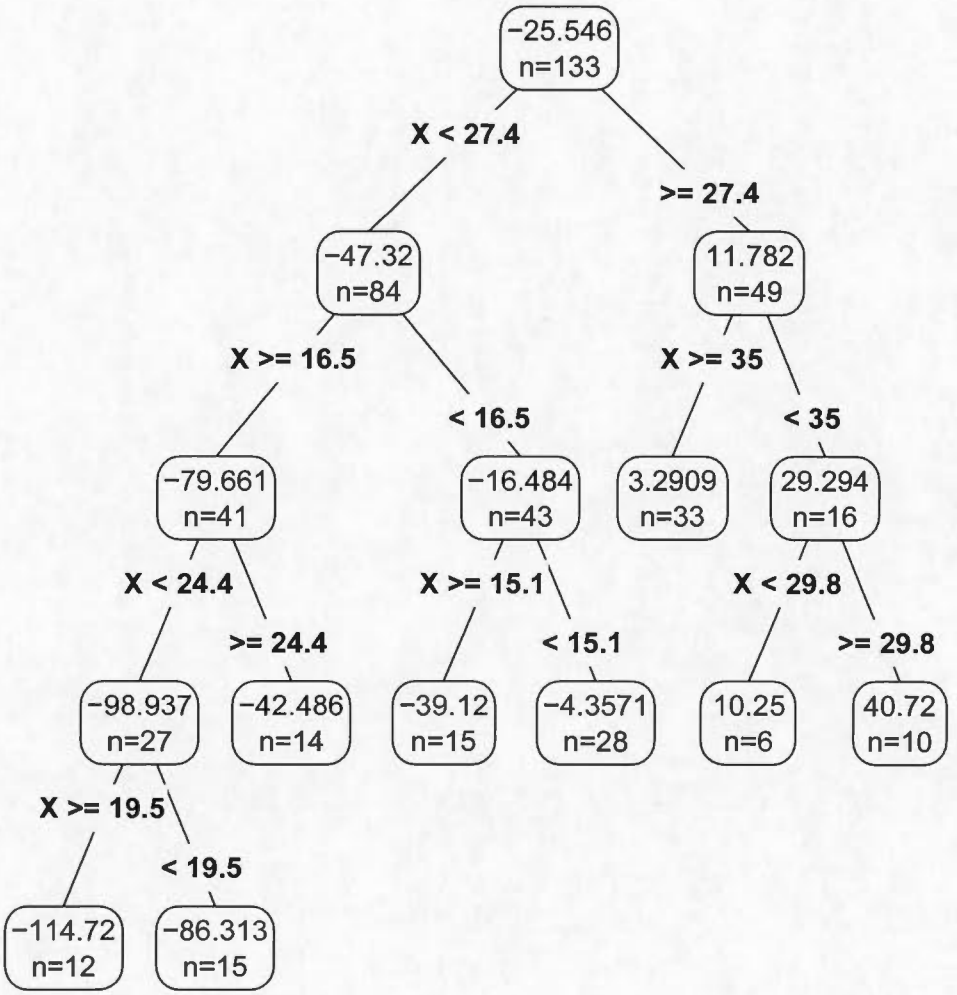


Figure 1.7 : Exemple d'arbre de régression (simulation d'accidents de motos)

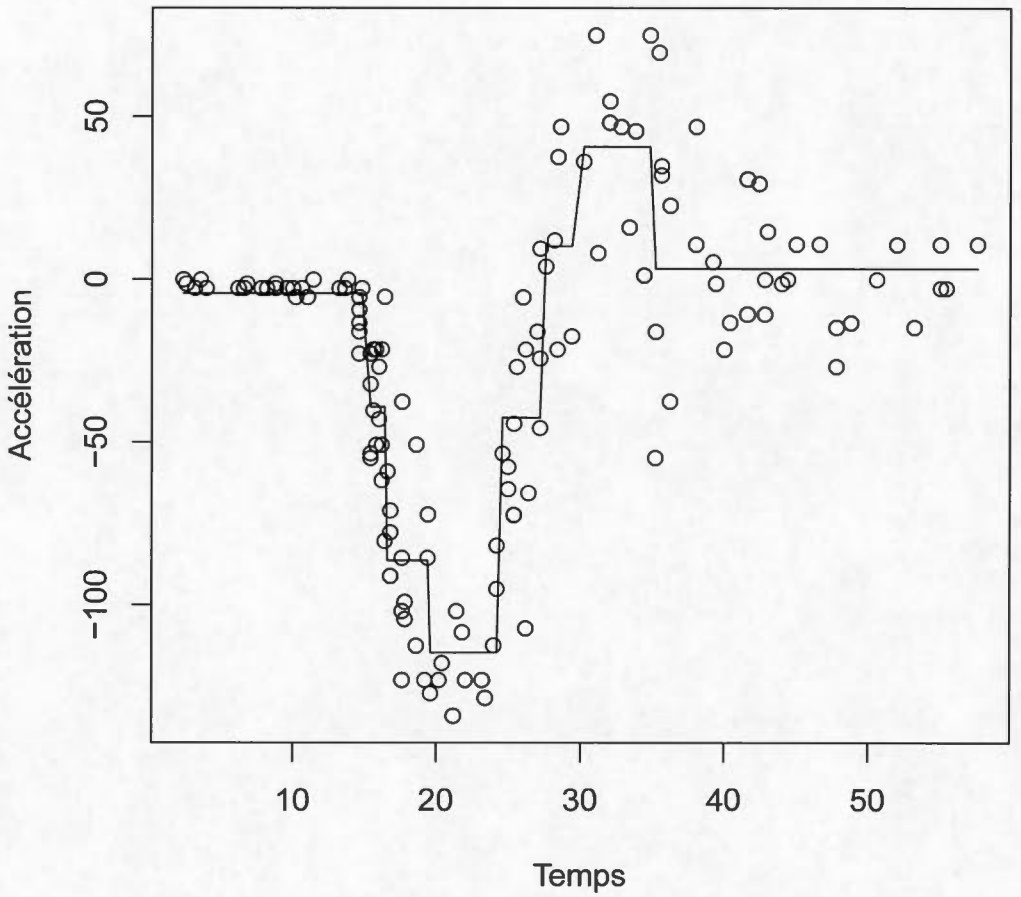


Figure 1.8 : Ajustement de courbe par arbre de régression (simulation d'accidents de motos)

1.2.3 Arbre de régression bayésien

L'arbre de régression bayésien représente une alternative à l'arbre de régression classique utilisant des algorithmes gloutons (Chipman *et al.*, 1998). En spécifiant une loi *a priori* sur l'arbre de régression, on obtient une loi *a posteriori* pour celui-ci. L'algorithme de Metropolis-Hastings (Hastings, 1970) peut être utilisé pour permettre l'exploration de la loi *a posteriori* et la recherche des arbres de régression les plus probables.

La méthode bayésienne permet d'explorer un plus grand nombre d'arbres potentiels que les algorithmes gloutons. Il est aussi possible de déterminer la moyenne pondérée des arbres de la loi *a posteriori*. Ainsi, il est généralement le cas qu'on obtient des prédictions beaucoup plus précises que ce que l'on obtiendrait par l'utilisation d'un seul arbre (Chipman *et al.*, 1998). La méthode bayésienne pourrait donc permettre d'avoir une meilleure performance que par la méthode fréquentiste.

Un arbre de régression bayésien contient un vecteur paramétrique $\Theta = (\theta_1, \theta_2, \ldots, \theta_b)$ incluant les paramètres associés à chaque feuille de l'arbre. Pour chaque observation $(y_{kl}, \mathbf{x}_{kl})$, où k représente la k-ème feuille de l'arbre et l représente la l-ème observation de la k-ème feuille, on a que $y_{kl}|\mathbf{x}_{kl}, \Theta, T \sim f(y_{kl}|\theta_k)$. La loi de la variable expliquée est choisie pour être une loi normale avec variance constante. Cette loi est définie de la façon suivante :

$$y_{k1}, y_{k2}, \ldots, y_{kn_k} \overset{i.i.d.}{\sim} N(u_k, \sigma^2), \qquad k = 1, \ldots, b. \tag{1.2}$$

Il est à noter qu'il est aussi possible d'utiliser une loi normale avec une variance spécifique à chaque feuille de l'arbre.

La loi d'un arbre de régression bayésien a la forme suivante :

$$P(Y|X,\Theta,T) = \prod_{k=1}^{b}\prod_{l=1}^{n_k} f(y_{kl}|\theta_k), \tag{1.3}$$

où $Y = (y_1, y_2, \ldots, y_n)$, $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, $n = \sum_{k=1}^{b} n_k$ et T est l'arbre de régression. La loi *a priori* $P(\Theta, T)$ est décomposée selon la relation conditionnelle suivante :

$$P(\Theta, T) = P(\Theta|T)P(T). \tag{1.4}$$

Ainsi, on spécifie $P(T)$ et $P(\Theta|T)$ séparément.

La loi *a priori* $P(T)$ ne possède pas de forme fermée. Cette loi est plutôt spécifiée indirectement par un processus stochastique. Pour générer un arbre selon la loi $P(T)$, on utilise le processus récursif suivant :

1. On construit T, un arbre vide contenant un seul noeud η.
2. On divise le noeud η avec probabilité $p_{SPLIT}(\eta, T)$.
3. Si on divise le noeud η, on construit les enfants selon la règle de séparation ρ échantillonnée de la distribution $p_{RULE}(\rho|\eta, T)$. Ensuite, on réapplique les étapes 2 et 3 aux enfants venant d'être créés, s'il y en a.

Le choix le plus naturel pour $p_{RULE}(\rho|\eta, T)$ est simplement de choisir uniformément une variable x_j de l'ensemble des variables explicatives $(x_1, \ldots, x_p)$. Si x_j est continue, on choisit uniformément une observation x_{ij} parmi $(x_{1j}, \ldots, x_{nj})$. Autrement, si x_j est une variable catégorique, on choisit uniformément une catégorie C disponible et incluse dans l'ensemble des catégories de x_j. La matrice des variables explicatives X est donc fixe dans $p_{RULE}(\rho|\eta, T)$; c'est la structure de l'arbre T qui change cette probabilité d'un noeud à l'autre. En principe, $p_{RULE}(\rho|\eta, T)$ dépend donc conditionnellement de X mais pour demeurer cohérent avec la littérature, on utilise la même notation que Chipman *et al.* (1998).

Pour ce qui est de $p_{SPLIT}(\eta, T)$, on utilise la définition suivante :

$$p_{SPLIT}(\eta, T) = \alpha(1 + d_\eta)^{-\beta}, \tag{1.5}$$

où d_η est la profondeur du noeud η, $\alpha \in [0, 1]$ et $\beta \geq 0$. Cette définition permet de contrôler la taille et la forme des arbres associées à $P(T)$. Le paramètre α dicte le nombre moyen de feuilles dans les arbres, tandis que β est vu comme un paramètre qui dicte la profondeur moyenne des arbres (Chipman *et al.*, 1998). Plus de détails sur la probabilité de séparation $p_{SPLIT}(\eta, T)$ et ses effets sur la loi *a priori* $P(T)$ sont présentés dans le prochain chapitre.

Pour le choix de la loi *a priori* $P(\Theta|T)$, on utilise la conjugée d'une loi normale

$$\mu_1, \ldots, \mu_b | \sigma, T \sim N(\bar{\mu}, \sigma^2/a), \tag{1.6}$$

$$\sigma^2 | T \sim \text{Inv-Gamma}(\nu/2, \nu\lambda/2). \tag{1.7}$$

Si l'on possède de l'information *a priori* sur u_k $(k = 1, \ldots, b)$ et σ^2, on peut choisir les hyperparamètres de façon à représenter cette information, mais Chipman *et al.* (1998) utilise la méthode qui suit. Soit s_* l'écart-type combiné de l'arbre fréquentiste produit par l'algorithme glouton, défini par l'équation suivante :

$$s_* = \sqrt{\frac{\sum_{k=1}^{b}(n_k - 1)s_k^2}{\sum_{k=1}^{b}(n_k - 1)}},$$

où n_k est le nombre d'observations dans la k-ème feuille de l'arbre et s_k est l'écart-type échantillonnal de Y dans la k-ème feuille. Soit s^* l'écart-type échantillonnal de Y. On choisit les hyperparamètres ν et λ de façon à inclure σ dans l'intervalle (s_*, s^*) avec très grande probabilité. Ensuite, les hyperparamètres a et $\bar{\mu}$ sont choisis de façon à ce que la loi *a priori* des u_k $(k = 1, \ldots, b)$ inclut la majeure partie des valeurs de Y.

Tel que nous le verrons, il est pratique d'éliminer la dépendance de $P(Y|X,\Theta,T)$ à Θ. Par intégration, on obtient $P(Y|X,T) = \int P(Y|X,\Theta,T)P(\Theta|T)d\Theta$. Sur la base des définitions (1.6) et (1.7), l'expression simplifiée pour $P(Y|X,T)$ est la suivante :

$$P(Y|X,T) = \frac{ca^{b/2}}{\prod_{k=1}^{b}(n_k+a)^{1/2}} \left(\sum_{k=1}^{b}(s_k+t_k) + \nu\lambda \right)^{-(n+\nu)/2},$$

où c est une constante qui ne dépend pas de T, k est l'indice de la k-ème feuille, $s_k = \sum_{l=1}^{n_k}(y_{kl}-\bar{y}_k)^2$, où $\bar{y}_k = \sum_{l=1}^{n_k} y_{kl}$ et $t_k = [n_k a/(n_k+a)](\bar{y}_k - \bar{\mu})^2$.

La loi *a posteriori* de T est déterminée par la relation suivante :

$$P(T|X,Y) \propto P(Y|X,T)P(T).$$

À cause du très grand nombre d'arbres possibles, une évaluation complète de tous les arbres T est difficile, sauf pour des exemples triviaux (Chipman *et al.*, 1998). Il est donc ardu de calculer la constante de normalisation et de déterminer exactement quels arbres ont la plus grande probabilité *a posteriori* $P(T|X,Y)$. Ainsi, on utilise plutôt un algorithme de Metropolis-Hasting pour l'échantillonnage de la loi *a posteriori* $P(T|X,Y)$. Pour ce faire, on génère une suite d'arbres $T^0, T^1, T^2, \ldots$, distribués selon une loi qui converge vers $P(T|X,Y)$. La transition de T^i à T^{i+1} se fait par les deux étapes suivantes :

1. On génère un nouvel arbre T^* avec probabilité de proposition $q(T^i, T^*)$.
2. On choisit $T^{i+1} = T^*$ avec probabilité $\alpha(T^i, T^*) = \min\{\frac{q(T^*,T^i)}{q(T^i,T^*)} \frac{P(Y|X,T^*)P(T^*)}{P(Y|X,T^i)P(T^i)}, 1\}$, sinon on choisit $T^{i+1} = T^i$.

Chaque nouvel arbre généré dans la suite est une modification simple de l'arbre précédent. La loi de proposition $q(T^i, T^*)$ est donc déterminée en choisissant au hasard l'une des quatre modifications suivantes :

— GROW (probabilité de 25%) : On choisit uniformément une feuille de l'arbre. Ensuite, on sépare la feuille en deux en choisissant une règle de séparation selon p_{RULE}.

— PRUNE (probabilité de 25%) : On choisit uniformément un parent de deux feuilles et on détruit les deux feuilles.

— CHANGE (probabilité de 40%) : On choisit uniformément un noeud interne et on change la règle de séparation pour une nouvelle règle admissible qu'on choisit au hasard.

— SWAP (probabilité de 10%) : On choisit uniformément un couple parent-enfant qui consiste en deux noeuds internes. Ensuite, on échange leurs règles de séparation si les enfants du parent n'ont pas la même règle de séparation. Sinon, on échange les règles de séparation des deux enfants contre celle de leur parent et vice-versa.

La suite d'arbres générée par cette méthode tend à être concentrée sur les régions à haute probabilité *a posteriori* de T ; on peut donc déterminer approximativement quels sont les arbres à haute probabilité *a posteriori* $P(T|X,Y)$. Les probabilités *a posteriori* des arbres générés de la suite sont approximées par la formule suivante :

$$\frac{P(Y|X,T^i)P(T^i)}{\sum_i P(Y|X,T^i)P(T^i)}.$$

Chaque modification est réversible. Cela implique que la chaîne de Markov des arbres est réversible. Il peut donc être démontré que la suite d'arbres de régression, générée par l'algorithme de Metropolis-Hasting, converge vers $P(T|X,Y)$ (Chipman *et al.*, 1998).

On introduit un exemple pour mieux comprendre le fonctionnement d'une étape CHANGE et d'une étape SWAP. La Figure 1.9 est l'arbre construit par l'algorithme glouton du paquet `rpart` en utilisant la base de données sur la qualité de l'air à New York en 1973, de mai jusqu'à septembre (Chambers *et al.*, 1983). La variable expliquée est la quantité d'ozone (ppb) et les variables explicatives $Wind$ et $Temp$ sont respectivement la vitesse du vent (mph) et la température moyenne

(degré F). Dans la Figure 1.10, on illustre l'effet de l'étape CHANGE sur l'arbre de la Figure 1.9 : change la règle de séparation $Temp \leq 77.5$ pour la règle de séparation $Temp \leq 60$. Dans la Figure 1.11, on illustre l'effet de l'étape SWAP sur l'arbre de la Figure 1.9 : échange la règle de séparation du parent $Wind \leq 7.15$ pour la règle de séparation de l'enfant $Temp \leq 77.5$ et vice-versa.

Pour éviter que l'algorithme reste coincé dans un mode local de la loi *a posteriori* de l'arbre, il est recommandé de redémarrer l'algorithme plusieurs fois. De cette façon, il est possible d'obtenir un plus grand éventail d'arbres de régression avec une grande probabilité *a posteriori*. Il est de plus recommandé de commencer l'algorithme avec un arbre ayant une unique feuille (Chipman *et al.*, 1998).

La méthode bayésienne offre donc certains avantages en comparaison à la méthode traditionnelle par algorithme glouton. La méthode bayésienne permet la priorisation de certains types d'arbres désirables par le choix des paramètres de la loi *a priori* $P(T)$. En effet, on pourrait vouloir des petits arbres ou des grands arbres. Un deuxième avantage de la méthode bayésienne est que l'algorithme de Metropolis-Hastings permet d'explorer une beaucoup plus grande quantité d'arbres que la méthode fréquentiste par algorithme glouton.

La méthode bayésienne à quand même un désavantage; À cause de sa nature probabiliste, l'algorithme de Metropolis-Hastings génère une suite infinie d'arbres variés qui change à chaque réinitialisation de l'algorithme. Les algorithmes gloutons, quant à eux, choisissent toujours la meilleure règle de séparation (selon un certain critère) à chaque étape, jusqu'à atteindre une certaine taille (selon le critère d'arrêt). Les algorithmes gloutons permettent donc d'obtenir une suite finie d'arbres qui reste la même à chaque réinitialisation de l'algorithme.

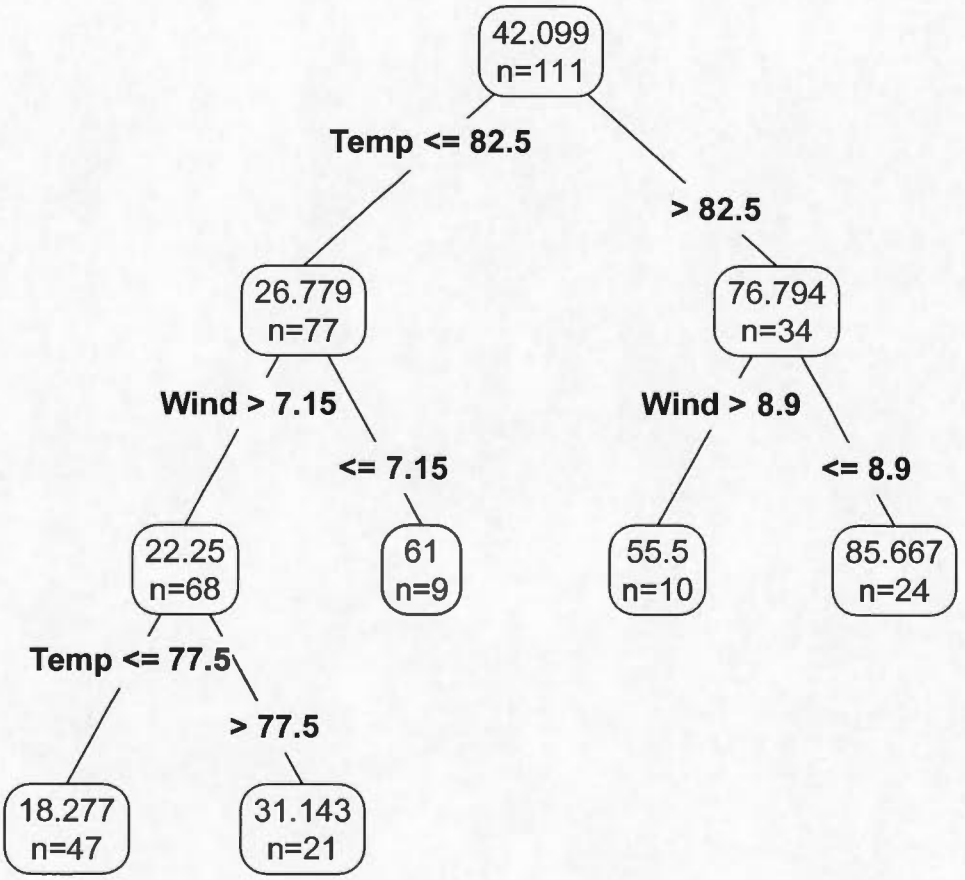


Figure 1.9 : Arbre de régression original

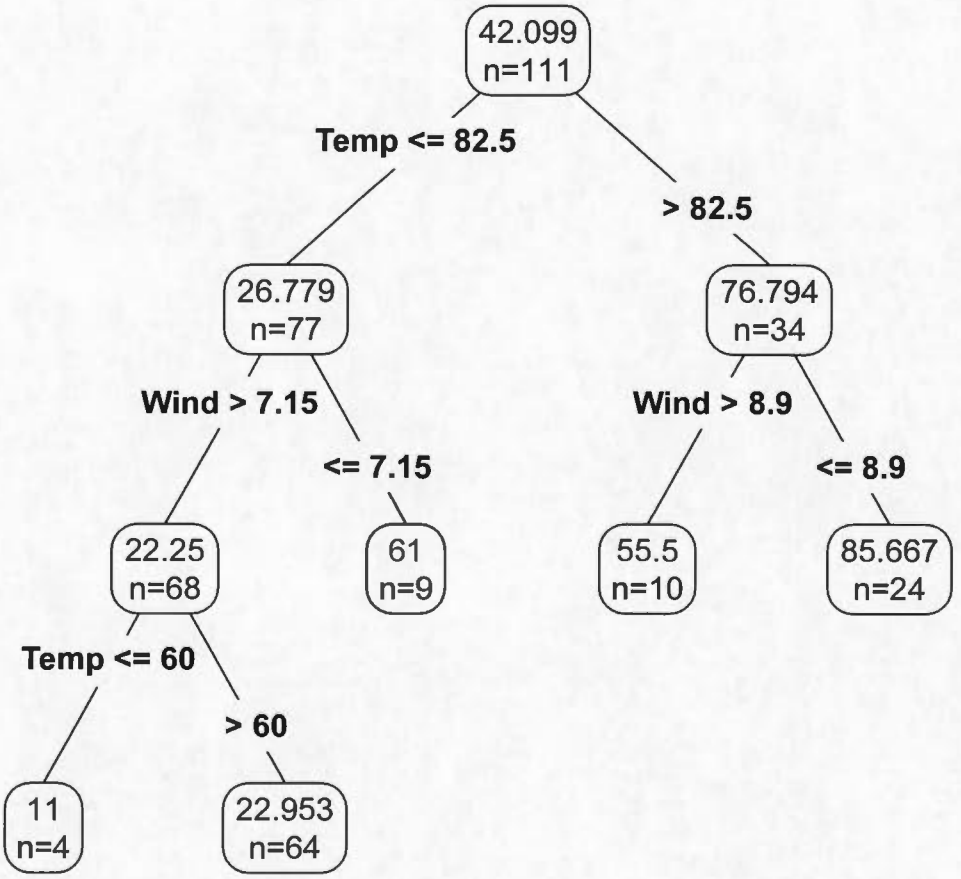


Figure 1.10 : Arbre de régression de la Figure 1.9 après étape CHANGE où l'on change la règle de séparation $Temp \leq 77.5$ pour la règle de séparation $Temp \leq 60$

