

Université de Montréal

**Sélection de modèles robuste : régression linéaire et algorithme
à sauts réversibles**

par

Philippe Gagnon

Département de mathématiques et de statistique
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Statistique

13 octobre 2017

SOMMAIRE

Dans cette thèse, deux aspects incontournables de l'analyse statistique sont traités, soient la sélection de modèles et l'estimation des paramètres. Ceci est effectué dans un contexte bayésien par l'intermédiaire de trois articles. Dans le premier, ces aspects sont traités d'un point de vue computationnel. L'algorithme à sauts réversibles, une méthode Monte Carlo par chaînes de Markov permettant simultanément la sélection de modèles et l'estimation des paramètres, est analysé dans l'objectif d'indiquer à l'utilisateur la façon optimale de l'implémenter. Un algorithme implémenté optimalement correspond à un algorithme engendrant des chaînes de Markov qui explorent leur espace d'états de façon optimale. L'objectif est atteint par l'intermédiaire de l'optimisation d'un processus stochastique correspondant à la limite (en distribution) de la suite des processus stochastiques engendrés par cet algorithme. Dans le deuxième article, une stratégie menant à l'estimation robuste des paramètres d'un modèle de régression linéaire en présence de valeurs aberrantes est présentée. La stratégie consiste à poser des hypothèses plus adaptées à cette éventualité de présence de valeurs aberrantes, comparativement au modèle traditionnel basé sur l'hypothèse de normalité des erreurs. Il s'agit de remplacer cette hypothèse de normalité par une hypothèse de distribution à ailes extrêmement relevées. La robustesse, se traduisant par la convergence de la distribution *a posteriori* des paramètres (basée sur l'échantillon entier) vers celle excluant les valeurs aberrantes, est garantie lorsque le nombre de valeurs aberrantes ne dépasse pas un certain seuil. Finalement, les résultats présentés dans les deux premiers articles sont combinés afin d'introduire une approche bayésienne de régression robuste sur composantes principales faisant intervenir la sélection de modèles dans le processus de prédiction. Ces caractéristiques de robustesse et d'incorporation de la sélection de modèles dans l'analyse contribuent à l'amélioration de la précision des prédictions produites.

Mots-clés : algorithme Metropolis de type marche aléatoire ; analyse en composantes principales ; distributions à ailes extrêmement relevées ; inférence bayésienne ; méthodes de Monte Carlo par chaînes de Markov ; robustesse ; valeurs aberrantes.

SUMMARY

Model selection and parameter estimation are two main aspects of statistical analysis. This thesis discusses these aspects from a Bayesian point of view via three papers. The first one deals with a computational procedure, named the reversible jump algorithm, that allows to simultaneously select models and estimate parameters. This sampler being difficult to tune in practice, we aim at providing guidelines to users for an optimal implementation. An optimally tuned sampler corresponds to a sampler that generates Markov chains that optimally explore their state space. Our goal is achieved through the optimisation of a stochastic process that corresponds to the limit (in distribution) of the sequence of stochastic processes engendered by the algorithm. In the second paper, a strategy leading to robust estimation of the parameters of a linear regression model in presence of outliers is presented. The strategy is to make assumptions that are more adapted to the eventual presence of outliers, compared with the traditional model assuming normality of errors. This normality assumption is indeed replaced by a super heavy-tailed distribution assumption. Robustness, which is represented by the convergence of the posterior distribution of the parameters (based on the whole sample) towards that arising from the nonoutliers only, is guaranteed when the number of outliers does not exceed a given threshold. Finally, the results presented in the first two papers are combined to introduce a Bayesian robust principal component regression approach that involves model selection in the prediction process. The characteristics of this approach contribute to increase the accuracy of the predictions produced.

Keywords: Bayesian inference; Markov chain Monte Carlo methods; Outliers; Principal component analysis; Random walk Metropolis algorithm; Robustness; Super heavy-tailed distributions.

Table des matières

Sommaire	iii
Summary	v
Liste des tableaux	xi
Table des figures	xiii
Liste des sigles et des abréviations	xv
Dédicace	xvii
Remerciements	xix
Introduction	1
Références.....	3
Chapitre 1. Notions préliminaires	5
1.1. Les méthodes Monte Carlo par chaînes de Markov.....	5
1.1.1. Exemple d'utilisation de l'algorithme MH.....	9
1.1.2. L'algorithme à sauts réversibles.....	11
1.2. L'analyse en composantes principales.....	15
Références.....	17
Chapitre 2. Weak Convergence and Optimisation of the Reversible Jump Algorithm	19
Résumé.....	19
Abstract.....	19
2.1. Introduction.....	20
2.2. Sampling Context.....	21
2.3. Towards Optimal Implementation of the Reversible Jump.....	25

2.3.1. Weak Convergence Results	25
2.3.2. Optimisation	26
2.4. Practical Considerations	31
2.4.1. Optimal Implementation and Generalisation	31
2.4.2. Simulation Study	32
2.5. Conclusion	35
2.6. Proof of Theorem 2.1	35
2.6.1. Pseudo-Generator	36
2.6.2. Proof of the Convergence of the Finite-Dimensional Distributions	37
2.7. Results Used in the Proof of Theorem 2.1	39
2.8. Proofs of Propositions 2.1 and 2.2	42
References	45
Chapitre 3. Annexe du chapitre 2	47
3.1. Vérification du critère de réversibilité	47
3.2. Théorème 8.2 du chapitre 4 de Ethier et Kurtz (1986)	48
3.2.1. Vérification des conditions du théorème 8.2	49
3.3. Complément de la démonstration du théorème 2.1 : $\mathbb{E} \left[\left \varphi_{1,n}(t) - G_1 h(\mathbf{Z}_{1,2}^n(t)) \right \right] \rightarrow 0$	52
3.4. Complément de la démonstration du théorème 2.1 : vérification des conditions du résultat (c) du théorème 3.1	71
3.5. Corollaire 8.6 du chapitre 4 de Ethier et Kurtz (1986)	78
3.5.1. Vérification des conditions du corollaire 3.1	79
3.6. Autres résultats utilisés dans la démonstration du théorème 2.1	83
3.7. Démonstration du corollaire 2.1	84
Références	85
Chapitre 4. Bayesian Robustness to Outliers in Linear Regression	87
Résumé	87
Abstract	87

4.1. Introduction	88
4.2. Resolution of Conflicts in Linear Regression	90
4.2.1. Model	90
4.2.2. Log-Regularly Varying Distributions	92
4.2.3. Resolution of conflicts	92
4.3. Numerical Study	94
4.3.1. Analysis of the Relationship Between Two Stock Market Indexes	95
4.3.2. Convergence of the Posterior	98
4.3.3. Simulation Study	99
4.4. Conclusion	100
4.5. Proofs	101
4.5.1. Proof of Proposition 4.1	101
4.5.2. Proof of Theorem 4.1	103
References	113
Chapitre 5. An Efficient Bayesian Robust Principal Component Regression	115
Résumé	115
Abstract	115
5.1. Introduction	116
5.2. Principal Component Regression	118
5.2.1. First Situation: No Outliers	119
5.2.2. Second Situation: Possible Presence of Outliers	120
5.2.3. Reversible Jump Algorithm	121
5.2.4. Optimal Implementation	124
5.3. Simulation Study	126
5.4. Real Data Analysis: Prediction of Returns for the S&P 500	128
5.5. Conclusion and Further Remarks	132
References	132
5.6. Proofs	133
5.7. Appendix	136

Conclusion	137
Appendice : fonctions R	139

Liste des tableaux

1. I	Estimations de μ et σ^2 basées sur les chaînes présentées à la figure 1.1, en utilisant la moyenne et la variance échantillonales	11
4. I	Returns for day i of January 2011 for S&P 500 (y_i) and S&P/TSX (x_i) in percentages, $i = 1, \dots, 19$	95
4. II	Posterior medians and 95% HPD intervals under the three models based on the analysis of the data set presented in Table 4. I	97
4. III	Posterior medians and 95% HPD intervals under the three models based on the analysis of the data set presented in Table 4. I, but excluding observation 18	97
4. IV	Sum of the MSE of the estimators of β_1, β_2 and β_3 under the three scenarios and the three assumptions of f	100
4. V	MSE of the estimators of σ under the three scenarios and the three assumptions of f	100
5. I	Cumulative explained variations for the four first principal components	127
5. II	Posterior probabilities of models and parameter estimates under the super heavy-tailed distribution assumption based on the data set without outliers	129
5. III	Posterior probabilities of models and parameter estimates under the super heavy-tailed distribution assumption based on the data set with the outlier	129
5. IV	Cumulative explained variations for the first eighth to eleventh components	130
5. V	Posterior probabilities of models and parameter estimates under the super heavy-tailed distribution assumption based on the returns in %	131
5. VI	Posterior probabilities of models and parameter estimates under the super heavy-tailed distribution assumption based on the returns in %, without the second observation	132
5. VII	Names of the companies, funds, and indicators used in Section 5.4 with their ticker symbol (if available)	136


Table des figures

1.1	Trois chaînes de Markov découlant du RWM avec $q(x, \cdot) = \mathcal{N}(x, \ell^2)$ et $\pi(\cdot \mathbf{D}) = \mathcal{N}(0, 1)$, où (a) $\ell = 0.1$, (b) $\ell = 2.43$, et (c) $\ell = 10$	11
2.1	Sum of the ACFs as a function of s for different values of τ and $A = 2, 5, 25$, when $f = \mathcal{N}(\mu, \sigma^2)$ and ℓ is set to its optimal value (for each graph, the solid line represents the function arising from the optimal value of the constant τ)	30
2.2	(a) Integral of the sum of the ACFs as a function of τ , for $A = 2, 5, 25$ (the diamonds represent the optimal value for τ), (b) Optimal value for τ as a function of A ; for both graphs, $f = \mathcal{N}(\mu, \sigma^2)$ and ℓ is set to its optimal value	30
2.3	MADs around k^* , μ and σ , and the global measure, for every $\tau \in (0, 1)$, in the cases where $A = 2, 5, 25$ (the vertical lines represent the optimal values for τ , which are 0.415, 0.334 and 0.194, when $A = 2, 5, 25$, respectively)	34
4.1	January 2011 daily returns of S&P 500 and S&P/TSX with posterior medians under the nonrobust (blue solid line), partially robust (orange dashed line), and robust (green dotted line) models, respectively	95
4.2	Densities of the standard normal (blue solid line), Student with 10 degrees of freedom and a known scale parameter of 0.88 (orange dashed line), and log-Pareto-tailed standard normal with $\alpha = 1.96$ and $\varphi = 4.08$ (green dotted line)	96
4.3	Estimation of β and σ when $ y_{18} = -y_{18}$ increases from 0.5 to 6 under three different assumptions on f : standard normal density (blue solid line), Student density (orange dashed line) and log-Pareto-tailed standard normal density (green dotted line)	98
5.1	Sample variance for each of the principal components	127
5.2	Sample variance for each of the principal components	130
5.3	Errors of Model 4 under the super heavy-tailed assumption computed using posterior medians	131

LISTE DES SIGLES ET DES ABRÉVIATIONS

ACF	<i>Autocorrelation function</i>
ACP	Analyse en composantes principales
BIC	<i>Bayesian information criterion</i>
HPD	<i>Highest posterior density</i>
MAD	<i>Mean absolute deviation</i>
MAP	<i>Maximum a posteriori probability</i>
MCMC	<i>Markov chain Monte Carlo</i>
MH	Metropolis-Hastings
MSE	<i>Mean square error</i>
PCA	<i>Principal component analysis</i>
PCR	<i>Principal component regression</i>
PDF	<i>Probability density function</i>
PMF	<i>Probability mass function</i>
p.s.	Presque sûrement
RWM	<i>random walk Metropolis</i>
S&P 500	<i>Standard & Poor's 500</i>
S&P/TSX	<i>Standard & Poor's/Toronto Stock Exchange</i>

DÉDICACE

À Shazia, mon amour 

REMERCIEMENTS

Le doctorat aura été pour moi une expérience très agréable, et surtout, très enrichissante. Je ne peux passer sous silence le nom des nombreuses personnes qui ont contribué à cela. J'essayerai par contre d'être bref, parce que je dois avouer que j'ai fait beaucoup de belles rencontres. D'ailleurs, je m'excuse à l'avance auprès de ceux que j'aurai oubliés.

Il va s'en dire que mes premiers remerciements vont à Shazia, à qui je dédie cette thèse. Je te remercie d'avoir été à mes côtés tout au long de ce périple. Tu étais là quand ça allait, et aussi quand ça n'allait pas. Merci encore !

Je remercie ensuite ma directrice de recherche, Mylène Bédard, ainsi que mon codirecteur, Alain Desgagné. Je me rappellerai toujours du moment où Mylène m'avait dit : « tu voulais un projet théorique, et bien en voilà un ! » C'est en effet ce que je souhaitais, et c'est ce que j'ai eu. Ce fut un projet très intéressant, très *challengeant*, qui m'aura donné l'opportunité de m'ouvrir sur le monde des MCMC (un merveilleux monde, soit dit en passant). Ce fut un plaisir de travailler avec toi ; en particulier, j'ai apprécié ton efficacité dans les derniers mois où je « rushais » pour terminer la thèse. Merci ! Alain, le projet que tu m'as confié était aussi intéressant que celui à la maîtrise. En fait, je l'ai trouvé encore plus intéressant puisque j'avais, après la maîtrise et les années de scolarité au doctorat, la profondeur nécessaire pour m'y plonger complètement. Je suis on ne peut plus satisfait du résultat. Encore une fois, j'ai apprécié travailler avec toi. Ta rigueur m'a toujours épatée. Elle aura contribué à augmenter mon talent de chercheur. Merci !

Je remercie maintenant les professeurs dont j'ai reçu l'enseignement au doctorat. Je remercie en particulier Louis-Pierre qui m'a enseigné les probabilités. Il s'agit du cours où j'ai le plus appris. J'ai beaucoup apprécié le fait que tu étais très disponible. Je remercie aussi Christian pour les cours de consultation. Ça ne paraissait peut-être pas toujours, mais j'ai beaucoup apprécié. J'ai aussi beaucoup apprécié nos nombreuses discussions. Merci Christian pour ça et pour tes précieux conseils. Je remercie Lea Popovic pour son cours sur les *Stochastic Differential Equations* à Concordia. Je remercie aussi Claudia Gagné qui m'a donné l'opportunité d'enseigner à l'Université de Montréal.

Je remercie tout le personnel administratif à l'Université de Montréal. En particulier, je remercie Jean-François, Églantine, Marie-Claude, Anne-Marie, Émilie, Lise, Julie, Marielle et Michele. Je remercie aussi Guillaume pour le support informatique.

Je remercie tous mes « partners » à l'Université de Montréal : mon « partner » de doctorat, Aurélien, mes partners de bureau, Laurence, Younes et Serge, et mes partners dans les cours, Bianca, Marc-Olivier, Benoît, Alexandre, Geneviève, Joël, Guillaume, Victor et Ewen. Je remercie aussi les gens que j'ai eu le plaisir de côtoyer à l'université. En particulier, je remercie Michaël, Vincent, Justin, Louis-Xavier, Guillaume, Jean, Mouloud, Janie, Steve et David.

Finalement, je remercie ma famille proche : Pierre, Suzanne, Patrice, Rachel, Béatrice, Romain, Isabelle, Émilie, Dany et Bernard. Je remercie aussi les familles Poonja et Merali (et Phil) que j'apprécie beaucoup. Et juste avant de terminer, je remercie mes amis proches : Marco, Dave, Sarra, Jean Bertrand, Gabriel et Miko.

Un gros merci tout le monde !

INTRODUCTION

Dans plusieurs domaines, l'analyse statistique typique est composée des deux étapes primordiales suivantes : la sélection de modèle et l'estimation des paramètres. Lorsque l'analyse est effectuée sous le paradigme bayésien, ces étapes sont habituellement basées sur la distribution *a posteriori* conjointe des modèles et de leurs paramètres. Cette distribution a souvent une forme complexe, notamment lorsque les modèles sont sophistiqués, ce qui requiert l'aide de méthodes d'approximations numériques. Les méthodes de Monte Carlo par chaînes de Markov (MCMC, de l'anglais *Markov chain Monte Carlo*) représentent la solution la plus fréquemment employée. Ces méthodes indiquent comment générer un échantillon provenant de la distribution *a posteriori* qui nous intéresse (que l'on appelle distribution cible) par l'intermédiaire de chaînes de Markov. Ces chaînes ont une distribution limite qui correspond à la distribution cible, et c'est donc suite à l'atteinte de la stationnarité que les réalisations forment un échantillon ayant la qualité souhaitée. Cette propriété, combinée à certaines conditions de régularité, nous assurent que l'échantillon peut être utilisé pour approximer des probabilités et espérances (autrement dit, elles nous assurent que la loi des grands nombres s'applique), ce qui mène à la sélection de modèles et l'estimation des paramètres. La plupart des méthodes MCMC ne permettent toutefois pas d'effectuer ces deux tâches simultanément (voir [Green \(1995\)](#) pour plus de détails).

L'algorithme à sauts réversibles est une méthode MCMC très utile qui permet les changements de sous-espaces (où les sous-espaces de départ et d'arrivée peuvent être de dimensions différentes, représentant les espaces paramétriques de deux modèles différents). Ainsi, à partir d'un seul échantillon produit par cet algorithme, il est possible de sélectionner des modèles et d'estimer leurs paramètres. Il y a cependant un prix à payer : plusieurs fonctions doivent être spécifiées par l'utilisateur afin de l'implémenter. Ces fonctions servent, par exemple, à générer des candidats pour le prochain état de la chaîne de Markov lors d'une itération donnée. Le prix est d'autant plus élevé considérant qu'il n'existe pas de directive claire afin d'assister l'utilisateur dans cette étape de spécification et que des fonctions mal ajustées diminuent la qualité de l'échantillon produit (car la chaîne se déplace de façon inappropriée). Dans le [chapitre 2](#), la question de spécification des fonctions requises est traitée dans un contexte relativement simple pouvant correspondre à une analyse typique (sélection de modèles et estimation des paramètres) d'un groupe de modèles qui sont emboîtés et qui possèdent des caractéristiques précises. Cette spécificité du contexte permet de diminuer le

nombre de fonctions à ajuster, ce qui rend possible la démonstration de résultats théoriques. Ceux-ci donnent lieu à l'élaboration d'une stratégie pour l'étape de spécification. Les fonctions ajustées selon cette procédure engendrent un algorithme produisant des échantillons de qualité optimale. La procédure permet, par le fait même, une implémentation simple de l'algorithme à sauts réversibles.

Certaines démonstrations des résultats théoriques sont inspirées de celles de résultats similaires, et sont plutôt présentées en annexe. Leur nombre étant relativement élevé, un chapitre leur est dédié ; il s'agit du chapitre 3. Notez que l'algorithme à sauts réversible est considéré comme l'une des méthodes MCMC les plus complexes. C'est pourquoi une introduction aux méthodes MCMC est faite à la section 1.1. Elle permettra aux non spécialistes d'avoir un bon aperçu de ces méthodes avant l'étude de ce cas particulier.

Dans le chapitre 4, nous traitons de l'aspect estimation des paramètres dans un contexte plus spécifique : la modélisation basée sur la régression linéaire en présence possible de valeurs aberrantes. Le modèle de régression linéaire est habituellement utilisé en supposant la normalité des erreurs. Il est bien connu que cette hypothèse peut mener à des conclusions qui ne sont pas en ligne avec la majorité des observations lorsque le jeu de données contient des valeurs aberrantes. En fait, les conclusions ne sont souvent en ligne ni avec les valeurs aberrantes, ni avec les valeurs non aberrantes. Elles représentent plutôt un juste milieu, et sont donc peu utiles. Nous croyons que la façon appropriée de régler le problème est de limiter l'impact des valeurs aberrantes afin d'obtenir des conclusions cohérentes avec la majorité des observations (les valeurs non aberrantes). La stratégie pour y arriver est simple. L'hypothèse de normalité sur le terme d'erreur est remplacée par une hypothèse plus adaptée à la présence possible de valeurs aberrantes, et plus précisément, par une hypothèse de distribution à ailes extrêmement relevées. Cette stratégie a pour avantage de permettre aux utilisateurs d'effectuer les analyses selon la procédure habituelle (en estimant les paramètres par les médianes *a posteriori* par exemple). Des résultats théoriques de robustesse sont démontrés. Ceux-ci garantissent un impact limité (voire inexistant) des valeurs aberrantes.

Finalement, au chapitre 5, les stratégies présentées aux chapitres 2 et 4 sont combinées afin d'introduire une approche bayésienne de régression robuste sur composantes principales. Cette modélisation étant basée sur une régression linéaire où les variables explicatives sont des composantes principales, elle possède donc comme caractéristique que l'information véhiculée par une de ces variables est (linéairement) indépendante de celle apportée par n'importe quelle autre. Cette caractéristique, quoique attrayante, n'indique en rien que l'information apportée est pertinente. Il est ainsi proposé d'identifier les variables véhiculant de l'information importante par l'entremise d'une sélection de modèles. C'est à ce moment qu'entre en jeu la stratégie de spécification des fonctions requises à l'implémentation d'un algorithme à sauts réversibles. Il n'est cependant *a priori* pas assuré que les stratégies empruntées aux chapitres 2 et 4 soient efficaces. En effet, les résultats de robustesse sont présentés dans un contexte d'estimation de paramètres d'un modèle donné (plutôt que dans un contexte d'estimation de paramètres et de sélection de modèles). De plus, les modèles étudiés au chapitre 5 sont plus complexes que ceux analysés au chapitre 2. Dans le chapitre 5,

il est démontré empiriquement que ces stratégies s'avèrent efficaces dans ce contexte. Ainsi, ce travail permet, en plus de la présentation d'une nouvelle approche bayésienne de régression sur composantes principales, de faire un premier pas vers la généralisation des résultats présentés aux chapitres 2 et 4. Notez qu'une introduction à l'analyse en composantes principales est faite à la section 1.2 afin de faciliter la compréhension de la méthode présentée au chapitre 5.

RÉFÉRENCES

Green, P. J. 1995, «Reversible jump Markov chain Monte Carlo computation and Bayesian model determination», *Biometrika*, vol. 82, n° 4, p. 711–732.

Chapitre 1

NOTIONS PRÉLIMINAIRES

Dans ce chapitre, une introduction aux méthodes MCMC et à l'analyse en composantes principales est faite. La partie sur les MCMC se trouve à la section 1.1, puis celle sur l'analyse en composantes principales à la section 1.2.

1.1. LES MÉTHODES MONTE CARLO PAR CHAÎNES DE MARKOV

J'affectionne le paradigme bayésien pour plusieurs raisons. La principale est la suivante : les analyses statistiques effectuées sous celui-ci sont composées d'étapes formant un processus intuitif et les résultats engendrés sont faciles à interpréter. Le processus est le suivant : on évalue *a priori* la probabilité $\pi(A)$ d'un certain événement A ($\pi(\cdot)$ est appelée la distribution *a priori*), ensuite une expérience relative à cet événement se réalise et des données \mathbf{D} sont collectées, puis finalement on met à jour *a posteriori* notre évaluation de la probabilité de cet événement en calculant $\pi(A | \mathbf{D})$ ($\pi(\cdot | \mathbf{D})$ est appelée la distribution *a posteriori*). Le résultat $\pi(A | \mathbf{D})$ est donc la version « mise à jour » de la probabilité de l'événement A reflétant notre opinion (*a priori*) et les résultats de l'expérience. Par exemple, supposons que dans le domaine de l'assurance automobile on évalue à $\pi(A) = 10\%$ la probabilité qu'un assuré n'ayant aucune année d'expérience soit « dangereux ». Suite à quelques années d'expérience sans accident, on calculerait $\pi(A | \mathbf{D})$ qui refléterait cet historique, et ainsi, serait inférieur à 10%. Les probabilités *a priori* et le processus de mise à jour sont habituellement basés sur des modèles mathématiques probabilistes.

Malgré ces aspects rendant l'analyse statistique bayésienne attirante, ce n'est pourtant que relativement récemment qu'elle est devenue pratique courante, et ceci pour plusieurs raisons, dont la complexité des calculs requis dans la production de l'analyse. En effet, les calculs font souvent intervenir des intégrales par rapport à la distribution *a posteriori* qui est souvent connue qu'à une constante près et qui a habituellement une structure complexe, requérant l'aide de méthodes d'approximations numériques. C'est l'arrivée des méthodes MCMC et d'ordinateurs possédant des capacités calculatoires exceptionnelles qui a partiellement résolu le problème. L'idée derrière les méthodes MCMC est de construire une chaîne de Markov ayant comme distribution stationnaire la

distribution *a posteriori* par rapport à laquelle nous souhaitons calculer des intégrales (appelée distribution cible). Ainsi, la simulation d'une telle chaîne permet, suite à l'atteinte de la stationnarité (et moyennant la satisfaction de certaines conditions), d'obtenir un échantillon pouvant être utilisé pour calculer des approximations des intégrales qui nous intéressent. Cette idée fut introduite par [Metropolis *et al.* \(1953\)](#) dans un contexte physique pour le calcul d'intégrales sur différentes configurations de systèmes de particules par rapport à un poids fonction de l'énergie de chaque configuration. [Hastings \(1970\)](#) a généralisé et amélioré cette méthode en présentant une version pouvant être utile face à des problèmes numériques survenant en statistique. C'est ce qui a donné naissance à l'algorithme maintenant bien connu appelé Metropolis-Hastings (MH). [Hastings \(1970\)](#) a souligné le fait que la distribution cible devait être connue seulement à une constante près, et que la méthode nécessitait strictement qu'il soit possible d'évaluer la distribution cible en n'importe quel point de son domaine (plutôt que des intégrales par rapport à celle-ci). Ces avantages pointaient directement vers l'utilisation de cette méthode d'approximation numérique dans un contexte d'analyses statistiques bayésiennes.

L'algorithme MH permet la construction de la chaîne de Markov et la simulation de celle-ci. Nous donnons maintenant les étapes de l'algorithme en se concentrant sur la situation où la distribution cible a une densité strictement positive notée elle aussi $\pi(\mathbf{x} \mid \mathbf{D})$ avec $\mathbf{x} \in \mathbb{R}^d, d \in \{1, 2, \dots\}$. Cette hypothèse et cet abus de notation ont pour seul objectif de simplifier l'explication. Les étapes de l'algorithme se résument comme suit :

- Soit \mathbf{x} , l'état de la chaîne à l'itération $m \in \{1, 2, \dots\}$.
1. Un candidat $\mathbf{y} \in \mathbb{R}^d$ pour le prochain état de la chaîne (l'état à l'itération $m + 1$) est simulé d'une distribution instrumentale (sélectionnée par l'utilisateur) ayant une densité $q(\mathbf{x}, \cdot)$. Nous supposons que $q(\mathbf{x}, \cdot)$ est strictement positive, encore une fois pour simplifier l'explication.
 2. Le candidat est accepté selon une probabilité $\alpha(\mathbf{x}, \mathbf{y})$ que nous définirons sous peu. Si le candidat est accepté, l'état de la chaîne à l'itération $m + 1$ est donné par \mathbf{y} , sinon la chaîne demeure au même état, c'est-à-dire que l'état de la chaîne à l'itération $m + 1$ est donné par \mathbf{x} .

Ce processus engendre bel et bien une chaîne de Markov. La fonction α joue un rôle central : c'est elle qui garantit que la distribution stationnaire est la distribution cible (nous expliquerons pourquoi un peu plus tard). Sa forme est donnée plus précisément par

$$\alpha(\mathbf{x}, \mathbf{y}) := \frac{s(\mathbf{x}, \mathbf{y})}{1 + \frac{\pi(\mathbf{x}|\mathbf{D}) q(\mathbf{x}, \mathbf{y})}{\pi(\mathbf{y}|\mathbf{D}) q(\mathbf{y}, \mathbf{x})}},$$

où s est une fonction symétrique telle que $0 \leq \alpha \leq 1$.

Une première question peut nous venir en tête : quelle fonction s devrions-nous choisir ? Cette question fut traitée rapidement après l'article de [Hastings \(1970\)](#) par l'un de ses étudiants, un dénommé P. H. Peskun. Dans son article, [Peskun \(1973\)](#) indique que le « meilleur » choix de

fonction s est donné par

$$s(\mathbf{x}, \mathbf{y}) := \begin{cases} 1 + \frac{\pi(\mathbf{x}|\mathbf{D}) q(\mathbf{x}, \mathbf{y})}{\pi(\mathbf{y}|\mathbf{D}) q(\mathbf{y}, \mathbf{x})} & \text{si } \frac{\pi(\mathbf{y}|\mathbf{D}) q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x}|\mathbf{D}) q(\mathbf{x}, \mathbf{y})} \geq 1, \\ 1 + \frac{\pi(\mathbf{y}|\mathbf{D}) q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x}|\mathbf{D}) q(\mathbf{x}, \mathbf{y})} & \text{si } \frac{\pi(\mathbf{y}|\mathbf{D}) q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x}|\mathbf{D}) q(\mathbf{x}, \mathbf{y})} < 1. \end{cases}$$

Il s'agit du meilleur choix au sens où la précision des approximations des intégrales qui nous intéressent est la meilleure lorsque la fonction s ci-haut est utilisée. Plus précisément, [Peskun \(1973\)](#) a démontré que la variance asymptotique (lorsque le nombre d'itérations tend vers l'infini) de l'estimateur empirique est minimale lorsque cette fonction s est utilisée. [Peskun \(1973\)](#) indique aussi que ce résultat est dû au fait que c'est cette fonction qui engendre le plus de mouvement dans la chaîne. Depuis ce temps, on associe bonne capacité d'exploration de l'espace d'états à bonne précision des approximations. Le résultat de [Peskun \(1973\)](#) est démontré dans un contexte d'espaces dénombrables, mais il a été généralisé aux espaces quelconques par [Tierney \(1998\)](#). Lorsque la fonction s est utilisée, la probabilité d'acceptation est donnée par

$$\alpha(\mathbf{x}, \mathbf{y}) = 1 \wedge \frac{\pi(\mathbf{y} | \mathbf{D})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x} | \mathbf{D})q(\mathbf{x}, \mathbf{y})}, \quad (1.1)$$

où « \wedge » représente le minimum entre deux quantités. En fait, l'algorithme MH est plutôt connu comme étant le processus décrit plus haut, mais où la probabilité d'acceptation est donnée par l'expression (1.1).