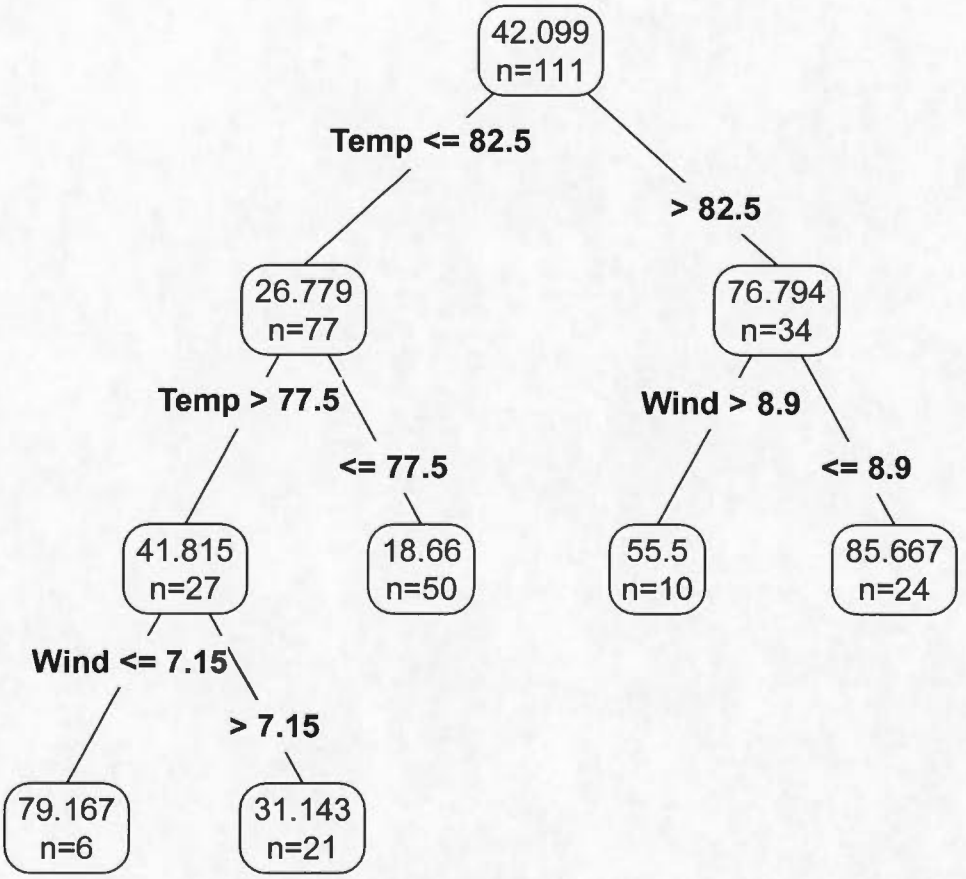


Figure 1.11 : Arbre de régression de la Figure 1.9 après étape SWAP où l'on échange la règle de séparation du parent $Wind \leq 7.15$ avec celle de son enfant

1.2.4 Exemple d'arbre de régression bayésien

Pour ajuster la courbe de l'exemple de simulation d'accidents de motos, par arbre bayésien, on utilise le paquet `tgp` (Gramacy, 2007) du langage de programmation `R`. Ce paquet permet d'estimer l'arbre maximum *a posteriori* (MAP) et il peut aussi être utilisé afin de calculer l'estimation de la prédiction moyenne des arbres *a posteriori*.

L'arbre MAP est défini comme étant l'arbre ayant la plus grande probabilité *a posteriori*. On ne peut déterminer exactement quel est l'arbre MAP, mais on peut l'estimer en choisissant l'arbre qui a la plus grande grande probabilité *a posteriori* de tous les arbres générés par un certain nombre d'itérations de l'algorithme de Metropolis-Hastings.

Pour estimer la prédiction moyenne des arbres *a posteriori* à un certain $\mathbf{x}$, on génère des arbres de la loi *a posteriori* et pour chaque arbre, on prédit $f(\mathbf{x})$ par $\hat{f}(\mathbf{x}) = \sum_{k=1}^{b} \hat{c}_k \mathbb{1}(\mathbf{x} \in R_k)$. Ensuite, on calcule la moyenne des prédictions $\hat{f}(\mathbf{x})$ des arbres générés de la loi *a posteriori*.

Les paramètres du paquet `tgp` sont laissés à leurs valeurs par défaut. Plus précisément, la valeur des hyperparamètres α et β est par défaut, 0.50 et 2, respectivement. Les 2000 premières réalisations sont écartées, 7000 réalisations sont observées et chaque deuxième réalisation est enlevée. Pour déterminer le meilleur arbre de régression (arbre MAP), l'algorithme est redémarré cinq fois.

La Figure 1.12 représente l'arbre de régression ayant la plus grande probabilité *a posteriori* après cinq redémarrages de l'algorithme. Cet arbre bayésien est semblable à l'arbre de régression fréquentiste de la Figure 1.7. Les branches sont similaires, la seule différence notable est que l'arbre bayésien est moins profond et possède moins de feuilles.

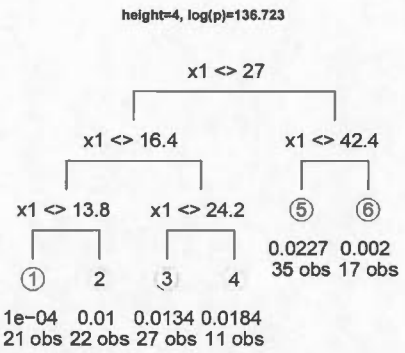


Figure 1.12 : Estimation de l'arbre de régression bayésien MAP (simulation d'accidents de motos)

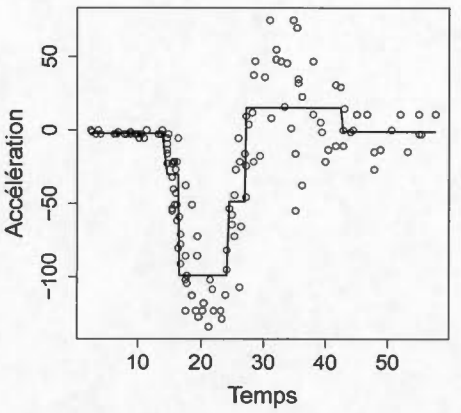


Figure 1.13 : Ajustement de courbe par estimation de l'arbre de régression bayésien MAP (simulation d'accidents de motos)

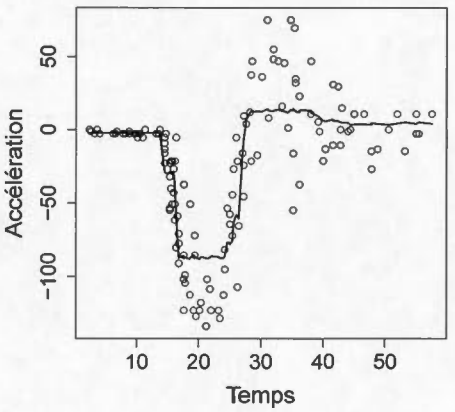


Figure 1.14 : Ajustement de courbe par estimation de la prédiction moyenne des arbres *a posteriori* à chaque observation $\mathbf{x}_i$ $(i = 1, \ldots, n)$ (simulation d'accidents de motos)

Dans la Figure 1.13, on voit l'ajustement de la courbe par estimation de l'arbre MAP. Ayant peu de feuilles, l'ajustement de la courbe de cet arbre n'est pas très précis. On constate la même chose avec la courbe d'estimation des prédictions moyennes des arbres *a posteriori* dans la Figure 1.14. Dans le prochain chapitre, on montre qu'il est possible d'obtenir de meilleurs arbres lorsqu'on choisit plus adéquatement les hyperparamètres de la loi *a priori* $P(T)$. Ainsi, il est possible de rivaliser avec la méthode fréquentiste pour le meilleur ajustement des données.

CHAPITRE II

IMPORTANCE DE LA LOI *A PRIORI* $P(T)$

Dans ce chapitre, on discute des hyperparamètres de la loi *a priori* $P(T)$. En effet, le choix des hyperparamètres α et β est très important, car il permet de réduire la probabilité *a posteriori* des arbres jugés improbables et de se concentrer sur les arbres plus importants. Tel que mentionné au Chapitre 1, l'hyperparamètre α est un paramètre qui change en moyenne le nombre total de feuilles des arbres (Chipman *et al.*, 1998). L'hyperparamètre β, quant à lui, est un paramètre qui change en moyenne la profondeur des arbres (Chipman *et al.*, 1998).

2.1 Retour sur l'exemple des accidents de motos

Pour simplicité, on a précédemment utilisé les hyperparamètres par défaut ($\alpha = 0.50$ et $\beta = 2$) pour implanter le modèle d'arbre de régression bayésien dans l'exemple d'accidents de motos. Pour illustrer l'importance de bien choisir ces paramètres, une nouvelle implantation du modèle est effectuée, mais cette fois avec des hyperparamètres α et β choisis en fonction de notre connaissance de la courbe. On trouve que l'on obtient une meilleure estimation de la courbe par l'arbre MAP avec ce nouveau choix d'hyperparamètres de la loi *a priori* $P(T)$. De plus, on obtient que les prédictions moyennes *a posteriori* des arbres sont meilleures lorsque les hyperparamètres α et β sont bien choisis. Pour déterminer

la qualité de l'ajustement de la courbe par l'arbre MAP, on utilise le coefficient de détermination (R^2). Pour déterminer la qualité des moyennes prédictives *a posteriori*, on utilise le coefficient de détermination de validation croisé (R^2_{CV}). Ces coefficients sont présentés dans les deux paragraphes suivants.

Le coefficient de détermination (R^2) peut être défini de la façon suivante :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{2.1}$$

où $\overline{y}$ est la moyenne échantillonnale de la variable expliquée et $\hat{f}(\mathbf{x}_i)$ est la prédiction de la variable expliquée par le modèle statistique en question (dans ce cas, l'arbre de régression) de la i-ème observation. Le R^2 représente le degré d'ajustement de la courbe par le modèle statistique. Lorsque $R^2 = 1$, on a un ajustement parfait, tandis que lorsque $R^2 = 0$, on a un ajustement équivalent à l'utilisation de $\overline{y}$ pour prédire la variable expliquée y_i ($i = 1, ..., n$). Tel que discuté au chapitre précédent, l'estimation d'un arbre CART fréquentiste consiste généralement à minimiser $\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2$ (Hastie *et al.*, 2009) et par conséquent, à maximiser le R^2. Il est à noter que certains algorithmes utilisent plutôt le concept d'entropie pour estimer les arbres fréquentistes, mais utilisent quand même le R^2 comme mesure de qualité de l'estimation (Therneau *et al.*, 2014). Le R^2 est donc considéré une mesure valide pour déterminer le degré d'ajustement des arbres de régression.

La validation croisée est une technique permettant d'estimer la capacité prédictive d'un modèle statistique (Hastie *et al.*, 2009). Pour se faire, on sépare nos données en un certain nombre K d'échantillons. Pour chaque échantillon, on construit le modèle en question à partir des $K - 1$ autres échantillons et on prédit la variable expliquée associée à chacune des observations de cet échantillon. Ainsi, on obtient des prédictions de la variable expliquée pour chaque observation de notre jeu de données. Avec ces prédictions, il est possible de calculer le coefficient de

détermination de validation croisée (R^2_{CV}) par la formule suivante :

$$R^2_{CV} = 1 - \frac{\sum_{r=1}^{K} \sum_{s=1}^{n_r} (y_{rs} - \hat{f}_{-r}(\mathbf{x}_{rs}))^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2},$$

où r ($r = 1, \ldots, K$) représente le r-ème échantillon, n_r est le nombre d'observations dans le r-ème échantillon, s ($s = 1, \ldots, n_r$) représente la s-ème observation du r-ème échantillon et $\hat{f}_{-r}(\mathbf{x}_{rs})$ est la prédiction de la variable expliquée de la s-ème observation du r-ème échantillon par le modèle statistique obtenu en utilisant tous les échantillons sauf le r-ème. Dans le cas de la validation croisée, cette mesure représente le degré d'ajustement des observations hors échantillon plutôt que le degré d'ajustement des observations de l'échantillon. Le R^2_{CV} est donc un bon indicateur de la capacité prédictive d'un modèle statistique (Hastie *et al.*, 2009).

2.2 Choix de $P(T)$ pour l'exemple d'accidents de motos

En examinant la Figure 1.1, on constate que la fonction à estimer contient plusieurs pentes très abruptes. La fonction décroît rapidement à 14 ms jusqu'à l'atteinte de son minimum à 21 ms. Ensuite, la fonction augmente rapidement jusqu'à 31 ms et décroît graduellement jusqu'à 40 ms. La fonction passe donc rapidement d'une valeur à une autre ; on peut donc s'attendre à ce que la profondeur optimale de l'arbre soit plutôt grande. De plus, l'arbre optimal fréquentiste ayant 8 feuilles (Figure 1.3), on doit s'assurer de ne pas attribuer une probabilité *a priori* trop petite aux arbres possédant ce nombre de feuilles.

Les arbres associés au choix des hyperparamètres $\alpha = 0.50$ et $\beta = 2$ ont, en théorie, peu de feuilles (moins de 2.1 feuilles en moyenne [Chipman *et al.*, 1998]) et sont peu profonds. Avec ce choix d'hyperparamètres, les arbres avec 8 feuilles, tel que l'arbre optimal fréquentiste, sont improbables. En utilisant les hyperparamètres $\alpha = 0.95$ et $\beta = 0.50$, on s'attend à avoir des arbres avec beaucoup plus de feuilles (7 feuilles en moyenne [Chipman *et al.*, 1998]) et plus profonds, car β est plus

petit.

2.2.1 Comparaison des arbres MAP avec l'arbre fréquentiste optimal

La Figure 2.1 montre l'estimation de l'arbre de régression MAP après cinq redémarrages lorsque $\alpha = 0.95$ et $\beta = 0.50$. On constate que l'arbre est très semblable à l'arbre fréquentiste de la Figure 1.3. Les règles de séparation sont similaires et le nombre de feuilles dans les deux arbres est le même. L'arbre MAP obtenu lorsque $\alpha = 0.95$ et $\beta = 0.50$ est donc beaucoup plus semblable à l'arbre fréquentiste que ne l'est l'arbre MAP obtenu lorsque $\alpha = 0.50$ et $\beta = 2$. Ceci amène à émettre l'hypothèse que cet arbre est meilleur que l'arbre obtenu avec les hyperparamètres par défaut.

Dans la Figure 2.2, on peut observer l'ajustement de la courbe par l'arbre fréquentiste, par l'arbre bayésien MAP lorsque $\alpha = 0.50$ et $\beta = 2$ et par l'arbre bayésien MAP lorsque $\alpha = 0.95$ et $\beta = 0.50$. La courbe semble plutôt bien ajustée par l'arbre fréquentiste et l'arbre bayésien MAP lorsque $\alpha = 0.95$ et $\beta = 0.50$. Ces deux arbres ajustent beaucoup mieux le minimum de la courbe à 20 ms et le maximum à 35 ms. Par contre, l'arbre fréquentiste semble tout de même plus précis au niveau de l'ajustement de la courbe au maximum que l'arbre bayésien lorsque $\alpha = 0.95$ et $\beta = 0.50$. Ces observations peuvent être confirmées en inspectant la valeur du R^2 associé à chacun des arbres. Pour l'arbre fréquentiste, le R^2 est de 0.81. Pour les arbres bayésiens, on a que $R^2 = 0.75$ lorsque $\alpha = 0.50$ et $\beta = 2$ et $R^2 = 0.80$ lorsque $\alpha = 0.95$ et $\beta = 0.50$. Cela montre que le nouveau choix d'hyperparamètres ($\alpha = 0.95$ et $\beta = 0.50$) amène à un meilleur ajustement que précédemment, mais que celui-ci reste légèrement inférieur à l'ajustement obtenu de l'arbre fréquentiste.

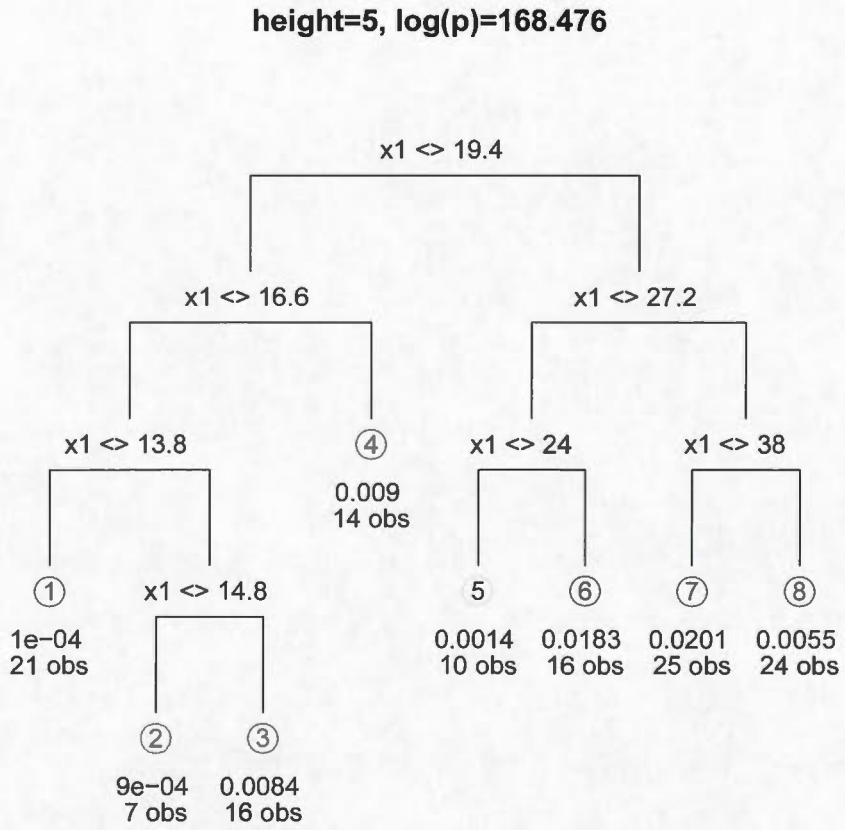


Figure 2.1 : Estimation de l'arbre de régression bayésien MAP avec $\alpha = 0.95$ et $\beta = 0.50$ (simulation d'accidents de motos)

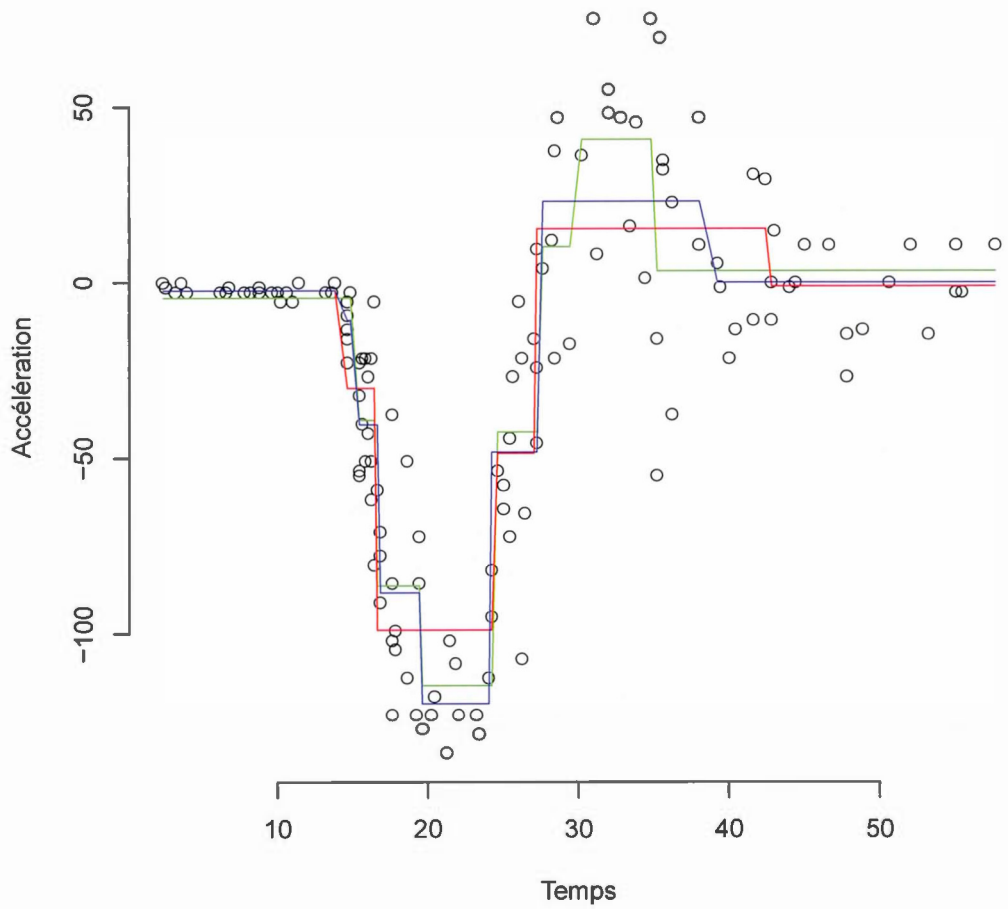


Figure 2.2 : Ajustement de courbe par arbre de régression fréquentiste (en vert) et par arbre de régression bayésien MAP estimé pour deux différents choix d'hyperparamètres de la loi *a priori* de T, $(\alpha = 0.50, \beta = 2)$ (en rouge) et $(\alpha = 0.95, \beta = 0.50)$ (en bleu) (simulation d'accidents de motos).

Il est important de noter qu'il est toutefois possible que l'arbre bayésien MAP possède une aussi bonne capacité prédictive que l'arbre fréquentiste, malgré le fait que son R^2 soit plus petit. L'estimation de l'arbre MAP est utile pour l'interprétation et la représentation graphique mais la meilleur prédiction obtenue d'un modèle bayésien est celle définie comme étant la moyenne de la loi prédictive *a posteriori* (Hoeting *et al.*, 1999). Dans la prochaine section, on compare donc la capacité prédictive de l'arbre fréquentiste optimal à celle des arbres bayésiens sur la base de leurs prédictions moyennes *a posteriori*.

2.2.2 Comparaison des capacités prédictives moyennes des arbres bayésiens et arbre optimal fréquentiste

Précédemment, on a observé directement l'effet des hyperparamètres sur les arbres obtenus et la qualité des ajustements à travers le R^2. Pour vérifier la capacité prédictive des arbres, on utilise maintenant la validation croisée en 5 échantillons. Pour chaque échantillon, on détermine la prédiction moyenne *a posteriori* des arbres.

Lorsque $\alpha = 0.50$ et $\beta = 2$, le R^2_{CV} est de 0.62 tandis que lorsque $\alpha = 0.95$ et $\beta = 0.50$, le R^2_{CV} est de 0.68. La validation croisée a été répétée plusieurs fois et on constate que les estimations du R^2_{CV} sont robustes. Il apparaît donc que lorsqu'on choisit les hyperparamètres de $P(T)$ en fonction de notre connaissance de la courbe, on obtient des arbres qui prédisent mieux la courbe.

Pour l'arbre fréquentiste, on obtient un R^2_{CV} de 0.7. L'arbre fréquentiste a donc une capacité prédictive légèrement supérieure à celle de l'arbre bayésien lorsque $\alpha = 0.95$ et $\beta = 0.50$. On peut suspecter que cela est dû en grande partie à la variabilité des arbres obtenue avec ce choix d'hyperparamètres. Avec un choix encore plus approprié de α et β, il serait possible de rivaliser encore plus avec l'arbre

fréquentiste. C'est pourquoi il est important de bien savoir choisir les hyperparamètres de la loi *a priori* de T.

Ainsi, choisir adéquatement les hyperparamètres α et β peut permettre d'obtenir des arbres ajustant beaucoup mieux la variable expliquée et peut permettre d'avoir une meilleure capacité de prédiction. Dans le prochain chapitre, on approfondie l'étude de la loi *a priori* de T et dans le chapitre subséquent, on explique comment bien choisir les hyperparamètres α et β.

CHAPITRE III

LA LOI *A PRIORI* $P(T)$

Puisque $P(T)$ dépend de $p_{RULE}(\rho|\eta, T)$, qui dépend elle-même de la matrice X, il serait davantage correct de désigner la loi a priori de l'arbre par $P(T|X)$ plutôt que par $P(T)$, tel que fait dans Chipman et al. (1998). Dans ce chapitre, on discute en détail de la loi *a priori* $P(T)$ et de l'influence de la matrice des variables explicatives X sur cette loi. Il est important de noter que, à notre connaissance, aucune discussion n'a été amorcée sur la dépendance de $P(T)$ sur X dans la littérature des modèles d'arbres bayésiens. Pourtant, l'effet de X est très important. La loi $P(T)$ dépend non seulement des hyperparamètres α et β, mais aussi du nombre de variables et d'observations distinctes représentées dans la matrice X. On discute ensuite en détail du choix des hyperparamètres de la loi *a priori* de T. Plus particulièrement, on démontre comment déterminer la moyenne et la variance du nombre de feuilles idéales des arbres et on explique comment choisir α et β de façon à obtenir une loi *a priori* avec la moyenne et la variance du nombre de feuilles désirés.

3.1 Exemple illustratif

On présente maintenant un exemple simple pour illustrer l'effet de la matrice X sur $P(T)$. Imaginons une seule variable explicative mesurée sur 10 sujets, $x_1 =$

$(1, 1, 1, 2, 2, 3, 3, 3, 4, 4)$; les choix possibles de règles de séparation sont donc $x_1 \leq 1$, $x_1 \leq 2$ et $x_1 \leq 3$. Étant donné qu'il n'y a que quatre observations distinctes, il est impossible d'avoir plus de quatre feuilles dans un arbre. De plus, sachant qu'il n'y a que trois règles de séparation possibles, il est impossible d'avoir un arbre avec une profondeur plus grande que trois. Rappelons que la profondeur d'un arbre est définie comme étant le nombre maximal de règles de séparation consécutives amenant à une des feuilles dans l'arbre construit. Ainsi, peu importe le choix des hyperparamètres α et β, la loi *a priori* de l'arbre attribue implicitement une probabilité nulle aux arbres avec plus de quatre feuilles et avec une profondeur plus grande que trois. L'influence de X sur $P(T)$ est donc très grande dans ce cas-ci : moins on a d'observations distinctes, plus l'influence de X est importante.

Tel qu'illustré dans l'exemple précédent, l'effet de X est grand lorsque le nombre d'observations distinctes est petit. On peut donc se demander si l'effet de la matrice X devient négligeable avec un nombre assez grand d'observations. Dans ce chapitre, on constate que l'influence de X sur $P(T)$ peut être petite ou grande dépendamment du choix de α et β.

L'influence du nombre d'observations distinctes sur $P(T)$, lorsqu'on a une seule variable, est expliquée en détail d'abord dans le cas où $\beta = 0$ et ensuite dans le cas général où $\beta \neq 0$ (séparé en une section sur le choix conventionnel de $\alpha \in [0, 1]$ et une autre section sur $\alpha > 1$).

3.2 Cas $\beta = 0$

La loi *a priori* de T introduite par Chipman *et al.* (1998) est définie par un processus récursif tel quel spécifié dans le premier chapitre. La forme et la grandeur de l'arbre induites par ce processus récursif dépend entièrement de la probabilité de séparation $p_{SPLIT}(\eta, T) = \alpha(1+d_\eta)^{-\beta}$ et de la matrice des variables explicatives

X. Si on présume que $\beta = 0$, alors $p_{SPLIT}(\eta, T) = \alpha$, où $\alpha \in [0, 1]$. Chaque arbre bayésien avec b feuilles a donc une probabilité *a priori* égale à $\alpha^{b-1}(1-\alpha)^b$. De plus, il existe C_{b-1} arbres avec b feuilles, où C_{b-1} est le $b-1$-ième nombre de Catalan (Koshy, 2008).

Définition 3.1. *Les nombres de Catalan (Koshy, 2008) sont définis par la relation de récurrence*

$$C_0 = 1 \quad et \quad C_{n+1} = \sum_{b=0}^{n} C_b C_{n-b} \ \ lorsque\ n \geq 0.$$

Lorsque $\beta = 0$, la probabilité du nombre de feuilles dans un arbre est donc définie de la façon suivante.

$$P(N = b) = C_{b-1}\alpha^{b-1}(1-\alpha)^b \ \ si\ 1 \leq b < \infty, \tag{3.1}$$

où b est le nombre de feuilles dans un arbre. Sachant que $P(1 \leq N \leq \infty)$ doit être égal à 1, on peut déduire que $P(N = \infty) = \max\{0, 2-1/\alpha\}$ (la démonstration est incluse dans l'Appendice A). On peut maintenant amener la proposition suivante :

Proposition 3.2. *Soit T un arbre défini de façon récursive avec probabilité de séparation α. Si N est le nombre de feuilles de T, alors $P(N = b) = p(b)$, où*

$$p(b) = \begin{cases} C_{b-1}\alpha^{b-1}(1-\alpha)^b & si\ 1 \leq b < \infty, \\ max\{0, 2-1/\alpha\} & si\ b = \infty \end{cases} \quad pour\ tout\ \alpha \in [0,1].$$

On peut ensuite déterminer analytiquement l'espérance et la variance du nombre de feuilles selon α (les démonstrations sont incluses dans l'Appendice A).

Proposition 3.3. *Soit T un arbre défini de façon récursive avec probabilité de séparation α. Si N est le nombre de feuilles de T, alors*

$$E(N) = \begin{cases} \frac{(1-\alpha)}{(1-2\alpha)} & si\ \alpha \leq \frac{1}{2}; \\ \infty & si\ \alpha > \frac{1}{2}. \end{cases}$$

Proposition 3.4. *Soit T un arbre défini de façon récursive avec probabilité de séparation de α. Si N est le nombre de feuilles de T, alors*

$$Var(N) = \begin{cases} \frac{(1-\alpha)\alpha}{(1-2\alpha)^3} & si\ \alpha \leq \frac{1}{2}; \\ \infty & si\ \alpha > \frac{1}{2}. \end{cases}$$

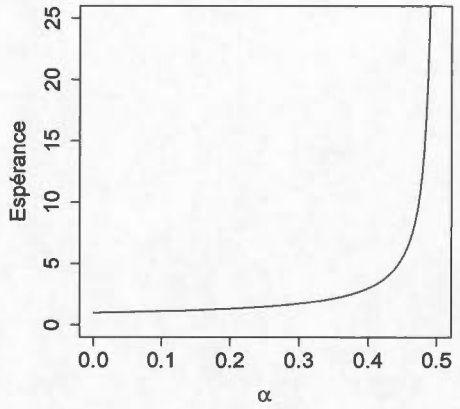


Figure 3.1 : Espérance du nombre de feuilles en fonction de α dans l'intervalle $[0, 0.50)$ lorsque $\beta = 0$

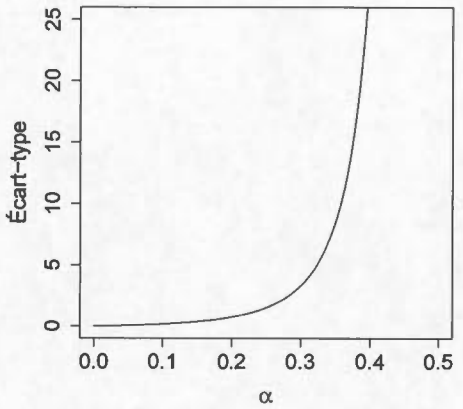


Figure 3.2 : Écart-type du nombre de feuilles en fonction de α dans l'intervalle $[0, 0.50)$ lorsque $\beta = 0$

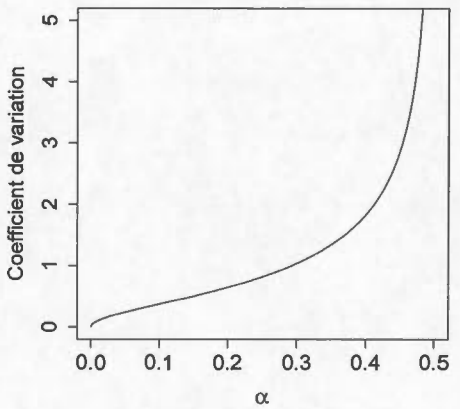


Figure 3.3 : Coefficient de variation du nombre de feuilles en fonction de α dans l'intervalle $[0, 0.50)$ lorsque $\beta = 0$

Dans les Figures 3.1, 3.2 et 3.3, on peut observer l'effet de la valeur de α sur l'espérance, l'écart-type et le coefficient de variation du nombre de feuilles dans un arbre, lorsque $\alpha \in [0, 0.5)$. Le coefficient de variation est défini comme étant l'écart-type divisé par l'espérance. On constate que l'espérance, l'écart-type et le coefficient de variation convergent tous vers l'infini lorsque α s'approche de 0.50. Ceci peut être vérifié analytiquement en calculant la limite de ces quantités lorsque α s'approche de 0.50.

Proposition 3.5.

$$\lim_{\alpha \to 0.50^-} E(N) = \infty$$

Démonstration.

$$\lim_{\alpha \to 0.50^-} E(N) = \lim_{\alpha \to 0.50^-} \frac{(1-\alpha)}{(1-2\alpha)} = \frac{1-0.50}{1-1} = \frac{0.50}{0} = \infty$$

□

Proposition 3.6.

$$\lim_{\alpha \to 0.50^-} Var(N) = \infty$$

Démonstration.

$$\lim_{\alpha \to 0.50^-} Var(N) = \lim_{\alpha \to 0.50^-} \frac{\alpha(1-\alpha)}{(1-2\alpha)^3} = \frac{0.50(1-0.50)}{(1-1)^3} = \frac{0.25}{0} = \infty$$

□

Proposition 3.7.

$$\lim_{\alpha \to 0.50^-} \frac{\sqrt{Var(N)}}{E(N)} = \infty$$

Démonstration.

$$\begin{aligned}
\lim_{\alpha \to 0.50^-} \frac{\sqrt{Var(N)}}{E(N)} &= \lim_{\alpha \to 0.50^-} \frac{\sqrt{\frac{\alpha(1-\alpha)}{(1-2\alpha)^3}}}{\frac{(1-\alpha)}{(1-2\alpha)}} \\
&= \lim_{\alpha \to 0.50^-} \frac{\sqrt{\alpha(1-\alpha)}(1-2\alpha)}{(1-2\alpha)^{3/2}(1-\alpha)} \\
&= \lim_{\alpha \to 0.50^-} \frac{\sqrt{\alpha}}{\sqrt{(1-2\alpha)(1-\alpha)}} \\
&= \frac{\sqrt{0.50}}{\sqrt{(1-1)(1-0.50)}} \\
&= \frac{\sqrt{0.50}}{0} \\
&= \infty
\end{aligned}$$

□

Dans les Figures 3.1, 3.2 et 3.3, on remarque aussi que lorsque α est plus petit que 0.40, la courbe de l'espérance augmente très légèrement, de façon presque linéaire, tandis que lorsque α est plus grand que 0.40, la courbe augmente très rapidement, de façon exponentielle. Si on veut un arbre avec beaucoup de feuilles, il faut donc choisir un α entre 0.40 et 0.50. Le coefficient de variation devient assez grand lorsque α est supérieur à 0.40. Lorsque qu'on choisit un α de façon à obtenir des arbres avec beaucoup de feuilles en moyenne, il faut donc s'attendre à voir une très grande variété de nombres de feuilles dans les arbres *a priori* générés.

Il est à noter que la loi de probabilité du nombre de feuilles dans un arbre est seulement valide si l'on présume qu'on a un nombre infini d'observations distinctes pour chaque variable explicative. En pratique, la probabilité *a priori* des arbres est influencée par X. Ce sont donc les résultats théoriques dans un cas idéal. En effet, le processus récursif de création d'un arbre bayésien peut facilement être adapté pour prendre en compte l'impact de X en ajoutant une restriction

sur la croissance de l'arbre. Il s'agit d'empêcher un noeud de se séparer s'il n'y a qu'une seule observation unique dans ce noeud. En produisant un algorithme représentant ce processus récursif, on peut estimer l'espérance et la variance du nombre de feuilles dans un arbre lorsqu'on a une seule variable explicative. Cela nous permet d'observer l'effet du nombre d'observations distinctes sur l'espérance et la variance du nombre de feuilles.

Dans la Figure 3.4 et la Figure 3.5, on observe l'effet du nombre d'observations distinctes sur l'espérance estimée du nombre de feuilles dans un arbre.

Lorsque $\alpha = 0.25$, la valeur théorique de l'espérance du nombre de feuilles qu'on obtient, en présumant qu'il y a un nombre infini d'observations distinctes, est de 1.5 feuilles. On voit dans la Figure 3.4 qu'après seulement 50 observations distinctes, on obtient une moyenne d'environ 1.45 feuilles. Avec un peu plus d'observations distinctes, la moyenne du nombre de feuilles est très près de 1.5 feuilles. On atteint donc un résultat plus ou moins équivalent au résultat attendu, lorsque $\alpha = 0.25$, peu importe le nombre d'observations distinctes.

Lorsque $\alpha = 0.45$, la valeur théorique de l'espérance du nombre de feuilles qu'on obtient, en présumant un nombre infini d'observations distinctes, est de 5.5 feuilles. C'est un nombre de feuilles plus grand que lorsque $\alpha = 0.25$, mais qui reste tout de même relativement petit. On remarque qu'on n'atteint pas la valeur théorique du cas idéal même lorsque le nombre d'observations distinctes est égal à 20 000. La courbe croît très lentement, de façon linéaire, à partir d'environ 2000 observations. Lorsqu'on a 1 000 000 d'observations distinctes, l'espérance estimée du nombre de feuilles est 4.49. On constate ainsi que l'espérance augmente avec le nombre d'observations distinctes, mais cette augmentation est tellement graduelle qu'il n'est pas réaliste d'imaginer qu'on puisse avoir assez d'observations pour rendre l'effet de X négligeable. Le choix de $\alpha = 0.45$ apparaît donc inapproprié pour

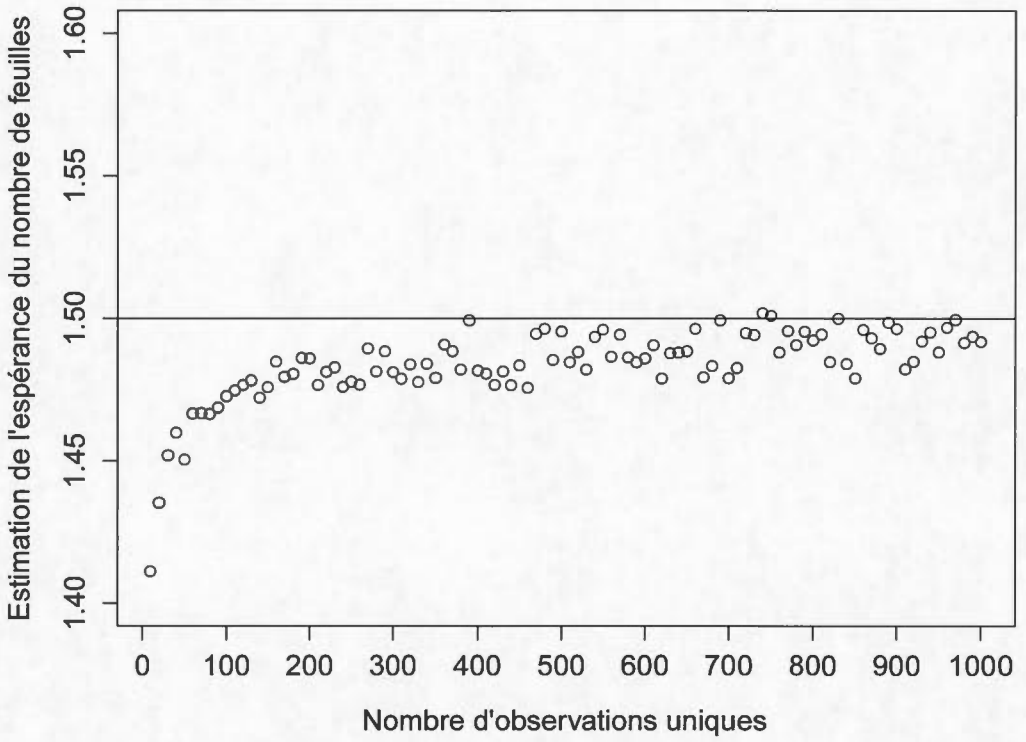


Figure 3.4 : Estimation de l'espérance du nombre de feuilles en fonction du nombre d'observations distinctes par 50 000 arbres générés lorsque $\alpha = 0.25$. Le trait horizontal représente la valeur de l'espérance lorsqu'on présume une infinité d'observations distinctes.

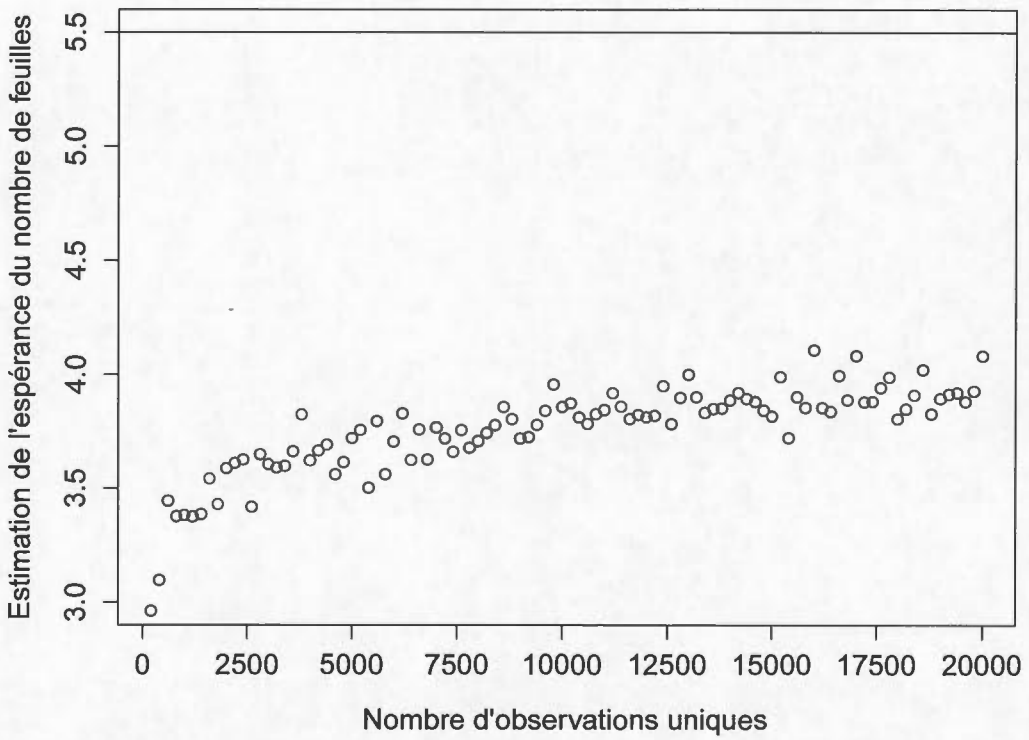


Figure 3.5 : Estimation de l'espérance du nombre de feuilles en fonction du nombre d'observations distinctes pour 5000 arbres générés lorsque $\alpha = 0.45$. Le trait horizontal représente la valeur de l'espérance lorsqu'on présume infinité d'observations distinctes.

construire des arbres avec une moyenne de 5.5 feuilles.

Ces résultats montrent que lorsque l'on cherche à avoir des arbres de taille moyenne ou grande, l'effet de X devient très important sur le nombre moyen de feuilles des arbres. Il faut donc prendre en compte le nombre d'observations distinctes afin de bien choisir α pour atteindre le nombre de feuilles moyen désiré. Les tableaux présentés dans la prochaine section permettent de déterminer l'espérance approximative du nombre de feuilles dans un arbre en fonction du nombre d'observations distinctes, de α et de β.

Jusqu'à maintenant, on n'a discuté que du cas simple avec $\beta = 0$. Dans la section qui suit on discute du cas plus général où $\beta \neq 0$.

3.3 Cas $\beta \neq 0$ et $\alpha \in [0, 1]$

Le paramètre β sert à imposer une certaine restriction sur la profondeur des arbres distribués selon la loi a priori $P(T)$. Si on choisit un grand β ($\beta \geq 0$), on devrait empêcher les arbres de devenir profonds et ainsi obtenir des arbres plus équilibrés (Chipman *et al.*, 1998). Indirectement, β semble aussi influencer le nombre de feuilles moyen des arbres obtenus, justement à cause de la restriction qu'il impose sur la profondeur des arbres (Chipman *et al.*, 1998). Il est donc intéressant d'étudier l'effet de α et β sur le nombre de feuilles moyen et sur la profondeur moyenne dans un arbre.