Il est à noter que lorsque la fonction q est symétrique, la probabilité d'acceptation est donnée par $\alpha(\mathbf{x}, \mathbf{y}) = 1 \wedge \pi(\mathbf{y} | \mathbf{D})/\pi(\mathbf{x} | \mathbf{D})$. En observant cette forme de probabilité d'acceptation, on remarque que les candidats \mathbf{y} se trouvant dans les régions à haute densité sont acceptés fréquemment, comparativement aux candidats se trouvant dans les régions à faible densité. Le design de la fonction q est donc un aspect primordial de l'implémentation de l'algorithme MH considérant le lien étroit entre ce design et la capacité d'exploration de l'espace d'états de la chaîne de Markov.

Considérons à partir de maintenant que la probabilité d'acceptation a la forme donnée en (1.1). La probabilité que la chaîne de Markov passe d'un état contenu dans l'ensemble A à un état contenu dans l'ensemble B est donnée par

$$\int_A \pi(\mathbf{x} | \mathbf{D}) \left(\int_B q(\mathbf{x}, \mathbf{y}) \alpha(\mathbf{x}, \mathbf{y}) d\mathbf{y} + \mathbb{1}_B(\mathbf{x}) \int_{\mathbb{R}^d} q(\mathbf{x}, \mathbf{y}) (1 - \alpha(\mathbf{x}, \mathbf{y})) d\mathbf{y} \right) d\mathbf{x}.$$

La partie $\int_B q(\mathbf{x}, \mathbf{y}) \alpha(\mathbf{x}, \mathbf{y}) d\mathbf{y}$ représente la probabilité qu'un candidat appartenant à B soit proposé et accepté, et la partie $\mathbb{1}_B(\mathbf{x}) \int_{\mathbb{R}^d} q(\mathbf{x}, \mathbf{y}) (1 - \alpha(\mathbf{x}, \mathbf{y})) d\mathbf{y}$ représente la probabilité de rejeter le candidat, considérant que la chaîne se trouve à un état contenu dans B . Ce sont les deux façons d'arriver à un état dans B . La probabilité ci-haut peut être réécrite de la façon suivante :

$$\begin{aligned} & \int_A \int_B \pi(\mathbf{x} | \mathbf{D}) q(\mathbf{x}, \mathbf{y}) \alpha(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x} + \int_A \pi(\mathbf{x} | \mathbf{D}) \mathbb{1}_B(\mathbf{x}) \left(\int_{\mathbb{R}^d} q(\mathbf{x}, \mathbf{y}) (1 - \alpha(\mathbf{x}, \mathbf{y})) d\mathbf{y} \right) d\mathbf{x} \\ &= \int_A \int_B \pi(\mathbf{x} | \mathbf{D}) q(\mathbf{x}, \mathbf{y}) \wedge \pi(\mathbf{y} | \mathbf{D}) q(\mathbf{y}, \mathbf{x}) d\mathbf{y} d\mathbf{x} + \int_{A \cap B} \pi(\mathbf{x} | \mathbf{D}) \left(\int_{\mathbb{R}^d} q(\mathbf{x}, \mathbf{y}) (1 - \alpha(\mathbf{x}, \mathbf{y})) d\mathbf{y} \right) d\mathbf{x} \end{aligned}$$

$$= \int_B \int_A \pi(\mathbf{y} \mid \mathbf{D}) q(\mathbf{y}, \mathbf{x}) \alpha(\mathbf{y}, \mathbf{x}) d\mathbf{x} d\mathbf{y} + \int_B \pi(\mathbf{x} \mid \mathbf{D}) \mathbb{1}_A(\mathbf{x}) \left(\int_{\mathbb{R}^d} q(\mathbf{x}, \mathbf{y}) (1 - \alpha(\mathbf{x}, \mathbf{y})) d\mathbf{y} \right) d\mathbf{x},$$

où le théorème de Fubini a été utilisé à la dernière ligne. Il s'agit de la probabilité que la chaîne de Markov passe d'un état contenu dans l'ensemble B à un état contenu dans l'ensemble A . Il y a donc égalité entre la probabilité que la chaîne de Markov passe d'un état contenu dans l'ensemble A à un état contenu dans l'ensemble B et la probabilité que la chaîne de Markov passe d'un état contenu dans l'ensemble B à un état contenu dans l'ensemble A . L'égalité est vérifiée pour tout A, B , ce qui implique que la chaîne est réversible et que $\pi(\cdot \mid \mathbf{D})$ est une distribution stationnaire. Si la chaîne est par ailleurs irréductible, alors $\pi(\cdot \mid \mathbf{D})$ est l'unique distribution stationnaire. Si de plus la chaîne est apériodique, alors la loi des grands nombres s'applique. C'est ce principe qui nous garantit qu'on peut approximer les intégrales qui nous intéressent par l'intermédiaire de la simulation de la chaîne de Markov.

Pour fins de clarté, nous définissons formellement le concept d'irréductibilité et d'apériodicité, et nous énonçons le résultat concernant la loi des grands nombres.

Définition 1.1 (Irréductibilité). *Une chaîne de Markov est dite ϕ -irréductible si ϕ est une mesure σ -finie non nulle telle que pour tout ensemble A avec $\phi(A) > 0$ et pour tout \mathbf{x} , il existe un entier positif n tel que $P^n(\mathbf{x}, A) > 0$, où $P^n(\mathbf{x}, A)$ est la probabilité d'atteindre un état contenu dans l'ensemble A en partant de \mathbf{x} en n itération.*

L'irréductibilité permet de garantir que la chaîne peut atteindre n'importe quel endroit de son espace d'état en un nombre fini d'itérations.

Définition 1.2 (Apériodicité). *Une chaîne de Markov ϕ -irréductible est apériodique s'il n'existe pas $d \geq 2$ ensembles disjoints E_1, \dots, E_d avec $\phi(E_i) > 0$ pour tout i , tels que $P(\mathbf{x}, E_{i+1}) = 1$ pour tout $\mathbf{x} \in E_i$, $1 \leq i \leq d-1$, et $P(\mathbf{x}, E_1) = 1$ pour tout $\mathbf{x} \in E_d$, où $P(\mathbf{x}, E) := P^1(\mathbf{x}, E)$ est la probabilité d'atteindre un état contenu dans l'ensemble E en partant de \mathbf{x} en une itération.*

L'apériodicité permet de garantir que la chaîne n'a pas de comportement cyclique.

Théorème 1.1 (Loi des grands nombres pour l'algorithme MH - Tierney (1994)). *Soit $\{\mathbf{X}(n), n \in \mathbb{N}\}$, une chaîne de Markov découlant de l'algorithme MH avec distribution stationnaire $\pi(\cdot \mid \mathbf{D})$ (la densité est aussi notée $\pi(\cdot \mid \mathbf{D})$ par abus de notation). Si la chaîne est $\pi(\cdot \mid \mathbf{D})$ -irréductible et apériodique, alors*

$$\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}(i)) \rightarrow \mathbb{E}[f(\mathbf{X})] \text{ avec probabilité 1 lorsque } n \rightarrow \infty,$$

pour toute fonction f telle que $\mathbb{E}[f(\mathbf{X})]$ existe, où cette espérance est calculée par rapport à $\pi(\cdot \mid \mathbf{D})$.

L'irréductibilité et l'apériodicité ne peuvent être vérifiées en toute généralité ; elles dépendent de la distribution instrumentale utilisée. Cependant, si le choix de cette distribution est le moins raisonnable, l'algorithme MH donne habituellement lieu à des chaînes irréductibles et apériodiques. Un exemple est maintenant présenté afin d'illustrer comment ces propriétés peuvent être vérifiées en pratique et comment fonctionne concrètement l'algorithme MH.

1.1.1. Exemple d'utilisation de l'algorithme MH

Cet exemple est simpliste puisque l'utilisation des méthodes MCMC n'y est en fait pas nécessaire, mais nous croyons qu'il sera révélateur pour le lecteur. La distribution *a posteriori* par rapport à laquelle nous souhaitons calculer des intégrales (la distribution cible) est une normale d'espérance μ et de variance σ^2 , que l'on note $\mathcal{N}(\mu, \sigma^2)$. Cette situation survient lorsque, par exemple, $\mathbf{D} := (x_1, \dots, x_n) \in \mathbb{R}^n$, où X_1, \dots, X_n sont $n \in \{1, 2, \dots\}$ variables aléatoires conditionnellement indépendantes ayant chacune une distribution $\mathcal{N}(\theta, \tau^2)$ sachant θ (τ est une constante positive connue), et que la distribution *a priori* de θ est une $\mathcal{N}(\theta_0, \tau_0^2)$ ($\theta_0 \in \mathbb{R}$ et $\tau_0 > 0$ sont des constantes connues). Dans ce cas, la distribution *a posteriori* de θ est une $\mathcal{N}(\mu, \sigma^2)$, où

$$\mu := \frac{\tau^2}{\tau^2 + \tau_0^2 n} \times \theta_0 + \frac{\tau_0^2 n}{\tau^2 + \tau_0^2 n} \times \frac{1}{n} \sum_{i=1}^n x_i,$$

$$\sigma^2 := \left(\frac{1}{\tau_0^2} + \frac{n}{\tau^2} \right)^{-1}.$$

Afin d'approximer les intégrales, on utilise l'algorithme MH avec $q(x, \cdot) = \mathcal{N}(x, \ell^2)$, où ℓ est une constante positive. Lorsque, comme dans ce cas, la distribution q est une normale centrée à l'état présent de la chaîne, l'algorithme MH se nomme l'algorithme RWM (pour *random walk Metropolis*). La distribution q est symétrique, et tel qu'expliqué à la section 1.1, la probabilité d'acceptation est donnée par $\alpha(\mathbf{x}, \mathbf{y}) = 1 \wedge \pi(\mathbf{y} | \mathbf{D}) / \pi(\mathbf{x} | \mathbf{D})$. Dans le présent contexte, nous avons donc

$$\alpha(x, y) = 1 \wedge \exp\left(-\frac{1}{2\sigma^2} \left[(y - \mu)^2 - (x - \mu)^2 \right]\right).$$

On sait que la chaîne engendrée par l'algorithme RWM est réversible. On sait aussi que $\int_A \pi(x | \mathbf{D}) dx > 0$ pour tout $A \subset \mathbb{R}$ non vide, et que la chaîne peut atteindre théoriquement n'importe quel ensemble A en partant de n'importe quel état x , pour tout $x \in \mathbb{R}$, et ceci en une itération car $q(x, \cdot)$ est strictement positive. Ainsi, la chaîne est $\pi(\cdot | \mathbf{D})$ -irréductible selon la définition 1.1 (notez qu'il y a un léger abus de notation ici). De façon plus générale, si $\pi(\cdot | \mathbf{D})$ est strictement positive sur son domaine et qu'on utilise une densité $q(\mathbf{x}, \cdot)$ strictement positive sur ce même domaine pour tout \mathbf{x} , alors la chaîne est irréductible.

Une chaîne de Markov est apériodique s'il y a une probabilité non nulle que la chaîne demeure au même état deux itérations de suite. Pour l'algorithme MH (et RWM), cela se traduit par une probabilité non nulle de rejeter un candidat \mathbf{y} . Plus formellement, considérons une chaîne $\pi(\cdot | \mathbf{D})$ -irréductible telle que l'ensemble $A := \{\mathbf{x} : 1 - \int_{\mathbb{R}^d} q(\mathbf{x}, \mathbf{y}) \alpha(\mathbf{x}, \mathbf{y}) d\mathbf{y} > 0\}$ est tel que $\int_A \pi(\mathbf{x} | \mathbf{D}) d\mathbf{x} > 0$, où A représente l'ensemble des \mathbf{x} tel qu'il y a une probabilité non nulle de rejeter \mathbf{y} en partant de \mathbf{x} . Alors, la chaîne atteindra l'ensemble A en un nombre fini d'itérations (parce qu'elle est irréductible), ce qui implique qu'il y a en effet une probabilité non nulle que la chaîne demeure au même état deux itérations de suite. Dans ce cas, nous sommes assurés que la chaîne est apériodique.

Cette propriété est donc habituellement vérifiée pour les chaînes découlant de l'algorithme MH puisqu'il y a généralement une probabilité non nulle de rejeter un candidat lorsque la chaîne se trouve à un état appartenant à un ensemble de probabilité non nulle.

Pour notre exemple, nous allons considérer pour la suite que $\mu = 0$ et $\sigma^2 = 1$, donc $\pi(\cdot | \mathbf{D}) = \mathcal{N}(0, 1)$. Si la chaîne est à l'état x et que $x \in [0, 0.5)$ (une région probable sous la distribution *a posteriori*) et $\ell = 1$, par exemple, alors la probabilité de rejeter un candidat y est donnée par

$$1 - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-x)^2\right) \left(1 \wedge \exp\left(-\frac{1}{2}[y^2 - x^2]\right)\right) dy.$$

L'intégrale est bornée inférieurement par 0 et

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-x)^2\right) \left(1 \wedge \exp\left(-\frac{1}{2}[y^2 - x^2]\right)\right) &\leq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-x)^2\right) \exp\left(-\frac{1}{2}[y^2 - x^2]\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp(-y^2 + yx) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2(1/2)}\left(y - \frac{x}{2}\right)^2\right) \exp\left(\frac{x^2}{4}\right). \end{aligned}$$

Alors, la probabilité de rejeter un candidat est bornée supérieurement par 1 et inférieurement par

$$\begin{aligned} 1 - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-x)^2\right) \left(1 \wedge \exp\left(-\frac{1}{2\sigma^2}[(y-\mu)^2 - (x-\mu)^2]\right)\right) dy \\ \geq 1 - \sqrt{1/2} \exp\left(\frac{x^2}{4}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1/2)}} \exp\left(-\frac{1}{2(1/2)}\left(y - \frac{x}{2}\right)^2\right) dy \\ = 1 - \sqrt{1/2} \exp\left(\frac{x^2}{4}\right) \geq 1 - \sqrt{1/2} \exp\left(\frac{(1/2)^2}{4}\right) \approx 0.25 \end{aligned}$$

Ainsi, il y a une probabilité non nulle que la chaîne demeure au même état deux itérations de suite. Donc, la chaîne est apériodique. On a alors la confirmation que la loi des grands nombres s'applique, et ainsi, qu'on peut approximer les intégrales qui nous intéressent par l'intermédiaire de la simulation de la chaîne de Markov. L'algorithme RWM dans ce cas est donné par :

- Soit $x \in \mathbb{R}$, l'état de la chaîne à l'itération $m \in \{1, 2, \dots\}$.
- 1. Un candidat $y \in \mathbb{R}$ pour le prochain état de la chaîne (l'état à l'itération $m + 1$) est simulé de la distribution $\mathcal{N}(x, \ell^2)$.
- 2. Le candidat est accepté avec probabilité

$$\alpha(x, y) = 1 \wedge \exp\left(-\frac{1}{2\sigma^2}[(y-\mu)^2 - (x-\mu)^2]\right) = 1 \wedge \exp\left(-\frac{1}{2}[y^2 - x^2]\right). \quad (1.2)$$

Si le candidat est accepté, l'état de la chaîne à l'itération $m + 1$ est donné par y , sinon la chaîne demeure au même état, c'est-à-dire que l'état de la chaîne à l'itération $m + 1$ est donné par x .

Tel qu’expliqué à la section 1.1, le design de la fonction q est très important car il influence la qualité des candidats et par conséquent la précision des approximations obtenues suite à la simulation de la chaîne de Markov. Dans cet exemple, le design de la fonction q se traduit par un choix de valeur pour la constante ℓ . Une petite valeur de ℓ engendrera une chaîne dont les déplacements sont modestes (mais où les candidats sont souvent acceptés, voir la probabilité d’acceptation (1.2)), tandis qu’une grande valeur de ℓ engendrera une chaîne dont les déplacements sont rares (car les candidats sont fréquemment rejetés). Ces deux comportements sont indésirables. Sherlock et Roberts (2009) indiquent que dans le présent exemple, la valeur optimale pour ℓ est 2.43. La recherche présentée dans ma thèse (surtout au chapitre 2) s’inscrit de cet axe de design optimal des méthodes MCMC.

L’impact du choix de valeur pour la constante ℓ est illustré à la figure 1.1 et au tableau 1. I.

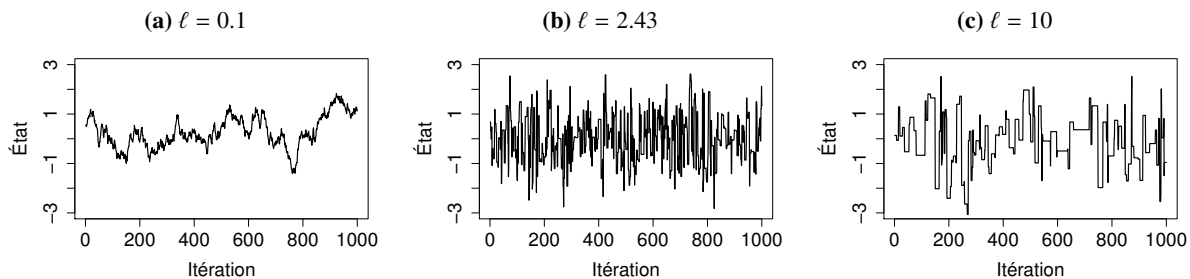


FIGURE 1.1. Trois chaînes de Markov découlant du RWM avec $q(x, \cdot) = \mathcal{N}(x, \ell^2)$ et $\pi(\cdot | \mathbf{D}) = \mathcal{N}(0, 1)$, où (a) $\ell = 0.1$, (b) $\ell = 2.43$, et (c) $\ell = 10$

ℓ	Estimations pour	
	$\mu = 0$	$\sigma^2 = 1$
0.1	0.27	0.41
2.43	-0.02	0.98
10	-0.09	1.13

Tableau 1. I. Estimations de μ et σ^2 basées sur les chaînes présentées à la figure 1.1, en utilisant la moyenne et la variance échantillonnales

1.1.2. L’algorithme à sauts réversibles

L’algorithme MH, quoique polyvalent, a ses propres limites. En particulier, il ne permet pas de générer d’échantillons provenant de distributions cibles dont la dimension varie, comme les distributions *a posteriori* conjointes de différents modèles et de leurs paramètres. Ceci implique que la sélection de modèles, effectuée par l’intermédiaire du calcul des probabilités *a posteriori* des différents modèles, ne peut être faite directement en utilisant cet algorithme. Green (1995) est venu corriger la situation en introduisant une nouvelle méthode MCMC : l’algorithme à sauts

réversibles. Cette méthode a la particularité d'engendrer des chaînes de Markov visitant des sous-espaces de différentes dimensions, ceux-ci représentant les espaces paramétriques des différents modèles.

L'algorithme MH permet d'approximer des intégrales par rapport à la distribution *a posteriori* $\pi(\mathbf{x}^k | k, \mathbf{D})$ où $k \in \{1, 2, \dots\}$ représente l'indicateur de modèle ($k = 1$ indique que le modèle 1 est considéré par exemple), $\mathbf{x}^k \in \mathbb{R}^{d^k}$ représente le vecteur des paramètres du modèle k , et d^k représente le nombre de paramètres du modèle k . Tout au long de ce document, lorsque plusieurs modèles sont évalués ou comparés, l'exposant indique lequel de ceux-ci est considéré. Par exemple, si l'on souhaite comparer différents modèles de régression, le modèle 1 pourrait représenter le modèle avec la constante seulement. Alors, $k = 1$ indique qu'on modélise la variable dépendante avec le modèle 1, et \mathbf{x}^1 est le vecteur contenant le paramètre d'échelle du terme d'erreur de ce modèle et la constante, avec $d^1 = 2$. Donc autrement dit, pour un modèle donné, l'algorithme MH permet d'approximer des intégrales ayant la forme suivante :

$$\int_{\mathbb{R}^{d^k}} f(\mathbf{x}^k) \pi(\mathbf{x}^k | k, \mathbf{D}) d\mathbf{x}^k,$$

où la densité a aussi été notée $\pi(\mathbf{x}^k | k, \mathbf{D})$, et en supposant que cette intégrale soit finie et que \mathbf{X}^k soit une variable aléatoire continue. Pour cette section, nous allons encore une fois faire l'hypothèse que \mathbf{X}^k est une variable aléatoire continue, et pour simplifier les explications, nous allons supposer que la $\pi(\cdot | k, \mathbf{D})$ est une densité de probabilité strictement positive sur \mathbb{R}^{d^k} , et ceci pour tout k . Ces hypothèses et l'abus de notation ont pour seul objectif de simplifier l'explication.

L'objectif est maintenant d'approximer des intégrales par rapport à la distribution *a posteriori* $\pi(k, \mathbf{x}^k | \mathbf{D})$ de façon efficace, ce que l'algorithme à sauts réversibles nous permet d'accomplir. Revenons à l'idée de base des méthodes MCMC un instant : construire une chaîne de Markov ayant comme distribution stationnaire la distribution *a posteriori* par rapport à laquelle nous souhaitons calculer des intégrales. L'idée derrière l'algorithme à sauts réversibles est de construire une chaîne de Markov ayant comme distribution stationnaire $\pi(k, \mathbf{x}^k | \mathbf{D})$, plutôt que $\pi(\mathbf{x}^k | k, \mathbf{D})$ comme c'était le cas avec l'algorithme MH. Ainsi, la simulation d'une chaîne engendrée par l'algorithme à sauts réversibles permet d'atteindre notre objectif, soit d'approximer des intégrales par rapport à $\pi(k, \mathbf{x}^k | \mathbf{D})$, ce qui mène à la sélection de modèles et l'estimation des paramètres. Les étapes de l'algorithme se résument comme suit :

- Soit $(k, \mathbf{x}^k)(m)$, l'état de la chaîne à l'itération $m \in \{1, 2, \dots\}$.
1. Une valeur k' est simulé de la distribution $g(k(m), \cdot)$, où $g(k(m), \cdot)$ est une fonction de probabilité. Ceci signifie qu'un déplacement de type $k(m) \mapsto k'$ sera tenté. Le cas $k' = k(m)$ implique qu'il n'y a pas de changement de modèle et qu'un candidat pour $\mathbf{x}^{k'}(m+1)$ sera proposé, ce qui correspond à une itération de l'algorithme MH habituel. On suppose que si $g(k(m), k') > 0$, alors $g(k', k(m)) > 0$; c'est-à-dire que s'il y a une probabilité non nulle

qu'un déplacement $k(m) \mapsto k'$ soit proposé, alors il y a une probabilité non nulle qu'un déplacement $k' \mapsto k(m)$ soit proposé.

2. Un vecteur $\mathbf{u}_{k(m) \mapsto k'}$ est simulé d'une densité de probabilité strictement positive $q_{k(m) \mapsto k'}$.
3. Une fonction $T_{k(m) \mapsto k'}$ est appliquée à $(\mathbf{x}^k(m), \mathbf{u}_{k(m) \mapsto k'})$, et $T_{k(m) \mapsto k'}(\mathbf{x}^k(m), \mathbf{u}_{k(m) \mapsto k'}) := (\mathbf{y}^{k'}, \mathbf{u}_{k' \mapsto k(m)})$. Le vecteur $\mathbf{y}^{k'}$ représente le candidat pour les paramètres du modèle k' , et $\mathbf{u}_{k' \mapsto k(m)}$ représente les variables qui permettent, lorsque combinées avec $\mathbf{y}^{k'}$ et suite à l'application de $T_{k(m) \mapsto k'}^{-1}$, de retrouver $(\mathbf{x}^{k'}(m), \mathbf{u}_{k(m) \mapsto k'})$. On suppose ainsi que $T_{k \mapsto k'}$ est un difféomorphisme pour tout k, k' . Donc, cette fonction est bijective et telle que les composantes de $T_{k \mapsto k'}$ et $T_{k' \mapsto k} := T_{k \mapsto k'}^{-1}$ admettent des dérivées partielles de premier ordre continues. Ainsi, les espaces de départ et d'arrivée de $T_{k(m) \mapsto k'}$ sont isomorphes et ceci implique que $(\mathbf{x}^k(m), \mathbf{u}_{k(m) \mapsto k'})$ et $(\mathbf{y}^{k'}, \mathbf{u}_{k' \mapsto k(m)})$ sont de mêmes tailles. Les variables $\mathbf{u}_{k' \mapsto k(m)}$ ont une utilité théorique plutôt que pratique.
4. Le déplacement est accepté, c'est-à-dire $(k, \mathbf{x}^k)(m+1) = (k', \mathbf{y}^{k'})$, avec probabilité

$$\alpha((k, \mathbf{x}^k)(m), (k', \mathbf{y}^{k'})) := 1 \wedge \frac{\pi(k', \mathbf{y}^{k'} \mid \mathbf{D})g(k', k(m))q_{k' \mapsto k(m)}(\mathbf{u}_{k' \mapsto k(m)})|J_{T_{k(m) \mapsto k'}}|}{\pi((k, \mathbf{x}^k)(m) \mid \mathbf{D})g(k(m), k')q_{k(m) \mapsto k'}(\mathbf{u}_{k(m) \mapsto k'})},$$

où $|J_{T_{k(m) \mapsto k'}}|$ est le déterminant du jacobien de la fonction $T_{k(m) \mapsto k'}$. Si le déplacement est rejeté, la chaîne demeure au même état, c'est-à-dire $(k, \mathbf{x}^k)(m+1) = (k, \mathbf{x}^k)(m)$.

Remarque 1.1. *En utilisant l'algorithme à sauts réversibles, on peut obtenir une chaîne de Markov qui ne se déplace que selon des mouvements de la forme $k(m) \mapsto k(m)$ en posant $g(k(m), k(m)) = 1$. Dans ce cas $T_{k(m) \mapsto k(m)}(\mathbf{x}^k(m), \mathbf{u}_{k(m) \mapsto k(m)}) := (\mathbf{u}_{k(m) \mapsto k(m)}, \mathbf{x}^k(m)) =: (\mathbf{y}^{k'}, \mathbf{u}_{k' \mapsto k(m)})$, et donc $|J_{T_{k(m) \mapsto k(m)}}| = 1$. L'algorithme MH est alors un cas particulier de l'algorithme à sauts réversibles.*

Comme l'algorithme MH, l'algorithme à sauts réversibles engendre bel et bien une chaîne de Markov. Il est cependant beaucoup plus complexe que l'algorithme MH. C'est le prix à payer pour que la chaîne de Markov visite l'espace des modèles et de leurs paramètres, plutôt que l'espace des paramètres pour un modèle donné. Cette augmentation du niveau de complexité se traduit essentiellement par l'augmentation du nombre de fonctions à spécifier pour implémenter l'algorithme. Considérant que seulement un type de déplacement se produit sous l'algorithme MH, soient les déplacements du type $k(m) \mapsto k(m)$, seulement une fonction devait être spécifiée, soit $q_{k(m) \mapsto k(m)}$ qui était notée q à la section 1.1. Pour l'algorithme à sauts réversibles, on doit spécifier la fonction $g(k, \cdot)$ pour tout k , la fonction $q_{k \mapsto k'}$ pour tout k, k' , et la fonction $T_{k \mapsto k'}$ pour tout k, k' . Afin d'effectuer cette tâche de façon efficace, il est préférable de miser sur la simplicité en tirant avantage de la structure des modèles et de leurs paramètres. Par exemple, considérons la situation où k représente le nombre de paramètres. Nous pourrions déterminer que $g(k, i) = 1/3$ pour $i \in \{k-1, k, k+1\}$ peu importe la valeur de k . Ainsi, nous proposerions de passer d'un modèle à k paramètres au modèle à $k-1$ ou $k+1$ paramètres avec probabilité $1/3$. Un contexte simple est étudié au chapitre 2 et un design optimal des fonctions requises pour l'implémentation de l'algorithme à sauts réversibles dans cette situation est proposé. L'algorithme à sauts réversibles est par ailleurs utilisé dans une

situation concrète de modélisation de données à l'aide de régressions sur composantes principales au chapitre 5.

Nous allons maintenant vérifier que la chaîne de Markov engendrée par l'algorithme à sauts réversibles satisfait bien la propriété de réversibilité par rapport à la distribution cible. On doit donc vérifier que la probabilité que la chaîne de Markov passe d'un état contenu dans l'ensemble A à un état contenu dans l'ensemble B est égale à la probabilité que la chaîne de Markov passe d'un état contenu dans l'ensemble B à un état contenu dans l'ensemble A . Considérant que la chaîne ne peut pas visiter simultanément deux modèles, on vérifie que la probabilité de passer du modèle k , où $\mathbf{x}^k \in A$, au modèle k' , où $\mathbf{y}^{k'} \in B$, est égale à la probabilité de passer du modèle k' , où $\mathbf{y}^{k'} \in B$, au modèle k , où $\mathbf{x}^k \in A$, pour tout k, k' et $A \subset \mathbb{R}^{d^k}$, $B \subset \mathbb{R}^{d^{k'}}$.

Premièrement, si k' a une probabilité nulle, le déplacement sera accepté avec probabilité 0. De plus, la chaîne de Markov a une probabilité nulle de se trouver à k' . On peut alors se limiter au cas où k et k' ont tous les deux une probabilité non nulle. La probabilité de passer du modèle k , où $\mathbf{x}^k \in A$, au modèle k' , où $\mathbf{y}^{k'} \in B$, est donnée par

$$\begin{aligned} & \int_A \pi(k, \mathbf{x}^k | \mathbf{D}) g(k, k') \int_{\{(\mathbf{x}^k, \mathbf{u}_{k \rightarrow k'}) : \mathbf{y}^{k'} \in B\}} q_{k \rightarrow k'}(\mathbf{u}_{k \rightarrow k'}) \alpha((k, \mathbf{x}^k), (k', \mathbf{y}^{k'})) d\mathbf{u}_{k \rightarrow k'} d\mathbf{x}^k \\ & + \int_A \pi(k, \mathbf{x}^k | \mathbf{D}) g(k, k') \mathbb{1}_{\{k'\} \times B}(k, \mathbf{x}^k) \left(\int_{\mathbb{R}^k} q_{k \rightarrow k'}(\mathbf{u}_{k \rightarrow k'}) (1 - \alpha((k, \mathbf{x}^k), (k', \mathbf{y}^{k'}))) d\mathbf{u}_{k \rightarrow k'} \right) d\mathbf{x}^k. \end{aligned} \quad (1.3)$$

Il est à noter que pour \mathbf{x}^k et $\mathbf{u}_{k \rightarrow k'}$ donnés, on connaît la valeur de $\mathbf{y}^{k'}$ et $\mathbf{u}_{k' \rightarrow k}$ par l'intermédiaire de la fonction $T_{k \rightarrow k'}$. Le premier terme correspond à un déplacement de la forme $k \mapsto k'$. Ceci implique qu'on simule $\mathbf{u}_{k \rightarrow k'}$, on calcule le candidat $\mathbf{y}^{k'}$ en utilisant $T_{k \rightarrow k'}$, et on l'accepte. Le deuxième terme correspond au cas où on refuse le candidat, mais où $k' = k$ et \mathbf{x}^k se trouvait déjà dans B . Ce terme est donc différent de 0 seulement lorsque $k = k'$. Dans ce qui suit, on traite les deux termes séparément, en commençant par le premier.

Le premier terme peut être réécrit de la façon suivante :

$$\begin{aligned} & \int_{\{(\mathbf{x}^k, \mathbf{u}_{k \rightarrow k'}) : \mathbf{x}^k \in A, \mathbf{y}^{k'} \in B\}} \pi(k, \mathbf{x}^k | \mathbf{D}) g(k, k') q_{k \rightarrow k'}(\mathbf{u}_{k \rightarrow k'}) \alpha((k, \mathbf{x}^k), (k', \mathbf{y}^{k'})) d(\mathbf{x}^k, \mathbf{u}_{k \rightarrow k'}) \\ & = \int_{\{(\mathbf{x}^k, \mathbf{u}_{k \rightarrow k'}) : \mathbf{x}^k \in A, \mathbf{y}^{k'} \in B\}} \pi(k', \mathbf{y}^{k'} | \mathbf{D}) g(k', k) q_{k' \rightarrow k}(\mathbf{u}_{k' \rightarrow k}) |J_{T_{k \rightarrow k'}}| \alpha((k', \mathbf{y}^{k'}), (k, \mathbf{x}^k)) d(\mathbf{x}^k, \mathbf{u}_{k \rightarrow k'}) \\ & = \int_{\{(\mathbf{y}^{k'}, \mathbf{u}_{k' \rightarrow k}) : \mathbf{y}^{k'} \in B, \mathbf{x}^k \in A\}} \pi(k', \mathbf{y}^{k'} | \mathbf{D}) g(k', k) q_{k' \rightarrow k}(\mathbf{u}_{k' \rightarrow k}) \alpha((k', \mathbf{y}^{k'}), (k, \mathbf{x}^k)) d(\mathbf{y}^{k'}, \mathbf{u}_{k' \rightarrow k}), \end{aligned}$$

en utilisant la définition de la fonction α à la première égalité, puis en effectuant un changement de variables de la forme $(\mathbf{y}^{k'}, \mathbf{u}_{k' \rightarrow k}) = T_{k \rightarrow k'}(\mathbf{x}^k, \mathbf{u}_{k \rightarrow k'})$ à la deuxième égalité. La dernière expression correspond à la probabilité, en partant de k' avec $\mathbf{y}^{k'} \in B$, de proposer et d'accepter un déplacement vers k , où $\mathbf{x}^k \in A$. Ceci permet de conclure que, lorsqu'on se limite à la situation où le candidat est accepté, la probabilité, en partant de k avec $\mathbf{x}^k \in A$, de se déplacer vers k' , où $\mathbf{y}^{k'} \in B$, est la même que celle du trajet inverse. Autrement dit, il ne reste qu'à traiter le cas où le candidat est rejeté, ce

qui correspond à analyser le deuxième terme dans (1.3). La seule façon que ce terme soit non nul est que $k = k'$ et \mathbf{x}^k appartienne à la fois à A et à B . Nous pouvons donc établir la réversibilité de la chaîne de Markov engendrée par l'algorithme à sauts réversible puisque

$$\begin{aligned} & \int_A \pi(k, \mathbf{x}^k | \mathbf{D}) g(k, k') \mathbb{1}_{\{k'\} \times B}(k, \mathbf{x}^k) \left(\int_{\mathbb{R}^k} q_{k \rightarrow k}(\mathbf{u}_{k \rightarrow k}) (1 - \alpha((k, \mathbf{x}^k), (k', \mathbf{y}^{k'}))) d\mathbf{u}_{k \rightarrow k} \right) d\mathbf{x}^k \\ &= \int_{A \cap B} \pi(k, \mathbf{x}^k | \mathbf{D}) g(k, k') \mathbb{1}_{\{k'\}}(k) \left(\int_{\mathbb{R}^k} q_{k \rightarrow k}(\mathbf{u}_{k \rightarrow k}) (1 - \alpha((k, \mathbf{x}^k), (k', \mathbf{y}^{k'}))) d\mathbf{u}_{k \rightarrow k} \right) d\mathbf{x}^k, \end{aligned}$$

qui est la probabilité, en partant de $k' = k$ avec $\mathbf{y}^{k'} \in A \cap B$, de rejeter le candidat.

Comme pour l'algorithme MH, si la chaîne est par ailleurs irréductible et apériodique, alors la loi des grands nombres s'applique. Tel qu'expliqué à la section 1.1, l'irréductibilité et l'apériodicité ne peuvent être vérifiées en toute généralité. Cependant, si le choix des fonctions nécessaires à l'implémentation de l'algorithme à sauts réversibles est le moins raisonnable, l'irréductibilité et l'apériodicité seront habituellement vérifiées en pratique, comme il pourra être constaté aux chapitres 2 et 5.

1.2. L'ANALYSE EN COMPOSANTES PRINCIPALES

L'analyse en composantes principales (ACP) est une technique qui permet de réduire la dimension d'un jeu de données tout en préservant sa structure au maximum. La notion de préservation maximale de la structure sera explicitée un peu plus tard dans cette section. Il est utile de réduire la dimension d'un jeu de données lorsque, par exemple, celle-ci fait en sorte qu'il est impossible d'estimer les paramètres d'un modèle. C'est le cas lorsque le nombre de variables explicatives est supérieur au nombre d'observations dans un contexte de régression linéaire. Il s'agit d'ailleurs d'un des problèmes étudiés au chapitre 5.

L'idée, introduite par [Pearson \(1901\)](#) puis développée par [Hotelling \(1933\)](#), est d'appliquer une transformation linéaire \mathbf{A}_q à la matrice contenant le jeu de données \mathbf{X} de manière à obtenir une nouvelle matrice de plus petite dimension dont les colonnes, nommées « composantes principales », sont orthogonales. Plus précisément, la multiplication matricielle $\mathbf{X}\mathbf{A}_q =: \mathbf{Z}_q$ est effectuée. La matrice \mathbf{X} est de dimension $n \times p$ où n représente habituellement la taille de l'échantillon et p le nombre de variables. La matrice \mathbf{A}_q a des colonnes orthonormales et est de dimension $p \times q$, où $q \leq r$ et r est le rang de la matrice \mathbf{X} . La matrice \mathbf{Z}_q est donc de dimension $n \times q$. Une première façon de préserver la structure de \mathbf{X} est de conserver au maximum la variabilité présente dans ce jeu de données. C'est l'idée qu'on eu ces auteurs. Les colonnes de \mathbf{A}_q sont donc choisies de manière à ce que les colonnes \mathbf{Z}_q soient les plus variables possible. La construction est plus précisément la suivante :

1. On débute par la première colonne, c'est-à-dire la maximisation de

$$\frac{1}{n-1} \sum_{i=1}^n (z_{i1} - \bar{z}_1)^2,$$

en considérant qu'il s'agit d'une fonction de \mathbf{a}_1 , soit la première colonne de \mathbf{A}_q , où z_{ij} est l'élément (i, j) de la matrice \mathbf{Z}_q et $\bar{z}_1 := (1/n) \sum_{i=1}^n z_{i1}$. Maintenant, supposons que les colonnes de \mathbf{X} sont centrées, c'est-à-dire que $(1/n) \sum_{i=1}^n x_{ij} = 0$ pour tout j . Alors, $\bar{z}_1 = 0$ puisque $z_{i1} = \sum_{j=1}^p x_{ij} a_{j1}$, où a_{ij} est l'élément (i, j) de la matrice \mathbf{A}_q . Donc, on peut de manière équivalente maximiser $\sum_{i=1}^n z_{i1}^2 = \mathbf{z}_1^T \mathbf{z}_1 = \mathbf{a}_1^T \mathbf{X}^T \mathbf{X} \mathbf{a}_1$, où \mathbf{z}_1 est la première colonne de \mathbf{Z}_q . On maximise alors cette quantité sous la contrainte : $\mathbf{a}_1^T \mathbf{a}_1 = 1$ (la matrice \mathbf{A}_q a des colonnes orthonormales). On utilise la technique du multiplicateur de Lagrange, ce qui nous mène à l'équation suivante suite à la dérivée par rapport à \mathbf{a}_1 :

$$(\mathbf{X}^T \mathbf{X} - \lambda \mathcal{I}_p) \mathbf{a}_1 = \mathbf{0},$$

où λ est le multiplicateur de Lagrange et \mathcal{I}_p est la matrice identité de dimension $p \times p$. Ainsi, λ est une valeur propre de $\mathbf{X}^T \mathbf{X}$ et \mathbf{a}_1 est le vecteur propre correspondant. Afin de déterminer à laquelle des valeurs propres on fait référence, on se rappelle que la quantité à maximiser est $\mathbf{a}_1^T \mathbf{X}^T \mathbf{X} \mathbf{a}_1 = \mathbf{a}_1^T \lambda \mathbf{a}_1 = \lambda$. Donc, λ (qui est à un facteur près la variance échantillonnale de la première composante) doit être maximale, ce qui implique que \mathbf{a}_1 est le vecteur propre correspondant à la valeur propre la plus élevée que l'on note λ_1 .

2. Maintenant que nous avons trouvé \mathbf{a}_1 , on trouve \mathbf{a}_2 de la même façon en ajoutant la contrainte que $\mathbf{z}_1^T \mathbf{z}_2 = 0$. On peut démontrer que la solution est similaire, c'est-à-dire que \mathbf{a}_2 est le vecteur propre de $\mathbf{X}^T \mathbf{X}$ correspondant à la valeur propre la plus élevée, mais en considérant que $\mathbf{z}_2 \neq \mathbf{z}_1$. Donc, en supposant qu'il n'y a pas répétition dans les valeurs propres, \mathbf{a}_2 est le vecteur propre correspondant à la deuxième valeur propre la plus élevée, que l'on note λ_2 . La répétition de valeurs propres est un phénomène inhabituel et nous ne devrions donc pas trop se soucier de ceci en pratique (voir [Jolliffe \(1986\)](#), section 2.4 pour de plus amples explications).
3. On procède ainsi jusqu'à \mathbf{a}_q .

Il est généralement de bonne pratique de travailler avec un jeu de données dont les colonnes sont standardisées, c'est-à-dire que $(1/(n-1)) \sum_{i=1}^n x_{ij}^2 = 1$ pour tout j (rappel : $(1/n) \sum_{i=1}^n x_{ij} = 0$ pour tout j). Ainsi, les résultats de l'ACP ne sont pas influencés par l'échelle par rapport à laquelle les données sont mesurées. Notons que dans ce cas, la matrice $(1/(n-1)) \mathbf{X}^T \mathbf{X}$ est la matrice de corrélation. Cette matrice joue donc un rôle central dans l'ACP.

À ce point, on pourrait se demander si la construction décrite plus haut préserve la structure du jeu de données au maximum en considérant d'autres critères que la variabilité. C'est en fait le cas, ce qui représente une propriété importante de l'ACP. Afin de fournir un exemple d'un de ces critères, nous introduisons un résultat qui se nomme « décomposition en valeurs singulières ». Il indique qu'une matrice \mathbf{X} de dimension $n \times p$ peut s'écrire

$$\mathbf{X} = \mathbf{U} \mathbf{L} \mathbf{A}^T, \tag{1.4}$$

où

- (i) \mathbf{U} et \mathbf{A} sont des matrices de dimensions respectives $n \times r$ et $p \times r$, ayant des colonnes orthonormales et étant telles que $\mathbf{U}^T \mathbf{U} = \mathcal{I}_r$ et $\mathbf{A}^T \mathbf{A} = \mathcal{I}_r$,
- (ii) \mathbf{L} est une matrice diagonale de dimension $r \times r$,
- (iii) r est le rang de \mathbf{X} .

Considérons que \mathbf{X} est, comme plus haut, la matrice contenant le jeu de données, et que ses colonnes sont centrées. On peut démontrer que \mathbf{X} satisfait l'équation (1.4) lorsque

- (i) $\mathbf{A} := \mathbf{A}_r$, soit la matrice de dimension $p \times r$ ayant comme colonnes les vecteurs $\mathbf{a}_1, \dots, \mathbf{a}_r$ construits de la façon décrite précédemment,
- (ii) \mathbf{U} est la matrice dont les colonnes sont données par $\mathbf{u}_j = \lambda_j^{-1/2} \mathbf{X} \mathbf{a}_j$ où λ_j est la j -ième valeur propre de $\mathbf{X}^T \mathbf{X}$, pour $j = 1, \dots, r$,
- (iii) \mathbf{L} est la matrice diagonale de dimension $r \times r$ dont les éléments sur la diagonale sont donnés par $\lambda_j^{1/2}$.

Ce résultat implique que $\mathbf{X} \mathbf{A}_r = \mathbf{U} \mathbf{L} = \mathbf{Z}_r$, nous procurant une méthode alternative pour trouver les composantes principales. Par ailleurs, ce résultat indique que les éléments de la matrice \mathbf{X} peuvent s'écrire de la façon suivante :

$$x_{ij} = \sum_{k=1}^r u_{ik} \lambda_k^{1/2} a_{jk} = \sum_{k=1}^r z_{ik} a_{jk},$$

où u_{ik} est l'élément (i, k) de la matrice \mathbf{U} . En ne considérant que les $q < r$ premières composantes principales (ou les q premières colonnes de \mathbf{U} et \mathbf{L}), on obtient

$${}_q \hat{x}_{ij} := \sum_{k=1}^q z_{ik} a_{jk}.$$

Il est légitime de se demander si ${}_q \hat{x}_{ij}$ est une bonne approximation de x_{ij} . Il se trouve qu'il s'agit de la meilleure approximation obtenue à partir de matrices de rang q . En effet, la matrice de rang q ${}_q \mathbf{X}$ qui minimise $\|{}_q \mathbf{X} - \mathbf{X}\|$ est celle composée des ${}_q \hat{x}_{ij}$, où $\|\cdot\|$ est la norme euclidienne. Les transformations linéaires de \mathbf{X} en utilisant les valeurs et les vecteurs propres de $\mathbf{X}^T \mathbf{X}$ contiennent, en ce sens, beaucoup d'information sur la structure de \mathbf{X} .

Plus la valeur de q augmente, plus la structure de \mathbf{X} est préservée. La valeur de q est parfois fixée arbitrairement. Par exemple, l'utilisateur pourrait vouloir se restreindre à 2 composantes principales. Elle est plus souvent déterminée par le biais de critères plus objectifs. L'un de ceux-ci sera décrit au chapitre 5.

RÉFÉRENCES

- Green, P. J. 1995, «Reversible jump Markov chain Monte Carlo computation and Bayesian model determination», *Biometrika*, vol. 82, n° 4, p. 711–732.
- Hastings, W. K. 1970, «Monte Carlo sampling methods using Markov chains and their applications», *Biometrika*, vol. 57, n° 1, p. 97–109.

- Hotelling, H. 1933, «Analysis of a complex of statistical variables into principal components», *J. Educ. Psychol.*, vol. 24, p. 417–441, 498–520.
- Jolliffe, I. T. 1986, «Principal component analysis», Springer.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller et E. Teller. 1953, «Equation of state calculations by fast computing machines», *J. Chem. Phys.*, vol. 21, p. 1087.
- Pearson, K. 1901, «On lines and planes of closest fit to systems of points in space», *Phil. Mag.*, vol. 2, n° 6, p. 559–572.
- Peskun, P. 1973, «Optimum Monte-Carlo sampling using Markov chains», *Biometrika*, vol. 60, n° 3, p. 607–612.
- Sherlock, C. et G. O. Roberts. 2009, «Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets», *Bernoulli*, vol. 15, n° 3, p. 774–798.
- Tierney, L. 1994, «Markov chains for exploring posterior distributions», *Ann. Statist.*, p. 1701–1728.
- Tierney, L. 1998, «A note on Metropolis-Hastings kernels for general state spaces», *Ann. Appl. Probab.*, vol. 8, p. 1–9.

Chapitre 2

WEAK CONVERGENCE AND OPTIMISATION OF THE REVERSIBLE JUMP ALGORITHM

Dans ce chapitre, un article soumis pour publication conjointement écrit avec ma directrice, Mylène Bédard, et mon co-directeur, Alain Desgagné, est présenté.

RÉSUMÉ

L'algorithme à sauts réversibles est une méthode de Monte Carlo par chaînes de Markov introduite par [Green \(1995\)](#) qui permet de passer d'un sous-espace, ayant une certaine dimension, à un autre, ayant une dimension différente, et ainsi, de sélectionner des modèles. Bien que cette méthode est de plus en plus utilisée dans les secteurs clés de l'activité humaine (par exemple, en biologie et en finance), il demeure un défi de l'implémenter de façon efficace. Dans cet article, un contexte d'échantillonnage relativement simple est étudié afin d'obtenir des résultats théoriques qui mènent à l'optimisation de l'algorithme à sauts réversibles et à une implémentation simplifiée. Le résultat théorique le plus important est la convergence faible de la suite des processus stochastiques engendrés par l'algorithme. Il représente la contribution principale apportée par cet article puisqu'il s'agit, à notre connaissance, du premier résultat de convergence faible pour l'algorithme à sauts réversibles.

ABSTRACT

The reversible jump algorithm is a useful Markov chain Monte Carlo method introduced by [Green \(1995\)](#) that allows switches between subspaces of differing dimensionality, and therefore, model selection. Although this method is now increasingly used in key areas of human activity (e.g. finance and biology), it remains a challenge to practically and efficiently implement it. In this paper, we focus on a simple sampling context in order to obtain theoretical results that lead to optimisation of the reversible jump algorithm, and consequently, to easy implementation. The key result is the weak convergence of the sequence of stochastic processes engendered by the

algorithm. This represents the main contribution of this paper as this is, to our knowledge, the first weak convergence result for the reversible jump algorithm.

MSC 2010 subject classifications: Primary 65C05; secondary 62F15.

Keywords: Bayesian inference; Integrated autocorrelation time; Markov chain Monte Carlo methods; Model selection; Optimal implementation; Random walk Metropolis.

2.1. INTRODUCTION

Markov chain Monte Carlo (MCMC) methods are most commonly applied in Bayesian analysis of complex statistical models to compute estimates. They are also used to solve problems approached from a frequentist perspective, for instance in clustering (see [Kang \(2013\)](#)). In this paper, we however explain the different concepts by focusing on the primary use of these methods.

The principle of MCMC methods is to construct a Markov chain with an invariant measure that corresponds to the distribution from which we are interested in obtaining a sample (usually called the target distribution). The implementation of such samplers usually requires the specification of some functions. For instance, at each step of the Metropolis-Hastings (MH) algorithm ([Metropolis et al. \(1953\)](#) and [Hastings \(1970\)](#)), the most commonly used method, a candidate for the next state of the Markov chain is generated from a proposal distribution (which has to be specified) and accepted according to a probability function (which is provided by the method). In a Bayesian context, this means that at each step, an attempt to update the parameters is made using the proposal distribution. The specification of the required functions can be challenging for non-specialists (and even for specialists), which makes them doubt the quality of their outputs. Indeed, a poor design of these functions can lead to an inefficient algorithm, in the sense that the resulting Markov chain explores its state space slowly, thus producing an inadequate sample (see [Peskun \(1973\)](#) and [Tierney \(1998\)](#) for a detailed explanation).

[Roberts et al. \(1997\)](#) studied the MH algorithm in the situation where the proposal distribution is a normal centered around the current state of the chain (this algorithm is a random walk Metropolis (RWM)). In this case, the specification step consists in selecting the variance of the normal distribution. This task is however not trivial as small variances lead to tiny movements of the Markov chain, while large variances induce high rejection rates of candidates. In their paper, the authors prove the existence of an optimal variance for the random walk, assuming that the algorithm is used to sample from a distribution of n independent and identically distributed (i.i.d.) random variables. They also provide a simple strategy to determine this optimal variance, which leads to a straightforward implementation of the algorithm. A lot of research has been carried out to generalise this result to more elaborate target distributions (e.g. [Roberts and Rosenthal \(2001\)](#), [Neal and Roberts \(2006\)](#), [Bédard \(2007\)](#), [Bédard \(2008\)](#), [Beskos et al. \(2009\)](#), [Bédard et al. \(2012\)](#), [Mattingly et al. \(2012\)](#) and [Beskos et al. \(2013\)](#)).

A flaw of RWM algorithms (and MH algorithms in general) is that they do not allow switches between subspaces of differing dimensionality, and therefore, model selection. This gap was corrected by [Green \(1995\)](#) with the introduction of the reversible jump algorithm. This method has a tremendous potential because of its capability to deliver information on both the “good” model and its parameters, simultaneously. For instance, [Richardson and Green \(1997\)](#) used it to estimate the number of components and the parameters of mixtures. This advantage comes with a downside: many functions have to be specified in order to do the implementation. In this paper, we study the reversible jump algorithm in the same mindset as [Roberts et al. \(1997\)](#); we aim at providing guidelines to users and open new research directions towards an automatic reversible jump algorithm.

Existing research on the reversible jump algorithm has mainly focused on ways to facilitate subspace switchings (e.g. [Brooks et al. \(2003\)](#), [Hastie \(2005\)](#), [Al-Awadhi et al. \(2004\)](#) and [Karagiannis and Andrieu \(2013\)](#)), and therefore, the exploration of the entire state space. The main drawback of the proposed approaches is the difficulty to implement them. This justifies the need for practical guidelines that will promote accessibility of the reversible jump algorithm. As a first step towards automated implementation of this method, we focus on a simple sampling context in order to obtain theoretical results. The context is defined in [Section 2.2](#). The key result, which is the weak convergence of the sequence of stochastic processes engendered by the algorithm, is presented in [Section 2.3.1](#). In [Section 2.3.2](#), this result is used to propose an optimal design for the sampler. This is followed in [Section 2.4.1](#) by a detailed procedure to implement an efficient reversible jump algorithm. In that section, we also discuss extensions of our results to more elaborate target distributions which include, for instance, posterior distributions arising from a robust principal component regression. In [Section 2.4.2](#), we illustrate the impact of the design of the sampler via a simulation study. Finally, the conclusion is given in [Section 2.5](#). The proof of the weak convergence is substantial and can be found in [Section 2.6](#). Results used in this proof that also have substantial demonstrations are presented in [Section 2.7](#) to ease the reading. For the same reason, the proofs of propositions included in the text can be found in [Section 2.8](#).

2.2. SAMPLING CONTEXT

Let

$$\pi_n(k, \mathbf{x}^k) = p(k) \prod_{i=1}^{n+k} f(x_i^k)$$

be the joint posterior distribution of (K^n, \mathbf{X}^{K^n}) , where $K^n \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$ ($\lfloor \cdot \rfloor$ is the floor function), $\mathbf{X}^{K^n} := (X_1^{K^n}, \dots, X_{n+K^n}^{K^n}) \in \mathbb{R}^{n+K^n}$, $n \in \{7, 8, \dots\}$, f is a strictly positive one-dimensional probability density function (PDF) with respect to Lebesgue measure, and p is a probability mass function (PMF) such that $p(k) > 0$ for all $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$. We consider n to be an integer greater than or equal to 7 just to avoid technical complications in the proofs. The random variable K^n represents the model indicator ($K^n = 1$ implies that Model 1 is considered, for instance), and

\mathbf{X}^{K^n} is the parameter vector of Model K^n . Therefore, X_1^1, \dots, X_{n+1}^1 are the $n+1$ parameters of Model 1, X_1^2, \dots, X_{n+2}^2 are the $n+2$ parameters of Model 2, etc. To simplify the notation, we will denote $K := K^n$ and then $\mathbf{X}^K := \mathbf{X}^{K^n}$. Note that the random variables X_1^K, \dots, X_{n+K}^K are conditionally i.i.d. given K , and that the random variables X_i^j and X_i^l have the same distribution, for all $i \leq n+j$, when $j \leq l$.

The objective is to obtain a representative sample from the joint posterior distribution of (K, \mathbf{X}^K) through MCMC methods in order to estimate probabilities, expectations, or any other quantity we might be interested in. MCMC users look for a simple and efficient way to attain this goal and the purpose of this paper is to provide guidelines. The following reversible jump algorithm is applied to sample from π_n :

- Considering that the time- m state of the chain is $(K(m), \mathbf{X}^{K(m)}(m))$, $m \in \mathbb{N}$, the type of movement that will be attempted $J(m+1) \in \{1, 2, 3\}$ is generated from g , a PMF such that $g(j) > 0$ for $j \in \{1, 2, 3\}$.
- If $J(m+1) = 1$, an attempt to update the parameters of the current model is made using a random walk. More precisely, $\mathbf{Y}^{K(m)}(m+1) \sim \mathcal{N}(\mathbf{X}^{K(m)}(m), (\ell^2/(n+K(m)))\mathcal{I}_{n+K(m)})$ is generated, where $\mathbf{Y}^{K(m)}(m+1) := (Y_1^{K(m)}(m+1), \dots, Y_{n+K(m)}^{K(m)}(m+1))$, $\mathcal{I}_{n+K(m)}$ is the identity matrix of size $n+K(m)$ and ℓ is a positive constant. This candidate is accepted, i.e. $(K(m+1), \mathbf{X}^{K(m+1)}(m+1)) = (K(m), \mathbf{Y}^{K(m)}(m+1))$, with probability

$$1 \wedge \frac{\prod_{i=1}^{n+K(m)} f(Y_i^{K(m)}(m+1))}{\prod_{i=1}^{n+K(m)} f(X_i^{K(m)}(m))}. \quad (2.1)$$

- If $J(m+1) = 2$, an attempt to add a parameter to switch from Model $K(m)$ to Model $K(m)+1$ is made. More precisely, $U(m+1) \sim q$ is generated and this candidate is accepted, i.e. $(K(m+1), \mathbf{X}^{K(m+1)}(m+1)) = (K(m)+1, (\mathbf{X}^{K(m)}(m), U(m+1)))$, with probability

$$1 \wedge \frac{f(U(m+1))p(K(m)+1)g(3)}{q(U(m+1))p(K(m))g(2)}, \quad (2.2)$$

where q is a strictly positive PDF.

- If $J(m+1) = 3$, an attempt to withdraw the last parameter to switch from Model $K(m)$ to Model $K(m)-1$ is made, i.e. $(K(m+1), \mathbf{X}^{K(m+1)}(m+1)) = (K(m)-1, \mathbf{X}^{K(m)-}(m))$, and this is accepted with probability

$$1 \wedge \frac{q(X_{n+K(m)}^{K(m)}(m))p(K(m)-1)g(2)}{f(X_{n+K(m)}^{K(m)}(m))p(K(m))g(3)}, \quad (2.3)$$

where $\mathbf{X}^{K(m)-}(m)$ is the vector $\mathbf{X}^{K(m)}(m)$ without the last component (more precisely $\mathbf{X}^{K(m)-}(m) := (X_1^{K(m)}, \dots, X_{n+K(m)-1}^{K(m)})$).

- In case of rejection, the chain remains at the same state, i.e. $(K(m+1), \mathbf{X}^{K(m+1)}(m+1)) = (K(m), \mathbf{X}^{K(m)}(m))$.