Contrairement au cas précédent avec $\beta = 0$, lorsque $\beta \neq 0$, les valeurs de $\alpha \in [0.50, 1]$ n'amènent pas à une espérance infinie du nombre de feuilles (Chipman *et al.*, 1998). On peut donc choisir un α entre 0 et 1. Chipman *et al.* (1998) recommandent quatre choix de (α, β) : (0.50, 0.50), (0.95, 0.50), (0.95, 1) et (0.95, 1.5). De plus, dans les paquets `tgp` (Gramacy, 2007) et `BayesTree` (Chipman et McCulloch, 2014) du langage de programmation `R`, le choix d'hyperparamètres

(α, β) est respectivement (0.50, 2) et (0.95,2) par défaut. Lorsque $\alpha < 0.50$ ou lorsque $\beta > 1$, on pénalise grandement le nombre moyen de feuilles dans les arbres ; c'est pourquoi on ne considère pas les α plus petits que 0.50 ou les β plus grands que 1 dans les analyses qui suivent.

Lorsque $\beta \neq 0$, on ne peut définir la distribution du nombre de feuilles ou de la profondeur de façon explicite. On doit donc avoir recours à l'algorithme récursif de création d'arbres pour approximer l'espérance et l'écart-type du nombre de feuilles et de la profondeur. Comme précédemment, cet algorithme peut facilement être adapté pour prendre en compte l'effet d'un nombre fini d'observations distinctes sur les résultats. Pour chaque choix de α et β, on simule 5000 arbres à partir de l'algorithme récursif. Ainsi, on obtient une estimation précise de l'espérance et de l'écart-type du nombre de feuilles et de la profondeur.

Dans les Tableaux 3.1 et 3.2, on peut observer l'effet du choix de α et β sur le nombre moyen de feuilles et sur la profondeur moyenne des arbres, dans un cas idéal où l'on a un nombre infini d'observations distinctes avec une seule variable. On remarque que β pénalise grandement le nombre de feuilles moyen dans les arbres. Ainsi, si on cherche à diminuer la profondeur moyenne des arbres, tout en gardant le même nombre de feuilles moyen, il faut non seulement augmenter β de façon à atteindre la profondeur désirée, mais aussi augmenter α de façon à retrouver le nombre de feuilles moyen de départ.

Il existe deux façons d'obtenir un certain nombre moyen de feuilles visé en spécifiant les valeurs des hyperparamètres α et β. On peut choisir un petit α afin d'obtenir des arbres avec peu de feuilles en moyenne et choisir un petit β pour ne pas trop pénaliser le nombre moyen de feuilles dans les arbres. Autrement, on peut choisir un grand α de façon à obtenir des arbres avec beaucoup de feuilles en moyenne et pénaliser grandement le nombre moyen de feuilles en choisissant un

Tableau 3.1 : Espérance et écart-type du nombre de feuilles lorsqu'on a une infinité d'observations distinctes pour $\alpha = (0.60, 0.80, 0.90, 0.95)$ et $\beta = (0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1)$

	$\alpha = 0.60$		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$	
β	μ	σ	μ	σ	μ	σ	μ	σ
0.30	3.91	4.50	12.39	14.59	28.70	30.20	47.27	45.31
0.35	3.56	3.70	9.30	10.16	17.43	17.28	25.02	23.19
0.40	3.22	3.21	7.31	7.32	12.40	11.90	16.47	14.50
0.45	3.19	3.01	6.15	5.84	9.50	8.36	12.15	10.25
0.50	2.81	2.47	5.42	4.77	7.86	6.69	9.65	7.72
0.75	2.38	1.71	3.63	2.57	4.51	3.03	5.12	3.35
1	2.17	1.32	2.99	1.77	3.53	2.02	3.89	2.12

grand β. En choisissant de petits α et β, on peut donc obtenir le même nombre moyen de feuilles qu'en choisissant de grands α et β. On peut observer ce phénomène dans le Tableau 3.1. Par exemple, si l'on choisit $(\alpha, \beta) = (0.60, 0.30)$, on a un nombre moyen de feuilles de 3.91 et si l'on choisit $(\alpha, \beta) = (0.95, 1)$, on obtient un nombre de feuilles moyen presque identique (3.89).

On remarque que lorsque $(\alpha, \beta) = (0.60, 0.30)$, la profondeur moyenne est de 2.14, avec écart-type de 2.72, et lorsque $(\alpha, \beta) = (0.95, 1)$, la profondeur moyenne est de 2.42, avec écart-type de 1.45. Pour un nombre moyen de feuilles équivalent, on obtient donc une moyenne de profondeur légèrement plus grande lorsque α et β sont grands. Pour tous les choix de α et β amenant au même nombre moyen de feuilles dans le Tableau 3.1, on constate que la moyenne de la profondeur est semblable, mais légèrement plus élevée lorsque α et β sont grands (voir Tableau 3.2). Chipman *et al.* (1998) expliquent que le paramètre β sert à contrôler la

Tableau 3.2 : Espérance et écart-type de la profondeur lorsqu'on a une infinité d'observations distinctes pour $\alpha = (0.60, 0.80, 0.90, 0.95)$ et $\beta = (0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1)$

	$\alpha = 0.60$		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$	
β	μ	σ	μ	σ	μ	σ	μ	σ
0.30	2.14	2.72	5.60	5.25	9.65	6.99	12.92	7.87
0.35	1.96	2.38	4.64	4.25	7.32	5.36	9.20	5.80
0.40	1.77	2.17	3.95	3.50	5.88	4.34	7.22	4.62
0.45	1.76	2.09	3.43	3.06	4.96	3.57	5.94	3.81
0.50	1.52	1.79	3.14	2.70	4.31	3.12	5.08	3.24
0.75	1.24	1.39	2.16	1.80	2.74	1.90	3.12	1.96
1	1.09	1.14	1.77	1.39	2.17	1.46	2.42	1.45

profondeur des arbres *a priori*, mais le Tableau 3.2 révèle qu'en pénalisant pour la profondeur, on pénalise également le nombre de feuilles. Si on veut garder le même nombre moyen de feuilles, mais réduire la profondeur moyenne, on doit réduire α et β. Cette réduction de profondeur est généralement très petite et est maximale lorsque $\beta = 0$. Il ne semble donc pas possible de modifier de façon significative la profondeur des arbres tout en gardant le même nombre de feuilles.

On remarque aussi, dans les Tableaux 3.1 et 3.2, que lorsqu'on choisit de grands α et β, on obtient de plus petits écarts-types sur le nombre de feuilles dans un arbre et sur la profondeur que lorsqu'on choisit de petits α et β. Par exemple, lorsque $(\alpha, \beta) = (0.60, 0.30)$, l'écart-type du nombre de feuilles est de 4.50 et l'écart-type de la profondeur est de 2.72. Toutefois, lorsque $(\alpha, \beta) = (0.95, 1)$, l'écart-type du nombre de feuilles est de 2.12 et l'écart-type de la profondeur est de 1.45. En fait, on peut observer ce phénomène avec tous les choix de α

et β auxquels correspondent le même nombre de feuilles moyen. Ainsi, β semble être principalement un paramètre affectant la variabilité des arbres plutôt qu'un paramètre réduisant la profondeur des arbres.

On va maintenant vérifier l'hypothèse qui stipule que d'augmenter α et β, de façon à obtenir le même nombre moyen de feuilles, réduit la variabilité des arbres et augmente de façon très modeste la moyenne de la profondeur lorsqu'on a un nombre fini d'observations distinctes (500 et 2000). Ensuite, on va analyser plus en détail la distribution du nombre de feuilles et celle de la profondeur des arbres lorsqu'on a de petits α et β comparés à lorsqu'on a de grands α et β.

Dans les Tableaux 3.3 et 3.4, on observe l'effet du choix de α et β sur le nombre moyen de feuilles et sur la profondeur moyenne des arbres, lorsqu'on a 500 observations distinctes avec une seule variable. Similairement, on constate que si l'on choisit de petits (α, β) et de grands (α, β) ayant approximativement la même moyenne du nombre de feuilles, la moyenne de la profondeur est légèrement plus grande lorsque α et β sont grands. Par exemple, lorsque $(\alpha, \beta) = (0.80, 0.30)$, la moyenne et l'écart-type du nombre de feuilles sont de 8.19 et 7.54 respectivement, et la moyenne et l'écart-type de la profondeur sont de 4.12 et 3.22 respectivement. Lorsqu'on a $(\alpha, \beta) = (0.95, 0.50)$, la moyenne et l'écart-type du nombre de feuilles sont de 8.05 et 5.50 respectivement, et la moyenne et l'écart-type de la profondeur sont de 4.39 et 2.46 respectivement.

Dans les Tableaux 3.5 et 3.6, on observe l'effet du choix de α et β sur le nombre moyen de feuilles et sur la profondeur moyenne des arbres, lorsqu'on a 2000 observations distinctes avec une seule variable. On arrive encore au même constat que précédemment. En exemple, quand $(\alpha, \beta) = (0.70, 0.45)$, la moyenne et l'écart-type du nombre de feuilles sont de 4.04 et 3.60 respectivement, et la moyenne et l'écart-type de la profondeur sont de 2.32 et 2.30 respectivement. Lorsque

Tableau 3.3 : Espérance et écart-type du nombre de feuilles lorsqu'on a 500 observations distinctes pour $\alpha = (0.80, 0.90, 0.95, 0.99)$ et $\beta = (0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1)$

	$\alpha = 0.8$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.99$	
β	μ	σ	μ	σ	μ	σ	μ	σ
0.30	8.19	7.54	13.21	10.64	17.15	12.50	20.85	13.63
0.35	7.01	6.31	10.63	8.46	13.24	9.73	16.23	10.79
0.40	5.99	5.19	8.91	6.92	11.05	7.94	13.20	8.58
0.45	5.44	4.60	7.65	5.84	9.23	6.47	10.70	7.05
0.50	4.96	3.93	6.77	4.91	8.05	5.50	9.21	5.92
0.75	3.54	2.38	4.33	2.77	5.00	3.05	5.40	3.11
1	2.92	1.73	3.49	1.96	3.80	1.99	4.08	2.14

Tableau 3.4 : Espérance et écart-type de la profondeur lorsqu'on a 500 observations distinctes pour $\alpha = (0.80, 0.90, 0.95, 0.99)$ et $\beta = (0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1)$

	$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.99$	
β	μ	σ	μ	σ	μ	σ	μ	σ
0.30	4.12	3.22	5.79	3.47	6.79	3.30	7.65	3.10
0.35	3.71	3.02	5.09	3.16	5.93	3.09	6.76	2.88
0.40	3.30	2.69	4.58	2.90	5.35	2.86	6.08	2.73
0.45	3.06	2.47	4.13	2.67	4.82	2.66	5.37	2.53
0.50	2.89	2.33	3.82	2.46	4.39	2.46	4.91	2.40
0.75	2.10	1.69	2.63	1.77	3.04	1.79	3.30	1.75
1	1.70	1.34	2.13	1.41	2.37	1.37	2.54	1.38

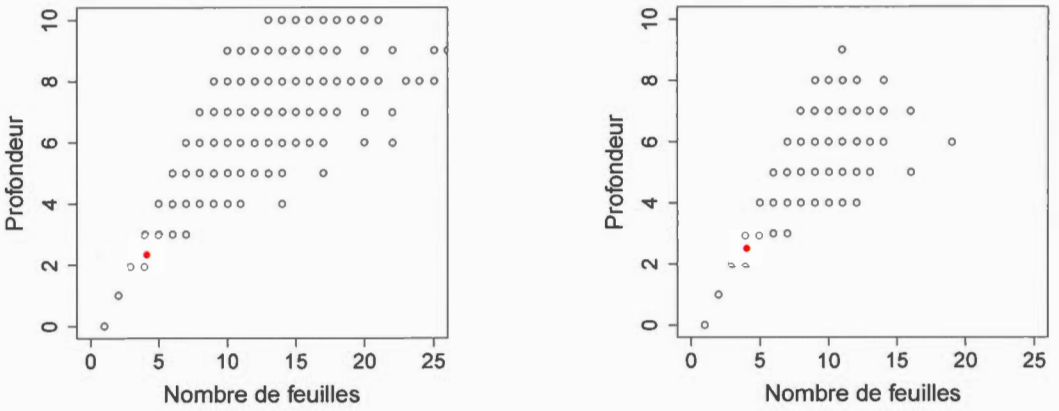
Tableau 3.5 : Espérance et écart-type du nombre de feuilles lorsqu'on a 2000 observations distinctes pour $\alpha = (0.70, 0.90, 0.95, 0.995)$ et $\beta = (0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1)$

	$\alpha = 0.70$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.995$	
β	μ	σ	μ	σ	μ	σ	μ	σ
0.30	5.67	5.94	16.16	13.67	21.11	16.34	27.96	18.88
0.35	5.00	4.88	12.54	10.41	15.80	12.39	20.10	14.28
0.40	4.41	4.20	9.94	8.14	12.75	9.70	15.39	10.84
0.45	4.04	3.60	8.43	6.60	10.38	7.76	12.32	8.69
0.50	3.71	3.16	7.21	5.61	8.45	6.17	10.21	7.04
0.75	2.92	2.05	4.49	2.94	5.05	3.14	5.52	3.29
1	2.52	1.53	3.49	1.95	3.79	2.05	4.13	2.15

$(\alpha, \beta) = (0.995, 1)$, la moyenne et l'écart-type du nombre de feuilles sont de 4.13 et 2.15 respectivement, et la moyenne et l'écart-type de la profondeur sont de 2.59 et 1.39 respectivement.

Tableau 3.6 : Espérance et écart-type de la profondeur lorsqu'on a 2000 observations distinctes pour $\alpha = (0.70, 0.90, 0.95, 0.995)$ et $\beta = (0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1)$

	$\alpha = 0.70$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.995$	
β	μ	σ	μ	σ	μ	σ	μ	σ
0.30	3.00	3.04	6.54	3.95	7.66	3.81	8.96	3.53
0.35	2.77	2.73	5.70	3.54	6.63	3.52	7.67	3.33
0.40	2.47	2.49	4.95	3.22	5.92	3.26	6.71	3.10
0.45	2.32	2.30	4.49	2.91	5.24	2.97	5.89	2.90
0.50	2.13	2.09	4.00	2.69	4.56	2.69	5.26	2.63
0.75	1.66	1.56	2.73	1.86	3.07	1.85	3.38	1.84
1	1.38	1.26	2.14	1.42	2.36	1.43	2.59	1.39

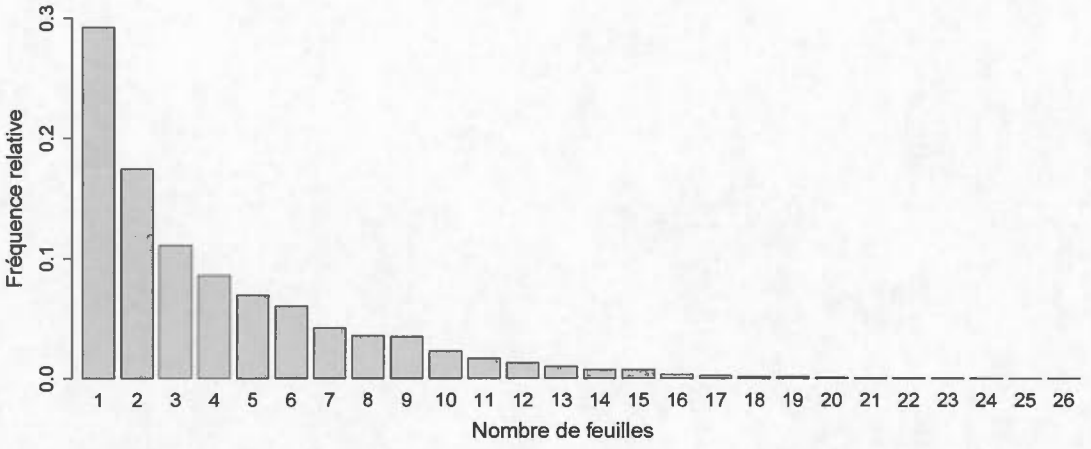


(a) (b)

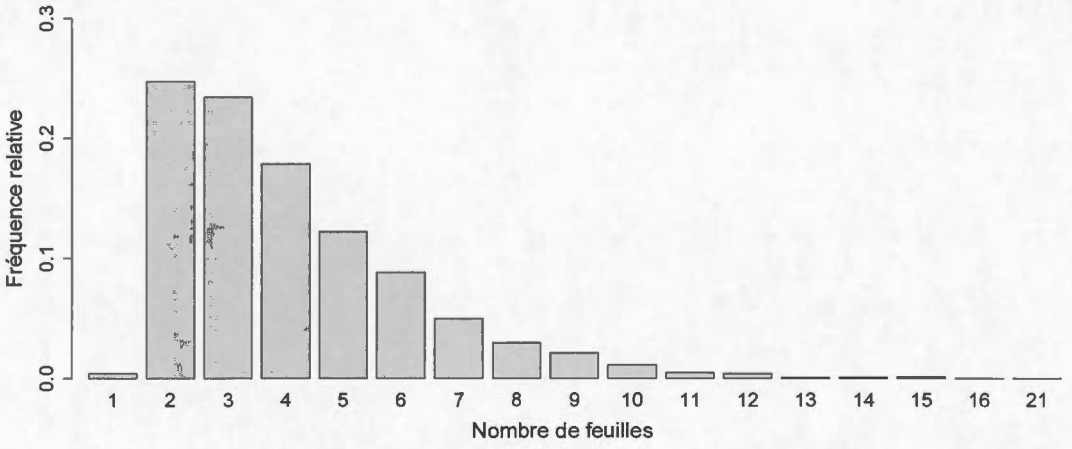
Figure 3.6 : Profondeur en fonction du nombre de feuilles observée dans un échantillon de 5000 arbres générés lorsqu'on a 2000 observations distinctes avec : (a) $(\alpha, \beta) = (0.70, 0.45)$; (b) $(\alpha, \beta) = (0.995, 1)$. Le point rouge représente la moyenne du nombre de feuilles et de la profondeur.

Dans les Figures 3.6(a) et 3.6(b), on inspecte plus en détail le cas lorsqu'on a 2000 observations avec $(\alpha, \beta) = (0.70, 0.45)$ et $(\alpha, \beta) = (0.995, 1)$. Tel que spécifié dans notre hypothèse, la moyenne de la profondeur est légèrement plus grande lorsque $(\alpha, \beta) = (0.995, 1)$. De plus, on constate qu'il y a beaucoup moins d'arbres profonds et d'arbres avec beaucoup de feuilles lorsque $(\alpha, \beta) = (0.995, 1)$. Choisir de grands α et β semble donc empêcher les arbres d'être très grands (avec beaucoup de feuilles et une grande profondeur). La moyenne de la profondeur étant tout de même légèrement plus grande lorsque α et β sont grands, ceci suggère qu'on doit aussi avoir une réduction du nombre de petits arbres (avec peu de feuilles et une petite profondeur).

Dans les Figures 3.7(a), 3.7(b), 3.8(a) et 3.8(b), on peut voir les fréquences relatives du nombre de feuilles et de la profondeur des 5000 arbres générés. On

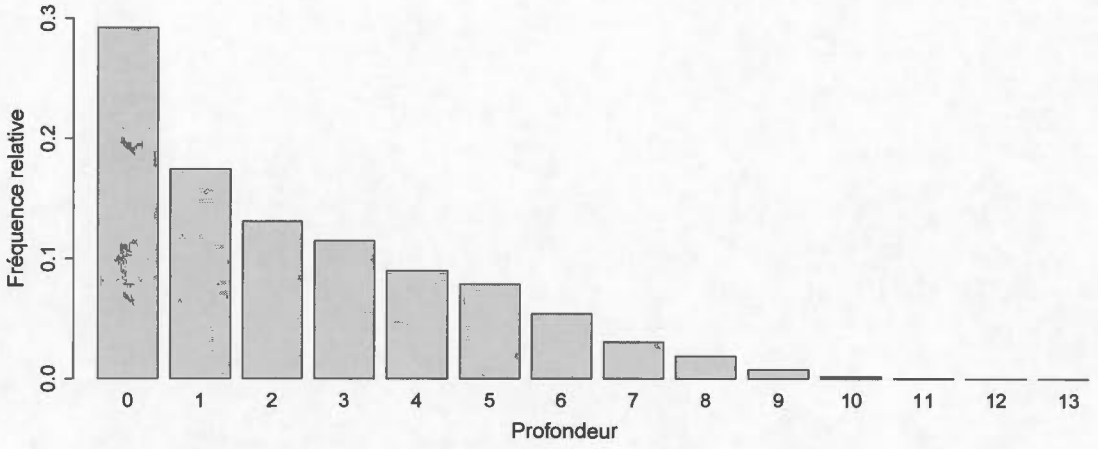


(a)

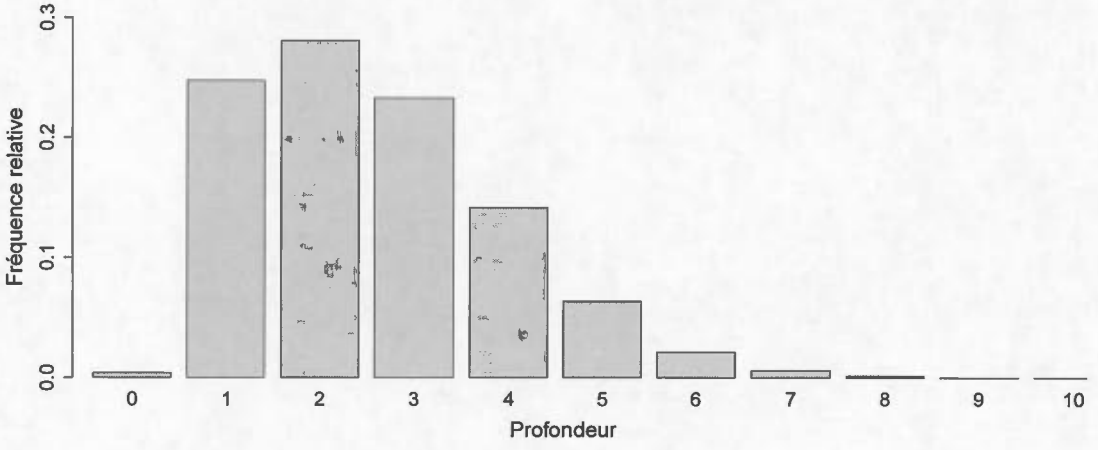


(b)

Figure 3.7 : Estimation de la fonction de masse du nombre de feuilles basée sur 5000 arbres générés avec : (a) $(\alpha, \beta) = (0.70, 0.45)$; (b) $(\alpha, \beta) = (0.995, 1)$



(a)



(b)

Figure 3.8 : Estimation de la fonction de masse de la profondeur basée sur 5000 arbres générés avec : (a) $(\alpha, \beta) = (0.70, 0.45)$; (b) $(\alpha, \beta) = (0.995, 1)$

remarque immédiatement que lorsque $(\alpha, \beta) = (0.995, 1)$, il y a beaucoup moins d'arbres avec une seul feuille et beaucoup moins d'arbres avec beaucoup de feuilles que lorsque $(\alpha, \beta) = (0.70, 0.45)$. Similairement, on constate que lorsque $(\alpha, \beta) = (0.995, 1)$, il y a beaucoup moins d'arbres de profondeur nulle et beaucoup moins d'arbres très profonds que lorsque $(\alpha, \beta) = (0.70, .45)$. Ainsi, on confirme donc l'hypothèse comme quoi augmenter α et β, de façon à obtenir le même nombre de feuilles, réduit la variabilité des arbres.

L'hyperparamètre β est donc un paramètre très important, mais il ne permet pas d'obtenir une pénalisation de la profondeur des arbres sans réduire en même temps le nombre de feuilles moyen. Lorsque $\beta = 0$, il est impossible de réduire la variabilité des arbres sans réduire le nombre moyen de feuilles des arbres. Le paramètre β permet de réduire la variabilité des arbres en ajustant α et β dans la même direction. Plus α et β sont grands, plus on réduit la variabilité des arbres. Lorsque α et β sont grands, on augmente aussi légèrement la profondeur moyenne des arbres, mais cet effet est très petit. La principale utilisation de β est donc de permettre la réduction de la variabilité des arbres et ainsi d'obtenir des arbres avec une plus grande similarité à l'arbre moyen *a priori* qu'on recherche.

3.4 Cas $\beta \neq 0$ et $\alpha > 1$

Tel que discuté précédemment, la loi *a priori* $P(T)$ est spécifiée indirectement par un processus récursif de création d'arbre basé sur la probabilité de séparation $p_{SPLIT}(\eta, T) = \alpha(1 + d_\eta)^{-\beta}$. Pour que la loi *a priori* $P(T)$ existe, on doit donc avoir que $p_{SPLIT}(\eta, T) = \alpha(1 + d_\eta)^{-\beta} \in [0, 1]$ pour tout choix de α et β. On a que $(1 + d_\eta)^{-\beta} \in (0, 1]$, pour tout $d_\eta \geq 0$ et $\beta \geq 0$. En utilisant la restriction $\alpha \in [0, 1]$, on s'assure donc que $p_{SPLIT}(\eta, T) \in [0, 1]$, pour tout $d_\eta \geq 0$ et $\beta \geq 0$. Cette restriction impose une limite au nombre de feuilles et à la variance que l'on

peut atteindre. Il est impossible d'obtenir de grands arbres avec une variabilité arbitrairement petite.

Pour permettre l'utilisation de $\alpha > 1$, on redéfinit la probabilité de séparation de la façon suivante :

$$p_{SPLIT^*}(\eta, T) = min[\alpha(1 + d_\eta)^{-\beta}, 1]. \tag{3.2}$$

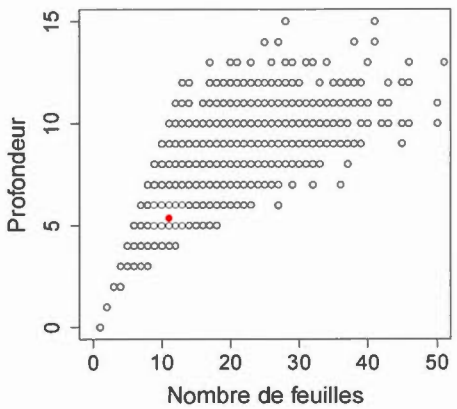
Ainsi, $p_{SPLIT^*}(\eta, T) \in [0, 1]$ pour tout $\alpha \geq 0$ et $\beta \geq 0$. Lorsque $\alpha \in [0, 1]$, on obtient les mêmes probabilités de séparation qu'auparavant. Les résultats précédents, de la sous-section 3.4 (cas $\beta \neq 0$ et $\alpha \in [0, 1]$), sont donc encore valides avec cette définition.

On présente un exemple simple pour comprendre les effets de cette définition alternative de la probabilité de séparation. Si l'on choisit $\alpha = 5$ et $\beta = 1$, on a que

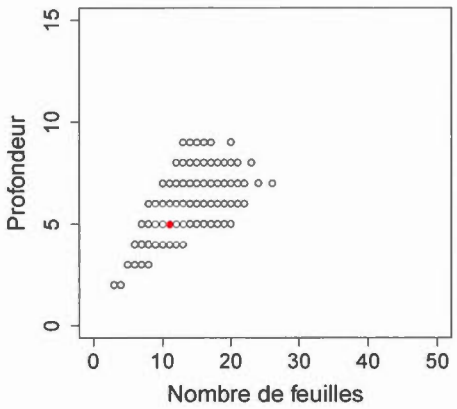
$$p_{SPLIT^*}(\eta, T) = \begin{cases} 1 & \text{si } d_\eta < 5; \\ \frac{5}{(1+d_\eta)} & \text{si } d_\eta \geq 5. \end{cases}$$

Pour qu'un arbre ait une probabilité *a priori* nulle, il faut que $p_{SPLIT^*}(\eta, T) = 1$ pour tout $\eta \in T$. Dans cet exemple, les arbres avec une profondeur plus petite que 5 ont donc une probabilité *a priori* nulle. Avec la définition originale de la probabilité de séparation, seul l'arbre à profondeur zéro ($d_\eta = 0$) lorsque $\alpha = 1$ permettait l'obtention de $p_{SPLIT^*}(\eta, T) = 1$ pour tout $\eta \in T$. On ne pouvait donc pas assigner une probabilité *a priori* nulle aux arbres avec une profondeur plus grande que zéro. Ainsi, la nouvelle définition de la probabilité de séparation permet un contrôle accru sur la loi *a priori* $P(T)$.

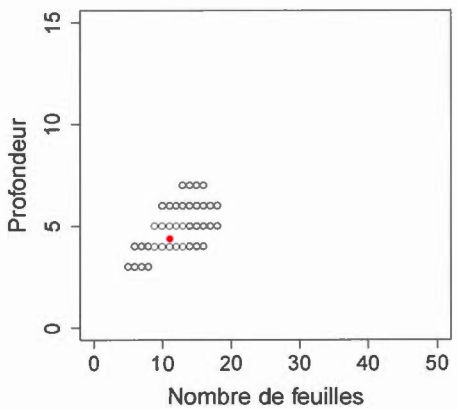
Dans les Figures 3.9(a), 3.9(b) et 3.9(c), on observe l'effet de (α, β) = (0.95, 0.43), (α, β) = (6.6, 2) et (α, β) = (325, 5) sur le nombre de feuilles et la profondeur des 5000 arbres *a priori* générés lorsqu'on a 2000 observations distinctes. Les trois ensembles de valeurs de α et β ont été choisis de façon à atteindre environ le même



(a)



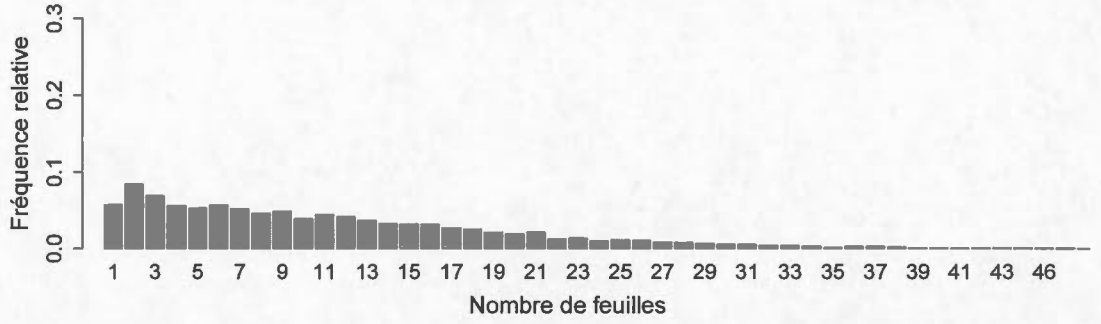
(b)



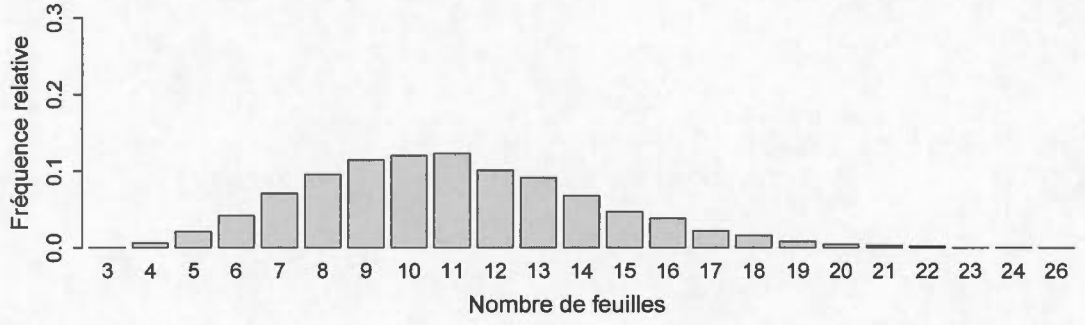
(c)

Figure 3.9 : Profondeur en fonction du nombre de feuilles observée dans un échantillon de 5000 arbres générés lorsqu'on a 2000 observations distinctes avec : (a) $(\alpha, \beta) = (0.95, 0.43)$; (b) $(\alpha, \beta) = (6.6, 2)$; (c) $(\alpha, \beta) = (325, 5)$. Le point rouge représente la moyenne du nombre de feuilles et de la profondeur.

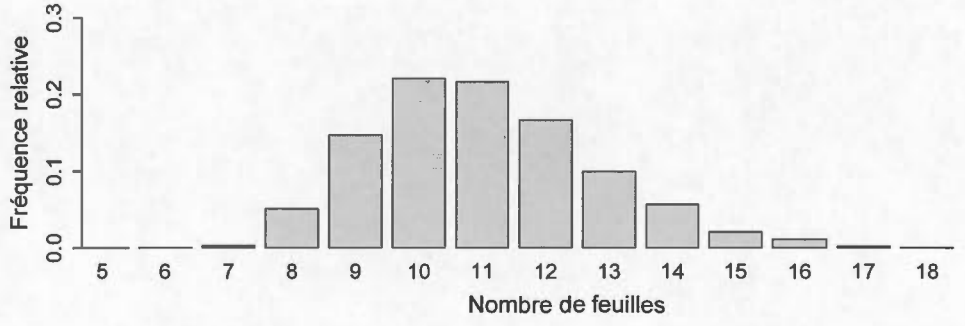
nombre de feuilles (approximativement 11 feuilles en moyenne). On constate que plus α et β sont grands, moins on obtient d'arbres avec peu de feuilles ou beaucoup de feuilles. On constate le même phénomène avec la profondeur. La définition alternative de la probabilité de séparation permet donc d'obtenir des arbres avec une taille aussi grande que l'on désire et une variance aussi petite que l'on désire.



(a)

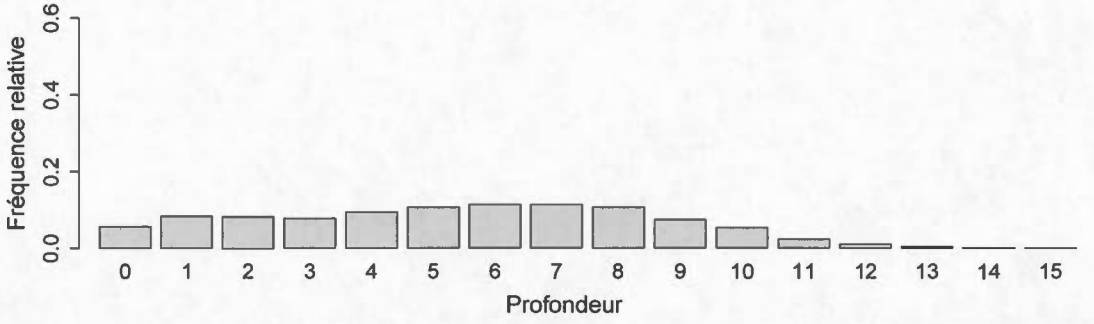


(b)

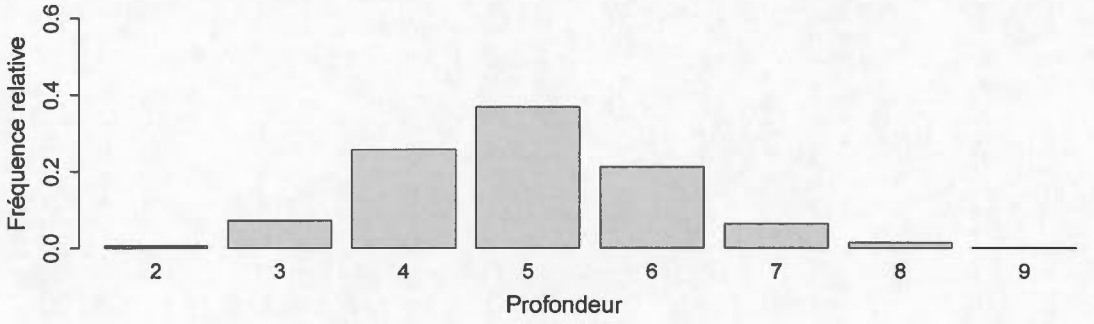


(c)

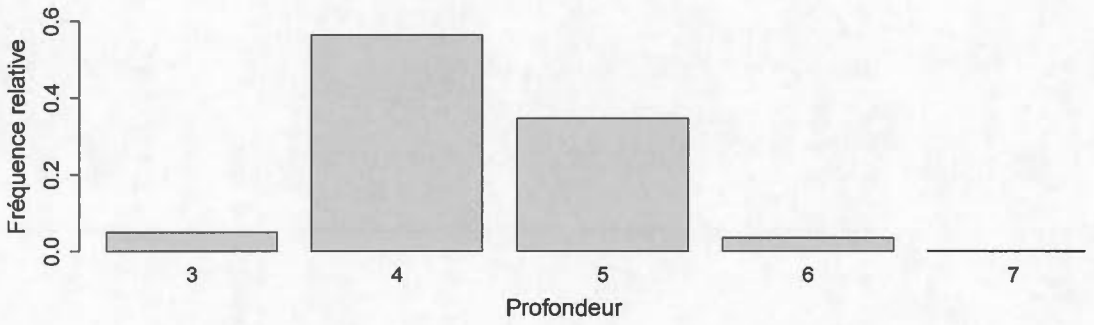
Figure 3.10 : Estimation de la fonction de masse du nombre de feuilles basée sur 5000 arbres générés avec : (a) $(\alpha, \beta) = (0.95, 0.43)$; (b) $(\alpha, \beta) = (6.6, 2)$; (c) $(\alpha, \beta) = (325, 5)$



(a)



(b)



(c)

Figure 3.11 : Estimation de la fonction de masse de la profondeur basée sur 5000 arbres générés avec : (a) $(\alpha, \beta) = (0.95, 0.43)$; (b) $(\alpha, \beta) = (6.6, 2)$; (c) $(\alpha, \beta) = (325, 5)$

Dans les Figures 3.10(a), 3.10(b) et 3.10(c) on peut voir les fréquences relatives du nombre de feuilles des 5000 arbres générés avec $(\alpha, \beta) = (0.95, 0.43)$, $(\alpha, \beta) = (6.6, 2)$ et $(\alpha, \beta) = (325, 5)$. Similairement, dans les Figures 3.11a, 3.11b et 3.11c, on peut voir les fréquences relatives de la profondeur des 5000 arbres générés avec les mêmes choix de α et β. On remarque que la forme de la distribution du nombre de feuilles et de la distribution de la profondeur change de façon considérable lorsque l'on choisit de grands α et β. Il y a beaucoup moins d'arbres avec peu de feuilles ou beaucoup de feuilles. Similairement, il y a beaucoup moins d'arbres profonds ou peu profonds. Les probabilités de masse deviennent de plus en plus centrées autour de la moyenne du nombre de feuilles et de la profondeur lorsque α et β sont grands.

La définition alternative de la probabilité de séparation (3.2) rend possible l'obtention d'un plus grand contrôle sur la forme des arbres selon la loi *a priori* de T. On peut ainsi obtenir précisément la moyenne et la variance du nombre de feuilles désirées *a priori*.

3.5 Choix des hyperparamètres de la loi *a priori* $P(T)$

On a vu dans le Chapitre 2 que le choix des hyperparamètres de $P(T)$ est très important. De plus, dans la partie précédente du Chapitre 3, on a constaté que les hyperparamètres α et β interagissent ensemble pour contrôler la moyenne et la variance du nombre de feuilles. On ne peut donc choisir α et β séparément. On discute de comment déterminer la moyenne et la variance du nombre de feuilles idéales des arbres. Ensuite, on explique comment choisir α et β de façon à obtenir une loi *a priori* avec la moyenne et la variance du nombre de feuilles désirés.

3.5.1 Moyenne du nombre de feuilles

Les grands arbres ont une meilleure précision que les petits arbres car ceux-ci séparent plus fortement les données. En effet, un arbre avec n observations et n feuilles a une précision parfaite ($R^2 = 1$), tandis qu'un arbre avec une seule feuille a une précision équivalente à l'utilisation de la moyenne de la variable expliquée ($R^2 = 0$). On peut donc interpréter le nombre de feuilles comme une mesure de balance entre le surapprentissage et le sous-apprentissage. Le surapprentissage est lorsqu'on a un modèle très complexe (un arbre avec beaucoup de feuilles) qui explique très bien les données présentes (grand R^2) mais qui prédit très mal la variable expliquée de nouvelles observations (petit R^2_{CV}). Cela veut donc dire que le modèle statistique se généralise mal à de nouvelles observations. Le sous-apprentissage est lorsqu'on a un modèle peu complexe (un arbre avec peu de feuilles) qui n'arrive pas à bien prédire les observations présentes et nouvelles (petit R^2 et petit R^2_{CV}). On recherche généralement une certaine balance de complexité dans les modèles statistiques pour prévenir le surapprentissage et le sous-apprentissage. Il faut donc trouver un nombre moyen de feuilles pas trop grand et pas trop petit. Il existe plusieurs façons de choisir le nombre moyen de feuilles *a priori* des arbres bayésiens.

On peut utiliser l'opinion d'un expert sur le choix du nombre de feuilles moyen. Un expert sur le sujet en question peut avoir une idée en avance du nombre de feuilles qui devrait être formé. On peut aussi s'attendre à ce qu'il y ait un certain nombre d'interactions entre les variables et avec cette information, on peut choisir de construire de petits ou grands arbres. Par exemple, dans un cas génétique, on peut suspecter que les gènes ont peu d'effet individuellement, mais que ça prend environ trois gènes qui interagissent pour prédire la variable expliquée. Dans ce cas, on présume que les arbres doivent avoir une profondeur de 3 et être

balancés ; ainsi on recherche des arbres avec une moyenne de 8 feuilles. Le nombre d'interactions nécessaires entre les variables explicatives permettant de prédire la variable expliquée peut donc être utilisé comme information pour décider du nombre de feuilles moyen désiré.

Une autre façon de choisir la moyenne du nombre de feuilles est de se baser sur l'arbre fréquentiste CART ayant le plus grand R^2_{CV}. On recommande de choisir des hyperparamètres α et β qui ne rendent pas trop improbable l'obtention du nombre de feuilles de l'arbre optimal fréquentiste.

3.5.2 Variance du nombre de feuilles

Le choix de la variance du nombre de feuilles est simplement déterminé selon la certitude que l'on a par rapport au nombre de feuilles moyen choisi. Il est possible qu'il n'y ait pas de nombre de feuilles idéal. Dans ce cas, on doit s'assurer que la variance du nombre de feuilles soit assez grande pour obtenir des arbres avec un nombre de feuilles assez varié.

3.5.3 Choisir α et β

Lorsqu'on cherche à maximiser la variance du nombre de feuilles, on peut simplement choisir $\beta = 0$ et ajuster α entre 0 et 0.50 de façon à obtenir le nombre de feuilles désiré. Autrement, si on cherche à restreindre la variabilité du nombre de feuilles des arbres, on doit choisir $\beta \neq 0$. Dans ce cas, pour obtenir la moyenne et la variance du nombre de feuilles désirées, on doit avoir recours à un processus d'essai-erreur. On doit commencer par choisir β et ensuite, ajuster α jusqu'à l'obtention de la moyenne du nombre de feuilles désirée. Si la variance est trop petite ou trop grande, on doit choisir β un peu plus petit ou un peu plus grand et réajuster α de façon à obtenir la moyenne du nombre de feuilles désirée. On

répète ce processus jusqu'à l'obtention de la moyenne et la variance du nombre de feuilles désirées. Pour se guider, il est recommander d'avoir en main un tableau de la moyenne et de l'écart-type du nombre de feuilles selon α, β et le nombre d'observations distinctes, tels que les Tableaux 3.1, 3.3 et 3.5.

Il est à noter que dû à la contrainte de $\alpha \leq 1$, lorsque $\beta \neq 0$, il y a une limite sur le nombre de feuilles moyen que l'on peut obtenir pour chaque choix de β. On ne peut donc obtenir de très grands arbres, en moyenne, avec une variance du nombre de feuilles très petite. En utilisant la définition alternative de la probabilité de séparation (3.2), il est possible de choisir $\alpha > 1$. Tel que constaté précédemment, cela permet à $P(T)$ d'avoir un nombre de feuilles moyen aussi grand que désiré et une variance associée aussi petite que désirée.

D'un point de vue non bayésien et plus centré sur la prédiction, on peut aussi choisir plusieurs α et β différents et utiliser la validation croisée ou toute autre méthode de validation pour déterminer quel choix de α et β amène à la meilleure capacité prédictive.

CHAPITRE IV

ÉTUDE DE SIMULATION

Dans ce chapitre, on effectue une étude de simulation pour l'estimation d'un arbre de régression connu à partir des modèles d'arbres de régression bayésiens et fréquentistes. On s'intéresse à trois facteurs en particulier : le choix des hyperparamètres α et β de la loi *a priori* de T, la taille échantillonnale et le rapport signal pour bruit. On cherche à analyser l'impact de chaque facteur, mais particulièrement celui du choix des hyperparamètres de $P(T)$.

4.1 Objectifs et hypothèses

Le principal objectif de l'étude est de vérifier si l'on obtient de meilleures prédictions et des arbres *a posteriori* plus semblables à l'arbre recherché lorsqu'on choisit les hyperparamètres α et β de façon à obtenir une loi *a priori* centrée sur l'arbre de régression à estimer. On émet l'hypothèse que le choix des hyperparamètres de la loi *a priori* de T est moins important lorsque le nombre d'observations est grand. De plus, on suspecte qu'il est plus difficile d'obtenir des arbres semblables à l'arbre désiré lorsque le rapport signal pour bruit est élevé, peu importe le choix de α et β. On tente donc de valider ces conjectures.