Note that the resulting process $\{(K(m), \mathbf{X}^K(m)), m \in \mathbb{N}\}$ is a π_n -irreducible and aperiodic Markov chain. In addition, it is easily shown that this Markov chain satisfies the reversibility condition with respect to π_n (it is nevertheless explicitly verified in Section 3.1), and therefore, that it is ergodic, which guarantees that the Law of Large Numbers holds.

Regularity conditions imposed on the different functions are now described. They allow to obtain the theoretical results stated in Section 2.3.

First, we assume that the following smoothness conditions on the function f are satisfied: $f \in C^2(\mathbb{R})$ (the space of real-valued functions on \mathbb{R} with continuous second derivative), $(\log f(x))'$ is Lipschitz continuous and $\mathbb{E}[(\log f(X))'^4] < \infty$, where the expectation is computed with respect to f . This last condition can be replaced by $\mathbb{E}[(f''(X)/f(X))^2] < \infty$, which is slightly stronger. We also assume that there exists a constant $A^* \geq 1$ such that

$$0 < \frac{f}{q} \leq A^* \text{ and therefore } \frac{1}{A^*} \leq \frac{q}{f} < \infty.$$

This condition corresponds to that required for the rejection sampling method. It ensures that the tails of q are at least as heavy as those of f , and thus, that q induces a good exploration of the state space. A small value for the constant A^* means that q is similar to f , and therefore, that it is a good choice of proposal distribution. Note that, when we can directly sample from f , we can set $q = f$.

The distribution p also fulfills some conditions. We assume that the mode of this distribution is in the middle of the set $\{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$ and that this PMF is symmetric with respect to this mode. Two distinct cases thus have to be considered: when $\lfloor \sqrt{n} \log n \rfloor$ is even or odd. When $\lfloor \sqrt{n} \log n \rfloor$ is odd, the mode is $(\lfloor \sqrt{n} \log n \rfloor + 1)/2$ and we assume that

$$\begin{aligned} p(k+1) &= a_{k,n} p(k), k \in \{(\lfloor \sqrt{n} \log n \rfloor + 1)/2, \dots, \lfloor \sqrt{n} \log n \rfloor - 1\}, \\ p(k-1) &= a_{k-1,n} p(k), k \in \{2, \dots, (\lfloor \sqrt{n} \log n \rfloor + 1)/2\}, \end{aligned}$$

where $a_{k,n} := (1 - b_{k,n}/\sqrt{n})$ with

$$b_{k,n} := \left| \frac{k - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}} \right|.$$

Note that $a_{k,n}$ decreases with the distance between k and the mode. This distribution is symmetric with respect to $(\lfloor \sqrt{n} \log n \rfloor + 1)/2$ and is such that

$$p\left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} + k\right) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} - k\right) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2}\right) \prod_{i=1}^k \left(1 - \frac{i-1/2}{n}\right),$$

where $k \in \{1, \dots, (\lfloor \sqrt{n} \log n \rfloor - 1)/2\}$.

When $\lfloor \sqrt{n} \log n \rfloor$ is even, the distribution p is bimodal with modes at $\lfloor \sqrt{n} \log n \rfloor / 2$ and $\lfloor \sqrt{n} \log n \rfloor / 2 + 1$. Using the same definitions as above for $a_{k,n}$ and $b_{k,n}$, we assume that

$$p(k+1) = a_{k,n} p(k), k \in \{\lfloor \sqrt{n} \log n \rfloor / 2 + 1, \dots, \lfloor \sqrt{n} \log n \rfloor - 1\},$$

$$p(k-1) = a_{k-1,n}p(k), k \in \{2, \dots, \lfloor \sqrt{n} \log n \rfloor / 2\},$$

which implies that

$$p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} + 1 + k\right) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - k\right) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2}\right) \prod_{i=1}^k \left(1 - \frac{i}{n}\right), \quad (2.4)$$

where $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor / 2 - 1\}$.

The assumptions on p imply that its mode is in the middle of its domain, with probabilities that decrease with the distance to this mode at an exponential rate which is bounded below by $1/2$ (the ratios $p(k+1)/p(k)$ and $p(k-1)/p(k)$ are essentially bounded below by $1/2$). Consider for instance the case where $\lfloor \sqrt{n} \log n \rfloor = 5$ ($n = 7$). The shape of p is such that the “best” model has $n + (\lfloor \sqrt{n} \log n \rfloor + 1)/2 = 10$ parameters, because the mode of p is $K = 3$ and the models have $n + K$ parameters. The model with an additional parameter is less appropriate (so is the model with one less parameter), in the sense that $p(4)/p(3) = p(2)/p(3) = 0.93$. The more parameters we add (or withdraw), the less appropriate the models are. This structure becomes natural when reversible jump users rank the models by number of parameters and believe the posterior distribution of models reflects the existence of a balance between a good fit (which involves a lot of parameters), and simplicity and stability of models, a principle aligned with Occam’s razor. As an example in which this structure arises, consider a principal component regression where the model indicator K represents the number of components included in the model (i.e. Model $K = k$ is the model with the first k components). It is easy to imagine situations where it would be optimal to include some, but not all, components. This would result in a posterior distribution for the various models featuring a structure as described above (see, e.g., the real data analysis in Chapter 5). Note that such a PMF p leads to an efficient exploration of the entire state space, because if the ratios $p(k+1)/p(k)$ and $p(k-1)/p(k)$ were close to 0 for some values of k , there would be more rejected attempts of switches from Model k to Model $k+1$ or $k-1$ (see (2.2) and (2.3)).

The hypothesised mathematical structure of the PMF p allows to obtain theoretical results, as (see Proposition 2.1 in Section 2.3.1 for the formal statement)

$$\mathbb{P}\left(\frac{K - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \leq x\right) \rightarrow \Phi(x), \forall x \in \mathbb{R}, \text{ as } n \rightarrow \infty, \quad (2.5)$$

where $K \sim p$ and Φ is the cumulative distribution function of the standard normal. It can be proved that, for all values of k such that $b_{k,n}$ is well-defined, $b_{k,n} \leq \log(n)/2$ and, therefore, that $b_{k,n}/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, which implies that $a_{k,n} \rightarrow 1$ as $n \rightarrow \infty$. Moreover, for all n and for all values of k such that $b_{k,n}$ is well-defined, $0 < b_{k,n}/\sqrt{n} \leq 1/2$, which implies that $1/2 \leq a_{k,n} < 1$. Thus, when $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor - 1\}$, we have that $1/2 \leq p(k+1)/p(k) \leq 2$ and $p(k+1)/p(k) \rightarrow 1$ as $n \rightarrow \infty$, because this ratio is essentially equal to $a_{k,n}$ or $a_{k,n}^{-1}$. To summarise, the assumptions on p and the standardisation of K in (2.5) are such that the resulting random variable is continuous and

takes values on the real line, in the limit, which makes this convergence in distribution possible. Note that the assumptions on p indeed imply that $p(k) > 0$ for all $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$.

Finally, we define the function g as follows:

$$g(j) := \begin{cases} \tau & \text{if } j = 1, \\ (1 - \tau)A/(A + 1) & \text{if } j = 2, \\ (1 - \tau)/(A + 1) & \text{if } j = 3, \end{cases} \quad (2.6)$$

where $0 < \tau < 1$ is a constant and $A := 2A^*$. Considering this definition, the acceptance probability arising from the inclusion of an extra parameter (see (2.2)) becomes the minimum between 1 and $f(U)/q(U) \times 1/A \times p(K + 1)/p(K)$. By assumption, $2f/q \leq A$ and $p(K + 1)/p(K) \leq 2$; therefore, this acceptance probability is simply $f(U)/q(U) \times 1/A \times p(K + 1)/p(K)$, which is easier to handle mathematically than the minimum function expressed in (2.2). Furthermore, the acceptance probability arising from the withdrawal of the last parameter (see (2.3)) becomes the minimum between 1 and $q(X_{n+K}^K)/f(X_{n+K}^K) \times A \times p(K - 1)/p(K) \geq 1$, which means that this type of movement is automatically accepted (whenever it is possible to withdraw a parameter, i.e. when $K > 1$).

2.3. TOWARDS OPTIMAL IMPLEMENTATION OF THE REVERSIBLE JUMP

In order to implement the reversible jump algorithm described in Section 2.2, we have to specify the PDF q and values for the constants A , τ and ℓ . In Section 2.3.1, we present weak convergence results that are used in Section 2.3.2 to find asymptotically optimal (as $n \rightarrow \infty$) values for τ and ℓ . In Section 2.4.1, we provide guidelines to suitably design q , which implicitly allows to determine the constant A .

2.3.1. Weak Convergence Results

In order to study the asymptotic behaviour of the algorithm, we consider the following rescaled stochastic process:

$$\mathbf{Z}^n(t) := \left(\frac{K(\lfloor nt \rfloor) - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}}, \mathbf{X}^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor) \right), \quad (2.7)$$

where $t \geq 0$. The continuous-time stochastic process $\{\mathbf{Z}^n(t), t \geq 0\}$ is a sped up and modified version of $\{(K(m), \mathbf{X}^K(m)), m \in \mathbb{N}\}$ (see Section 2.2 for the definition of this process) that admits jumps. In a given iteration, the average distance travelled by the parameters

$$\mathbf{X}^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor) := (X_1^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor), \dots, X_{n+K(\lfloor nt \rfloor)}^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor))$$

decreases with n because the variance of the random walk is proportional to $1/(n + K(\lfloor nt \rfloor))$. In addition, the distance travelled by $\{K(\lfloor nt \rfloor)/\sqrt{n}, t \geq 0\}$, each time it moves, is $1/\sqrt{n}$. The decreasing size of the jumps, combined with the acceleration of $\{\mathbf{Z}^n(t), t \geq 0\}$, result in a continuous and non-trivial limiting process. As explained in Section 2.2, we subtract $\lfloor \sqrt{n} \log n \rfloor / (2\sqrt{n})$ from

$\{K(\lfloor nt \rfloor) / \sqrt{n}, t \geq 0\}$ in order to obtain a limiting process with components that take values on the real line. The asymptotic behaviour of the first component of $\{\mathbf{Z}^n(t), t \geq 0\}$, denoted $\{Z_1^n(t), t \geq 0\}$, is described in Proposition 2.1.

Proposition 2.1. *Consider the context described in Section 2.2, the stochastic process $\{\mathbf{Z}^n(t), t \geq 0\}$ defined in (2.7), and assume that $\mathbf{Z}^n(0) \sim \pi_n$. Then, as $n \rightarrow \infty$, $Z_1^n(t)$ converges in distribution towards a standard normal random variable, for all $t \geq 0$.*

PROOF. See Section 2.8. ■

The main result is now stated.

Theorem 2.1. *Consider the context described in Section 2.2, the stochastic process $\{\mathbf{Z}^n(t), t \geq 0\}$ defined in (2.7), and assume that $\mathbf{Z}^n(0) \sim \pi_n$. Then, as $n \rightarrow \infty$, the first two components of $\{\mathbf{Z}^n(t), t \geq 0\}$ converge weakly towards a bidimensional Langevin diffusion, i.e.*

$$\{\mathbf{Z}_{1,2}^n(t), t \geq 0\} := \{(Z_1^n(t), Z_2^n(t)), t \geq 0\} \Rightarrow \{\mathbf{Z}(t), t \geq 0\} \text{ as } n \rightarrow \infty,$$

where the process $\{\mathbf{Z}(t), t \geq 0\}$ is comprised of two independent components such that $Z_1(0) \sim \mathcal{N}(0, 1)$, $Z_2(0) \sim f$,

$$\begin{aligned} dZ_1(t) &= \sqrt{2(1-\tau)/(A+1)} dB_1(t) - (1-\tau)/(A+1) \times Z_1(t) dt, \\ dZ_2(t) &= \sqrt{2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)} dB_2(t) + \tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)(\log f(Z_2(t)))' dt, \end{aligned}$$

with $\{B_1(t), t \geq 0\}, \{B_2(t), t \geq 0\}$ being two independent Wiener processes and $\Upsilon := \mathbb{E}[(\log f(Z_2(0)))'^2]$.

PROOF. See Section 2.6. ■

The notation “ \Rightarrow ” represents weak convergence (or convergence in distribution) of processes in the Skorokhod topology (for more details about this type of convergence, see Section 3 of [Ethier and Kurtz \(1986\)](#)).

2.3.2. Optimisation

The sample paths of $\{\mathbf{Z}(t), t \geq 0\}$ depend on τ, A, ℓ, Υ and f . In this section, we optimise theoretically the state space exploration of $\{\mathbf{Z}(t), t \geq 0\}$ with respect to ℓ and τ . As a result, in the situation where the dimension of the models is large enough, i.e. for large enough n , reversible jump users will know how to choose the values of ℓ and τ in order to obtain the optimal algorithm. Indeed, optimising the asymptotic state space exploration of $\{(K(\lfloor nt \rfloor), X_1^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor)), t \geq 0\}$ is sufficient to optimise the asymptotic state space exploration of $\{(K(\lfloor nt \rfloor), \mathbf{X}^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor)), t \geq 0\}$. This is due to the fact that, in addition to optimising the exploration of models, we also optimise the state space exploration of the first parameter, and all parameters of a model share a similar behaviour.

During the theoretical optimisation, the constant A is considered to be fixed because its value cannot be arbitrarily chosen. Indeed, it is tied to the ratio $f/q \leq A^* = A/2$. In addition, everything

suggests that selecting a small value for this constant is desirable. The constant Υ and the function f are obviously fixed. Note that the PDF q only has an impact on the sample paths of $\{\mathbf{Z}(t), t \geq 0\}$ through the constant A .

We first optimise the algorithm with respect to ℓ . The stochastic process $\{Z_2(t), t \geq 0\}$ can be written as $Z_2(t) = V(2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2) \times t)$, where $\{V(t), t \geq 0\}$ is the following Langevin diffusion:

$$dV(t) = dB_2(t) + (\log f(V(t)))'/2 \times dt.$$

The term $2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)$ that multiplies the time index of $\{V(t), t \geq 0\}$ to obtain $\{Z_2(t), t \geq 0\}$ is sometimes called the “speed measure” of $\{Z_2(t), t \geq 0\}$ (hereafter, this shall be the meaning attributed to this expression). Viewed as a function of τ and ℓ , the process $\{Z_2(t), t \geq 0\}$ that optimally explores its state space is thus the one with the largest speed. We can optimise the algorithm with respect to ℓ by maximising the speed of $\{Z_2(t), t \geq 0\}$ with respect to this variable, because the value of ℓ does not have an impact on the sample paths of $\{Z_1(t), t \geq 0\}$ (see the definition of $\{Z_1(t), t \geq 0\}$ in Theorem 2.1). The function $2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)$ is maximised with respect to ℓ by $\ell = 2.38/\sqrt{\Upsilon}$, as stated in Corollary 1.2 of Roberts et al. (1997). We therefore obtain the same optimal value as these authors. This conclusion does not come as a surprise, since updating the parameters in our reversible jump algorithm (see Section 2.2) corresponds to the usual RWM step studied by these authors. Furthermore, the conditional distribution of the parameters given any model is essentially the same as their target distribution.

The optimisation with respect to ℓ tells us that the most efficient way to update the parameters is to set $\ell = 2.38/\sqrt{\Upsilon}$. Therefore, the optimal variance for the random walk is $(2.38^2/(\Upsilon(n + K(m))))\mathcal{I}_{n+K(m)}$. As mentioned in Roberts et al. (1997), Υ can be seen as a measure of “roughness” of f . Consider for instance the case $f = \mathcal{N}(\mu, \sigma^2)$, $\Upsilon = 1/\sigma^2$, representing the fact that small values for σ corresponds to “rough” f functions. In this situation, it is recommended to use small values for ℓ , because the optimal value for ℓ is 2.38σ . This also suggests that the thicker the tails of f , the larger should be the value for ℓ .

It could seem necessary to know $\Upsilon = \mathbb{E}[(\log f(Z_2(0)))'^2]$ in order to use this optimal scaling result. Fortunately, the practical 0.234 rule provided by Roberts et al. (1997) can be employed, as stated in Corollary 2.1. This corollary is an adapted version of Corollary 1.2 of Roberts et al. (1997). Its proof is similar to the one given by these authors, and is thus omitted (it can nevertheless be found in Section 3.7).

Corollary 2.1. *In the context described in Section 2.2, assume that $(K(0), \mathbf{X}^K(0)) \sim \pi_n$. Then, for all $m \in \mathbb{N}$,*

$$\mathbb{E} \left[1 \wedge \prod_{i=1}^{n+K(m)} \frac{f(Y_i^{K(m)}(m+1))}{f(X_i^{K(m)}(m))} \right] \rightarrow 2\Phi(-\ell\sqrt{\Upsilon}/2) \text{ as } n \rightarrow \infty.$$

In addition, setting $\ell = 2.38/\sqrt{\Upsilon}$ is equivalent to having $2\Phi(-\ell\sqrt{\Upsilon}/2) = 2\Phi(-2.38/2) = 0.234$.

Therefore, in order to reach optimal efficiency with respect to ℓ , reversible jump users can monitor the acceptance rate of candidates $\mathbf{Y}^{K(m)}(m+1)$, where $\mathbf{Y}^{K(m)}(m+1) \sim \mathcal{N}(\mathbf{X}^{K(m)}(m), (\ell^2/(n +$

$K(m))\mathcal{I}_{n+K(m)}$, and tune the value of ℓ such that this rate is approximately 0.234. Note that this rate must be computed by considering only iterations in which there has been an attempt at updating the parameters, i.e. iterations belonging to the set $\{m : J(m) = 1\}$.

Note that the speed measures of $\{Z_1(t), t \geq 0\}$ and $\{Z_2(t), t \geq 0\}$, given respectively by $2(1 - \tau)/(A + 1)$ and $2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)$, theoretically confirm that users should select the smallest value for A satisfying $f/q \leq A^* = A/2$. Indeed, the speed measure of $\{Z_1(t), t \geq 0\}$ is maximised when A is small, and the one of $\{Z_2(t), t \geq 0\}$ does not depend on A .

We now optimise the algorithm with respect to τ . We need a measure that takes into account the fact that an increase in the value of τ results in an increase in the speed of $\{Z_2(t), t \geq 0\}$, but also in a decrease in the speed of $\{Z_1(t), t \geq 0\}$, and vice versa. Intuitively, when the value of τ is increased, more updates of the parameters (and therefore less model switchings) are proposed. It would seem natural to consider the total speed of these two processes to optimise the algorithm with respect to τ . The total speed is given by (using the optimal value for ℓ)

$$2 \left[\tau(2.38^2/\Upsilon)\Phi(-2.38/2) + (1 - \tau)/(A + 1) \right].$$

However, it is not a suitable measure because if, for instance $(2.38^2/\Upsilon)\Phi(-2.38/2) = (5.66/\Upsilon)\Phi(-1.19) > 1/(A + 1)$, it would be proposed to choose the value of τ as close as possible to 1. In such a situation, there would be very few model switchings, which would result in a slow exploration of the entire state space.

We thus need a measure that penalises such a behaviour. This is achieved using integrated linear combinations of the autocorrelation functions (ACFs) of $\{Z_1(t), t \geq 0\}$ and $\{Z_2(t), t \geq 0\}$. Indeed, if for instance we set τ close to 1, $\{Z_2(t), t \geq 0\}$ would be a “fast” process with an ACF that decreases rapidly towards 0, while $\{Z_1(t), t \geq 0\}$ would be a “slow” process with an almost constant ACF around the value 1 (which is not desirable). Therefore, the sum of these two functions would decrease rapidly towards 1, thereafter remaining almost constant around this value. There should then exist a value of τ between 0 and 1 that induces two relatively “fast” processes, with a sum of ACFs that decreases relatively rapidly towards 0. We thus consider the integral of the sum of the ACFs of $\{Z_1(t), t \geq 0\}$ and $\{Z_2(t), t \geq 0\}$ to optimise the algorithm with respect to τ :

$$\int_0^\infty \{\text{corr}[Z_1(t), Z_1(t + s)] + \text{corr}[Z_2(t), Z_2(t + s)]\} ds, \quad t \geq 0. \quad (2.8)$$

This measure is inspired from the effective sample size (see Section 12.3.5 of [Robert and Casella \(2004\)](#)) and can be viewed as the sum of the (infinitesimally) integrated autocorrelation times of $\{Z_1(t), t \geq 0\}$ and $\{Z_2(t), t \geq 0\}$. It therefore represents a measure of the total “inefficiency” of these processes and the optimal value of τ is the one that minimises it.

We need to compute the ACFs of $\{Z_1(t), t \geq 0\}$ and $\{Z_2(t), t \geq 0\}$ in order to optimise the algorithm with respect to τ . The process $\{Z_1(t), t \geq 0\}$ satisfies the conditions of Theorem 2.1

stated by Bibby et al. (2005), implying that

$$\text{corr}[Z_1(t), Z_1(t+s)] = \exp\{-(1-\tau)s/(A+1)\}, \quad s, t \geq 0.$$

The behaviour of $\{Z_2(t), t \geq 0\}$ depends on f and consequently its ACF cannot be computed in all generality. A particular situation is now studied in order to obtain general information about the optimal value of τ . It is natural to consider the case where $f = \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma > 0$, since it implies that the stationary distributions of $\{Z_1(t), t \geq 0\}$ and $\{Z_2(t), t \geq 0\}$ are respectively a standard normal and a $\mathcal{N}(\mu, \sigma^2)$. Therefore, it represents a situation where all the components of $\{\mathbf{Z}^n(t), t \geq 0\}$ have a similar behaviour when n is large enough. Using the optimal value for ℓ , we have that

$$dZ_2(t) = \sqrt{2\tau(5.66\sigma^2)\Phi(-1.19)}dB_2(t) - \tau(5.66\sigma^2)\Phi(-1.19) \times (Z_2(t) - \mu)/\sigma^2 dt.$$

This process also satisfies the conditions of Theorem 2.1 stated by Bibby et al. (2005), implying that

$$\text{corr}[Z_2(t), Z_2(t+s)] = \exp\{-\tau 5.66\Phi(-1.19)s\}, \quad s, t \geq 0.$$

Thus, when $f = \mathcal{N}(\mu, \sigma^2)$ and ℓ is set to its optimal value, the optimal value for τ is

$$\frac{\sqrt{5.66\Phi(-1.19)(A+1)} - 1}{5.66\Phi(-1.19)(A+1) - 1}. \quad (2.9)$$

The constant A clearly has an impact on the optimal value of the constant τ . In fact, the optimal value decreases as the value of A increases (this is depicted in Figures 2.1 and 2.2). Note that the sum of the ACFs arising from the optimal value of the constant τ represents the curve which decreases most rapidly towards 0, in the sense that the area under the curve is minimised. It indicates that the underlying process $\{\mathbf{Z}(t), t \geq 0\}$ optimally explores its state space. When $A = 2$, the optimal value for τ is 0.415. This situation corresponds to $f = q$, and therefore to the best choice of distribution q . When $A = 5$, the optimal value for τ is 0.334, and when $A = 25$, it is 0.194. The constant A therefore has an indirect impact on the sample paths of $\{Z_2(t), t \geq 0\}$ when τ is set to its optimal value. Indeed, the speed measure of this process, which is given by $2\tau\ell^2\Phi(-\ell\sqrt{Y}/2)$, decreases as A increases if τ is set to its optimal value. Again, selecting the smallest admissible value for A is desirable.

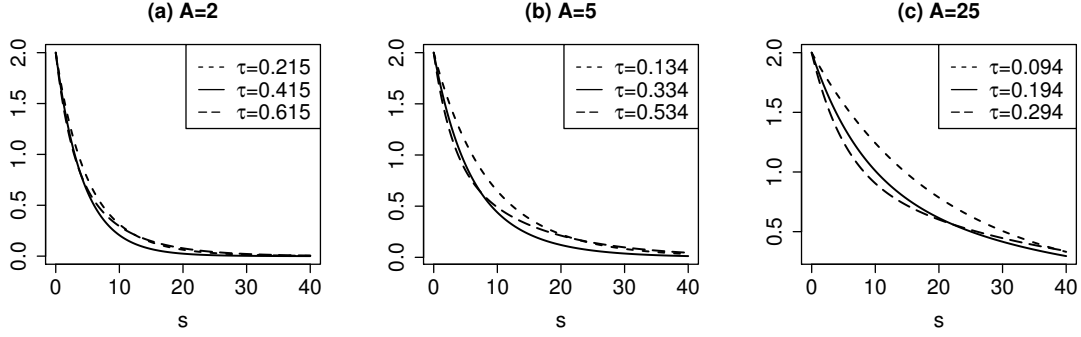


Figure 2.1. Sum of the ACFs as a function of s for different values of τ and $A = 2, 5, 25$, when $f = \mathcal{N}(\mu, \sigma^2)$ and ℓ is set to its optimal value (for each graph, the solid line represents the function arising from the optimal value of the constant τ)

In the situation where we cannot directly sample from f , the constant A represents in some way the level of precision in the design of the distribution q . For illustrative purpose, assume that $f = \mathcal{N}(\mu_1, \sigma_1^2)$ and that a user considers the proposal distribution $q = \mathcal{N}(\mu_2, \sigma_2^2)$. If he believes he might have overestimated the variability by a factor of at most 1.5 ($\sigma_2/\sigma_1 \leq 1.5$) and the location

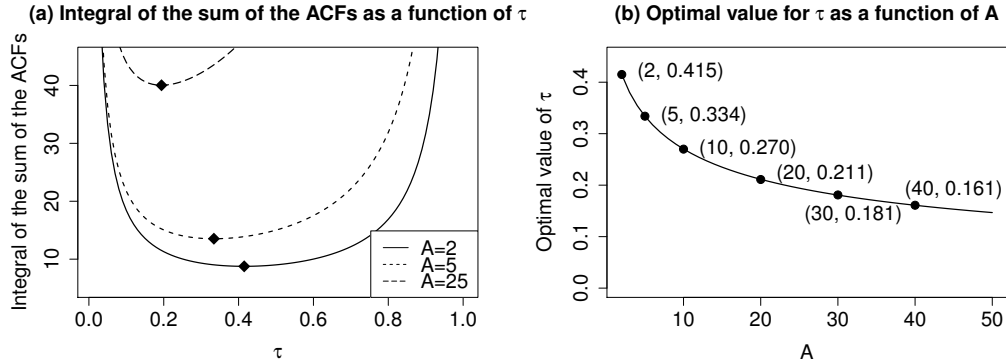


Figure 2.2. (a) Integral of the sum of the ACFs as a function of τ , for $A = 2, 5, 25$ (the diamonds represent the optimal value for τ), (b) Optimal value for τ as a function of A ; for both graphs, $f = \mathcal{N}(\mu, \sigma^2)$ and ℓ is set to its optimal value

by at most 1 ($0 \leq \mu_2 - \mu_1 \leq 1$), then, provided that $\sigma_2^2 \geq 1 + \sigma_1^2$, he should set the constant A to 5 (see Figure 2.1 (b) for the impact on the sum of the ACFs). Indeed,

$$\frac{f(x)}{q(x)} \leq \max_x \frac{f(x)}{q(x)} = \frac{\sigma_2}{\sigma_1} \exp \left\{ \frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2 - \sigma_1^2} \right\} \leq 2.47 \leq A^* =: A/2.$$

Note that this upper bound is valid only if $\sigma_2 > \sigma_1$; otherwise, the ratio f/q is unbounded. It means that users should always be conservative towards variability, in particular when they lack information about the location.

Suppose that a user sets ℓ to its optimal value and, say, $A = 5$ (while, theoretically, $A \geq 5$, which means that the smallest admissible value is selected). The optimisation with respect to τ

tells us that the optimal distribution g^* is given by

$$g^*(j) = \begin{cases} \tau = 0.33 & \text{if } j = 1, \\ (1 - \tau)A/(A + 1) = (1 - 0.33)5/(5 + 1) = 0.56 & \text{if } j = 2, \\ (1 - \tau)/(A + 1) = (1 - 0.33)/(5 + 1) = 0.11 & \text{if } j = 3, \end{cases}$$

assuming that $f = \mathcal{N}(\mu, \sigma^2)$. Recall that the random variable $J(m + 1), m \in \mathbb{N}$, is distributed according to g and this random variable indicates which movement type is attempted at iteration $m + 1$: update of the parameters ($J(m + 1) = 1$), inclusion of an extra parameter ($J(m + 1) = 2$) or withdrawal of the last parameter ($J(m + 1) = 3$).

Upon examination of the optimal distribution g^* given above, the probabilities $g^*(2)$ and $g^*(3)$ might appear unbalanced. We should however focus on the probabilities of the actual movements of $\{K(m), m \in \mathbb{N}\}$. Intuitively, it seems effective that movements of type $K(m) \mapsto K(m) + 1$ (inclusion of an extra parameter) be as frequent as those of type $K(m) \mapsto K(m) - 1$ (withdrawal of the last parameter). Given that there is an attempt to include an extra parameter (i.e. given that $J(m + 1) = 2$), Proposition 2.2 indicates that the average acceptance probability of the candidate $U(m + 1)$, where $U(m + 1) \sim q$, converges towards $1/A$ as $n \rightarrow \infty$.

Proposition 2.2. *Consider the context described in Section 2.2 and the function g defined in (2.6). If we assume that $(K(0), \mathbf{X}^K(0)) \sim \pi_n$, then for all $m \in \mathbb{N}$,*

$$\mathbb{E} \left[1 \wedge \frac{f(U(m + 1))p(K(m) + 1)g(3)}{q(U(m + 1))p(K(m))g(2)} \right] \rightarrow \frac{1}{A} \text{ as } n \rightarrow \infty.$$

PROOF. See Section 2.8. ■

It means that the average probability of a movement of type $K(m) \mapsto K(m) + 1$ is asymptotically $(1 - \tau)A/(A + 1) \times 1/A = (1 - \tau)/(A + 1)$. This is equal to the average probability of a movement of type $K(m) \mapsto K(m) - 1$. Indeed, the probability to withdraw the last parameter in a given iteration is $(1 - \tau)/(A + 1)$ for all n (this movement is automatically accepted, as explained in Section 2.2).

2.4. PRACTICAL CONSIDERATIONS

2.4.1. Optimal Implementation and Generalisation

When users are able to state that $\pi_n(k, \mathbf{x}^k) = p(k) \prod_{i=1}^{n+k} f(x_i^k)$, with $f = \mathcal{N}(\mu, \sigma^2)$ and p defined as in Section 2.2, they can directly construct the optimal algorithm setting $q = f$ (therefore $A = 2$), $\tau = 0.415$ (the optimal value for τ given by (2.9)), and $\ell = 2.38\sigma$ (the optimal value for ℓ). This situation is however unlikely to occur. Our recommendation for optimising the algorithm is the following: perform some trial runs to tune the value of ℓ (according to Corollary 2.1), and to obtain general information about f such as the location and scale (which is useful when we cannot directly sample from f). The information gathered about f enables to improve the design of the proposal distribution q , and thus to reduce the value of the upper bound of f/q , given by

$A^* =: A/2$, which in turn leads to a new optimal value for τ . The optimal value for τ given by (2.9) is theoretically valid when $f = \mathcal{N}(\mu, \sigma^2)$, but it should be suitable if f has a similar shape. If users lack information about f at the beginning of the process, they should start with conservative values for A and ℓ .

The optimal scaling result of Roberts et al. (1997) is known to be relatively robust, in the sense that it holds under weaker assumptions (see, e.g., Roberts and Rosenthal (2001) and Bédard (2007)). We believe that the results presented in this paper are also robust and we conjecture that they are valid when

$$\pi_n(k, \mathbf{x}^k) = p(k) \prod_{i=1}^{n+k} f_i(x_i^k), \quad (2.10)$$

where $f_i(x_i^k) := (1/\sigma_i)f((x_i^k - \mu_i)/\sigma_i)$, $\mu_i \in \mathbb{R}$ and $\sigma_i > 0$ are constants, and f and p satisfy the assumptions described in Section 2.2. The sampler described in Section 2.2 is applied here, the only difference being that a proposal distribution q_{K+1} is used to add a parameter to switch from Model K to Model $K + 1$, in order to accommodate for the different functions f_i . We assume that $f_i/q_i \leq A_i^* \leq A_n^*$ for all $i \in \{n+1, \dots, n + \lfloor \sqrt{n} \log n \rfloor\}$ with $A_n^* := \max_i A_i^*$, and $A_n^* \leq A^*$ for all n . The procedure described previously for optimising the algorithm is then applied with this new constant A^* . We thus expect the results in this paper to be applicable to models in which the parameters are independent but not identically distributed. In particular, they turn out to be useful in designing the reversible jump algorithm when the posterior distribution arises from a robust principal component regression, as shown in Chapter 5. This is explained by the fact that the posterior has a structure similar to that in (2.10). The authors are consequently able to design an efficient sampler that allows identification of the relevant components and estimation of the parameters for prediction purpose. Note that the generalisation of our results to these contexts is not trivial.

2.4.2. Simulation Study

Samples produced by the reversible jump algorithm provide information about the joint posterior distribution of K , the model indicator, and \mathbf{X}^K , the parameters of Model K . As a result, users can choose a model (usually the one with the highest frequency in the sample) and estimate its parameters (using sample means and intervals for instance). Ideally, the chosen model, along with the parameter estimates, would be the same as if the “true” posterior distribution had been used. In Section 2.3, we have explained how to implement an efficient reversible jump algorithm using optimisation results that have been derived from Theorem 2.1. In this section, we illustrate the impact of the design of the sampler on the estimation of the target distribution via a simulation study.

Implementing the reversible jump algorithm described in Section 2.2 comes down to specifying the PDF q and the values of the constants A , τ and ℓ . In Section 2.3, it has been shown that the constants τ and A have an impact on the estimation of the whole joint posterior distribution of

(K, \mathbf{X}^K) . The constant ℓ has an impact on the \mathbf{X}^K part only and this has been thoroughly studied by [Roberts and Rosenthal \(2001\)](#). In [Section 2.3](#), it has also been explained that, asymptotically, the PDF q only has an impact through the constant A . In this section, we therefore focus on showing the impact of the constants τ and A on the estimation of the target distribution. More precisely, considering that $f = \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$, we set $\ell = 2.38\sigma$ (its optimal value), $q = f$, and we evaluate the performance of the algorithm for every $\tau \in (0, 1)$, in the cases where $A = 2, 5, 25$. For a given A , the optimal value for τ can thus be determined using [\(2.9\)](#). Studying specific situations like this one will hopefully lead to a better understanding of the practical considerations for optimising the algorithm in general settings.

For fixed τ and A , we evaluate the performance of the reversible jump algorithm using mean absolute deviations (MADs) around quantities that are usually of interest for users: the posterior mode of K (denoted by k^*), and the posterior mean and standard deviation of X_i^K , $i \in \{1, \dots, n+K\}$ (we consider the first parameter X_1^K for the simulation study). For a given sample produced by the reversible jump algorithm, k^* is estimated by \widehat{k}^* , the mode of the sample related to the random variable K , and μ and σ are estimated by $\widehat{\mu}$ and $\widehat{\sigma}$, which are respectively the mean and standard deviation of the sample related to the random variable X_i^K . Representative samples lead to accurate estimates, thus resulting in small absolute deviations. For fixed τ and A , we approximate the MADs by independently running N times the reversible jump algorithm and by computing $\sum_{i=1}^N |\widehat{k}_i^* - k^*|/N$, $\sum_{i=1}^N |\widehat{\mu}_i - \mu|/N$ and $\sum_{i=1}^N |\widehat{\sigma}_i - \sigma|/N$, where \widehat{k}_i^* , $\widehat{\mu}_i$ and $\widehat{\sigma}_i$ are respectively the mode, mean and standard deviation based on the sample produced by the i th run. We also compute a global measure that mimics the one used in [Section 2.3](#) (and given in [\(2.8\)](#)) to optimise the algorithm with respect to τ . This global measure is a linear combination of a standardised version of the MADs:

$$\frac{\frac{1}{N} \sum_{i=1}^N |\widehat{k}_i^* - k^*|}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (\widehat{k}_i^* - k^*)^2}} + \frac{1}{2} \left(\frac{\frac{1}{N} \sum_{i=1}^N |\widehat{\mu}_i - \mu|}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (\widehat{\mu}_i - \mu)^2}} + \frac{\frac{1}{N} \sum_{i=1}^N |\widehat{\sigma}_i - \sigma|}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (\widehat{\sigma}_i - \sigma)^2}} \right).$$

In this simulation study, $N = 1,000$, $\mu = 0$, $\sigma = 1$, and each sample is of size 100,000. The results are presented in [Figure 2.3](#).

As expected, the performance of the algorithm regarding the estimation of the posterior distribution of K decreases as the value of τ increases due to fewer model switchings. An increase in τ has the opposite effect regarding the estimation of the posterior of \mathbf{X}^K . The vertical lines represent the optimal values for τ , which are 0.415, 0.334 and 0.194, when $A = 2, 5, 25$, respectively. These values are optimal in the sense that they allow to attain the appropriate balance between an efficient estimation of the posterior of K (but a poor estimation of the posterior of \mathbf{X}^K) and an efficient estimation of the posterior of \mathbf{X}^K (but a poor estimation of the posterior of K).

[Figure 2.3](#) also helps illustrate that reversible jump users should favor the smallest admissible value for A , an aspect that has been theoretically justified in [Section 2.3](#). Indeed, the value of A has a direct impact on the performance regarding the estimation of the posterior distribution of K (the performance decreases as the value of A increases), and it has an indirect impact on the performance regarding the estimation of the posterior of \mathbf{X}^K through the optimal value for τ .

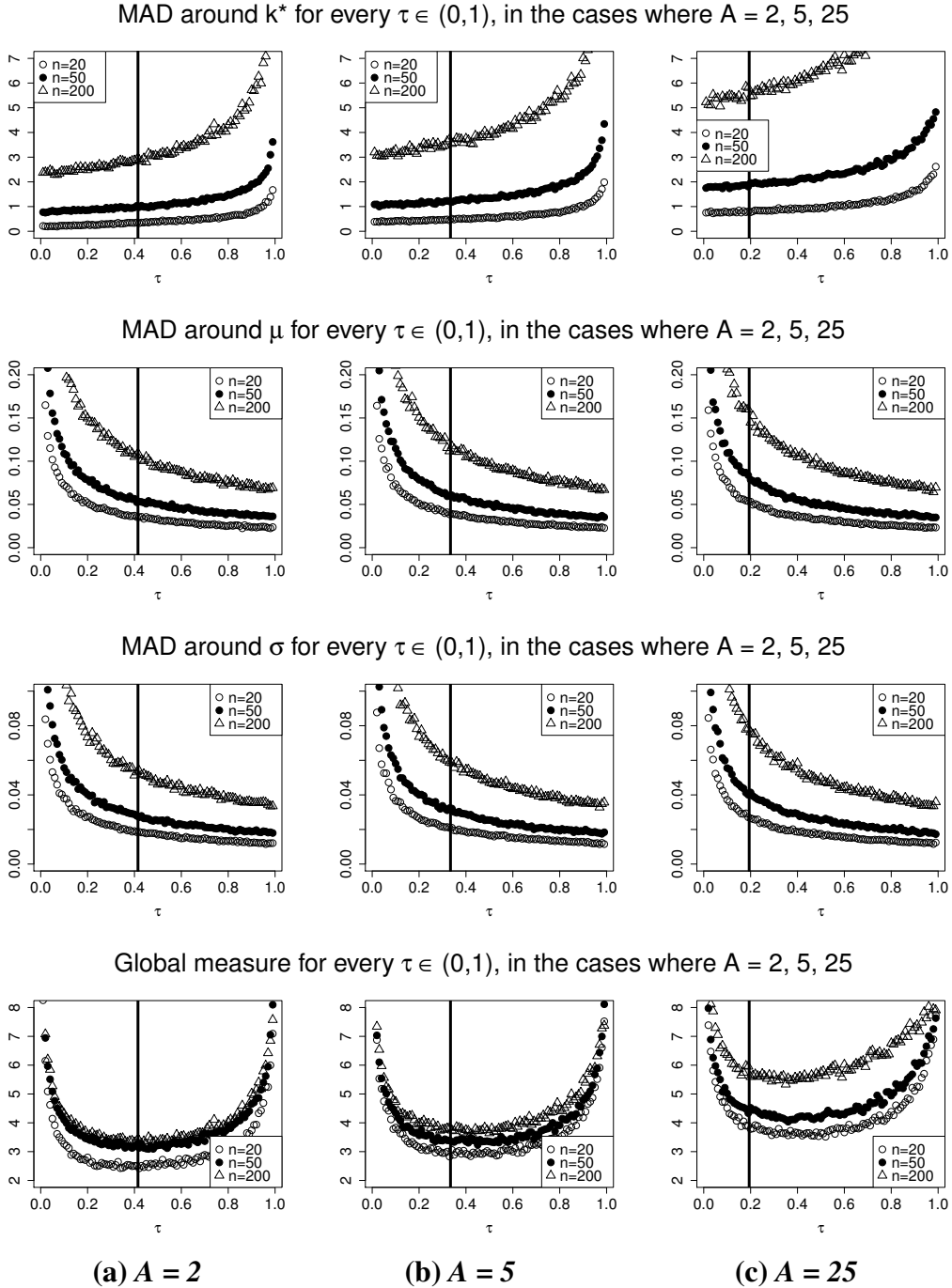


Figure 2.3. MADs around k^* , μ and σ , and the global measure, for every $\tau \in (0, 1)$, in the cases where $A = 2, 5, 25$ (the vertical lines represent the optimal values for τ , which are 0.415, 0.334 and 0.194, when $A = 2, 5, 25$, respectively)

We finally note that, as $n \rightarrow \infty$, the curves defined by the global measure as a function of τ look like those in Figure 2.2 (a). However, for moderate values of n , the curves defined by the global measure are almost flat between 0.2 and 0.6. It indicates that, for moderate values of n , selecting

any value in this range for τ is almost optimal. As n increases, users should narrow down to the optimal value of τ , especially for large value of A .

2.5. CONCLUSION

In this paper, we have provided guidelines to practically and efficiently implement the reversible jump algorithm described in Section 2.2. The performance of this algorithm depends on the inputs required for its implementation: the constants ℓ , τ and A , and the proposal distribution q . The theoretical results derived in Section 2.3 allow to choose values for ℓ and τ that are asymptotically optimal. The optimal value for ℓ is given by $2.38/\sqrt{\Upsilon}$, which corresponds to an acceptance rate of candidates for updating the parameters of approximately 0.234 (considering only iterations in which there has been an attempt to update the parameters). The optimal value for τ can be determined, for a given A , using (2.9). The practical guidelines given in Section 2.4.1 enable to suitably design the proposal distribution q , which implicitly allows to determine the constant A .

The theoretical results hold when the algorithm is applied to sample from a target distribution π_n that satisfies the assumptions provided in Section 2.2. Essentially, the target distribution π_n must be a product of the PMF p (the distribution of the model indicator K) and a $(n + K)$ -product of PDFs f (the optimal value for τ arising from (2.9) is theoretically valid when $f = \mathcal{N}(\mu, \sigma^2)$). Being aware that this sampling context is simple, our goal was to make a first step towards automated implementation of the reversible jump algorithm. The distribution π_n is more often comprised of a product of different functions f_i , as in a context of selection of the principal components when the principal component regression is used to model the data. In such contexts, the asymptotic situation $n \rightarrow \infty$, which implies that the number of parameters in all models approach infinity, can also be unrealistic. However, if we consider the numerical experiment conducted in Section 2.4.2, selecting any value between 0.2 and 0.6 for τ should lead to good performances for moderate values of n . In addition, the 0.234 rule should still indicate a suitable scaling for the random walk. We therefore propose to consider this when tuning the reversible jump algorithm in such contexts. We also propose to use the heuristic approach described in Section 2.4.1, which purpose is to design the algorithm when it is applied to sample from more elaborate target distributions.

2.6. PROOF OF THEOREM 2.1

This section is dedicated to the demonstration of the main result of this paper, the weak convergence $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\} \Rightarrow \{\mathbf{Z}(t), t \geq 0\}$ in the Skorokhod topology as $n \rightarrow \infty$ (the stochastic processes $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$ and $\{\mathbf{Z}(t), t \geq 0\}$ have been defined in Theorem 2.1). Thus consider the sampling context described in Section 2.2.

In order to prove the result, we demonstrate the convergence of the finite-dimensional distributions of $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$ to those of $\{\mathbf{Z}(t), t \geq 0\}$. To achieve this, we verify condition (c) of Theorem 8.2 from Chapter 4 of Ethier and Kurtz (1986). The weak convergence then follows from

Corollary 8.6 of Chapter 4 of [Ethier and Kurtz \(1986\)](#). The remaining conditions of Theorem 8.2 and the conditions specified in Corollary 8.6 are either straightforward or easily derived from the proof given in this section. They are nevertheless explicitly verified in Chapter 3 (the remaining conditions of Theorem 8.2 are verified in Sections 3.2 and 3.4, and those specified in Corollary 8.6 are verified in Section 3.5).

The proof of the convergence of the finite-dimensional distributions relies on the convergence of (what we call) the “pseudo-generator”, a quantity that we now introduce. The proof follows in Section 2.6.2.

2.6.1. Pseudo-Generator

In this section, we introduce a quantity that we call the “pseudo-generator” of $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$ due to its similarity with the infinitesimal generator of stochastic processes. It is defined as follows:

$$\varphi_n(t) := n\mathbb{E}[h(\mathbf{Z}_{1,2}^n(t + 1/n)) - h(\mathbf{Z}_{1,2}^n(t)) \mid \mathcal{F}^{\mathbf{Z}^n}(t)],$$

where $h \in C_c^\infty(\mathbb{R}^2)$, the space of infinitely differentiable functions on \mathbb{R}^2 with compact support. Theorem 2.5 from Chapter 8 of [Ethier and Kurtz \(1986\)](#) allows us to restrict our attention to this set of functions when studying the limiting behaviour of the pseudo-generator (see Section 3.2.1 for more details).

Let

$$R^K(\lfloor nt \rfloor) := \frac{K(\lfloor nt \rfloor) - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}} = Z_1^n(t).$$

The pseudo-generator $\varphi_n(t)$ can be decomposed into three parts, each associated with a specific type of movement, as follows:

$$\varphi_n(t) = \varphi_{1,n}(t) + \varphi_{2,n}(t) + \varphi_{3,n}(t),$$

where $\varphi_{1,n}(t)$ is associated with the update of the parameters, i.e.

$$\varphi_{1,n}(t) := n\tau\mathbb{E}\left[\left(h(R^K, Y_1^K) - h(R^K, X_1^K)\right)\left(1 \wedge \frac{\prod_{i=1}^{n+K} f(Y_i^K)}{\prod_{i=1}^{n+K} f(X_i^K)}\right) \mid R^K, \mathbf{X}^K\right],$$

$\varphi_{2,n}(t)$ is associated with the inclusion of an extra parameter, i.e.

$$\varphi_{2,n}(t) := \frac{n(1-\tau)A}{A+1}\mathbb{E}\left[\left(h(R^{K+1}, X_1^K) - h(R^K, X_1^K)\right)\left(1 \wedge \frac{f(U)p(K+1)}{q(U)p(K)A}\right) \mid R^K, \mathbf{X}^K\right], \quad (2.11)$$

and $\varphi_{3,n}(t)$ is associated with the withdrawal of the last parameter, i.e.

$$\varphi_{3,n}(t) := \frac{n(1-\tau)}{A+1}\mathbb{E}\left[\left(h(R^{K-1}, X_1^K) - h(R^K, X_1^K)\right)\left(1 \wedge \frac{q(X_{n+K}^K)p(K-1)A}{f(X_{n+K}^K)p(K)}\right) \mid R^K, \mathbf{X}^K\right]. \quad (2.12)$$

Note that the Markov process $\{(R^K(m), \mathbf{X}^{K(m)}(m)), m \in \mathbb{N}\}$ is time-homogeneous, and consequently, the time index has been omitted to simplify the notation. Also note that, when there is an update of the parameters, only the parameters \mathbf{X}^K move (the model indicator remains the same). When an

extra parameter is included or the last parameter withdrawn, only the model indicator changes, as a switch from Model K to Model $K + 1$ or $K - 1$ is made.

2.6.2. Proof of the Convergence of the Finite-Dimensional Distributions

Condition (c) of Theorem 8.2 essentially reduces to the following convergence:

$$\mathbb{E} \left[\left| \varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t)) \right| \right] \rightarrow 0 \text{ as } n \rightarrow \infty,$$

where G is the generator of a diffusion with $G = G_1 + G_2$ and

$$\begin{aligned} G_1 h(\mathbf{Z}_{1,2}^n(t)) &= \tau \ell^2 \Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) (\log f(Z_2^n(t)))' h_y(\mathbf{Z}_{1,2}^n(t)) + \tau \ell^2 \Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) h_{yy}(\mathbf{Z}_{1,2}^n(t)), \\ G_2 h(\mathbf{Z}_{1,2}^n(t)) &= \frac{1 - \tau}{A + 1} \times -Z_1^n(t) h_x(\mathbf{Z}_{1,2}^n(t)) + \frac{1 - \tau}{A + 1} h_{xx}(\mathbf{Z}_{1,2}^n(t)). \end{aligned}$$

The function h above is the same function h involved in the definition of the random variable $\varphi_n(t)$ (given in Section 2.6.1). In other words, the convergence has to be proved for an arbitrary function $h \in C_c^\infty(\mathbb{R}^2)$. The functions h_x and h_{xx} respectively represent the first and second derivatives of h with respect to its first argument. Analogously, the functions h_y and h_{yy} respectively represent the first and second derivatives of h with respect to its second argument. Note that it exists a positive constant M such that h and all its derivatives are bounded in absolute value by this constant.

Using the triangle inequality, we have

$$\mathbb{E} \left[\left| \varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t)) \right| \right] \leq \mathbb{E} \left[\left| \varphi_{1,n}(t) - G_1 h(\mathbf{Z}_{1,2}^n(t)) \right| \right] + \mathbb{E} \left[\left| \varphi_{2,n}(t) + \varphi_{3,n}(t) - G_2 h(\mathbf{Z}_{1,2}^n(t)) \right| \right].$$

In this paper, we show that the second term on the right-hand side (RHS) converges towards 0 as $n \rightarrow \infty$. The proof that the first term converges towards 0 is similar to that of Theorem 1.1 of [Roberts et al. \(1997\)](#), and is thus omitted (it can nevertheless be found in Section 3.3).

The key here is the use of Taylor expansions in order to obtain derivatives of h as in generators of diffusions.

We first analyse $\varphi_{2,n}(t)$ as defined in (2.11). As explained in Section 2.2, $0 \leq p(K+1)/p(K) \leq 2$, and therefore,

$$\frac{f(U)p(K+1)}{q(U)p(K)A} \leq \frac{2f(U)}{q(U)A} \leq 1.$$

Consequently, since $h(R^{K+1}, X_1^K) = h(R^K + 1/\sqrt{n}, X_1^K)$,

$$\begin{aligned} \varphi_{2,n}(t) &= \frac{n(1-\tau)}{A+1} \left(h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \frac{p(K+1)}{p(K)} \mathbb{E} \left[\frac{f(U)}{q(U)} \mid R^K, \mathbf{X}^K \right] \\ &= \frac{n(1-\tau)}{A+1} \left(h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \frac{p(K+1)}{p(K)}. \end{aligned}$$

In the last equality, we use the fact that U is independent of (K, \mathbf{X}^K) , and therefore,

$$\mathbb{E} \left[\frac{f(U)}{q(U)} \mid R^K, \mathbf{X}^K \right] = \mathbb{E} \left[\frac{f(U)}{q(U)} \right] = \int_{-\infty}^{\infty} \frac{f(u)}{q(u)} q(u) du = 1.$$

Note that $\varphi_{2,n}(t) = 0$ when $K = \lfloor \sqrt{n} \log n \rfloor$ since $p(\lfloor \sqrt{n} \log n \rfloor + 1) = 0$.

We now study $\varphi_{3,n}(t)$ as defined in (2.12). As explained in Section 2.2, when $2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor$ we have that $p(K-1)/p(K) \geq 1/2$. Therefore, when $2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor$,

$$\frac{q(X_{n+K}^K)p(K-1)A}{f(X_{n+K}^K)p(K)} \geq \frac{q(X_{n+K}^K)A}{2f(X_{n+K}^K)} \geq 1.$$

This means that the acceptance probability of withdrawing the last parameter is 1, when it is possible to withdraw a parameter. Consequently, since $h(R^{K-1}, X_1^K) = h(R^K - 1/\sqrt{n}, X_1^K)$,

$$\varphi_{3,n}(t) = \frac{n(1-\tau)}{A+1} \left(h(R^K - 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor).$$

Note that $\varphi_{3,n}(t) = 0$ when $K = 1$ since $p(0) = 0$.

By using Taylor expansions of h around R^K , we obtain that

$$\begin{aligned} h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) &= \frac{1}{\sqrt{n}} h_x(R^K, X_1^K) + \frac{1}{2n} h_{xx}(R^K, X_1^K) + \frac{1}{6n^{3/2}} h_{xxx}(W, X_1^K), \\ h(R^K - 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) &= -\frac{1}{\sqrt{n}} h_x(R^K, X_1^K) + \frac{1}{2n} h_{xx}(R^K, X_1^K) - \frac{1}{6n^{3/2}} h_{xxx}(T, X_1^K), \end{aligned}$$

where W belongs to $(R^K, R^K + 1/\sqrt{n})$, T belongs to $(R^K - 1/\sqrt{n}, R^K)$, and h_{xxx} represents the third derivative of h with respect to its first argument.

Therefore,

$$\begin{aligned} \varphi_{2,n}(t) + \varphi_{3,n}(t) - G_2 h(\mathbf{Z}_{1,2}^n(t)) &= \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \\ &\quad \times \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \\ &\quad + \mathbb{1}(K=1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left(\sqrt{n} \times \frac{p(K+1)}{p(K)} + R^K \right) \\ &\quad - \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) (\sqrt{n} - R^K) \\ &\quad + \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left(\frac{p(K+1)}{p(K)} - 1 \right) \\ &\quad + \mathbb{1}(K=1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left(\frac{p(K+1)}{p(K)} - 2 \right) \\ &\quad - \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \\ &\quad + \frac{1-\tau}{6\sqrt{n}(A+1)} h_{xxx}(W, X_1^K) \frac{p(K+1)}{p(K)} \mathbb{1}(1 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \\ &\quad - \frac{1-\tau}{6\sqrt{n}(A+1)} h_{xxx}(T, X_1^K) \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor). \end{aligned} \tag{2.13}$$

We now show that the expectation of the absolute value of each term on the RHS in (2.13) converges towards 0 as $n \rightarrow \infty$. Consequently, using the triangle inequality we will obtain

$$\mathbb{E} \left[\left| \varphi_{2,n}(t) + \varphi_{3,n}(t) - G_2 h(\mathbf{Z}_{1,2}^n(t)) \right| \right] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

We start with the last terms in (2.13) and make our way up. It is clear that the expectation of the absolute value of each of the last two terms converges towards 0 as $n \rightarrow \infty$ since $|h_{xx}| \leq M$ and $0 \leq p(K+1)/p(K) \leq 2$.

We now analyse the fourth one (starting from the bottom). As $n \rightarrow \infty$,

$$\begin{aligned} & \mathbb{E} \left[\left| \mathbb{1}(K=1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left(\frac{p(K+1)}{p(K)} - 2 \right) \right| \right] \\ & \leq \frac{M(1-\tau)}{A+1} \times \mathbb{E} [\mathbb{1}(K=1)] = \frac{M(1-\tau)}{A+1} \times \mathbb{P}(K=1) \rightarrow 0, \end{aligned}$$

using $|h_{xx}| \leq M$ and $0 \leq |p(K+1)/p(K) - 2| \leq 2$ in the first inequality. Proposition 2.3 in Section 2.7 is then used to conclude that $\mathbb{P}(K=1) \rightarrow 0$ as $n \rightarrow \infty$. The proof for the third term (starting from the bottom) is similar.

Applying Lemmas 2.1 to 2.3 from Section 2.7, each of the remaining terms is seen to converge towards 0 in L^1 as $n \rightarrow \infty$, and thus

$$\mathbb{E} \left[\left| \varphi_{2,n}(t) + \varphi_{3,n}(t) - G_2 h(\mathbf{Z}_{1,2}^n(t)) \right| \right] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

2.7. RESULTS USED IN THE PROOF OF THEOREM 2.1

Lemma 2.1. *As $n \rightarrow \infty$, we have*

$$\mathbb{E} \left[\left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left(\frac{p(K+1)}{p(K)} - 1 \right) \right| \right] \rightarrow 0.$$

PROOF OF LEMMA 2.1. First,

$$\begin{aligned} & \mathbb{E} \left[\left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left(\frac{p(K+1)}{p(K)} - 1 \right) \right| \right] \\ & \leq \frac{M(1-\tau)}{2(A+1)} \mathbb{E} \left[\mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right], \end{aligned}$$

because $|h_{xx}| \leq M$.