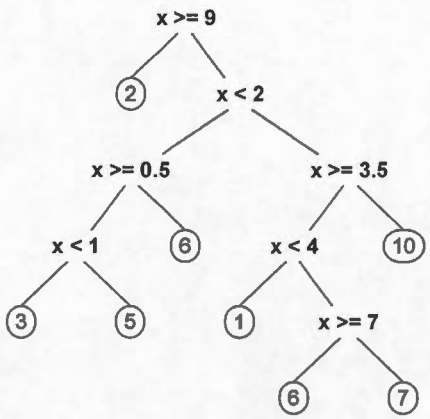


Figure 4.1 : Arbre de régression de l'exemple du Chapitre 5

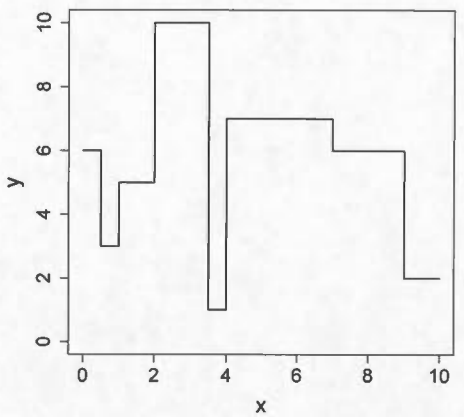


Figure 4.2 : Courbe de séparation de l'arbre de régression de l'exemple du Chapitre 5

4.2 Méthodes

4.2.1 Génération des données

Comme base pour la simulation, on considère l'arbre de régression univarié représenté dans la Figure 4.1. La courbe de séparation produite par cet arbre est représentée dans la Figure 4.2. Pour simuler des observations de cet arbre, on commence par générer n observations de la variable explicative, $x_1, \ldots, x_n \overset{i.i.d.}{\sim} U(0, 10)$. Ensuite, on génère les observations de la variable expliquée en ajoutant un terme d'erreur

$$y_i = f(x_i) + \epsilon_i,$$

où $f(x_i)$ est la valeur de la feuille déterminée par x_i dans l'arbre de la Figure 4.1, $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$ et σ_ϵ est l'écart-type du terme d'erreur aléatoire.

4.2.2 Facteurs

Le premier facteur étudié est la taille échantillonnale. Les échantillons avec moins d'observations contiennent moins d'information, il est donc plus difficile de faire de bonnes prédictions. De plus, lorsqu'il y a moins d'observations, on s'attend à voir une différence plus marquée dans les résultats associés aux différents choix des hyperparamètres de la loi *a priori* de T. On considère donc trois nombres d'observations possibles dans les analyses : petit ($n = 50$), moyen ($n = 100$) et grand ($n = 500$).

Le deuxième facteur d'intérêt est le rapport signal sur bruit (SNR). Ce facteur permet de comparer la magnitude du signal ($f(X)$) par rapport au bruit (le terme d'erreur). Sachant que le terme d'erreur utilisé dans cet exemple est centré à 0 et est indépendant d'une observation à l'autre, on définit le rapport signal sur bruit (Dicker, 2012) de la façon suivante :

$$SNR = \frac{\sigma^2_{f(X)}}{\sigma^2_\epsilon},$$

où $\sigma^2_{f(X)}$ est la variance de $f(X)$ et σ^2_ϵ est la variance du terme d'erreur. Lorsque $SNR > 1$, le signal est plus fort que le bruit. On considère deux niveaux de ce facteur dans les analyses : un rapport signal sur bruit faible et un rapport signal sur bruit élevé. On suspecte que le choix de la loi *a priori* de T a un effet plus important sur les résultats lorsque le rapport signal sur bruit est faible. On calcule $\sigma^2_{f(X)}$, la variance de $f(X)$, et l'on choisit deux différentes valeurs pour l'écart-type du terme d'erreur.

Proposition 4.2.1.

$$SNR = \frac{6.1}{\sigma^2_\epsilon}.$$

Démonstration.

$$\begin{aligned}
E(f(X)) &= \int_0^{10} \frac{f(x)}{10} dx \\
&= \int_0^{0.5} \frac{6}{10} dx + \int_{0.5}^{1} \frac{3}{10} dx + \int_1^2 \frac{5}{10} dx + \int_2^{3.5} \frac{1}{10} dx + \int_{3.5}^4 \frac{10}{10} dx + \int_4^7 \frac{7}{10} dx \\
&\quad + \int_7^9 \frac{6}{10} dx + \int_9^{10} \frac{2}{10} dx \\
&= 6
\end{aligned}$$

$$\begin{aligned}
E(f(X)^2) &= \int_0^{10} \frac{f(x)^2}{10} dx \\
&= \int_0^{0.5} \frac{6^2}{10} dx + \int_{0.5}^{1} \frac{3^2}{10} dx + \int_1^2 \frac{5^2}{10} dx + \int_2^{3.5} \frac{1^2}{10} dx + \int_{3.5}^4 \frac{10^2}{10} dx + \int_4^7 \frac{7^2}{10} dx \\
&\quad + \int_7^9 \frac{6^2}{10} dx + \int_9^{10} \frac{2^2}{10} dx \\
&= 42.1
\end{aligned}$$

Ainsi, $\sigma^2_{f(X)} = Var(f(X)) = E(f(X)^2) - E(f(X))^2 = 42.1 - 6^2 = 6.1.$

□

Lorsque $\sigma_\epsilon = 2$, on a que $SNR = 1.53$ et lorsque $\sigma_\epsilon = 0.5$, on a que $SNR = 24.40$. On choisit donc $\sigma_\epsilon = 2$ pour représenter le cas d'un rapport signal sur bruit faible et $\sigma_\epsilon = 0.5$ pour représenter le cas d'un rapport signal sur bruit fort.

Le troisième et principal facteur d'intérêt est le choix des hyperparamètres de $P(T)$. Ce facteur est le plus important des trois dû à la nature de ce mémoire. On veut comparer les résultats associés à un choix d'hyperparamètres par défaut ($\alpha = 0.50$, $\beta = 2$ dans le paquet `tgp`) à ceux obtenus d'un choix d'hyperparamètres qui centre la loi *a priori* autour de l'arbre de régression de la Figure 4.1. Pour ce faire, la sélection des paramètres α et β non standards est réalisée en choisissant $\alpha = 1$ et β de façon à obtenir environ 8 feuilles en moyenne (déterminé

à partir de simulations des arbres *a priori*). En choisissant $\alpha = 1$, on minimise la variance du nombre de feuilles. On tente de minimiser cette variance à cause de la grande certitude que l'on a sur le nombre de feuilles désiré étant donné que l'on connait l'arbre de régression véritable. Le choix $\alpha = 1$ permet aussi de donner une probabilité *a priori* nulle aux arbres avec une seule feuille. Les hyperparamètres choisis sont donc : ($\alpha = 1$, $\beta = .45$) lorsque $n = 50$, ($\alpha = 1$, $\beta = .50$) lorsque $n = 100$ et ($\alpha = 1$, $\beta = .56$) lorsque $n = 500$.

Pour caractériser la loi *a priori* de T selon les différents choix d'hyperparamètres, on génère 50 000 arbres de régression pour chaque choix et on calcule la moyenne, l'écart-type, le 10^e percentile et le 90^e percentile du nombre de feuilles des arbres pour chaque niveau d'observations ($n = 50$, $n = 100$, $n = 500$). Le Tableau 4.1 montre les statistiques associées au choix d'hyperparamètres par défaut tandis que le Tableau 4.2 est associé au choix d'hyperparamètres sélectionnés selon la stratégie mentionnée dans le paragraphe précédent.

Dans le Tableau 4.1, qui décrit la loi $P(T)$ avec les hyperparamètres par défaut, on constate que, peu importe le nombre d'observations, la moyenne du nombre de feuilles est de 1.64 et son écart-type est de 0.75. De plus, le 10^e percentile est de 1 et le 90^e percentile est de 3, indépendamment du nombre d'observations. Tel que discuté dans le Chapitre 3, le nombre d'observations distinctes a une influence minimale sur $P(T)$ lorsque les hyperparamètres mènent à un nombre de feuilles moyen très petit. Ainsi, les arbres avec 8 feuilles ont une probabilité *a priori* extrêmement faible, ce qui rend ce choix d'hyperparamètres inadéquat pour le problème en question.

Les détails sur les choix d'hyperparamètres non standards, menant à un nombre moyen de feuilles coïncidant avec le nombre de feuilles du vrai arbre (8), sont fournis dans le Tableau 4.2. On constate que l'on obtient une grande variance sur

le nombre de feuilles malgré le fait que $\alpha = 1$. En effet, le 10^e percentile est de 2 dans le cas $n = 500$ et de 3 dans les autres cas, tandis que le 90^e percentile est de 15 lorsque $n = 500$ et de 14 dans les autres cas. Bien que le nombre moyen de feuilles soit adéquat pour notre problème, la grande variance sur le nombre de feuilles indique que la loi a priori de T attribue une probabilité non négligeable aux arbres ayant le mauvais nombre de feuilles.

Tableau 4.1 : Espérance, écart-type, 10^e percentile et 90^e percentile du nombre de feuilles de l'arbre bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$) en fonction du nombre d'observations (n=50, 100, 500). Les valeurs sont estimées à partir de 50 000 arbres générés.

	μ	σ	10^e percentile	90^e percentile
$n = 50$	1.64	0.75	1	3
$n = 100$	1.64	0.75	1	3
$n = 500$	1.64	0.75	1	3

Tableau 4.2 : Espérance, écart-type, 10^e percentile et 90^e percentile du nombre de feuilles de l'arbre bayésien avec $P(T)$ non standard en fonction du nombre d'observations (n=50, 100, 500). Les valeurs sont estimées à partir de 50 000 arbres générés.

	μ	σ	10^e percentile	90^e percentile
$n = 50$, ($\alpha = 1$, $\beta = 0.45$)	7.92	4.31	3	14
$n = 100$, ($\alpha = 1$, $\beta = 0.50$)	7.99	4.65	3	14
$n = 500$, ($\alpha = 1$, $\beta = 0.56$)	7.99	5.02	2	15

4.2.3 Outils

Pour ajuster les arbres de régression bayésiens, on utilise le paquet `tgp` (Gramacy, 2007) du langage de programmation `R`. Il est important de noter que le paquet `tgp` utilise par défaut une loi *a priori* impropre pour la variance des observations, σ^2. Il est recommandé de ne pas utiliser cette loi *a priori* lorsque le jeu de données considéré contient beaucoup de bruit (Gramacy, 2007). Nous avons effectivement constaté que la loi *a priori* propre pour σ^2 performait beaucoup mieux que la loi impropre, lorsque le signal sur bruit est petit. Pour uniformiser les résultats, on a suivi cette recommandation non seulement lorsque le signal sur bruit est petit, mais aussi lorsque le signal sur bruit est grand. La loi *a priori* propre de σ^2 a ses propres hyperparamètres par défaut ; on a donc tenté de les modifier. Il a été constaté que de modifier le choix des hyperparamètres de la loi *a priori* propre de la variance ne changeait pas la qualité des résultats des arbres bayésiens de façon importante tant que la loi *a priori* accorde une probabilité non négligeable à la vraie valeur de la variance. Les hyperparamètres de la loi *a priori* propre pour σ^2 ont donc été laissés à leur défaut.

Tel que mentionné dans le Chapitre 1, on utilise le paquet `rpart` pour ajuster les arbres fréquentistes. Dans le cadre de cet exemple, on utilise les paramètres par défaut sauf pour la valeur de *minsplit* qui est remplacé par 2. Ce paramètre définit le nombre d'observations minimal requis dans une feuille pour que l'algorithme tente de séparer celle-ci. Similairement, la valeur du paramètre *minpart* du paquet `tgp` est remplacé par 2. Ce paramètre assigne une probabilité nulle aux arbres avec des feuilles contenant moins d'observations que *minpart*. Il est à noter que le modèle d'arbres fréquentiste de `rpart` génère un seul arbre, tandis que le modèle d'arbres bayésien de `tgp` génère une suite d'arbres distribués approximativement selon la loi *a posteriori* de T. Pour générer une suite d'arbres, le paquet `tgp` utilise

l'algorithme de Metropolis-Hastings. Les 2000 premières réalisations sont écartées, 7000 réalisations sont observées et chaque deuxième réalisation est enlevée. Tel que mentionné dans le Chapitre 1, les probabilités *a posteriori* des arbres générés de la suite sont approximées par la formule suivante :

$$\frac{P(Y|X,T^i)P(T^i)}{\sum_i P(Y|X,T^i)P(T^i)}.$$

On cherche à comparer les capacités prédictives des différents modèles d'arbres de régression (arbre bayésien avec loi *a priori* par défaut, arbre bayésien avec loi *a priori* centrée sur le nombre de feuilles recherché et arbre fréquentiste). Généralement, on ne connait pas la fonction $f(x)$, on ne peut donc pas générer de nouvelles observations $f(x) + \epsilon$. La validation croisée, telle que décrite dans le Chapitre 2, est donc utilisée en pratique pour estimer la capacité prédictive d'un modèle, car il n'est pas nécessaire de générer de nouvelles observations avec cette méthode. Dans le présent contexte, la fonction $f(x)$ est connue, ce qui implique qu'on peut directement générer des observations de celle-ci. Ainsi, plutôt que d'avoir recours à la validation croisée, on construit un jeu de données de validation en générant 1000 nouvelles observations (y_i, x_i). L'objectif est de prédire les nouvelles observations de la variable expliquée à l'aide du jeu de données utilisé pour ajuster les arbres de régression. On calcule ensuite le coefficient de détermination de validation (R^2_V), qui est basé sur les prédictions de ces nouvelles observations. Le R^2_V est défini de la même façon que le R^2 (équation 2.1) mais sur l'ensemble de validation plutôt que sur le jeu de données de formation. Cette mesure permet de juger de la capacité prédictive de l'arbre de régression ajusté.

On s'intéresse aussi au nombre de feuilles de l'arbre MAP et à l'erreur quadratique moyenne de $\hat{f}(x)$. L'arbre de la Figure 4.1 a 8 feuilles ; on s'attend donc à ce que l'arbre avec la plus grande probabilité *a posteriori* ait aussi 8 feuilles. L'arbre MAP est défini ici comme étant l'arbre de régression avec la plus grande

probabilité *a posteriori* parmi les 2500 itérations de l'algorithme de Metropolis-Hastings conservées. L'erreur quadratique moyenne de $\hat{f}(x)$ est calculée à l'aide de la formule suivante :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2}{n}}.$$

Plus cette mesure est petite, plus l'arbre estimé s'approche du véritable arbre.

Lorsque le nombre d'observations est petit, il y a beaucoup de variabilité dans les valeurs du R^2, du R^2_V, de l'erreur quadratique moyenne et du nombre de feuilles. On a donc construit 50 échantillons de 50 observations, 25 échantillons de 100 observations et 10 échantillons de 500 observations. Dans nos analyses, on rapporte la moyenne et l'écart-type du R^2, du R^2_V, du nombre de feuilles de l'arbre MAP et de l'erreur quadratique moyenne de $\hat{f}(x)$. De plus, pour améliorer la stabilité des estimations, on redémarre l'algorithme du modèle d'arbres bayésien 10 fois. Le redémarrage est recommandé par Chipman *et al.* (1998) car l'algorithme de Metropolis-Hastings reste souvent coincé dans un mode local de la loi *a posteriori* de T.

4.3 Résultats

On a considéré six scénarios pour effectuer l'étude par simulation ; ils sont présentés dans le Tableau 4.3.

Les résultats de chaque scénario sont présentés respectivement dans les Tableaux 4.4 à 4.9. Pour plus de détails, les diagrammes en boîtes du R^2_V et du nombre de feuilles sont présentés dans les Figures 4.3 à 4.14.

Tableau 4.3 : Description des paramètres des scénarios de l'étude de simulation (taille échantillonnale, écart-type du terme d'erreur, rapport signal sur bruit, choix d'hyperparamètres de la loi *a priori* $P(T)$ par défaut et non standard)

Scénario	n	σ_ϵ	SNR	$P(T)$ par défaut	$P(T)$ non standard
1	50	0.50	24.43	$(\alpha = 0.50, \beta = 2)$	$(\alpha = 1, \beta = 0.45)$
2	100	0.50	24.43	$(\alpha = 0.50, \beta = 2)$	$(\alpha = 1, \beta = 0.50)$
3	500	0.50	24.43	$(\alpha = 0.50, \beta = 2)$	$(\alpha = 1, \beta = 0.56)$
4	50	2	1.50	$(\alpha = 0.50, \beta = 2)$	$(\alpha = 1, \beta = 0.45)$
5	100	2	1.50	$(\alpha = 0.50, \beta = 2)$	$(\alpha = 1, \beta = 0.50)$
6	500	2	1.50	$(\alpha = 0.50, \beta = 2)$	$(\alpha = 1, \beta = 0.56)$

4.3.1 Rapport signal sur bruit grand (scénarios 1 à 3)

Les résultats du scénario 1 sont présentés dans le Tableau 4.4, la Figure 4.3 et la Figure 4.4. La plus grande valeur de l'espérance du R^2_V est obtenue par l'arbre fréquentiste suivi par le modèle d'arbres bayésien avec $P(T)$ non standard. Similairement, la plus petite valeur de l'espérance de l'erreur quadratique moyenne de $\hat{f}(x)$ est obtenue par l'arbre fréquentiste suivi par le modèle d'arbres bayésien avec $P(T)$ non standard. De plus, l'arbre fréquentiste et le modèle d'arbres bayésien avec $P(T)$ non standard s'approchent plus souvent du bon nombre de feuilles que le modèle d'arbres bayésien avec $P(T)$ par défaut. On a donc que le modèle d'arbres bayésien avec $P(T)$ non standard est un meilleur modèle prédictif que le modèle d'arbres bayésien avec $P(T)$ par défaut, mais est légèrement inférieur à l'arbre fréquentiste.

Les résultats du scénario 2 sont présentés dans le Tableau 4.5, la Figure 4.5 et la

Figure 4.6. Similairement au scénario 1, la plus grande valeur de l'espérance du R_V^2 est obtenue par l'arbre fréquentiste suivi par le modèle d'arbres bayésien avec $P(T)$ non standard et la plus petite espérance de l'erreur quadratique moyenne de $\hat{f}(x)$ est obtenue par l'arbre fréquentiste suivi par le modèle d'arbres bayésien avec $P(T)$ non standard. On remarque que le modèle d'arbres bayésien avec $P(T)$ non standard performe à peu près au même niveau que le modèle d'arbres bayésien avec $P(T)$ par défaut. Par contre, le modèle d'arbres bayésien avec $P(T)$ non standard a une médiane et un 75^e percentile du R_V^2 plus élevés que le modèle d'arbres bayésien avec $P(T)$ par défaut. Le modèle d'arbres bayésien avec $P(T)$ non standard a donc tendance à un peu mieux performer que le modèle d'arbres bayésien avec $P(T)$ par défaut, mais le gain de performance est petit. De plus, on remarque que contrairement aux deux autres modèles, le modèle d'arbres bayésien avec $P(T)$ non standard tend à surestimer légèrement le vrai nombre de feuilles. L'arbre fréquentiste, quant à lui, est supérieur aux modèles d'arbres bayésiens en terme de capacité prédictive.

Les résultats du scénario 3 sont présentés dans le Tableau 4.6, la Figure 4.7 et la Figure 4.8. À première vue, en se basant strictement le R_V^2, on pourrait croire que chaque modèle performe similairement en terme de prédiction, car l'espérance du R_V^2 est approximativement la même pour chaque modèle. Par contre, en regardant le diagramme en boîtes du R_V^2, on constate que la médiane et l'extrémité de la moustache inférieure du modèle d'arbres bayésien avec $P(T)$ par défaut sont plus petits que ceux dans le modèle d'arbres bayésien avec $P(T)$ non standard. De plus, on remarque que l'arbre fréquentiste tend à obtenir de plus grands R_V^2 que les modèles d'arbres bayésiens. La seule exception est dans un échantillon, où l'arbre fréquentiste obtient un R_V^2 légèrement plus petit que 0.88 et donc inférieur aux R_V^2 des modèles bayésiens. Cette valeur à l'écart explique pourquoi le modèle fréquentiste obtient le même R_V^2 moyen que les modèles d'arbres bayésiens. Fina-

lement, on remarque que l'arbre MAP du modèle d'arbres bayésien avec $P(T)$ non standard surestime le nombre de feuilles véritable, contrairement aux autres modèles. Les modèles sont donc semblables en terme de prédiction, mais le modèle d'arbres bayésien avec $P(T)$ non standard est légèrement préférable au modèle d'arbres bayésien avec $P(T)$ par défaut.

Lorsque le rapport signal sur bruit est grand, on a donc que l'arbre fréquentiste performe mieux que le modèle d'arbres bayésien avec $P(T)$ non standard et que ce dernier performe mieux que le modèle d'arbres bayésien avec $P(T)$ par défaut, peu importe le nombre d'observations et le choix d'hyperparamètres. De plus, bien choisir les valeurs des hyperparamètres α et β, de façon à obtenir une loi *a priori* de T centrée sur le bon nombre de feuilles, amène à un gain de performance, par rapport au choix d'hyperparamètres par défaut, qui devient de moins en moins important à mesure que le nombre d'observations augmente.