Considering the case where $\lfloor \sqrt{n} \log n \rfloor$ is odd, we have

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right] \\ & = \mathbb{E} \left[\mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right] \\ & \quad + \mathbb{E} \left[\mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right]. \end{aligned}$$

We analyse each term separately. The first one corresponds to the case where K is at the mode or at its right. Therefore, $p(K+1)/p(K) = a_{K,n}$ and

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right] \\ &= \mathbb{E} \left[\mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) |a_{K,n} - 1| \right] \\ &= \mathbb{E} \left[\mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \frac{|K - \lfloor \sqrt{n} \log n \rfloor / 2|}{n} \right] \\ &\leq \frac{\lfloor \sqrt{n} \log n \rfloor / 2 - 1}{n} \leq \frac{\log n}{2\sqrt{n}} \rightarrow 0, \end{aligned}$$

because $a_{K,n} = 1 - b_{K,n}/\sqrt{n}$ and $b_{K,n} = |K - \lfloor \sqrt{n} \log n \rfloor / 2| / \sqrt{n}$. We now study the case where K is at the left of the mode. Therefore, $p(K+1)/p(K) = a_{K,n}^{-1}$ and

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right] \\ &= \mathbb{E} \left[\mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) |a_{K,n}^{-1} - 1| \right] \\ &= \mathbb{E} \left[\mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{b_{K,n}}{\sqrt{n} - b_{K,n}} \right| \right] \leq \frac{(\log n)/2}{\sqrt{n} - (\log n)/2} \rightarrow 0, \end{aligned}$$

using similar mathematical arguments as above and the fact that $1/(2\sqrt{n}) \leq b_{K,n} \leq (\log n)/2$ when $2 \leq K \leq (\lfloor \sqrt{n} \log n \rfloor - 1)/2$. The proof for the case where $\lfloor \sqrt{n} \log n \rfloor$ is even is similar. \blacksquare

Lemma 2.2. *As $n \rightarrow \infty$, we have*

$$\mathbb{E} \left[\left| \mathbb{1}(K=1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left(\sqrt{n} \times \frac{p(K+1)}{p(K)} + R^K \right) \right| \right] \rightarrow 0,$$

and

$$\mathbb{E} \left[\left| \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) (\sqrt{n} - R^K) \right| \right] \rightarrow 0.$$

PROOF OF LEMMA 2.2. Using Proposition 2.3, $|h_x| \leq M$, $0 \leq p(K+1)/p(K) \leq 2$ and $R^K := (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$, we have

$$\begin{aligned} & \mathbb{E} \left[\left| \mathbb{1}(K=1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left(\sqrt{n} \times \frac{p(K+1)}{p(K)} + R^K \right) \right| \right] \\ &\leq \frac{(1-\tau)M}{A+1} \left(2\sqrt{n} + \frac{\log n}{2} \right) \mathbb{P}(K=1) \rightarrow 0, \end{aligned}$$

since

$$\mathbb{1}(K=1) \left| \sqrt{n} \times \frac{p(K+1)}{p(K)} + R^K \right| \leq \mathbb{1}(K=1) \left(2\sqrt{n} + \left| \frac{1 - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}} \right| \right)$$

$$\leq \mathbb{1}(K = 1) \left(2\sqrt{n} + \frac{\log n}{2} \right).$$

The proof that

$$\mathbb{E} \left[\left| \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) (\sqrt{n} - R^K) \right| \right] \rightarrow 0$$

is similar. ■

Lemma 2.3. *As $n \rightarrow \infty$, we have*

$$\mathbb{E} \left[\left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \rightarrow 0.$$

PROOF OF LEMMA 2.3. First,

$$\begin{aligned} & \mathbb{E} \left[\left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & \leq \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right], \end{aligned}$$

because $|h_x| \leq M$. Considering the case where $\lfloor \sqrt{n} \log n \rfloor$ is odd, we have

$$\begin{aligned} & \mathbb{E} \left[\left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & = \mathbb{E} \left[\left| \mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & \quad + \mathbb{E} \left[\left| \mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right]. \end{aligned}$$

We analyse each term separately. The first one corresponds to the case where K is at the mode or at its right. Therefore, $p(K+1)/p(K) = a_{K,n}$ and

$$\begin{aligned} & \mathbb{E} \left[\left| \mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & = \mathbb{E} \left[\left| \mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left[\sqrt{n} (a_{K,n} - 1) + R^K \right] \right| \right] \\ & = \mathbb{E} \left[\left| \mathbb{1} \left(\frac{1}{2\sqrt{n}} \leq R^K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 2}{2\sqrt{n}} \right) \left[-b_{K,n} + R^K \right] \right| \right] = 0, \end{aligned}$$

because $a_{K,n} = 1 - b_{K,n}/\sqrt{n}$ and $b_{K,n} = |K - \lfloor \sqrt{n} \log n \rfloor / 2| / \sqrt{n} = R^K$ when $R^K \geq 0$ ($R^K := (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$). We now study the case where K is at the left of the mode. Therefore, $p(K+1)/p(K) = a_{K,n}^{-1}$ and

$$\mathbb{E} \left[\left| \mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left[\sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right]$$

$$\begin{aligned}
&= \mathbb{E} \left[\mathbf{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \sqrt{n} (a_{K,n}^{-1} - 1) + R^K \right| \right] \\
&= \mathbb{E} \left[\mathbf{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{\sqrt{n}}{\sqrt{n} - b_{K,n}} \times b_{K,n} + R^K \right| \right] \\
&= \mathbb{E} \left[\mathbf{1} \left(\frac{4 - \lfloor \sqrt{n} \log n \rfloor}{2\sqrt{n}} \leq R^K \leq \frac{-1}{2\sqrt{n}} \right) \times -R^K \left| \frac{\sqrt{n}}{\sqrt{n} - b_{K,n}} - 1 \right| \right] \\
&\leq \mathbb{E} \left[\mathbf{1} \left(\frac{4 - \lfloor \sqrt{n} \log n \rfloor}{2\sqrt{n}} \leq R^K \leq \frac{-1}{2\sqrt{n}} \right) \frac{\log n}{2} \times \frac{b_{K,n}}{\sqrt{n} - b_{K,n}} \right] \\
&\leq \frac{\log n}{2} \times \frac{(\log n)/2}{\sqrt{n} - (\log n)/2} \rightarrow 0,
\end{aligned}$$

using similar mathematical arguments as above, the fact that $b_{K,n} = -R^K$ when $R^K < 0$, and that $1/(2\sqrt{n}) \leq -R^K \leq (\log n)/2$ when $(4 - \lfloor \sqrt{n} \log n \rfloor)/(2\sqrt{n}) \leq R^K \leq -1/(2\sqrt{n})$. The proof for the case where $\lfloor \sqrt{n} \log n \rfloor$ is even is similar. \blacksquare

Proposition 2.3. *The random variable K with PMF p defined in Section 2.2 is such that, for all $\rho \in \mathbb{R}$,*

$$\lim_{n \rightarrow \infty} n^\rho \mathbb{P}(K = 1) = \lim_{n \rightarrow \infty} n^\rho \mathbb{P}(K = \lfloor \sqrt{n} \log n \rfloor) = 0.$$

PROOF OF PROPOSITION 2.3. Consider the case where $\lfloor \sqrt{n} \log n \rfloor$ is even. Using equation (2.4), we have

$$p(1) = p(\lfloor \sqrt{n} \log n \rfloor) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2}\right) \prod_{i=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \left(1 - \frac{i}{n}\right).$$

In the proof of Proposition 2.1 in Section 2.8, we show that $p(\lfloor \sqrt{n} \log n \rfloor/2) \rightarrow 1/\sqrt{2\pi}$ as $n \rightarrow \infty$. Also, using the fact that $1 - x \leq \exp\{-x\}$ for all $x \in \mathbb{R}$, we have for all $\rho \in \mathbb{R}$

$$\begin{aligned}
n^\rho \prod_{i=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \left(1 - \frac{i}{n}\right) &\leq n^\rho \exp\left\{-\sum_{i=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \frac{i}{n}\right\} \\
&= n^\rho \exp\left\{-\frac{1}{2} \left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1\right)^2 \frac{1}{\sqrt{n}} - \frac{\lfloor \sqrt{n} \log n \rfloor}{2} \frac{1}{2n}\right\} \\
&\leq n^\rho \exp\left\{-\frac{1}{2} \left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2\sqrt{n}} - 1\right)^2\right\} \rightarrow 0,
\end{aligned}$$

as $n \rightarrow \infty$. Similarly, we can show the result for the case where $\lfloor \sqrt{n} \log n \rfloor$ is odd. \blacksquare

2.8. PROOFS OF PROPOSITIONS 2.1 AND 2.2

PROOF OF PROPOSITION 2.1. The random variable $Z_1^n(t)$ is defined as $(K(\lfloor nt \rfloor) - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$, and $K(\lfloor nt \rfloor) \sim p$ for all t and for all n (see Section 2.2 for the assumptions on p). Therefore, to simplify the notation, the time index is omitted for the rest of the proof. Consider the constant $z < 0$ and the case where $\lfloor \sqrt{n} \log n \rfloor$ is even. We have

$$\begin{aligned} \mathbb{P}((K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n} \leq z) &= \mathbb{P}(K \leq z \sqrt{n} + \lfloor \sqrt{n} \log n \rfloor / 2) \\ &= \sum_{k=\lceil -z \sqrt{n} \rceil}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} p(\lfloor \sqrt{n} \log n \rfloor / 2 - k) \\ &= p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2}\right) \sum_{k=\lceil -z \sqrt{n} \rceil}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right), \end{aligned}$$

using $\lfloor z \sqrt{n} + \lfloor \sqrt{n} \log n \rfloor / 2 \rfloor = \lfloor z \sqrt{n} \rfloor + \lfloor \sqrt{n} \log n \rfloor / 2 = \lfloor \sqrt{n} \log n \rfloor / 2 - \lceil -z \sqrt{n} \rceil$ in the second equality ($\lceil \cdot \rceil$ is the ceiling function), and equation (2.4) in the last equality. The sum above is well-defined if $n \geq \exp(-2z + 5)$ and we select n large enough to ensure this. Using again equation (2.4) and the fact that $\sum_{k=1}^{\lfloor \sqrt{n} \log n \rfloor} p(k) = 1$, we have

$$p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2}\right) = \left(2 \left(1 + \sum_{k=1}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right)\right)\right)^{-1}.$$

Therefore,

$$\mathbb{P}\left(\frac{K - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}} \leq z\right) = \frac{(1/\sqrt{n}) \sum_{k=\lceil -z \sqrt{n} \rceil}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} \prod_{i=1}^k (1 - i/n)}{\frac{2}{\sqrt{n}} + \frac{2}{\sqrt{n}} \sum_{k=1}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} \prod_{i=1}^k (1 - i/n)}.$$

Using the fact that $1 - x \leq \exp\{-x\}$ for all $x \in \mathbb{R}$, we have

$$\prod_{i=1}^k \left(1 - \frac{i}{n}\right) \leq \exp\left\{-\sum_{i=1}^k \frac{i}{n}\right\} = \exp\left\{-\frac{1}{2} \left(\frac{k}{\sqrt{n}}\right)^2 - \frac{k}{2n}\right\}.$$

In addition, for all $\delta > 0$, there exists $\epsilon > 0$ such that $\exp\{-(1 + \delta)x\} \leq 1 - x$ for $0 \leq x < \epsilon$. Therefore, since $0 \leq i/n \leq \lfloor \sqrt{n} \log n \rfloor / (2n) - 1/n \leq \log n / (2\sqrt{n}) \rightarrow 0$ as $n \rightarrow \infty$ when $1 \leq i \leq k \leq \lfloor \sqrt{n} \log n \rfloor / 2 - 1$, for all $\delta > 0$, there exists a constant $N > 0$ such that for all $n \geq N$,

$$\prod_{i=1}^k \left(1 - \frac{i}{n}\right) \geq \exp\left\{-\frac{(1 + \delta)}{2} \left(\frac{k}{\sqrt{n}}\right)^2 - \frac{k(1 + \delta)}{2n}\right\}.$$

The objective is to use a ‘‘Riemann sum’’ argument, where the length of the subintervals of the partition is $1/\sqrt{n}$, to study the asymptotic behaviour of the numerator and denominator of $\mathbb{P}((K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n} \leq z)$. More precisely, we now prove that the numerator of $\mathbb{P}((K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n} \leq z)$ converges towards $\int_{-z}^{\infty} \exp(-x^2/2) dx$ and that the denominator

converges towards $\int_{-\infty}^{\infty} \exp(-x^2/2)dx = \sqrt{2\pi}$. To achieve this, we use Lebesgue's dominated convergence theorem. First, we rewrite the numerator as

$$\frac{1}{\sqrt{n}} \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) = \int_{-z}^{\infty} \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) dx.$$

Now, we analyse the integrand. For all $x \in (-z, \infty)$ and for large enough n , there exists a unique $k' \in \{\lceil -z\sqrt{n} \rceil, \dots, \lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1\}$ with $\mathbb{1}_{\left[\frac{k'}{\sqrt{n}}, \frac{k'+1}{\sqrt{n}}\right)}(x) = 1$. Also, $0 \leq x - k'/\sqrt{n} < 1/\sqrt{n}$, which implies that $k'/\sqrt{n} \rightarrow x$ as $n \rightarrow \infty$. Consequently, using the upper bound and the lower bound on $\prod_{i=1}^k (1 - i/n)$,

$$\sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) = \prod_{i=1}^{k'} \left(1 - \frac{i}{n}\right) \rightarrow \exp\{-x^2/2\},$$

as $n \rightarrow \infty$, because $k'/n \leq \lfloor \frac{\sqrt{n} \log n}{2} \rfloor / (2n) - 1/n \leq \log n / (2\sqrt{n}) \rightarrow 0$. Now, we prove that the integrand is bounded by an integrable function that does not depend on n . For all $x \in (-z, \infty)$,

$$\begin{aligned} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) &\leq \exp\left\{-\frac{1}{2} \left(\frac{k}{\sqrt{n}}\right)^2 - \frac{k}{2n}\right\} \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) \\ &\leq \exp\left\{-\frac{1}{2}(x-1)^2\right\} \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x), \end{aligned}$$

using the upper bound on $\prod_{i=1}^k (1 - i/n)$ in the first inequality, and then $x \leq (k+1)/\sqrt{n} \leq (k/\sqrt{n}) + 1$. As a result,

$$\begin{aligned} \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) &\leq \exp\left\{-\frac{1}{2}(x-1)^2\right\} \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) \\ &= \exp\left\{-\frac{1}{2}(x-1)^2\right\} \mathbb{1}_{\left[\frac{\lceil -z\sqrt{n} \rceil}{\sqrt{n}}, \frac{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor}{\sqrt{n}}\right)}(x) \leq \exp\left\{-\frac{1}{2}(x-1)^2\right\}, \end{aligned}$$

which is integrable. Similarly, we can prove that the denominator of $\mathbb{P}((K - \lfloor \frac{\sqrt{n} \log n}{2} \rfloor) / \sqrt{n} \leq z)$ converges towards $\int_{-\infty}^{\infty} \exp(-x^2/2)dx = \sqrt{2\pi}$, and we can show the result for $z \geq 0$ and for the case where $\lfloor \frac{\sqrt{n} \log n}{2} \rfloor$ is odd. ■

PROOF OF PROPOSITION 2.2. The random variables $K(m)$ and $U(m+1)$ are independent and such that $K(m) \sim p$ and $U(m+1) \sim q$ for all $m \in \mathbb{N}$ (see Section 2.2 for the assumptions on p and q). Therefore, to simplify the notation, the time index is omitted for the rest of the proof. As explained in Section 2.6.2,

$$\mathbb{E} \left[1 \wedge \frac{f(U)p(K+1)}{q(U)p(K)A} \right] = \mathbb{E} \left[\frac{f(U)p(K+1)}{q(U)p(K)A} \right],$$

and $\mathbb{E}[f(U)/q(U)] = 1$. Finally, using Proposition 2.3, we have

$$\frac{1}{A} \mathbb{E} \left[\frac{p(K+1)}{p(K)} \right] = \frac{1}{A} \sum_{k=1}^{\lfloor \sqrt{n \log n} \rfloor - 1} p(k+1) = \frac{1}{A} (1 - p(1)) \rightarrow \frac{1}{A} \text{ as } n \rightarrow \infty.$$

■

REFERENCES

- Al-Awadhi, F., M. Hurn and C. Jennison. 2004, Improving the acceptance rate of reversible jump MCMC proposals, *Statist. Probab. Lett.*, vol. 69, n° 2, p. 189–198.
- Bédard, M. 2007, Weak convergence of Metropolis algorithms for non-i.i.d. target distributions, *Ann. Appl. Probab.*, vol. 17, p. 1222–1244.
- Bédard, M. 2008, Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234, *Stochastic Process. Appl.*, vol. 118, n° 12, p. 2198–2222.
- Bédard, M., R. Douc and E. Moulines. 2012, Scaling analysis of multiple-try MCMC methods, *Stochastic Process. Appl.*, vol. 122, n° 3, p. 758–786.
- Beskos, A., N. Pillai, G. Roberts, Sanz-Serna, Jesus-Maria and A. Stuart. 2013, Optimal tuning of the hybrid Monte Carlo algorithm, *Bernoulli*, vol. 19, n° 5A, p. 1501–1534.
- Beskos, A., G. Roberts and A. Stuart. 2009, Optimal scalings for local Metropolis-Hastings chains on non-product targets in high dimensions, *Ann. Appl. Probab.*, vol. 19, n° 3, p. 863–898.
- Bibby, B. M., I. M. Skovgaard and M. Sørensen. 2005, Diffusion-type models with given marginal distribution and autocorrelation function, *Bernoulli*, vol. 11, n° 2, p. 191–220.
- Brooks, S. P., P. Giudici and G. O. Roberts. 2003, Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 65, n° 1, p. 3–39.
- Ethier, S. N. and T. G. Kurtz. 1986, *Markov processes: Characterization and convergence*, Wiley.
- Green, P. J. 1995, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, vol. 82, n° 4, p. 711–732.
- Hastie, D. 2005, *Towards automatic reversible jump markov chain monte carlo*, Ph.D. Thesis, University of Bristol.
- Hastings, W. K. 1970, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, vol. 57, n° 1, p. 97–109.
- Kang, B. 2013, Fast determinantal point process sampling with application to clustering, in *Advances in Neural Information Processing Systems*, p. 2319–2327.
- Karagiannis, G. and C. Andrieu. 2013, Annealed importance sampling reversible jump MCMC algorithms, *J. Comp. Graph. Stat.*, vol. 22, n° 3, p. 623–648.
- Mattingly, J. C., N. S. Pillai and A. Stuart. 2012, Diffusion limits of the random walk Metropolis algorithm in high dimensions, *Ann. Appl. Probab.*, vol. 22, n° 3, p. 881–930.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. 1953, Equation of state calculations by fast computing machines, *J. Chem. Phys.*, vol. 21, p. 1087.

- Neal, P. and G. O. Roberts. 2006, Optimal scaling for partially updating MCMC algorithms, *Ann. Appl. Probab.*, vol. 16, n° 2, p. 475–515.
- Peskun, P. 1973, Optimum Monte-Carlo sampling using Markov chains, *Biometrika*, vol. 60, n° 3, p. 607–612.
- Richardson, S. and P. J. Green. 1997, On Bayesian analysis of mixtures with an unknown number of components (with discussion), *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 59, n° 4, p. 731–792.
- Robert, C. and G. Casella. 2004, Monte Carlo statistical methods, Springer Science & Business Media.
- Roberts, G. O., A. Gelman and W. R. Gilks. 1997, Weak convergence and optimal scaling of random walk Metropolis algorithms, *Ann. Appl. Probab.*, vol. 7, n° 1, p. 110–120.
- Roberts, G. O. and J. S. Rosenthal. 2001, Optimal scaling for various Metropolis-Hastings algorithms, *Statist. Sci.*, vol. 16, n° 4, p. 351–367.
- Tierney, L. 1998, A note on Metropolis-Hastings kernels for general state spaces, *Ann. Appl. Probab.*, vol. 8, p. 1–9.

Chapitre 3

ANNEXE DU CHAPITRE 2

3.1. VÉRIFICATION DU CRITÈRE DE RÉVERSIBILITÉ

Dans la section 2.2, il est indiqué que la chaîne de Markov engendrée par l'algorithme à sauts réversibles satisfait le critère de réversibilité par rapport à π_n . Ceci est vérifié explicitement dans cette section.

Comme il est expliqué dans l'article de Green (1995), il suffit de vérifier pour n'importe quels ensembles A et B que la probabilité de passer d'un état dans A à un état dans B est égale à la probabilité de passer d'un état dans B à un état dans A dans les deux situations suivantes : lors d'une mise à jour des paramètres et lors d'un changement de modèle. Green (1995) explique aussi qu'il suffit de considérer seulement le cas où le candidat est accepté (et donc, qu'il y a déplacement). Pour ce qui est de la mise à jour des paramètres, la probabilité de passer d'un état dans A à un état dans B , lorsque le candidat \mathbf{y}^k est accepté, est donnée par

$$p(k) \int_A \prod_{i=1}^{n+k} f(x_i^k) g(1) \int_B \prod_{i=1}^{n+k} \sqrt{\frac{n+k}{2\pi\ell^2}} \exp\left\{-\frac{n+k}{2\ell^2}(y_i^k - x_i^k)^2\right\} \left(1 \wedge \frac{\prod_{i=1}^{n+k} f(y_i^k)}{\prod_{i=1}^{n+k} f(x_i^k)}\right) d\mathbf{y}^k d\mathbf{x}^k,$$

et, en utilisant le théorème de Fubini, cette probabilité est égale à

$$p(k) \int_B \prod_{i=1}^{n+k} f(y_i^k) g(1) \int_A \prod_{i=1}^{n+k} \sqrt{\frac{n+k}{2\pi\ell^2}} \exp\left\{-\frac{n+k}{2\ell^2}(x_i^k - y_i^k)^2\right\} \left(1 \wedge \frac{\prod_{i=1}^{n+k} f(x_i^k)}{\prod_{i=1}^{n+k} f(y_i^k)}\right) d\mathbf{x}^k d\mathbf{y}^k,$$

qui est la probabilité de passer d'un état dans B à un état dans A lorsque le candidat \mathbf{x}^k est accepté. On note que cette égalité est valide pour toute valeur de $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$.

Pour ce qui est du changement de modèle, la probabilité de passer d'un modèle $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor - 1\}$, où les $n+k$ paramètres se trouvent dans l'ensemble A , au modèle $k+1$, où les $n+k+1$ paramètres se trouvent dans l'ensemble $A \times B$, est donnée par

$$p(k) \int_A \prod_{i=1}^{n+k} f(x_i^k) g(2) \int_B q(u) \left(1 \wedge \frac{f(u)p(k+1)g(3)}{q(u)p(k)g(2)}\right) du d\mathbf{x}^k.$$

Notez que lorsqu'on passe du modèle k au modèle $k + 1$, les $n + k$ paramètres présents dans le modèle k ne se déplacent pas. En utilisant le théorème de Fubini, cette probabilité est égale à

$$p(k + 1) \int_{A \times B} \prod_{i=1}^{n+k} f(x_i^k) f(u) g(3) \left(1 \wedge \frac{q(u) p(k) g(2)}{f(u) p(k + 1) g(3)} \right) d\mathbf{x}^k du,$$

qui est la probabilité de passer du modèle $k + 1$, où les paramètres se trouvent dans l'ensemble $A \times B$, au modèle k car, lorsque ce déplacement est effectué, on retire seulement le dernier paramètre (les $n + k$ premiers paramètres ne se déplacent pas). On note que lorsque $k = \lfloor \sqrt{n} \log n \rfloor$, la probabilité de passer au modèle $k + 1$ est égale à 0 car $p(k + 1) = 0$. Pour la même raison, la probabilité de se trouver au modèle $k = \lfloor \sqrt{n} \log n \rfloor + 1$ est égale à 0. Ainsi, pour toute valeur de $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$, la probabilité de passer du modèle k au modèle $k + 1$ est égale à la probabilité de passer du modèle $k + 1$ au modèle k . Le critère de réversibilité est donc vérifié.

3.2. THÉORÈME 8.2 DU CHAPITRE 4 DE [ETHIER ET KURTZ \(1986\)](#)

Comme il est mentionné dans la section 2.6, le théorème 8.2 du chapitre 4 de [Ethier et Kurtz \(1986\)](#) joue un rôle important dans la démonstration du théorème 2.1. L'énoncé du théorème 8.2 fait appel à la définition suivante provenant de la page 226 de [Ethier et Kurtz \(1986\)](#) :

Définition 3.1. *For $n = 1, 2, \dots$, let $\{\mathcal{G}_t^n\}$ be a complete filtration, and let \mathcal{L}_n be the space of real-valued $\{\mathcal{G}_t^n\}$ -progressive processes $\{\xi_n(t), t \geq 0\}$ satisfying*

$$\sup_{t \leq T} \mathbb{E}[|\xi_n(t)|] < \infty,$$

for each $T > 0$. Let $\hat{\mathcal{A}}_n$ be the collection of pairs $(\{\xi_n(t), t \geq 0\}, \{\varphi_n, t \geq 0\}) \in \mathcal{L}_n \times \mathcal{L}_n$ such that

$$\xi_n(t) = \int_0^t \varphi_n(s) ds$$

is a $\{\mathcal{G}_t^n\}$ -martingale.

Voici l'énoncé du théorème 8.2 (notez que tous les concepts et notations qui sont utilisés dans l'énoncé, mais qui ne sont pas encore définis, le seront ultérieurement) :

Théorème 3.1 (8.2). *Let (E, r) be complete and separable. Let $B \subset \overline{C}(E) \times \overline{C}(E)$ be linear, and suppose the closure of B generates a strongly continuous contraction semigroup $\{T(t)\}$ on $L \equiv \overline{\mathcal{D}(B)}$. Suppose $\{\mathbf{X}_n(t), t \geq 0\}, n = 1, 2, \dots$, is a $\{\mathcal{G}_t^n\}$ -progressive E -valued process, $\{\mathbf{X}(t), t \geq 0\}$ is a Markov process corresponding to $\{T(t)\}$, and $\mathbf{X}_n(0) \Rightarrow \mathbf{X}(0)$. Let $\mathcal{M} \subset \overline{C}(E)$ be separating and suppose either L is separating and $\{\mathbf{X}_n(t), n = 1, 2, \dots\}$ is relatively compact for each $t \geq 0$, or L is convergence determining. Then the following are equivalent:*

- (a) *The finite-dimensional distributions of $\{\mathbf{X}_n(t), t \geq 0\}$ converge weakly to those of $\{\mathbf{X}(t), t \geq 0\}$.*
- (b) *For each $(h, g) \in B$,*

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\left(h(\mathbf{X}_n(t + s)) - h(\mathbf{X}_n(t)) - \int_t^{t+s} g(\mathbf{X}_n(u)) du \right) \prod_{i=1}^k f_i(\mathbf{X}_n(t_i)) \right] = 0,$$

for all $k \geq 0, 0 \leq t_1 < t_2 < \dots < t_k \leq t < t + s$, and $f_1, \dots, f_k \in \mathcal{M}$.

(c) For each $(h, g) \in B$ and $T > 0$, there exist $(\{\xi_n(t), t \geq 0\}, \{\varphi_n, t \geq 0\}) \in \hat{\mathcal{A}}_n$ such that

$$\sup_n \sup_{s \leq T} \mathbb{E}[|\xi_n(s)|] < \infty,$$

$$\sup_n \sup_{s \leq T} \mathbb{E}[|\varphi_n(s)|] < \infty,$$

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[(\xi_n(t) - h(\mathbf{X}_n(t))) \prod_{i=1}^k f_i(\mathbf{X}_n(t_i)) \right] = 0,$$

and

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[(\varphi_n(t) - g(\mathbf{X}_n(t))) \prod_{i=1}^k f_i(\mathbf{X}_n(t_i)) \right] = 0,$$

for all $k \geq 0, 0 \leq t_1 < t_2 < \dots < t_k \leq t \leq T$, and $f_1, \dots, f_k \in \mathcal{M}$.

Nous utilisons l'équivalence entre le résultat (a) et le résultat (c) afin de démontrer le théorème 2.1. Afin d'utiliser cette équivalence, les conditions mentionnées dans l'énoncé du théorème 3.1 doivent donc être vérifiées. La section 3.2.1 est dédiée à ceci. Notez que le résultat (a) indique que la convergence en distribution suivante est valide : $(\mathbf{X}_n(t_1), \dots, \mathbf{X}_n(t_k))^T \Rightarrow (\mathbf{X}(t_1), \dots, \mathbf{X}(t_k))^T$ pour n'importe quel ensemble $\{t_1, \dots, t_k\}$, où k est un entier positif.

3.2.1. Vérification des conditions du théorème 8.2

Tout d'abord, on vérifie que l'espace (E, r) considéré au chapitre 2 est complet et séparable. En fait, il s'agit de l'espace euclidien de dimension 2, soit un espace complet et séparable.

Ensuite, on détermine la forme de B en utilisant le théorème 2.5 du chapitre 8 de Ethier et Kurtz (1986). Voici l'énoncé du théorème 2.5 :

Théorème 3.2 (2.5). Let $a : \mathbb{R}^d \rightarrow S^d$ satisfy $a_{ij} \in C^2(\mathbb{R}^d)$ with $\partial_k \partial_l a_{ij}$ bounded for $i, j, k, l = 1, \dots, d$ and let $b : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be Lipschitz continuous, i.e.

$$\sup_{\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{y}} |b(\mathbf{x}) - b(\mathbf{y})| / |\mathbf{x} - \mathbf{y}| < \infty.$$

Then, with G defined by

$$G = \frac{1}{2} \sum_{i,j} a_{ij}(\mathbf{x}) \partial_i \partial_j + \sum_{i=1}^d b_i(\mathbf{x}) \partial_i,$$

the closure of $\{(h, Gh) : h \in C_c^\infty(\mathbb{R}^d)\}$ is single-valued and generates a Feller semi-group on $\hat{C}(\mathbb{R}^d)$.

La notation S^d représente l'espace des matrices de taille $d \times d$ définies non-négatives et $\hat{C}(\mathbb{R}^d)$ représente l'espace des fonctions qui tendent vers 0 aux bornes du domaine \mathbb{R}^d .

On vérifie les hypothèses du théorème 3.2 afin d'utiliser son résultat. Le processus limite analysé au chapitre 2 est $\{\mathbf{Z}(t), t \geq 0\}$ (représentant le processus noté $\{\mathbf{X}(t), t \geq 0\}$ dans le théorème 3.1). Il s'agit d'un processus markovien qui a été défini dans l'énoncé du théorème 2.1. Son générateur

G est tel que $G = G_1 + G_2$, où

$$G_1 h(\mathbf{Z}(t)) = \tau \ell^2 \Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) (\log f(Z_2(t)))' h_y(\mathbf{Z}(t)) + \tau \ell^2 \Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) h_{yy}(\mathbf{Z}(t)),$$

$$G_2 h(\mathbf{Z}(t)) = \frac{1-\tau}{A+1} \times -Z_1(t) h_x(\mathbf{Z}(t)) + \frac{1-\tau}{A+1} h_{xx}(\mathbf{Z}(t)).$$

Donc, $a(\mathbf{x})$ est une matrice diagonale telle que le premier élément sur la diagonale est $2(1-\tau)/(A+1)$ et le deuxième est $2\tau \ell^2 \Phi(-\ell \sqrt{\Upsilon}/2)$. Ainsi, $a(\mathbf{x})$ est une matrice de taille 2×2 , non-négative qui ne dépend pas de \mathbf{x} . Alors, les éléments de $a(\mathbf{x})$ ont des dérivées deuxièmes continues et bornées (elles sont égales à 0).

Le premier élément de la fonction $b(\mathbf{x})$ est $(1-\tau)/(A+1) \times -x_1$ et le deuxième est $\tau \ell^2 \Phi(-\ell \sqrt{\Upsilon}/2) (\log f(x_2))'$ (en considérant que $\mathbf{x}^T := (x_1, x_2)$). Ces deux fonctions sont Lipschitz continues (la fonction $(\log f(x_2))'$ est Lipschitz continue par hypothèse), donc b est Lipschitz continue. Ainsi, les conditions du théorème 2.5 sont vérifiées, et la fermeture de $B := \{(h, Gh) : h \in C_c^\infty(\mathbb{R}^2)\}$ est à valeur unique et génère un semi-groupe de Feller (donc en particulier elle génère un semi-groupe de contractions fortement continues) $\{T(t)\}$ sur $\hat{C}(\mathbb{R}^2)$.

Maintenant que nous avons un candidat pour B , on s'assure qu'il respecte les conditions énoncées au théorème 3.1. On vérifie tout d'abord que B est linéaire. Premièrement, on vérifie que $(h, Gh) + (g, Gg) = (h+g, G(h+g)) \in B$ si $(h, Gh), (g, Gg) \in B$. Pour la première composante, si $h, g \in C_c^\infty(\mathbb{R}^2)$, il est clair que $h+g$ est à support compact. Aussi, la dérivée d'ordre quelconque de $h+g$ est la somme des dérivées de h et g . Donc, $h+g$ est infiniment dérivable. Pour la deuxième composante,

$$\begin{aligned} Gh(x) + Gg(x) &= \frac{1}{2} \sum_{i,j}^2 a_{ij}(x) \partial_i \partial_j h(x) + \sum_{i=1}^2 b_i(x) \partial_i h(x) \\ &\quad + \frac{1}{2} \sum_{i,j}^2 a_{ij}(x) \partial_i \partial_j g(x) + \sum_{i=1}^2 b_i(x) \partial_i g(x) \\ &= \frac{1}{2} \sum_{i,j}^2 a_{ij}(x) \partial_i \partial_j (h(x) + g(x)) + \sum_{i=1}^2 b_i(x) \partial_i (h(x) + g(x)) \\ &= G(h(x) + g(x)). \end{aligned}$$

Ensuite, on vérifie que $c(h, Gh) = (ch, Gch) \in B$ si $(h, Gh) \in B$ et $c \in \mathbb{R}$. Pour la première composante, si $h \in C_c^\infty(\mathbb{R}^2)$, il est clair que ch est à support compact et infiniment dérivable. Pour la deuxième composante, on peut vérifier de la même façon que précédemment que $cGh = Gch$. Alors, B est linéaire.

On vérifie maintenant que $B \subset \overline{C}(\mathbb{R}^2) \times \overline{C}(\mathbb{R}^2)$. La notation $\overline{C}(\mathbb{R}^2)$ représente l'espace des fonctions continues bornées ayant comme domaine \mathbb{R}^2 . On sait que $C_c^\infty(\mathbb{R}^2) \subset \overline{C}(\mathbb{R}^2)$. On sait aussi

que les éléments de $a(\mathbf{x})$ sont des constantes et donc que

$$\frac{1}{2} \sum_{i,j}^2 a_{ij}(\mathbf{x}) \partial_i \partial_j h(\mathbf{x}) \in C_c^\infty(\mathbb{R}^2) \subset \overline{C}(\mathbb{R}^2).$$

Par ailleurs, on sait que $b_i(\mathbf{x})$ est Lipschitz continue pour $i = 1, 2$, donc

$$\sum_{i=1}^2 b_i(\mathbf{x}) \partial_i h(\mathbf{x}) \in \overline{C}(\mathbb{R}^2),$$

car $b_i(\mathbf{x}) \partial_i h(\mathbf{x})$ est une fonction continue à support compact. Ainsi, $\{(h, Gh) : h \in C_c^\infty(\mathbb{R}^2)\} \subset \overline{C}(\mathbb{R}^2) \times \overline{C}(\mathbb{R}^2)$. Alors, B respecte les conditions énoncées dans le théorème 3.1, ce qui implique qu'on peut se restreindre à l'ensemble des fonctions $h \in C_c^\infty(\mathbb{R}^2)$ lorsqu'on étudie la convergence du pseudo-générateur, tel que mentionné à la section 2.6.1.

Maintenant, on vérifie que $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$ (représentant le processus noté $\{\mathbf{X}_n(t), t \geq 0\}$ dans le théorème 3.1), $n = 1, 2, \dots$, est $\{\mathcal{F}^{\mathbf{Z}^n}(t)\}$ -progressif à valeur dans \mathbb{R}^2 . Ce processus est défini dans l'énoncé du théorème 2.1.

Définition 3.2. *Un processus $\{\mathbf{X}(t), t \geq 0\}$ est $\{\mathcal{F}_t\}$ -progressif si pour tout $t \geq 0$, la restriction de $\{\mathbf{X}(t), t \geq 0\}$ à $[0, t] \times \Omega$ est $\mathcal{B}[0, t] \times \mathcal{F}_t$ -mesurable.*

Il est donc clair que $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}, n = 1, 2, \dots$, est $\{\mathcal{F}^{\mathbf{Z}^n}(t)\}$ -progressif à valeur dans \mathbb{R}^2 .

À présent, on vérifie que $\mathbf{Z}_{1,2}^n(0) \Rightarrow \mathbf{Z}(0)$, c'est-à-dire que $\mathbf{Z}_{1,2}^n(0)$ converge vers $\mathbf{Z}(0)$ en distribution lorsque $n \rightarrow \infty$. Comme il est indiqué au théorème 2.1, on suppose que les processus $\{\mathbf{Z}^n(t), t \geq 0\}$ et $\{\mathbf{Z}(t), t \geq 0\}$ démarrent à stationnarité. Ainsi, $Z_1(0) \sim \mathcal{N}(0, 1), Z_2(0) \sim f, Z_2^n(0) = X_1^{K(0)}(0) \sim f$, et

$$Z_1^n(0) = \frac{K(0) - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}},$$

où $K(0) \sim p$. La variable aléatoire $K(0)$ n'a pas d'impact sur $X_1^{K(0)}(0)$ (voir section 2.2 pour les hypothèses sur la distribution stationnaire π_n de $\{\mathbf{Z}^n(t), t \geq 0\}$). Les variables aléatoires $K(0)$ et $X_1^{K(0)}(0)$ sont donc indépendantes. Alors,

$$\mathbb{P}(\mathbf{Z}_{1,2}^n(0) \leq \mathbf{z}) = \mathbb{P}(Z_1^n(0) \leq z_1) \mathbb{P}(Z_2^n(0) \leq z_2),$$

où $\mathbf{z} := (z_1, z_2)^T \in \mathbb{R}^2$. Puisque $Z_2^n(0) \sim f$ et $Z_2(0) \sim f$, on n'a qu'à vérifier que $Z_1^n(0) \Rightarrow Z_1(0)$ afin de démontrer que $\mathbf{Z}_{1,2}^n(0) \Rightarrow \mathbf{Z}(0)$, car $Z_1(0)$ et $Z_2(0)$ sont indépendantes. La proposition 2.1 permet de confirmer que $Z_1^n(0) \Rightarrow Z_1(0) \sim \mathcal{N}(0, 1)$.

Afin de terminer la vérification des conditions du théorème 3.1, on démontre que L est suffisant pour déterminer la convergence (*convergence determining*). En fait, ce résultat est obtenu directement en considérant que, tel qu'expliqué à la section 3.5.1, $C_c^\infty(\mathbb{R}^2) \subset L$ est *convergence determining*.

Les conditions mentionnées dans l'énoncé du théorème 3.1 sont donc vérifiées et l'équivalence entre le résultat (a) et le résultat (c) de ce théorème peut être utilisée. C'est ce qui est fait dans la section 2.6.

3.3. COMPLÉMENT DE LA DÉMONSTRATION DU THÉORÈME 2.1 :

$$\mathbb{E} \left[\left| \varphi_{1,n}(t) - G_1 h(\mathbf{Z}_{1,2}^n(t)) \right| \right] \rightarrow 0$$

Tel qu'expliqué à la section 2.6, la démonstration que

$$\mathbb{E} \left[\left| \varphi_{1,n}(t) - G_1 h(\mathbf{Z}_{1,2}^n(t)) \right| \right] \rightarrow 0 \text{ lorsque } n \rightarrow \infty,$$

est semblable à celle du théorème 1.1 de Roberts *et al.* (1997). Elle est néanmoins détaillée dans cette section. On sait que

$$\varphi_{1,n}(t) := n\tau \mathbb{E} \left[\left(h(R^K, Y_1^K) - h(R^K, X_1^K) \right) \left(1 \wedge \frac{\prod_{i=1}^{n+K} f(Y_i^K)}{\prod_{i=1}^{n+K} f(X_i^K)} \right) \middle| R^K, \mathbf{X}^K \right],$$

et que

$$G_1 h(\mathbf{Z}_{1,2}^n(t)) = \tau \ell^2 \Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) (\log f(\mathbf{Z}_2^n(t)))' h_y(\mathbf{Z}_{1,2}^n(t)) + \tau \ell^2 \Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) h_{yy}(\mathbf{Z}_{1,2}^n(t)).$$

Tout d'abord, on effectue un développement en série de Taylor d'ordre 1 de la probabilité d'acceptation

$$\left(1 \wedge \prod_{i=1}^{n+K} \frac{f(Y_i^K)}{f(X_i^K)} \right) = \left(1 \wedge \exp \left\{ \sum_{i=1}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right),$$

par rapport à Y_1^K , et ceci autour du point X_1^K . Le premier terme de la série est donné par

$$\left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right).$$

Le deuxième terme de la série est donné par

$$\begin{cases} (Y_1 - X_1^K) (\log f(X_1^K))' \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} & \text{si } \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0, \\ 0 & \text{si } \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} > 0. \end{cases}$$

La dérivée n'existe pas lorsque $\sum_{i=2}^{n+K} \log(f(Y_i^K)/f(X_i^K)) = 0$, mais cet événement a probabilité 0 car les variables aléatoires sont continues. Le dernier terme de la série est donné par

$$\begin{cases} \frac{1}{2} (Y_1 - X_1^K)^2 \left((\log f(\epsilon_1^K))'' + ((\log f(\epsilon_1^K))')^2 \right) \exp \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} & \text{si } \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0, \\ 0 & \text{si } \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} > 0, \end{cases}$$

où ϵ_1^K appartient à (X_1^K, Y_1^K) ou à (Y_1^K, X_1^K) . Ainsi,

$$\begin{aligned}
\varphi_{1,n}(t) &= \tau n \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K)) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right] \\
&+ \tau n \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K))(Y_1 - X_1^K)(\log f(X_1^K))' \right. \\
&\times \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \left. \right] \\
&+ \frac{\tau}{2} n \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K))(Y_1 - X_1^K)^2 \left((\log f(\epsilon_1^K))'' + ((\log f(\epsilon_1^K))')^2 \right) \right. \\
&\times \exp \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \left. \right]. \quad (3.1)
\end{aligned}$$

On démontre que le premier terme converge vers $\tau \ell^2 \Phi(-\ell \sqrt{Y}/2) h_{yy}(\mathbf{Z}_{1,2}^n(t))$ (la deuxième partie de $G_1 h(\mathbf{Z}_{1,2}^n(t))$), que le deuxième terme converge vers $\tau \ell^2 \times \Phi(-\ell \sqrt{Y}/2) (\log f(\mathbf{Z}_2^n(t)))' h_y(\mathbf{Z}_{1,2}^n(t))$ (la première partie de $G_1 h(\mathbf{Z}_{1,2}^n(t))$), et que le troisième terme converge vers 0, et ceci dans L^1 lorsque $n \rightarrow \infty$. Ainsi, en utilisant l'inégalité du triangle, on aura que

$$\mathbb{E} \left[\left| \varphi_{1,n}(t) - G_1 h(\mathbf{Z}_{1,2}^n(t)) \right| \right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$. Les démonstrations étant assez longues, elles sont données par l'intermédiaire de lemmes. On commence par la démonstration que le troisième terme converge vers 0 qui est donnée via le lemme 3.1.

Lemme 3.1. *L'espérance suivante :*

$$\begin{aligned}
&\mathbb{E} \left[\left| \frac{\tau}{2} n \mathbb{E} \left[h(R^K, Y_1^K) - h(R^K, X_1^K) (Y_1 - X_1^K)^2 \left((\log f(\epsilon_1^K))'' + ((\log f(\epsilon_1^K))')^2 \right) \right. \right. \right. \\
&\times \exp \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \left. \right] \left. \right| \right],
\end{aligned}$$

converge vers 0 lorsque $n \rightarrow \infty$.

DÉMONSTRATION DU LEMME 3.1. Tout d'abord, on peut faire entrer la valeur absolue dans l'espérance conditionnelle en utilisant l'inégalité de Jensen pour espérance conditionnelle. De plus,

$$0 \leq \exp \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} < 1.$$

Ainsi, l'espérance donnée dans le lemme 3.1 est bornée supérieurement par

$$\begin{aligned}
&\frac{n\tau}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 \left| (h(R^K, Y_1^K) - h(R^K, X_1^K)) \left((\log f(\epsilon_1^K))'' + ((\log f(\epsilon_1^K))')^2 \right) \right| \right] \\
&\leq \frac{n\tau}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 \left| (h(R^K, Y_1^K) - h(R^K, X_1^K)) (\log f(\epsilon_1^K))'' \right| \right] \\
&\quad + \frac{n\tau}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 \left| (h(R^K, Y_1^K) - h(R^K, X_1^K)) ((\log f(\epsilon_1^K))')^2 \right| \right],
\end{aligned}$$

en utilisant l'inégalité du triangle. On démontre maintenant que chacun de ces termes converge vers 0. On débute par le premier. Tel qu'expliqué à la section 2.6, il existe une constante positive M telle que h et toutes ses dérivées sont bornées par M (car $h \in C_c^\infty(\mathbb{R}^2)$). On choisit M de manière à ce que $|(\log f(\epsilon_1^K))'| \leq M$ (la fonction $(\log f(x))'$ est Lipschitz continue par hypothèse). Ainsi,

$$\begin{aligned} & \frac{n\tau}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 |(h(R^K, Y_1^K) - h(R^K, X_1^K))(\log f(\epsilon_1^K))'| \right] \\ & \leq \frac{n\tau M}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 |(h(R^K, Y_1^K) - h(R^K, X_1^K))| \right]. \end{aligned}$$

En effectuant un développement en série de Taylor de la fonction h par rapport à Y_1^K , et ceci autour du point X_1^K , on a

$$h(R^K, Y_1^K) - h(R^K, X_1^K) = \mathbb{1}\{N \geq 1\} \sum_{k=1}^N \frac{h_y^{(k)}(R^K, X_1^K)}{k!} (Y_1^K - X_1^K)^k + \frac{h_y^{(N+1)}(R^K, W)}{k!} (Y_1^K - X_1^K)^{N+1},$$

où W appartient à (X_1^K, Y_1^K) ou à (Y_1^K, X_1^K) et $N \in \{0, 1, \dots\}$. La notation $h_y^{(k)}$ représente la k -ième dérivée de la fonction h par rapport à son deuxième argument. Ainsi, en utilisant un développement avec $N = 0$,

$$\begin{aligned} & \frac{n\tau M}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 |(h(R^K, Y_1^K) - h(R^K, X_1^K))| \right] \\ & = \frac{n\tau M}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 |h_y^{(1)}(R^K, W)(Y_1^K - X_1^K)| \right] \\ & \leq \frac{n\tau M^2}{2} \mathbb{E} \left[|Y_1^K - X_1^K|^3 \right] \\ & = \frac{n\tau M^2}{2} \mathbb{E} \left[\mathbb{E} \left[|Y_1^K - X_1^K|^3 \mid K, \mathbf{X}^K \right] \right] \\ & = \frac{n\tau M^2}{2} \mathbb{E} \left[\frac{2\ell^3}{(n+K)^{3/2}} \sqrt{\frac{2}{\pi}} \right] \leq \frac{n\tau M^2}{2} \frac{2\ell^3}{(n+1)^{3/2}} \sqrt{\frac{2}{\pi}} \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, car $|h_y^{(1)}| \leq M$, $K \geq 1$ et $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$.

On démontre maintenant que le deuxième terme converge vers 0. Premièrement, puisque $(\log f(x))'$ est Lipschitz continue,

$$\begin{aligned} |(\log f(\epsilon_1^K))'| & = |(\log f(\epsilon_1^K))' - (\log f(X_1^K))' + (\log f(X_1^K))'| \\ & \leq |(\log f(\epsilon_1^K))' - (\log f(X_1^K))'| + |(\log f(X_1^K))'| \\ & \leq M|\epsilon_1^K - X_1^K| + |(\log f(X_1^K))'| \leq M|Y_1^K - X_1^K| + |(\log f(X_1^K))'|, \end{aligned}$$

car ϵ_1^K appartient à (X_1^K, Y_1^K) ou à (Y_1^K, X_1^K) . Alors,

$$\begin{aligned} & \frac{n\tau}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 |(h(R^K, Y_1^K) - h(R^K, X_1^K))(\log f(\epsilon_1^K))'|^2 \right] \\ & \leq \frac{n\tau}{2} \mathbb{E} \left[(Y_1^K - X_1^K)^2 |(h(R^K, Y_1^K) - h(R^K, X_1^K))(\log f(X_1^K))'|^2 \right] \end{aligned}$$

$$\begin{aligned}
& + n\tau M \mathbb{E} \left[(Y_1 - X_1^K)^2 |h(R^K, Y_1^K) - h(R^K, X_1^K)| |(\log f(X_1^K))'| |Y_1^K - X_1^K| \right] \\
& + \frac{n\tau M^2}{2} \mathbb{E} \left[(Y_1 - X_1^K)^4 |h(R^K, Y_1^K) - h(R^K, X_1^K)| \right].
\end{aligned}$$

On démontre que chacun de ces termes converge vers 0. On débute par le dernier et on termine avec le premier. En utilisant l'inégalité du triangle,

$$\begin{aligned}
& \frac{n\tau M^2}{2} \mathbb{E} \left[(Y_1 - X_1^K)^4 |h(R^K, Y_1^K) - h(R^K, X_1^K)| \right] \\
& \leq n\tau M^3 \mathbb{E} \left[(Y_1 - X_1^K)^4 \right] = n\tau M^3 \mathbb{E} \left[\mathbb{E} \left[(Y_1 - X_1^K)^4 \mid K, \mathbf{X}^K \right] \right] \\
& = n\tau M^3 \mathbb{E} \left[\frac{3\ell^4}{(n+K)^2} \right] \leq n\tau M^3 \frac{3\ell^4}{(n+1)^2} \rightarrow 0,
\end{aligned}$$

lorsque $n \rightarrow \infty$, car $|h| \leq M$, $K \geq 1$ et $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$.

Ensuite, en utilisant l'inégalité du triangle,

$$\begin{aligned}
& n\tau M \mathbb{E} \left[(Y_1 - X_1^K)^2 |h(R^K, Y_1^K) - h(R^K, X_1^K)| |(\log f(X_1^K))'| |Y_1^K - X_1^K| \right] \\
& \leq 2n\tau M^2 \mathbb{E} \left[|(\log f(X_1^K))'| |Y_1^K - X_1^K|^3 \right] \\
& = 2n\tau M^2 \mathbb{E} \left[|(\log f(X_1^K))'| \mathbb{E} \left[|Y_1^K - X_1^K|^3 \mid K, \mathbf{X}^K \right] \right] \\
& = 2n\tau M^2 \mathbb{E} \left[|(\log f(X_1^K))'| \frac{2\ell^3}{(n+K)^{3/2}} \sqrt{\frac{2}{\pi}} \right] \\
& \leq 2n\tau M^2 \frac{2\ell^3}{(n+1)^{3/2}} \sqrt{\frac{2}{\pi}} \mathbb{E} \left[|(\log f(X_1^K))'| \right] \rightarrow 0,
\end{aligned}$$

lorsque $n \rightarrow \infty$, car $|h| \leq M$, $K \geq 1$, $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$ et $\mathbb{E}[|(\log f(X_1^K))'|] < \infty$ ($\mathbb{E}[|(\log f(X_1^K))'|^4] < \infty$ par hypothèse).

Finalement, en utilisant un développement avec $N = 0$,

$$\begin{aligned}
& \frac{n\tau}{2} \mathbb{E} \left[(Y_1 - X_1^K)^2 |h(R^K, Y_1^K) - h(R^K, X_1^K)| |(\log f(X_1^K))'| \right] \\
& = \frac{n\tau}{2} \mathbb{E} \left[(Y_1 - X_1^K)^2 |h_y^{(1)}(R^K, W)(Y_1^K - X_1^K)| |(\log f(X_1^K))'| \right] \\
& \leq \frac{n\tau M}{2} \mathbb{E} \left[|Y_1 - X_1^K|^3 |(\log f(X_1^K))'| \right] \\
& = \frac{n\tau M}{2} \mathbb{E} \left[\mathbb{E} \left[|Y_1 - X_1^K|^3 \mid K, \mathbf{X}^K \right] |(\log f(X_1^K))'| \right] \\
& = \frac{n\tau M}{2} \mathbb{E} \left[\frac{2\ell^3}{(n+K)^{3/2}} \sqrt{\frac{2}{\pi}} |(\log f(X_1^K))'| \right] \\
& \leq \frac{n\tau M}{2} \frac{2\ell^3}{(n+1)^{3/2}} \sqrt{\frac{2}{\pi}} \mathbb{E} \left[|(\log f(X_1^K))'| \right] \rightarrow 0,
\end{aligned}$$

lorsque $n \rightarrow \infty$, car $|h_y^{(1)}| \leq M$, $K \geq 1$, $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$ et $\mathbb{E}[(\log f(X_1^K))'^2] < \infty$. ■

Maintenant, on démontre que le premier terme de (3.1), donné par

$$n\tau\mathbb{E}\left[\left(h(R^K, Y_1^K) - h(R^K, X_1^K)\right)\left(1 \wedge \exp\left\{\sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)}\right\}\right) \mid R^K, \mathbf{X}^K\right],$$

converge vers $\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)h_{yy}(R^K, X_1^K)$ dans L^1 lorsque $n \rightarrow \infty$ par l'intermédiaire des lemmes 3.2 à 3.4.

Lemme 3.2. *On a*

$$\mathbb{E}\left[\left|n\tau\mathbb{E}\left[\left(h(R^K, Y_1^K) - h(R^K, X_1^K)\right)\left(1 \wedge \exp\left\{\sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)}\right\}\right) \mid R^K, \mathbf{X}^K\right] - \frac{\tau\ell^2}{2}\mathbb{E}\left[h_{yy}(R^K, X_1^K)\left(1 \wedge \exp\left\{\sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)}\right\}\right) \mid R^K, \mathbf{X}^K\right]\right| \right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$.

Lemme 3.3. *On a*

$$\mathbb{E}\left[\left|\frac{\tau\ell^2}{2}h_{yy}(R^K, X_1^K)\mathbb{E}\left[1 \wedge \exp\left\{\sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)}\right\} \mid R^K, \mathbf{X}^K\right] - \frac{\tau\ell^2}{2}h_{yy}(R^K, X_1^K)\mathbb{E}\left[1 \wedge \exp\{W^K\} \mid R^K, \mathbf{X}^K\right]\right| \right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$, où

$$W^K := \sum_{i=2}^{n+K} (\log f(X_i^K))'(Y_i^K - X_i^K) - \frac{\ell^2}{2(n+K)}((\log f(X_i^K))')^2.$$

Lemme 3.4. *On a*

$$\mathbb{E}\left[\left|\frac{\tau\ell^2}{2}h_{yy}(R^K, X_1^K)\mathbb{E}\left[1 \wedge \exp\{W^K\} \mid R^K, \mathbf{X}^K\right] - \tau\ell^2\Phi\left(-\frac{\ell\sqrt{\Upsilon}}{2}\right)h_{yy}(R^K, X_1^K)\right| \right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$, où

$$W^K := \sum_{i=2}^{n+K} (\log f(X_i^K))'(Y_i^K - X_i^K) - \frac{\ell^2}{2(n+K)}((\log f(X_i^K))')^2.$$

DÉMONSTRATION DU LEMME 3.2. Tel qu'expliqué à la démonstration du lemme 3.1, en utilisant un développement en série de Taylor, on a

$$h(R^K, Y_1^K) - h(R^K, X_1^K) = \mathbb{1}\{N \geq 1\} \sum_{k=1}^N \frac{h_y^{(k)}(R^K, X_1^K)}{k!} (Y_1^K - X_1^K)^k + \frac{h_y^{(N+1)}(R^K, W)}{k!} (Y_1^K - X_1^K)^{N+1},$$

où W appartient à (X_1^K, Y_1^K) ou à (Y_1^K, X_1^K) et $N \in \{0, 1, \dots\}$. Donc, en utilisant un développement avec $N = 2$, on obtient

$$\begin{aligned}
& n\tau \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K)) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right] \\
&= n\tau \mathbb{E} \left[h_y(R^K, X_1^K)(Y_1^K - X_1^K) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right] \\
&\quad + n\tau \frac{1}{2} \mathbb{E} \left[h_{yy}(R^K, X_1^K)(Y_1^K - X_1^K)^2 \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right] \\
&\quad + n\tau \frac{1}{6} \mathbb{E} \left[h_y^{(3)}(R^K, W)(Y_1^K - X_1^K)^3 \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right].
\end{aligned}$$

On démontre que le premier terme est égal à 0, que le deuxième converge vers

$$\frac{\tau \ell^2}{2} \mathbb{E} \left[h_{yy}(R^K, X_1^K) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right]$$

puis que le troisième converge vers 0 lorsque $n \rightarrow \infty$, et ceci dans L^1 . On débute par le premier. En utilisant le fait que Y_1^K, \dots, Y_{n+K}^K sont des variables aléatoires conditionnellement indépendantes sachant (K, \mathbf{X}^K) ,

$$\begin{aligned}
& n\tau \mathbb{E} \left[h_y(R^K, X_1^K)(Y_1^K - X_1^K) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right] \\
&= n\tau h_y(R^K, X_1^K) \mathbb{E} \left[(Y_1^K - X_1^K) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] \\
&= n\tau h_y(R^K, X_1^K) \mathbb{E}[Y_1^K - X_1^K \mid K, \mathbf{X}^K] \mathbb{E} \left[1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mid K, \mathbf{X}^K \right] \\
&= 0,
\end{aligned}$$

car $R^K := (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$ et $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2 / (n + K))$. De la même façon,

$$\begin{aligned}
& n\tau \frac{1}{2} \mathbb{E} \left[h_{yy}(R^K, X_1^K)(Y_1^K - X_1^K)^2 \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right] \\
&= n\tau \frac{\ell^2}{2(n + K)} h_{yy}(R^K, X_1^K) \mathbb{E} \left[\left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right].
\end{aligned}$$

Donc, pour ce qui est du deuxième terme,

$$\mathbb{E} \left[\left| \frac{n\tau \ell^2}{2(n + K)} h_{yy}(R^K, X_1^K) \mathbb{E} \left[\left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] \right| \right]$$

$$\begin{aligned}
& \left| -\frac{\tau\ell^2}{2} h_{yy}(R^K, X_1^K) \mathbb{E} \left[\left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] \right| \\
&= \mathbb{E} \left[\frac{\tau\ell^2}{2} \left| h_{yy}(R^K, X_1^K) \mathbb{E} \left[\left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] \right| \left| \frac{n}{n+K} - 1 \right| \right] \\
&\leq \frac{\tau\ell^2 M}{2} \mathbb{E} \left[\frac{K}{n+K} \right] \leq \frac{\tau\ell^2 M}{2} \frac{\sqrt{n} \log n}{n+1} \rightarrow 0,
\end{aligned}$$

car $|h_{yy}| \leq M$, $0 \leq 1 \wedge \exp\{x\} \leq 1$, et $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$.