Tableau 4.4 : Scénario 1 ($n = 50$, $\sigma_\epsilon = 0.50$, basé sur 50 échantillons). Espérance et écart-type du R^2, du R^2_V, de l'erreur quadratique moyenne de $\hat{f}(x)$ (RMSE) et du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.45$) et pour l'arbre fréquentiste

	Bayésien par défaut		**Bayésien non standard**		**Fréquentiste**	
	μ	σ	μ	σ	μ	σ
R^2	0.64	0.15	0.76	0.10	0.95	0.04
R^2_V [a]	0.48	0.14	0.57	0.13	0.65	0.13
RMSE	1.38	0.32	1.12	0.25	0.39	0.25
Nombre de feuilles [b]	5.72	0.61	6.66	0.75	6.68	1.24

[a] basé sur un échantillon de 1000 nouvelles observations

[b] basé sur l'arbre MAP dans le cas bayésien

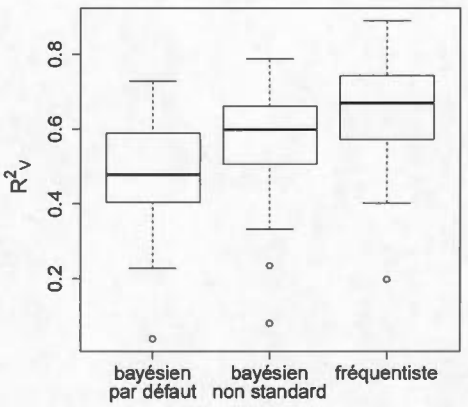


Figure 4.3 : Scénario 1 ($n = 50$, $\sigma_\epsilon = 0.50$, basé sur 50 échantillons). Diagramme en boîtes du R^2_V pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.45$) et pour l'arbre fréquentiste

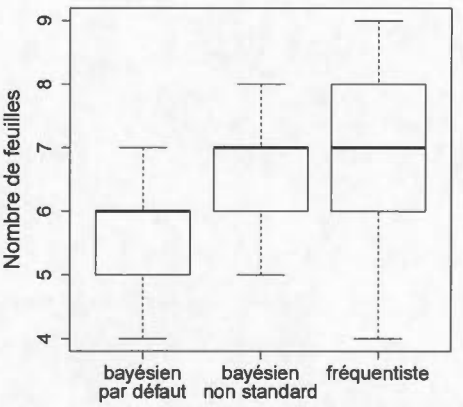


Figure 4.4 : Scénario 1 ($n = 50$, $\sigma_\epsilon = 0.50$, basé sur 50 échantillons). Diagramme en boîtes du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.45$) et pour l'arbre fréquentiste

Tableau 4.5 : Scénario 2 ($n = 100$, $\sigma_\epsilon = 0.50$, basé sur 25 échantillons). Espérance et écart-type du R^2, du R^2_V, de l'erreur quadratique moyenne de $\hat{f}(x)$ (RMSE) et du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.50$) et pour l'arbre fréquentiste

	Bayésien par défaut		**Bayésien non standard**		**Fréquentiste**	
	μ	σ	μ	σ	μ	σ
R^2	0.84	0.04	0.87	0.04	0.96	0.01
R^2_V [a]	0.74	0.08	0.75	0.07	0.81	0.07
RMSE	0.87	0.11	0.80	0.12	0.23	0.11
Nombre de feuilles [b]	7.16	0.75	8.6	0.96	7.52	0.71

[a] basé sur un échantillon de 1000 nouvelles observations

[b] basé sur l'arbre MAP dans le cas bayésien

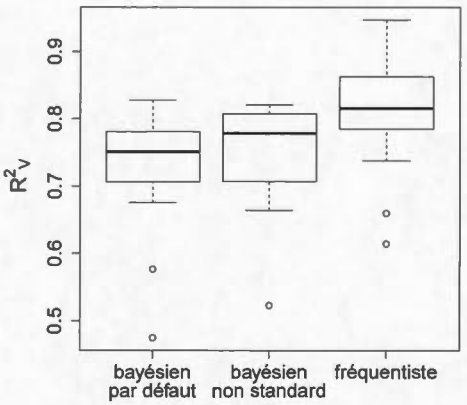


Figure 4.5 : Scénario 2 ($n = 100$, $\sigma_\epsilon = 0.50$, basé sur 25 échantillons). Diagramme en boîtes du R^2_V pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.50$) et pour l'arbre fréquentiste

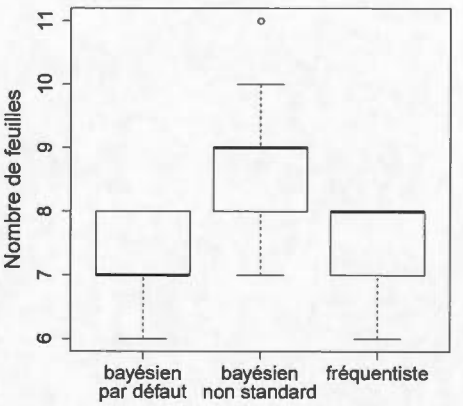


Figure 4.6 : Scénario 2 ($n = 100$, $\sigma_\epsilon = 0.50$, basé sur 25 échantillons). Diagramme en boîtes du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.50$) et pour l'arbre fréquentiste

Tableau 4.6 : Scénario 3 ($n = 500$, $\sigma_\epsilon = 0.50$, basé sur 10 échantillons). Espérance et écart-type du R^2, du R^2_V, de l'erreur quadratique moyenne de $\hat{f}(x)$ (RMSE) et du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.56$) et pour l'arbre fréquentiste

	Bayésien par défaut		**Bayésien non standard**		**Fréquentiste**	
	μ	σ	μ	σ	μ	σ
R^2	0.95	0.01	0.95	0.00	0.96	0.00
R^2_V [a]	0.93	0.01	0.93	0.01	0.93	0.02
RMSE	0.28	0.06	0.27	0.04	0.11	0.07
Nombre de feuilles [b]	8.6	0.97	9.7	1.16	8	0

[a] basé sur un échantillon de 1000 nouvelles observations

[b] basé sur l'arbre MAP dans le cas bayésien

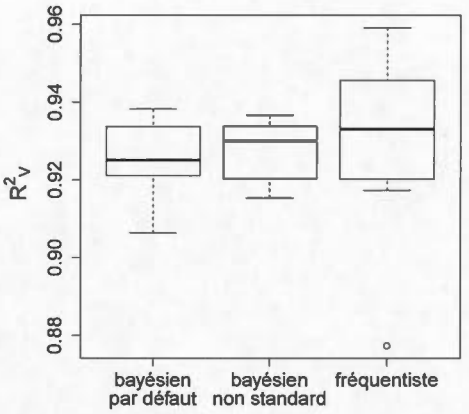


Figure 4.7 : Scénario 3 ($n = 500$, $\sigma_\epsilon = 0.50$, basé sur 10 échantillons). Diagramme en boîtes du R^2_V pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.56$) et pour l'arbre fréquentiste

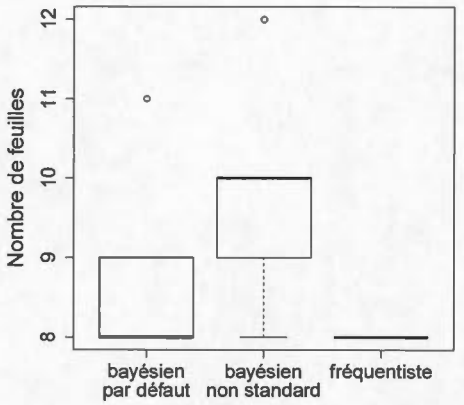


Figure 4.8 : Scénario 3 ($n = 500$, $\sigma_\epsilon = 0.50$, basé sur 10 échantillons). Diagramme en boîtes du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.56$) et pour l'arbre fréquentiste

4.3.2 Rapport signal sur bruit petit (scénarios 4 à 6)

Les résultats du scénario 4 sont présentés dans le Tableau 4.7, la Figure 4.9 et la Figure 4.10. La plus grande valeur de l'espérance du R_V^2 est obtenue par le modèle d'arbres bayésien avec $P(T)$ non standard suivi par l'arbre fréquentiste, tandis que la plus petite valeur de l'espérance de l'erreur quadratique moyenne de $\hat{f}(x)$ est obtenue par l'arbre fréquentiste suivi par le modèle d'arbres bayésien avec $P(T)$ non standard. De plus, on constate que le modèle d'arbres bayésien avec $P(T)$ non standard s'approche plus souvent du vrai nombre de feuilles que les autres modèles. Le modèle d'arbres bayésien avec $P(T)$ non standard est donc préférable au modèle d'arbres bayésien avec $P(T)$ par défaut, dans ce scénario.

Les résultats du scénario 5 sont présentés dans le Tableau 4.8, la Figure 4.11 et la Figure 4.12. On constate que le modèle d'arbres bayésien avec $P(T)$ non standard et l'arbre fréquentiste ont la même moyenne et le même écart-type sur le R_V^2. De plus, la moyenne du R_V^2 est la plus petite pour le modèle d'arbres bayésien avec $P(T)$ par défaut. Le modèle d'arbres bayésien avec $P(T)$ non standard a donc, en moyenne, la même performance prédictive que l'arbre fréquentiste. Pour ce qui est de l'erreur quadratique moyenne de $\hat{f}(x)$, on constate que la plus petite valeur est obtenue par l'arbre fréquentiste, suivi du modèle d'arbres bayésien avec $P(T)$ non standard. On conclut donc que le modèle d'arbres bayésien avec $P(T)$ non standard est préférable au modèle d'arbres bayésien avec $P(T)$ par défaut, dans ce scénario.

Les résultats du scénario 6 sont présentés dans le Tableau 4.9, la Figure 4.13 et la Figure 4.14. On observe que la plus grande valeur de l'espérance du R_V^2 est obtenue par le modèle d'arbres bayésien avec $P(T)$ non standard suivi par le modèle d'arbres bayésien avec $P(T)$ par défaut et que la plus petite valeur de l'erreur quadratique moyenne de $\hat{f}(x)$ est obtenue par le modèle d'arbres bayésien avec

$P(T)$ non standard suivi par le modèle d'arbres bayésien avec $P(T)$ par défaut. On constate aussi que l'écart-type du R^2_V et de l'erreur quadratique moyenne de l'arbre fréquentiste sont plus grands que ceux associés aux modèles d'arbres bayésiens. Finalement, on remarque que le modèle d'arbres bayésien avec $P(T)$ non standard a un nombre de feuilles moyen MAP très proche de 8 et une médiane égale à 8, contrairement aux deux autres modèles qui sous-estiment légèrement le vrai nombre de feuilles. Le modèle d'arbres bayésien avec $P(T)$ non standard est donc le plus approprié pour ce scénario. Par contre, le modèle d'arbres bayésien avec $P(T)$ non standard est seulement légèrement supérieur au modèle d'arbres bayésien avec $P(T)$ par défaut, en terme de capacité prédictive.

Lorsque le rapport signal sur bruit est petit, on a donc que le modèle d'arbres bayésien avec $P(T)$ non standard performe aussi bien (et même des fois mieux) que l'arbre fréquentiste, peu importe le nombre d'observations. De plus, on constate que de bien choisir ces hyperparamètres, de façon à obtenir une loi *a priori* de T centrée sur le bon nombre de feuilles, amène à un gain de performance, par rapport au choix d'hyperparamètres par défaut, mais que ce gain devient moins important à mesure que le nombre d'observations augmente.

Tableau 4.7 : Scénario 4 ($n = 50$, $\sigma_\epsilon = 2$, basé sur 50 échantillons). Espérance et écart-type du R^2, du R^2_V, de l'erreur quadratique moyenne de $\hat{f}(x)$ (RMSE) et du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.45$) et pour l'arbre fréquentiste

	Bayésien par défaut		**Bayésien non standard**		**Fréquentiste**	
	μ	σ	μ	σ	μ	σ
R^2	0.27	0.15	0.41	0.15	0.62	0.22
R^2_V [a]	0.14	0.09	0.23	0.10	0.21	0.14
RMSE	1.99	0.30	1.66	0.33	1.43	0.43
Nombre de feuilles[b]	3.86	0.88	5.68	0.86	5.76	2.71

[a] basé sur un échantillon de 1000 nouvelles observations

[b] basé sur l'arbre MAP dans le cas bayésien

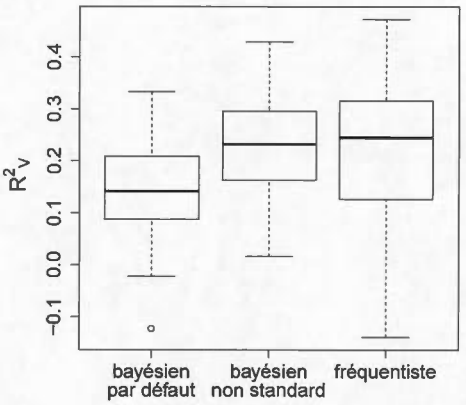


Figure 4.9 : Scénario 4 ($n = 50$, $\sigma_\epsilon = 2$, basé sur 50 échantillons). Diagramme en boîtes du R^2_V pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.45$) et pour l'arbre fréquentiste

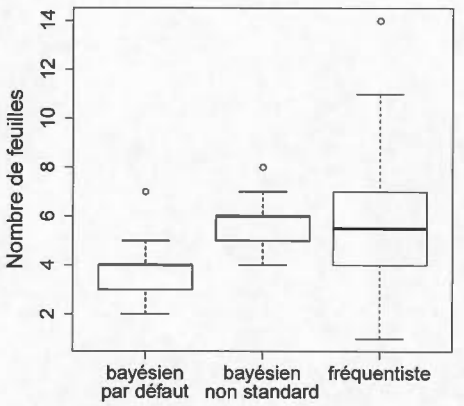


Figure 4.10 : Scénario 4 ($n = 50$, $\sigma_\epsilon = 2$, basé sur 50 échantillons). Diagramme en boîtes du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.45$) et pour l'arbre fréquentiste

Tableau 4.8 : Scénario 5 ($n = 100$, $\sigma_\epsilon = 2$, basé sur 25 échantillons). Espérance et écart-type du R^2, du R^2_V, de l'erreur quadratique moyenne de $\hat{f}(x)$ (RMSE) et du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.50$) et pour l'arbre fréquentiste

	Bayésien par défaut		**Bayésien non standard**		**Fréquentiste**	
	μ	σ	μ	σ	μ	σ
R^2	0.50	0.10	0.56	0.07	0.67	0.06
R^2_V [a]	0.35	0.08	0.40	0.06	0.40	0.06
RMSE	1.33	0.28	1.15	0.19	0.99	0.26
Nombre de feuilles[b]	5.4	0.50	6.6	1.04	7.08	2.60

[a] basé sur un échantillon de 1000 nouvelles observations

[b] basé sur l'arbre MAP dans le cas bayésien

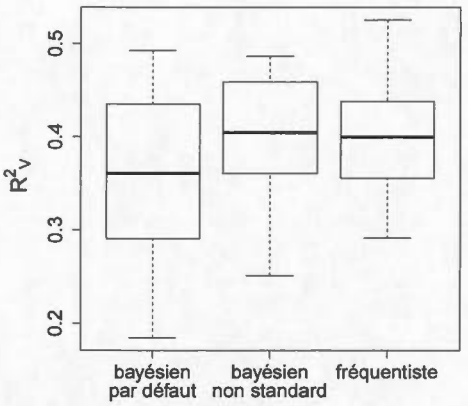


Figure 4.11 : Scénario 5 ($n = 100$, $\sigma_\epsilon = 2$, basé sur 25 échantillons). Diagramme en boîtes du R^2_V pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.50$) et pour l'arbre fréquentiste

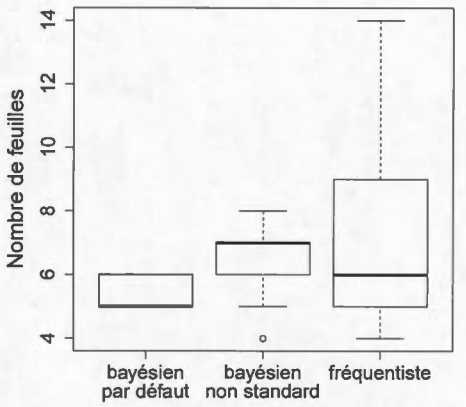


Figure 4.12 : Scénario 5 ($n = 100$, $\sigma_\epsilon = 2$, basé sur 25 échantillons). Diagramme en boîtes du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.50$) et pour l'arbre fréquentiste

Tableau 4.9 : Scénario 6 ($n = 500$, $\sigma_\epsilon = 2$, basé sur 10 échantillons). Espérance et écart-type du R^2, du R^2_V, de l'erreur quadratique moyenne de $\hat{f}(x)$ (RMSE) et du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.56$) et pour l'arbre fréquentiste

	Bayésien par défaut		**Bayésien non standard**		**Fréquentiste**	
	μ	σ	μ	σ	μ	σ
R^2	0.61	0.02	0.63	0.03	0.63	0.03
R^2_V [a]	0.56	0.02	0.57	0.02	0.55	0.04
RMSE	0.61	0.08	0.56	0.09	0.63	0.14
Nombre de feuilles [b]	6.7	0.95	8.2	1.62	6.7	0.95

[a] basé sur un échantillon de 1000 nouvelles observations

[b] basé sur l'arbre MAP dans le cas bayésien

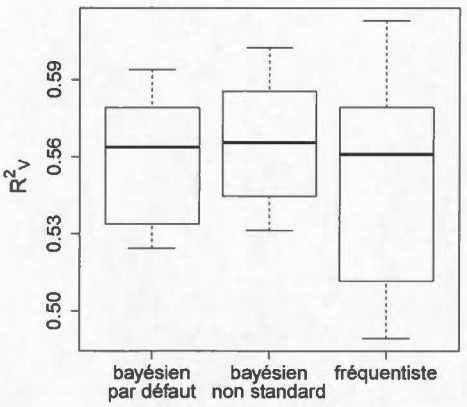


Figure 4.13 : Scénario 6 ($n = 500$, $\sigma_\epsilon = 2$, basé sur 10 échantillons). Diagramme en boîtes du R^2_V pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.56$) et pour l'arbre fréquentiste

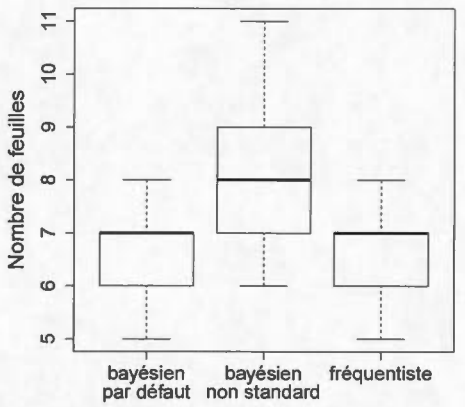


Figure 4.14 : Scénario 6 ($n = 500$, $\sigma_\epsilon = 2$, basé sur 10 échantillons). Diagramme en boîtes du nombre de feuilles pour le modèle d'arbres bayésien avec $P(T)$ par défaut ($\alpha = 0.50$ et $\beta = 2$), le modèle d'arbres bayésien avec $P(T)$ non standard ($\alpha = 1$ et $\beta = 0.56$) et pour l'arbre fréquentiste

4.4 Synthèse des résultats

Les résultats des scénarios 1 à 6 révèlent que le modèle d'arbres bayésien performe mieux lorsqu'on choisit les hyperparamètres de façon à obtenir une loi *a priori* centrée autour du vrai nombre de feuilles. De plus, le modèle d'arbres bayésien atteint au moins la même performance prédictive que l'arbre fréquentiste lorsque l'on choisit bien les hyperparamètres et que le rapport signal sur bruit est petit. Ce résultat est très important car dans les jeux de données basés sur de véritables phénomènes naturels (non simulés), le rapport signal sur bruit a tendance à être très petit. En effet, on a souvent besoin d'un grand nombre d'observations pour détecter le signal de jeux de données réelles. Avec le modèle d'arbres bayésien, on peut performer mieux dans les jeux de données contenant beaucoup de bruit en ayant seulement une idée imprécise du nombre de feuilles optimal de l'arbre à estimer. Les choix d'hyperparamètres, centrés sur le vrai nombre de feuilles, dans l'exemple de ce chapitre, amenaient à un nombre de feuilles *a priori* très variable, se situant entre 3 et 14 ou 2 et 15. On peut suspecter que la performance prédictive du modèle d'arbres bayésien aurait pu être encore meilleure si l'on avait utilisé la définition alternative de la probabilité de séparation (3.3). Avec la définition actuelle, tel qu'utilisée par les paquets `tgp` et `BayesTree`, il est impossible de réduire considérablement la probabilité *a priori* des arbres avec un nombre de feuilles très différent de 8. En conclusion, les résultats obtenus de cette étude de simulation suggèrent que le modèle d'arbres bayésien est un modèle à favoriser lorsque nous avons une connaissance *a priori* du nombre de feuilles de l'arbre (représentant la complexité désirée du modèle) et un jeu de données contenant beaucoup de bruit.

CONCLUSION

Dans ce mémoire, on a étudié en profondeur la loi *a priori* de l'arbre de régression bayésien définie par Chipman *et al.* (1998) dans le but de mieux comprendre son fonctionnement. On a également déterminé l'impact du choix d'hyperparamètres, du nombre de feuilles et du signal sur bruit sur l'estimation d'un arbre de régression univarié à l'aide du paquet `tgp` du langage de programmation `R`.

On a commencé par analyser un cas spécial et idéal, i.e., avec $\beta = 0$ et une infinité d'observations distinctes. Les formules de l'espérance et de la variance du nombre de feuilles, en fonction de α, ont été dérivées (Propositions 3.3 et 3.4). Ces analyses nous ont permis de constater que l'espérance du nombre de feuilles des arbres augmente avec α et atteint son maximum à l'infini, lorsque $\alpha \geq 0.50$.

On a ensuite analysé un cas plus réaliste, i.e., avec $\beta = 0$ et un nombre fini d'observations distinctes. Pour ce faire, on a utilisé un processus récursif de création d'arbre modifié qui on empêche la division d'un noeud menant à la création d'un enfant n'ayant aucune observation. Peu importe le nombre d'observations distinctes, on a constaté qu'on obtient approximativement la même moyenne du nombre de feuilles que dans le cas idéal avec une infinité d'observations distinctes lorsque α est petit. Par contre, lorsque α est grand, il est difficile d'obtenir des arbres avec autant de feuilles en moyenne que dans le cas idéal. Plus le nombre d'observations distinctes est grand, plus on s'approche du cas idéal ; si α est trop grand toutefois, il devient très improbable d'avoir suffisamment d'observations distinctes pour atteindre la valeur de l'espérance du nombre de feuilles du cas idéal.

Par la suite, on a étudié le cas général, i.e., avec $\beta \neq 0$ et un nombre varié d'observations distinctes ($n = \infty$, $n = 500$, $n = 2000$). Similairement au cas précédent avec $\beta = 0$, on a observé qu'il est difficile d'approcher des valeurs de l'espérance du nombre de feuilles et de la profondeur obtenues dans le cas idéal (avec une infinité d'observations distinctes) lorsque le nombre d'observations distinctes est petit. La matrice des variables explicatives X a donc un grand impact sur la loi *a priori* de T. En effet, le nombre d'observations distinctes influence $P(T)$ de façon à rendre plus difficile l'obtention d'arbres avec beaucoup de feuilles et très profonds. De plus, on a remarqué qu'augmenter α ou réduire β, peu importe le nombre d'observations distinctes (fini ou infini), réduit le nombre moyen de feuilles et la profondeur moyenne des arbres. Il existe donc une infinité de combinaisons possibles de α et β menant au même nombre de feuilles ou profondeur. L'hyperparamètre β ne permet donc pas de directement changer la profondeur moyenne sans changer le nombre de feuilles. Par contre, en choisissant de grands α et β, on peut réduire la variance du nombre de feuilles des arbres. Ainsi, les hyperparamètres α et β peuvent tous les deux contrôler la moyenne et la variance du nombre de feuilles de la loi $P(T)$.

À cause de la restriction $\alpha < 1$, il est impossible d'obtenir une loi *a priori* avec une moyenne arbitrairement grande et une variance arbitrairement petite du nombre de feuilles. En utilisant une définition alternative (Définition 3.2) de la probabilité de séparation (Définition 1.5), on permet l'utilisation de $\alpha > 1$. Cette nouvelle définition permet donc d'obtenir une moyenne du nombre de feuilles aussi grande que désirée (jusqu'au maximum imposé par le nombre d'observations distinctes dans X) et une variance du nombre de feuilles aussi petite que désirée.

L'ensemble des résultats mentionnés précédemment a permis de bien comprendre comment on peut choisir les hyperparamètres α et β de façon à définir $P(T)$ pour qu'elle reflète nos connaissances *a priori* sur la prédiction de la variable expliquée.

On a ensuite présenté une étude de simulation, où l'on a estimé un arbre de régression univarié de 8 feuilles avec différents choix d'hyperparamètres, de nombre d'observations et du rapport signal sur bruit. Il a été constaté qu'en choisissant les hyperparamètres de façon à positionner la moyenne *a priori* du nombre de feuille autour du véritable nombre de feuilles, on obtient des arbres *a posteriori* plus semblables au véritable arbre de régression et une capacité prédictive plus grande que lorsqu'on utilise les hyperparamètres par défaut. Cette amélioration de performance est particulièrement importante lorsque le rapport sur bruit est grand et lorsque le nombre d'observations est petit. De plus, on a remarqué que la performance prédictive des arbres de régression avec le choix plus adapté d'hyperparamètres est comparable à celle de l'arbre fréquentiste lorsque le rapport sur bruit est grand, peu importe le nombre d'observations.

Cette recherche comporte certaines limitations, la principale étant que l'on n'a pas étudié en détail les arbres multivariés avec plusieurs variables explicatives. L'effet de la matrice des variables explicatives X sur $P(T)$ est très complexe lorsqu'il y a plusieurs variables explicatives. Même sans avoir étudié l'effet de plusieurs variables explicatives sur $P(T)$, il est toutefois possible d'obtenir les valeurs de la moyenne et de l'écart-type du nombre de feuilles d'une matrice explicative X complexe. On peut simplement utiliser l'algorithme récursif de création d'arbres et simuler un grand nombre d'arbres possibles. Ensuite, on calcule la moyenne et l'écart-type approximatif basé sur cet échantillon. Ces simulations pourraient être plutôt longues à compléter, mais elles permettraient de bien identifier l'impact de la matrice X sur les arbres *a priori*. De plus, on n'a pas étudié les arbres de classification (avec une variable expliquée binaire ou catégorique). Il serait important de vérifier si les résultats de ce mémoire se généralisent aux arbres de classification. Finalement, la définition alternative de la probabilité de séparation, présentée dans ce mémoire, n'a pas été adaptée pour la méthode de Metropolis-

Hastings et testée réellement, elle demeure donc hypothétique.

Comme point de départ pour la recherche future, il serait intéressant de travailler sur les limitations de ce mémoire. Étudier les arbres de classification serait la tâche la plus simple. Il s'agirait simplement de construire un exemple synthétique avec une variable binaire ou catégorique et vérifier si la loi *a priori* de T influence les arbres *a posteriori* tel que prédit par ce mémoire. Analyser l'effet de la matrice des variables explicatives sur $P(T)$ serait une tâche beaucoup plus difficile. Il faudrait tester l'effet de différents types de matrices sur les lois *a priori* et *a posteriori*. Ainsi, on pourrait peut-être obtenir une vue d'ensemble et comprendre comment une matrice générale peut influencer $P(T)$. Implémenter la définition alternative de la probabilité de séparation ne serait probablement pas trop difficile, mais pourrait demander un grand temps de programmation pour s'assurer que la simulation fonctionne sans erreur.

Comme autre piste de recherche future, il serait intéressant de comprendre comment le choix de la loi *a priori* affecte la qualité des résultats des *bayesian additive regression trees* (BART) (Wu *et al.*, 2007). La méthode BART consiste en l'addition de plusieurs arbres bayésiens de façon à obtenir un modèle plus complexe qu'avec un seul arbre. Par défaut, les hyperparamètres du BART produisent en moyenne de très petits arbres (avec peu de feuilles et une petite profondeur), alors le nombre d'arbres additionnés doit être grand pour obtenir de bons résultats prédictifs. Alternativement, il serait possible de choisir des hyperparamètres qui produisent en moyenne de plus grands arbres (avec plus de feuilles et une plus grande profondeur) et ainsi on aurait besoin d'additionner moins d'arbres ensemble pour obtenir un bon modèle prédictif. On pourrait tenter de déterminer quel choix de α, β et du nombre d'arbres additionnés amène à la meilleure capacité prédictive.

APPENDICE A

DÉMONSTRATIONS

Définition 3.1. *Les nombres de Catalan sont définis par la relation de récurrence*

$$C_0 = 1 \quad et \quad C_{n+1} = \sum_{b=0}^{n} C_b C_{n-b} \ \ lorsque\ n \geq 0.$$

Lemme A.1. *Soit C_b le b-ième nombre de Catalan. On a que $\sum_{b=0}^{\infty} C_b x^b = \frac{1-\sqrt{1-4x}}{2x}$ pour tout $|x| \leq \frac{1}{4}$.*

Démonstration. On commence par calculer le rayon de convergence en utilisant la règle de d'Alembert (*ratio test*) (Adams, 1996). Sachant que $C_{b+1} = \left[\frac{2(2b+1)}{(b+2)}\right] C_b$ (Koshy, 2008),

$$\begin{aligned} r &= \lim_{b\to\infty} \left| \frac{C_{b+1}x^{b+1}}{C_b x^b} \right| \\ &= \lim_{b\to\infty} \left| \frac{\left[\frac{2(2b+1)}{(b+2)}\right] C_b x}{C_b} \right| \\ &= \lim_{b\to\infty} \left| \frac{2(2b+1)x}{b+2} \right| \\ &= \lim_{b\to\infty} \frac{(4b+2)|x|}{b+2} \\ &= 4|x|. \end{aligned}$$

Pour que la série converge, on doit avoir $r < 1$ et donc $|x| < \frac{1}{4}$.

On tente ensuite de calculer la somme de la série lorsque $|x| < \frac{1}{4}$. Soit $c(x) = \sum_{b=0}^{\infty} C_b x^b$, alors on a que

$$\begin{aligned}
c(x)^2 &= \left(\sum_{m=0}^{\infty} C_m x^m\right)\left(\sum_{n=0}^{\infty} C_n x^n\right) \\
&= \sum_{m=0}^{\infty} C_m \sum_{n=0}^{\infty} C_n x^{n+m} \\
&= \sum_{m=0}^{\infty} C_m \sum_{n=m}^{\infty} C_{n-m} x^n \\
&= \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} C_m C_{n-m} x^n \\
&= \sum_{n=0}^{\infty} \sum_{m=0}^{n} C_m C_{n-m} x^n \\
&= \sum_{n=0}^{\infty} C_{n+1} x^n \quad \text{par la définition 3.1} \\
&= \sum_{n=1}^{\infty} C_n x^{n-1}.
\end{aligned}$$

$$\begin{aligned}
c(x) &= \sum_{b=0}^{\infty} C_b x^b \\
&= C_0 + C_1 x + C_2 x^2 + C_3 x^3 + \ldots \\
&= C_0 + x[C_1 + C_2 x + C_3 x^2 + \ldots] \\
&= C_0 + x \sum_{n=1}^{\infty} C_n x^{n-1} \\
&= C_0 + x c(x)^2 \quad \text{par le résultat précédent} \\
&= 1 + x c(x)^2.
\end{aligned}$$

Alors, on a l'équation quadratique suivante

$$x c(x)^2 - c(x) + 1 = 0.$$

Cela implique que les solutions possibles sont

$$c(x) = \frac{1 \pm \sqrt{1-4x}}{2x}.$$

La solution doit satisfaire la condition suivante

$$\lim_{x \to 0^+} c(x) = C_0 = 1.$$

Dans le cas positif, on a que

$$\lim_{x \to 0^+} \frac{1+\sqrt{1-4x}}{2x} = \lim_{x \to 0^+} \frac{2}{2x} = \infty$$

et dans le cas négatif

$$\lim_{x \to 0^+} \frac{1-\sqrt{1-4x}}{2x} = \lim_{x \to 0^+} \frac{\frac{4}{2\sqrt{1-4x}}}{2} = \lim_{x \to 0^+} \frac{1}{\sqrt{1-4x}} = 1.$$

Il résulte donc que

$$c(x) = \frac{1-\sqrt{1-4x}}{2x} \text{ pour tout} |x| \leq \frac{1}{4}.$$

□

Proposition 3.2. *Soit T un arbre défini de façon récursive avec probabilité de séparation α. Si N est le nombre de feuilles de T, alors $P(N = b) = p(b)$, où*

$$p(b) = \begin{cases} C_{b-1}\alpha^{b-1}(1-\alpha)^b & \text{si } 1 \leq b < \infty, \\ max\{0, 2-1/\alpha\} & \text{si } b = \infty, \end{cases} \quad \text{pour tout } \alpha \in [0,1].$$

Démonstration. Il existe C_{b-1} arbres avec b feuilles et chacun de ces arbres a une probabilité *a priori* égale à $\alpha^{b-1}(1-\alpha)^b$. On a donc que $P(N = b) = C_{b-1}\alpha^{b-1}(1-\alpha)^b$ lorsque $1 \leq b < \infty$. Sachant que $P(1 \leq N \leq \infty)$ doit être égal à 1, on peut déduire $P(N = \infty)$ de la façon suivante :

$$
\begin{aligned}
P(N=\infty) &= P(1 \le N \le \infty) - P(1 \le N < \infty) \\
&= 1 - P(1 \le N < \infty) \\
&= 1 - \sum_{b=1}^{\infty} C_{b-1}\alpha^{b-1}(1-\alpha)^b \\
&= 1 - \sum_{b=0}^{\infty} C_b\alpha^b(1-\alpha)^{b+1} \\
&= 1 - (1-\alpha)\sum_{b=0}^{\infty} C_b[\alpha(1-\alpha)]^b.
\end{aligned}
$$

Sachant que $\alpha \in (0,1) \implies |\alpha(1-\alpha)| \le \frac{1}{4}$,

$$
\begin{aligned}
P(N=\infty) &= 1 - (1-\alpha)\frac{1-\sqrt{1-4(1-\alpha)\alpha}}{2(1-\alpha)\alpha} \quad \text{par le Lemme } A.1 \\
&= 1 - \frac{1-\sqrt{1-4(\alpha-\alpha^2)}}{2\alpha} \\
&= 1 - \frac{1-\sqrt{1-4\alpha+4\alpha^2}}{2\alpha} \\
&= 1 - \frac{1-\sqrt{(1-2\alpha)^2}}{2\alpha} \\
&= \begin{cases} 1 - \frac{1-(1-2\alpha)}{2\alpha} & \text{si } \alpha \le \frac{1}{2}; \\ 1 - \frac{1+(1-2\alpha)}{2\alpha} & \text{si } \alpha > \frac{1}{2}; \end{cases} \\
&= \begin{cases} 1-1 & \text{si } \alpha \le \frac{1}{2}; \\ 1 - \frac{(1-\alpha)}{\alpha} & \text{si } \alpha > \frac{1}{2}; \end{cases} \\
&= \begin{cases} 0 & \text{si } \alpha \le \frac{1}{2}; \\ 2 - 1/\alpha & \text{si } \alpha > \frac{1}{2}; \end{cases} \\
&= \max\{0, 2-1/\alpha\}.
\end{aligned}
$$

□

Lemme A.2. *Soit C_b le b-ième nombre de Catalan. On a que $\sum_{b=1}^{\infty} bC_{b-1}x^{b-1} = \frac{1}{\sqrt{1-4x}}$ pour tout $|x| \le \frac{1}{4}$.*

Démonstration. En présumant que $|x| \leq \frac{1}{4}$, on a que

$$\begin{aligned}
\frac{d}{dx}\sum_{b=0}^{\infty} C_b x^{b+1} &= \frac{d}{dx}\left[x\sum_{b=0}^{\infty} C_b x^b\right] \\
&= \frac{d}{dx}\left[x\frac{1-\sqrt{1-4x}}{2x}\right] \quad \text{par le Lemme } A.1 \\
&= \frac{d}{dx}\left[\frac{1-\sqrt{1-4x}}{2}\right] \\
&= \frac{d}{dx}\left[-\frac{\sqrt{1-4x}}{2}\right] \\
&= -\frac{-4}{4\sqrt{1-4x}} \\
&= \frac{1}{\sqrt{1-4x}}.
\end{aligned}$$

$$\begin{aligned}
\sum_{b=1}^{\infty} b C_{b-1} x^{b-1} &= \sum_{b=0}^{\infty} (b+1) C_b x^b \\
&= \sum_{b=0}^{\infty} \frac{d}{dx} C_b x^{b+1} \\
&= \frac{d}{dx}\sum_{b=0}^{\infty} C_b x^{b+1} \\
&= \frac{1}{\sqrt{1-4x}}.
\end{aligned}$$

□

Proposition 3.3. *Soit T un arbre défini de façon récursive avec probabilité de séparation α. Si N est le nombre de feuilles de T, alors*

$$E(N) = \begin{cases} \frac{(1-\alpha)}{(1-2\alpha)} & si\ \alpha \leq \frac{1}{2}; \\ \infty & si\ \alpha > \frac{1}{2}. \end{cases}$$

Démonstration.

$$E(N) = \sum_{b=1}^{\infty} bC_{b-1}\alpha^{b-1}(1-\alpha)^b + \left[\lim_{b\to\infty} b\right] \max\{0, 2-1/\alpha\}$$
$$= \begin{cases} (1-\alpha)\sum_{b=1}^{\infty} bC_{b-1}[\alpha(1-\alpha)]^{b-1} & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}. \end{cases}$$

Sachant que $\alpha \in (0,1) \implies |\alpha(1-\alpha)| \leq \frac{1}{4}$,

$$E(N) = \begin{cases} \frac{(1-\alpha)}{\sqrt{1-4(1-\alpha)\alpha}} & \text{si } \alpha \leq \frac{1}{2}; \quad \text{par le Lemme } A.2 \\ \infty & \text{si } \alpha > \frac{1}{2}; \end{cases}$$
$$= \begin{cases} \frac{(1-\alpha)}{\sqrt{(1-2\alpha)^2}} & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}; \end{cases}$$
$$= \begin{cases} \frac{(1-\alpha)}{(1-2\alpha)} & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}. \end{cases}$$

□

Lemme A.3. *Soit C_b le b-ième nombre de Catalan. On a que* $\sum_{b=1}^{\infty} b^2 C_{b-1} x^{b-1} = \frac{(1-2x)}{(1-4x)^{3/2}}$ *pour tout* $|x| \leq \frac{1}{4}$.

Démonstration. En présumant que $|x| \leq \frac{1}{4}$, on a que

$$\begin{aligned}
\frac{d^2}{dx^2}\sum_{b=0}^{\infty} C_b x^{b+2} &= \frac{d^2}{dx^2}\left[x^2 \sum_{b=0}^{\infty} C_b x^b\right] \\
&= \frac{d^2}{dx^2}\left[x^2\left(\frac{1-\sqrt{1-4x}}{2x}\right)\right] \quad \text{par le Lemme } A.1 \\
&= \frac{d^2}{dx^2}\left[\frac{x(1-\sqrt{1-4x})}{2}\right] \\
&= \frac{d}{dx}\left[\frac{1}{2} - \frac{1}{2}\left(\sqrt{1-4x} - \frac{4x}{2\sqrt{1-4x}}\right)\right] \\
&= \frac{d}{dx}\left[\frac{1}{2} - \frac{2(1-4x)-4x}{4\sqrt{1-4x}}\right] \\
&= \frac{d}{dx}\left[\frac{1}{2} - \frac{2-12x}{4\sqrt{1-4x}}\right] \\
&= \frac{d}{dx}\left[\frac{1}{2} - \frac{1-6x}{2\sqrt{1-4x}}\right] \\
&= \frac{d}{dx}\left[\frac{1}{2} + \frac{6x-1}{2\sqrt{1-4x}}\right] \\
&= \frac{6 \times 2\sqrt{1-4x} + \frac{4}{\sqrt{1-4x}}(6x-1)}{4(1-4x)} \\
&= \frac{12\sqrt{1-4x}}{4(1-4x)} + \frac{4(6x-1)}{4(1-4x)^{3/2}} \\
&= \frac{3}{\sqrt{1-4x}} + \frac{(6x-1)}{(1-4x)^{3/2}}.
\end{aligned}$$

$$\begin{aligned}
\frac{d^2}{dx^2}\sum_{b=0}^{\infty} C_b x^{b+2} &= \sum_{b=0}^{\infty} \frac{d^2}{dx^2} C_b x^{b+2} \\
&= \sum_{b=0}^{\infty} \frac{d}{dx}(b+2) C_b x^{b+1} \\
&= \sum_{b=0}^{\infty} (b+2)(b+1) C_b x^{b} \\
&= \sum_{b=1}^{\infty} (b+1) b C_{b-1} x^{b-1} \\
&= \sum_{b=1}^{\infty} (b^2+b) C_{b-1} x^{b-1} \\
&= \sum_{b=1}^{\infty} b^2 C_{b-1} x^{b-1} + \sum_{b=1}^{\infty} b C_{b-1} x^{b-1}.
\end{aligned}$$

$$\begin{aligned}
\sum_{b=1}^{\infty} b^2 C_{b-1} x^{b-1} &= \frac{d^2}{dx^2}\sum_{b=0}^{\infty} C_b x^{b+2} - \sum_{b=1}^{\infty} b C_{b-1} x^{b-1} \\
&= \frac{3}{\sqrt{1-4x}} + \frac{(6x-1)}{(1-4x)^{3/2}} - \frac{1}{\sqrt{1-4x}} \\
&= \frac{2}{\sqrt{1-4x}} + \frac{(6x-1)}{(1-4x)^{3/2}} \\
&= \frac{2(1-4x)+(6x-1)}{(1-4x)^{3/2}} \\
&= \frac{(1-2x)}{(1-4x)^{3/2}}.
\end{aligned}$$

□

Lemme A.4. *Soit T un arbre défini de façon récursive avec probabilité de séparation de α. Si N est le nombre de feuilles de T, alors*

$$E(N^2) = \begin{cases} \frac{(1-\alpha)(2\alpha^2-2\alpha+1)}{(1-2\alpha)^3} & si\ \alpha \leq \frac{1}{2}; \\ \infty & si\ \alpha > \frac{1}{2}. \end{cases}$$

Démonstration.

$$E(N^2) = \sum_{b=1}^{\infty} b^2 C_{b-1} \alpha^{b-1}(1-\alpha)^b + \left[\lim_{b\to\infty} b^2\right] \max\{0, 2-1/\alpha\}$$

$$= \begin{cases} (1-\alpha)\sum_{b=1}^{\infty} b^2 C_{b-1}[\alpha(1-\alpha)]^{b-1} & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}. \end{cases}$$

Sachant que $\alpha \in (0,1) \implies |\alpha(1-\alpha)| \leq \frac{1}{4}$,

$$E(N^2) = \begin{cases} (1-\alpha)\left[\frac{(1-2\alpha(1-\alpha))}{(1-4(1-\alpha)\alpha)^{3/2}}\right] & \text{si } \alpha \leq \frac{1}{2}; \quad \text{par le Lemme } A.3 \\ \infty & \text{si } \alpha > \frac{1}{2}; \end{cases}$$

$$= \begin{cases} (1-\alpha)\left[\frac{1-2\alpha+2\alpha^2}{(\sqrt{(1-2\alpha)^2})^3}\right] & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}; \end{cases}$$

$$= \begin{cases} \frac{(1-\alpha)(1-2\alpha+2\alpha^2)}{(1-2\alpha)^3} & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}. \end{cases}$$

□

Proposition 3.4. *Soit T un arbre défini de façon récursive avec probabilité de séparation de α. Si N est le nombre de feuilles de T, alors*

$$Var(N) = \begin{cases} \frac{(1-\alpha)\alpha}{(1-2\alpha)^3} & si\ \alpha \leq \frac{1}{2}; \\ \infty & si\ \alpha > \frac{1}{2}. \end{cases}$$

Démonstration.

$$
\begin{aligned}
Var(N) &= \begin{cases} E(N^2) - E(N)^2 & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}; \quad \text{car } E(N) = \infty, \text{ par Prop. 3.3} \end{cases} \\
&= \begin{cases} \frac{(1-\alpha)(1-2\alpha+2\alpha^2)}{(1-2\alpha)^3} - \frac{(1-\alpha)^2}{(1-2\alpha)^2} & \text{si } \alpha \leq \frac{1}{2}; \quad \text{par Prop. 3.3 et Lemme } A.4 \\ \infty & \text{si } \alpha > \frac{1}{2}; \end{cases} \\
&= \begin{cases} \frac{(1-\alpha)}{(1-2\alpha)^2}\left[\frac{2\alpha^2-2\alpha+1}{(1-2\alpha)} - (1-\alpha)\right] & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}; \end{cases} \\
&= \begin{cases} \frac{(1-\alpha)}{(1-2\alpha)^3}\left[2\alpha^2 - 2\alpha + 1 - (1-\alpha)(1-2\alpha)\right] & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}; \end{cases} \\
&= \begin{cases} \frac{(1-\alpha)}{(1-2\alpha)^3}\left[2\alpha^2 - 2\alpha + 1 - 1 + 3\alpha - 2\alpha^2\right] & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}; \end{cases} \\
&= \begin{cases} \frac{(1-\alpha)\alpha}{(1-2\alpha)^3} & \text{si } \alpha \leq \frac{1}{2}; \\ \infty & \text{si } \alpha > \frac{1}{2}. \end{cases}
\end{aligned}
$$

□

RÉFÉRENCES

Adams, R. A. (1996). *Calculus of several variables* (3^e éd.). Boston, USA : Addison-Wesley.

Breiman, L., Friedman, J., Stone, C. J. et Olshen, R. A. (1984). *Classification and regression trees* (1^e éd.). Wadsworth Statistics/Probability. London, UK : Chapman & Hall/CRC.

Chambers, J. M., Cleveland, W. S., Kleiner, B. et Tukey, P. A. (1983). *Graphical methods for data analysis* (1^e éd.). Wadsworth Statistics/Probability. London, UK : Chapman & Hall/CRC.

Chipman, H. et McCulloch, R. (2014). *BayesTree : Bayesian Additive Regression Trees* (Version 0.3-1.2). [Paquet R]. Récupéré de `http://CRAN.R-project.org/package=BayesTree`

Chipman, H. A., George, E. I. et McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, *93*(443), 935–948.

Denison, D. G., Mallick, B. K. et Smith, A. F. (1998). A Bayesian CART algorithm. *Biometrika*, *85*(2), 363–377.

Dicker, L. H. (2012). Residual variance and the signal-to-noise ratio in high-dimensional linear models. *arXiv preprint arXiv :1209.0012*.

Gramacy, R. B. (2007). tgp : an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models. *Journal of Statistical Software*, *19*(9), 6.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (2^e éd.). Springer Series in Statistics. New York City, USA : Springer.

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

Hoeting, J. A., Madigan, D., Raftery, A. E. et Volinsky, C. T. (1999). Bayesian model averaging : a tutorial. *Statistical science*, 382–401.

Knuth, D. E. (1998). *The art of computer programming : sorting and searching* (2^e éd.), volume 3. New York City, USA : Pearson Education.

Koshy, T. (2008). *Catalan numbers with applications* (1^e éd.). New York City, USA : Oxford University Press.

Montgomery, D. C., Peck, E. A. et Vining, G. G. (2012). *Introduction to linear regression analysis* (5^e éd.). Wiley Series in Probability and Statistics. Hoboken, USA : John Wiley & Sons.

Quinlan, J. R. (1993). *C4. 5 : programs for machine learning* (1^e éd.). Morgan Kaufmann Series in Machine Learning. San Francisco, USA : Morgan Kaufmann.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, *47*(1), 1–52.

Thathachar, M. A. et Sastry, P. S. (2004). *Networks of learning automata : Techniques for online stochastic optimization* (1^e éd.). New York City, USA : Springer US.

Therneau, T., Atkinson, B. et Ripley, B. (2014). *rpart : Recursive Partitioning and Regression Trees* (Version 4.1-10). [Paquet R]. Récupéré de `http://CRAN.R-project.org/package=rpart`

Wu, Y., Tjelmeland, H. et West, M. (2007). Bayesian CART : Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics*, *16*(1), 44–66.