Finalement, en utilisant l'inégalité de Jensen pour espérance conditionnelle,

$$\begin{aligned}
& \mathbb{E} \left[\left| \frac{n\tau}{6} \mathbb{E} \left[h_y^{(3)}(R^K, W)(Y_1^K - X_1^K)^3 \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] \right| \right] \\
&\leq \frac{n\tau}{6} \mathbb{E} \left[\left| \mathbb{E} \left[h_y^{(3)}(R^K, W)(Y_1^K - X_1^K)^3 \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] \right| \right] \\
&\leq \frac{n\tau M}{6} \mathbb{E} \left[\mathbb{E} \left[|Y_1^K - X_1^K|^3 \mid K, \mathbf{X}^K \right] \right] \\
&= \frac{n\tau M}{6} \mathbb{E} \left[\frac{2\ell^3}{(n+K)^{3/2}} \sqrt{\frac{2}{\pi}} \right] \leq \frac{n\tau M}{6} \frac{2\ell^3}{(n+1)^{3/2}} \sqrt{\frac{2}{\pi}} \rightarrow 0,
\end{aligned}$$

car $|h_y^{(3)}| \leq M$, $0 \leq 1 \wedge \exp\{x\} \leq 1$, $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$, et $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$. ■

DÉMONSTRATION DE LEMME 3.3. Tout d'abord, en utilisant l'inégalité de Jensen pour espérance conditionnelle,

$$\begin{aligned}
& \mathbb{E} \left[\left| \frac{\tau\ell^2}{2} h_{yy}(R^K, X_1^K) \mathbb{E} \left[1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mid R^K, \mathbf{X}^K \right] \right. \right. \\
&\quad \left. \left. - \frac{\tau\ell^2}{2} h_{yy}(R^K, X_1^K) \mathbb{E} \left[1 \wedge \exp \{W^K\} \mid R^K, \mathbf{X}^K \right] \right| \right] \\
&= \frac{\tau\ell^2}{2} \mathbb{E} \left[\left| h_{yy}(R^K, X_1^K) \left| \mathbb{E} \left[1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mid R^K, \mathbf{X}^K \right] - \mathbb{E} \left[1 \wedge \exp \{W^K\} \mid R^K, \mathbf{X}^K \right] \right| \right| \right] \\
&\leq \frac{\tau\ell^2 M}{2} \mathbb{E} \left[\left| 1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} - 1 \wedge \exp \{W^K\} \right| \right] \\
&\leq \frac{\tau\ell^2 M}{2} \mathbb{E} \left[\left| \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} - W^K \right| \right],
\end{aligned}$$

car $|h_{yy}| \leq M$ et la fonction $g(x) := 1 \wedge \exp(x)$ est Lipschitz continue avec un coefficient de 1, c'est-à-dire que $|g(x) - g(y)| \leq |x - y|$, $\forall x, y$ (tel qu'indiqué par la proposition A.4 de Bédard (2006)).

Ensuite, en effectuant un développement en série de Taylor de la fonction $\log f(Y_i^K)$ par rapport à Y_i^K autour de X_i^K ,

$$\log f(Y_i^K) - \log f(X_i^K) = (\log f(X_i^K))'(Y_i^K - X_i^K) + (1/2)(\log f(V_i^K))''(Y_i^K - X_i^K)^2,$$

où V_i^K appartient à (X_i^K, Y_i^K) ou à (Y_i^K, X_i^K) . Ainsi,

$$\sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} - W^K = \sum_{i=2}^{n+K} (1/2)(\log f(V_i^K))''(Y_i^K - X_i^K)^2 + \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2.$$

L'objectif est de démontrer que l'espérance de la valeur absolue de cette dernière expression converge vers 0 lorsque $n \rightarrow \infty$. Afin d'y arriver, on démontre, premièrement, que

$$\mathbb{E} \left[\frac{1}{2} \left| \sum_{i=2}^{n+K} (\log f(V_i^K))''(Y_i^K - X_i^K)^2 - \sum_{i=2}^{n+K} (\log f(X_i^K))''(Y_i^K - X_i^K)^2 \right| \right] \rightarrow 0,$$

puis que

$$\mathbb{E} \left[\left| \sum_{i=2}^{n+K} (1/2)(\log f(X_i^K))''(Y_i^K - X_i^K)^2 + \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2 \right| \right] \rightarrow 0.$$

Ainsi, on obtiendra le résultat en utilisant l'inégalité du triangle.

En utilisant l'inégalité de Cauchy-Schwartz pour espérance conditionnelle,

$$\begin{aligned} & \frac{1}{2} \mathbb{E} \left[\left| \sum_{i=2}^{n+K} (\log f(V_i^K))''(Y_i^K - X_i^K)^2 - \sum_{i=2}^{n+K} (\log f(X_i^K))''(Y_i^K - X_i^K)^2 \right| \right] \\ & \leq \frac{1}{2} \mathbb{E} \left[\sum_{i=2}^{n+K} (Y_i^K - X_i^K)^2 \left| (\log f(V_i^K))'' - (\log f(X_i^K))'' \right| \right] \\ & = \frac{1}{2} \mathbb{E} \left[\sum_{i=2}^{n+K} \mathbb{E} \left[(Y_i^K - X_i^K)^2 \left| (\log f(V_i^K))'' - (\log f(X_i^K))'' \right| \mid K, \mathbf{X}^K \right] \right] \\ & \leq \frac{1}{2} \mathbb{E} \left[\sum_{i=2}^{n+K} \mathbb{E} \left[(Y_i^K - X_i^K)^4 \mid K, \mathbf{X}^K \right]^{1/2} \right. \\ & \quad \left. \times \mathbb{E} \left[\left((\log f(V_i^K))'' - (\log f(X_i^K))'' \right)^2 \mid K, \mathbf{X}^K \right]^{1/2} \right] \\ & = \frac{\sqrt{3}\ell^2}{2} \mathbb{E} \left[\frac{1}{n+K} \sum_{i=2}^{n+K} \mathbb{E} \left[\left((\log f(V_i^K))'' - (\log f(X_i^K))'' \right)^2 \mid K, \mathbf{X}^K \right]^{1/2} \right] \\ & \leq \frac{\sqrt{3}\ell^2}{2} \mathbb{E} \left[\frac{1}{n+1} \sum_{i=2}^{n+K} \mathbb{E} \left[\left((\log f(V_i^K))'' - (\log f(X_i^K))'' \right)^2 \mid K, \mathbf{X}^K \right]^{1/2} \right], \end{aligned}$$

car $Y_i^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_i^K, \ell^2/(n+K))$ et $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$. On a

$$\mathbb{E} \left[\frac{1}{n+1} \sum_{i=2}^{n+K} \mathbb{E} \left[\left((\log f(V_i^K))'' - (\log f(X_i^K))'' \right)^2 \mid K, \mathbf{X}^K \right]^{1/2} \right]$$

$$\begin{aligned}
&= \mathbb{E} \left[\frac{1}{n+1} \sum_{i=2}^n \mathbb{E} \left[((\log f(V_i^K))'' - (\log f(X_i^K))'')^2 \mid K, \mathbf{X}^K \right]^{1/2} \right] \\
&\quad + \mathbb{E} \left[\frac{1}{n+1} \sum_{i=n+1}^{n+K} \mathbb{E} \left[((\log f(V_i^K))'' - (\log f(X_i^K))'')^2 \mid K, \mathbf{X}^K \right]^{1/2} \right].
\end{aligned}$$

On démontre maintenant que chacun de ces deux termes converge vers 0. En utilisant l'inégalité du triangle,

$$\begin{aligned}
&\mathbb{E} \left[\frac{1}{n+1} \sum_{i=n+1}^{n+K} \mathbb{E} \left[((\log f(V_i^K))'' - (\log f(X_i^K))'')^2 \mid K, \mathbf{X}^K \right]^{1/2} \right] \\
&\leq \mathbb{E} \left[\frac{1}{n+1} \sum_{i=n+1}^{n+K} 2M \right] \leq \frac{\sqrt{n} \log n}{n+1} 2M \rightarrow 0,
\end{aligned}$$

car $|(\log f(x))''| \leq M$ (la fonction $(\log f(x))'$ est Lipschitz continue par hypothèse), et $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$.

On peut démontrer que $|Y_i^K - X_i^K| \rightarrow 0$ presque sûrement (p.s.) lorsque $n \rightarrow \infty$, et ainsi $|V_i^K - X_i^K| \rightarrow 0$ p.s. car V_i^K appartient à (X_i^K, Y_i^K) ou à (Y_i^K, X_i^K) . Donc, $((\log f(V_i^K))'' - (\log f(X_i^K))'')^2 \rightarrow 0$ p.s. étant donné que $(\log f(x))''$ est une fonction continue ($(\log f(x))''$ est fonction de f, f' et f'' , et $f \in C^2(\mathbb{R})$ par hypothèse). Alors, en utilisant le théorème de convergence dominée pour espérance conditionnelle,

$$\frac{1}{n+1} \sum_{i=2}^n \mathbb{E} \left[\mathbb{E} \left[((\log f(V_i^K))'' - (\log f(X_i^K))'')^2 \mid K, \mathbf{X}^K \right]^{1/2} \right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$. En effet, $(1/n) \sum_{i=1}^n a_i \rightarrow 0$ lorsque $n \rightarrow \infty$ dans la situation où la suite $\{a_n\}_{n=1}^\infty$ est telle que $a_n \rightarrow 0$ lorsque $n \rightarrow \infty$ et $|a_n|$ est bornée par une constante qui ne dépend pas de n .

Ensuite, par la monotonie des normes L^p ,

$$\begin{aligned}
&\mathbb{E} \left[\left| \sum_{i=2}^{n+K} (1/2)(\log f(X_i^K))''(Y_i^K - X_i^K)^2 + \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2 \right| \right] \\
&\leq \frac{1}{2} \mathbb{E} \left[\left| \sum_{i=2}^{n+K} (\log f(X_i^K))''(Y_i^K - X_i^K)^2 + \frac{\ell^2}{n+K} ((\log f(X_i^K))')^2 \right|^2 \right]^{1/2}.
\end{aligned}$$

Par ailleurs,

$$\begin{aligned}
&\mathbb{E} \left[\left| \sum_{i=2}^{n+K} (\log f(X_i^K))''(Y_i^K - X_i^K)^2 + \frac{\ell^2}{n+K} ((\log f(X_i^K))')^2 \right|^2 \right] \\
&= \mathbb{E} \left[\left| \sum_{i=2}^{n+K} (\log f(X_i^K))''(Y_i^K - X_i^K)^2 \right|^2 \right] + \mathbb{E} \left[\left| \sum_{i=2}^{n+K} \frac{\ell^2}{n+K} ((\log f(X_i^K))')^2 \right|^2 \right]
\end{aligned}$$

$$+ 2\mathbb{E} \left[\left(\sum_{i=2}^{n+K} (\log f(X_i^K))'' (Y_i^K - X_i^K)^2 \right) \left(\sum_{i=2}^{n+K} \frac{\ell^2}{n+K} ((\log f(X_i^K))')^2 \right) \right].$$

On démontre que la somme de ces trois derniers termes converge vers 0.

Premièrement,

$$\begin{aligned} \mathbb{E} \left[\left(\sum_{i=2}^{n+K} (\log f(X_i^K))'' (Y_i^K - X_i^K)^2 \right)^2 \right] &= \mathbb{E} \left[\sum_{i=2}^{n+K} ((\log f(X_i^K))'')^2 (Y_i^K - X_i^K)^4 \right] \\ &+ \mathbb{E} \left[\sum_{i,j=2(i \neq j)}^{n+K} (\log f(X_i^K))'' (Y_i^K - X_i^K)^2 (\log f(X_j^K))'' (Y_j^K - X_j^K)^2 \right]. \end{aligned}$$

Mais, en utilisant le fait que $|(\log f(x))''| \leq M$, $Y_i^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_i^K, \ell^2/(n+K))$ et que $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=2}^{n+K} ((\log f(X_i^K))'')^2 (Y_i^K - X_i^K)^4 \right] &= \mathbb{E} \left[\sum_{i=2}^{n+K} ((\log f(X_i^K))'')^2 \mathbb{E}[(Y_i^K - X_i^K)^4 \mid K, \mathbf{X}^K] \right] \\ &\leq \mathbb{E} \left[M^2 \sum_{i=2}^{n+K} \frac{3\ell^4}{(n+K)^2} \right] \\ &= \mathbb{E} \left[M^2 \frac{3\ell^4(n+K-1)}{(n+K)^2} \right] \leq M^2 \frac{3\ell^4(n + \sqrt{n} \log n)}{(n+1)^2} \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$.

Ensuite, en considérant l'indépendance conditionnelle des variables aléatoires Y_1^K, \dots, Y_{n+K}^K sachant (K, \mathbf{X}^K) , puis celle des variables aléatoires X_1^K, \dots, X_{n+K}^K sachant K ,

$$\begin{aligned} &\mathbb{E} \left[\sum_{i,j=2(i \neq j)}^{n+K} (\log f(X_i^K))'' (Y_i^K - X_i^K)^2 (\log f(X_j^K))'' (Y_j^K - X_j^K)^2 \right] \\ &= \mathbb{E} \left[\sum_{i,j=2(i \neq j)}^{n+K} (\log f(X_i^K))'' (\log f(X_j^K))'' \mathbb{E}[(Y_i^K - X_i^K)^2 (Y_j^K - X_j^K)^2 \mid K, \mathbf{X}^K] \right] \\ &= \mathbb{E} \left[\sum_{i,j=2(i \neq j)}^{n+K} (\log f(X_i^K))'' (\log f(X_j^K))'' \left(\frac{\ell^2}{n+K} \right)^2 \right] \\ &= \mathbb{E} \left[\left(\frac{\ell^2}{n+K} \right)^2 \sum_{i,j=2(i \neq j)}^{n+K} \mathbb{E}[(\log f(X_i^K))'' (\log f(X_j^K))'' \mid K] \right] \\ &= \mathbb{E} \left[\left(\frac{\ell^2}{n+K} \right)^2 \sum_{i,j=2(i \neq j)}^{n+K} \mathbb{E}[(\log f(X_i^K))'']^2 \right] \\ &= \mathbb{E}[(\log f(X_1^K))']^2 \mathbb{E} \left[\frac{\ell^4(n+K-1)(n+K-2)}{(n+K)^2} \right]. \end{aligned}$$

L'hypothèse que les variables aléatoires X_1^K, \dots, X_{n+K}^K sont identiquement distribuées sachant K a été utilisée à la quatrième égalité, puis la proposition 3.1 a été appliquée à la cinquième.

Pour ce qui est du deuxième terme,

$$\begin{aligned} \mathbb{E} \left[\left(\sum_{i=2}^{n+K} \frac{\ell^2}{n+K} ((\log f(X_i^K))')^2 \right)^2 \right] &= \mathbb{E} \left[\sum_{i=2}^{n+K} \frac{\ell^4}{(n+K)^2} ((\log f(X_i^K))')^4 \right] \\ &\quad + \mathbb{E} \left[\sum_{i,j=2(i \neq j)}^{n+K} \frac{\ell^4}{(n+K)^2} ((\log f(X_i^K))')^2 ((\log f(X_j^K))')^2 \right]. \end{aligned}$$

Mais, en utilisant le fait que $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=2}^{n+K} \frac{\ell^4}{(n+K)^2} ((\log f(X_i^K))')^4 \right] &= \mathbb{E} \left[\frac{\ell^4}{(n+K)^2} \sum_{i=2}^{n+K} \mathbb{E}[(\log f(X_i^K))'^4 \mid K] \right] \\ &= \mathbb{E}[(\log f(X_1^K))'^4] \mathbb{E} \left[\frac{\ell^4(n+K-1)}{(n+K)^2} \right] \\ &\leq \mathbb{E}[(\log f(X_1^K))'^4] \frac{\ell^4(n + \sqrt{n} \log n)}{(n+1)^2} \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, car $\mathbb{E}[(\log f(X_1^K))'^4] < \infty$ par hypothèse. Ensuite, en utilisant le fait que les variables aléatoires X_1^K, \dots, X_{n+K}^K sont conditionnellement indépendantes et identiquement distribuées sachant K ,

$$\begin{aligned} &\mathbb{E} \left[\sum_{i,j=2(i \neq j)}^{n+K} \frac{\ell^4}{(n+K)^2} ((\log f(X_i^K))')^2 ((\log f(X_j^K))')^2 \right] \\ &= \mathbb{E} \left[\sum_{i,j=2(i \neq j)}^{n+K} \frac{\ell^4}{(n+K)^2} \mathbb{E}[(\log f(X_i^K))'^2 ((\log f(X_j^K))')^2 \mid K] \right] \\ &= \mathbb{E}[(\log f(X_1^K))'^2]^2 \mathbb{E} \left[\frac{\ell^4(n+K-1)(n+K-2)}{(n+K)^2} \right]. \end{aligned}$$

Finalement,

$$\begin{aligned} &2\mathbb{E} \left[\left(\sum_{i=2}^{n+K} (\log f(X_i^K))'' (Y_i^K - X_i^K)^2 \right) \left(\sum_{i=2}^{n+K} \frac{\ell^2}{n+K} ((\log f(X_i^K))')^2 \right) \right] \\ &= 2\mathbb{E} \left[\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} (\log f(X_i^K))'' ((\log f(X_i^K))')^2 (Y_i^K - X_i^K)^2 \right] \\ &\quad + 2\mathbb{E} \left[\frac{\ell^2}{n+K} \sum_{i,j=2(i \neq j)}^{n+K} (\log f(X_i^K))'' ((\log f(X_j^K))')^2 (Y_i^K - X_i^K)^2 \right]. \end{aligned}$$

Mais, en utilisant l'inégalité de Jensen,

$$\begin{aligned}
& \left| \mathbb{E} \left[\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} (\log f(X_i^K))'' ((\log f(X_i^K))')^2 (Y_i^K - X_i^K)^2 \right] \right| \\
&= \left| \mathbb{E} \left[\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} (\log f(X_i^K))'' ((\log f(X_i^K))')^2 \mathbb{E}[(Y_i^K - X_i^K)^2 \mid K, \mathbf{X}^K] \right] \right| \\
&= \left| \mathbb{E} \left[\frac{\ell^4}{(n+K)^2} \sum_{i=2}^{n+K} \mathbb{E}[(\log f(X_i^K))'' ((\log f(X_i^K))')^2 \mid K] \right] \right| \\
&= \left| \mathbb{E}[(\log f(X_1^K))'' ((\log f(X_1^K))')^2] \right| \mathbb{E} \left[\frac{\ell^4(n+K-1)}{(n+K)^2} \right] \\
&\leq \mathbb{E} \left[|(\log f(X_1^K))'' ((\log f(X_1^K))')^2| \right] \mathbb{E} \left[\frac{\ell^4}{n+K} \right] \\
&\leq \mathbb{E} \left[((\log f(X_1^K))')^2 \right] \frac{M\ell^4}{n+1} \rightarrow 0,
\end{aligned}$$

lorsque $n \rightarrow \infty$, car $Y_i^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_i^K, \ell^2/(n+K))$, $|(\log f(x))''| \leq M$ (la fonction $(\log f(x))'$ est Lipschitz continue par hypothèse), $\mathbb{E}[(\log f(X_1^K))']^2 < \infty$ ($\mathbb{E}[(\log f(X_1^K))']^4 < \infty$ par hypothèse), et $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$. Ensuite, en utilisant le fait que les variables aléatoires X_1^K, \dots, X_{n+K}^K sont conditionnellement indépendantes et identiquement distribuées sachant K , puis la proposition 3.1,

$$\begin{aligned}
& 2\mathbb{E} \left[\frac{\ell^2}{n+K} \sum_{i,j=2(i \neq j)}^{n+K} (\log f(X_i^K))'' ((\log f(X_j^K))')^2 (Y_i^K - X_i^K)^2 \right] \\
&= 2\mathbb{E} \left[\frac{\ell^2}{n+K} \sum_{i,j=2(i \neq j)}^{n+K} (\log f(X_i^K))'' ((\log f(X_j^K))')^2 \mathbb{E}[(Y_i^K - X_i^K)^2 \mid K, \mathbf{X}^K] \right] \\
&= 2\mathbb{E} \left[\frac{\ell^4}{(n+K)^2} \sum_{i,j=2(i \neq j)}^{n+K} \mathbb{E}[(\log f(X_i^K))'' ((\log f(X_j^K))')^2 \mid K] \right] \\
&= 2\mathbb{E}[(\log f(X_1^K))''] \mathbb{E}[(\log f(X_1^K))']^2 \mathbb{E} \left[\frac{\ell^4(n+K-1)(n+K-2)}{(n+K)^2} \right] \\
&= -2\mathbb{E}[(\log f(X_1^K))']^2 \mathbb{E} \left[\frac{\ell^4(n+K-1)(n+K-2)}{(n+K)^2} \right],
\end{aligned}$$

car $Y_i^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_i^K, \ell^2/(n+K))$.

La somme de tous les termes ci-dessus qui ne convergent pas vers 0, donnée par

$$\begin{aligned}
& \mathbb{E}[(\log f(X_1^K))']^2 \mathbb{E} \left[\frac{\ell^4(n+K-1)(n+K-2)}{(n+K)^2} \right] \\
&+ \mathbb{E}[(\log f(X_1^K))']^2 \mathbb{E} \left[\frac{\ell^4(n+K-1)(n+K-2)}{(n+K)^2} \right] \\
&- 2\mathbb{E}[(\log f(X_1^K))']^2 \mathbb{E} \left[\frac{\ell^4(n+K-1)(n+K-2)}{(n+K)^2} \right],
\end{aligned}$$

est, elle, égale à 0. ■

DÉMONSTRATION DU LEMME 3.4. Tout d'abord,

$$\begin{aligned} & \mathbb{E} \left[\left| \frac{\tau \ell^2}{2} h_{yy}(R^K, X_1^K) \mathbb{E} \left[1 \wedge \exp \{W^K\} \mid R^K, \mathbf{X}^K \right] - \tau \ell^2 \Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) h_{yy}(R^K, X_1^K) \right| \right] \\ &= \mathbb{E} \left[\left| \frac{\tau \ell^2}{2} h_{yy}(R^K, X_1^K) \right| \left| \mathbb{E} \left[1 \wedge \exp \{W^K\} \mid R^K, \mathbf{X}^K \right] - 2\Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) \right| \right] \\ &\leq \frac{\tau \ell^2 M}{2} \mathbb{E} \left[\left| \mathbb{E} \left[1 \wedge \exp \{W^K\} \mid K, \mathbf{X}^K \right] - 2\Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) \right| \right], \end{aligned}$$

car $|h_{yy}| \leq M$ et $R^K := (K - \lfloor \sqrt{n} \log \rfloor / 2) / \sqrt{n}$. On peut démontrer que

$$\begin{aligned} W^K \mid K, \mathbf{X}^K &= \sum_{i=2}^{n+K} (\log f(X_i^K))' (Y_i^K - X_i^K) - \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2 \mid K, \mathbf{X}^K \\ &\sim \mathcal{N} \left(-\frac{\ell^2}{2} \frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2, \frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2 \right). \end{aligned}$$

De plus, tel qu'indiqué par la proposition A.5 de [Bédard \(2006\)](#), si $W \sim \mathcal{N}(\mu, \sigma^2)$, alors

$$\mathbb{E}[1 \wedge e^W] = \Phi \left(\frac{\mu}{\sigma} \right) + \exp\{\mu + \sigma^2/2\} \Phi \left(-\sigma - \frac{\mu}{\sigma} \right).$$

Ainsi,

$$\begin{aligned} & \mathbb{E} \left[\left| \mathbb{E} \left[1 \wedge \exp \{W^K\} \mid K, \mathbf{X}^K \right] - 2\Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) \right| \right] \\ &= \mathbb{E} \left[\left| 2\Phi \left(-\frac{\ell}{2} \sqrt{\frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2} \right) - 2\Phi \left(-\frac{\ell \sqrt{\Upsilon}}{2} \right) \right| \right]. \end{aligned}$$

Aussi,

$$\frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2 = \frac{n-1}{n+K} \frac{1}{n-1} \sum_{i=2}^n ((\log f(X_i^K))')^2 + \frac{1}{n+K} \sum_{i=n+1}^{n+K} ((\log f(X_i^K))')^2.$$

On démontre maintenant que le premier terme converge vers Υ p.s. puis que le deuxième terme converge vers 0 en probabilité, et ceci lorsque $n \rightarrow \infty$. Ainsi, en utilisant le théorème de Slutsky, on saura que $(1/(n+K)) \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2 \rightarrow \Upsilon$ en probabilité et ainsi que $-(\ell/2) \sqrt{(1/(n+K)) \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2} \rightarrow -\ell \sqrt{\Upsilon} / 2$ en probabilité. Ensuite, en utilisant la proposition 3.2, on aura que l'espérance ci-dessus converge vers 0 car $0 \leq \Phi \leq 1$ est une fonction uniformément continue.

On sait, en utilisant la loi forte des grands nombres, que

$$\frac{1}{n-1} \sum_{i=2}^n ((\log f(X_i^K))')^2 \rightarrow \Upsilon \text{ p.s.,}$$

lorsque $n \rightarrow \infty$, car les variables aléatoires X_1^K, \dots, X_n^K sont indépendantes et identiquement distribuées (la variable aléatoire K n'a pas d'impact sur ces variables aléatoires). De plus,

$$\frac{n-1}{n+K} = \frac{1-1/n}{1+K/n} \rightarrow 1 \text{ p.s.},$$

lorsque $n \rightarrow \infty$, car $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$.

Pour ce qui est du deuxième terme, en utilisant que les variables aléatoires $X_{n+1}^K, \dots, X_{n+K}^K$ sont conditionnellement indépendantes et identiquement distribuées sachant K ,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{n+K} \sum_{i=n+1}^{n+K} ((\log f(X_i^K))')^2 \right] &= \mathbb{E} \left[\frac{1}{n+K} \sum_{i=n+1}^{n+K} \mathbb{E} [((\log f(X_i^K))')^2 \mid K] \right] \\ &= \mathbb{E} [((\log f(X_1^K))')^2] \mathbb{E} \left[\frac{K}{n+K} \right] \\ &\leq \mathbb{E} [((\log f(X_1^K))')^2] \frac{\lfloor \sqrt{n} \log n \rfloor}{n+1} \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, car $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$ et $\mathbb{E} [((\log f(X_1^K))')^2] < \infty$ (l'espérance $\mathbb{E} [((\log f(X_1^K))')^4] < \infty$ par hypothèse). Ainsi, on sait que $(1/(n+K)) \sum_{i=n+1}^{n+K} ((\log f(X_i^K))')^2 \rightarrow 0$ dans L^1 et donc en probabilité. ■

Maintenant, on démontre que le deuxième terme de (3.1), donné par

$$\begin{aligned} &\tau n \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K))(Y_1 - X_1^K)(\log f(X_1^K))' \right. \\ &\quad \left. \times \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \right], \end{aligned}$$

converge vers $\tau \ell^2 \Phi(-\ell \sqrt{Y}/2)(\log f(X_1^K))' h_y(R^K, X_1^K)$ dans L^1 lorsque $n \rightarrow \infty$ par l'intermédiaire des lemmes 3.5 à 3.7.

Lemme 3.5. *On a*

$$\begin{aligned} &\mathbb{E} \left[\left| n \tau \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K))(Y_1 - X_1^K)(\log f(X_1^K))' \right. \right. \right. \\ &\quad \left. \left. \times \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \right] \right. \\ &\quad \left. - \tau \mathbb{E} \left[\ell^2 h_y(R^K, X_1^K)(\log f(X_1^K))' \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right. \right. \\ &\quad \left. \left. \times \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \right] \right| \right] \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$.

Lemme 3.6. *On a*

$$\begin{aligned} & \mathbb{E} \left[\left| \tau \ell^2 h_y(R^K, X_1^K) (\log f(X_1^K))' \mathbb{E} \left[\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right. \right. \right. \\ & \quad \times \mathbb{1} \left. \left. \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \right] \right. \\ & \quad \left. \left. - \tau \ell^2 h_y(R^K, X_1^K) (\log f(X_1^K))' \mathbb{E} \left[\exp \{W^K\} \mathbb{1} \{W^K < 0\} \mid R^K, \mathbf{X}^K \right] \right] \right| \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, où

$$W^K := \sum_{i=2}^{n+K} (\log f(X_i^K))' (Y_i^K - X_i^K) - \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2.$$

Lemme 3.7. *On a*

$$\begin{aligned} & \mathbb{E} \left[\left| \tau \ell^2 h_y(R^K, X_1^K) (\log f(X_1^K))' \mathbb{E} \left[\exp \{W^K\} \mathbb{1} \{W^K < 0\} \mid K, \mathbf{X}^K \right] \right. \right. \\ & \quad \left. \left. - \tau \ell^2 \Phi \left(-\frac{\ell \sqrt{Y}}{2} \right) (\log f(X_1^K))' h_y(R^K, X_1^K) \right] \right| \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, où

$$W^K := \sum_{i=2}^{n+K} (\log f(X_i^K))' (Y_i^K - X_i^K) - \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2.$$

DÉMONSTRATION DU LEMME 3.5. Tout d'abord,

$$\begin{aligned} & \mathbb{E} \left[\left| n \tau \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K)) (Y_1 - X_1^K) (\log f(X_1^K))' \right. \right. \right. \\ & \quad \times \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \right] \\ & \quad - \tau \mathbb{E} \left[\ell^2 h_y(R^K, X_1^K) (\log f(X_1^K))' \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right. \\ & \quad \left. \left. \times \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \right] \right| \right] \\ & = \tau \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \mathbb{E} \left[\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \right. \right. \right. \\ & \quad \left. \left. \times (n(h(R^K, Y_1^K) - h(R^K, X_1^K)) (Y_1 - X_1^K) - \ell^2 h_y(R^K, X_1^K)) \mid K, \mathbf{X}^K \right] \right| \right], \end{aligned}$$

car $K := (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$. On remplace $n(h(R^K, Y_1^K) - h(R^K, X_1^K)) (Y_1 - X_1^K)$ par

$$n h_y(R^K, X_1^K) (Y_1^K - X_1^K)^2 + n h_{yy}(R^K, W) (Y_1^K - X_1^K)^3,$$

en utilisant un développement en série de Taylor (voir démonstration du lemme 3.1 pour plus de détails), et on analyse le premier terme. En utilisant l'indépendance conditionnelle des variables aléatoires Y_1, \dots, Y_{n+K} sachant (K, \mathbf{X}^K) ,

$$\begin{aligned} & h_y(R^K, X_1^K) \mathbb{E} \left[\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} n(Y_1^K - X_1^K)^2 \mid K, \mathbf{X}^K \right] \\ &= h_y(R^K, X_1^K) \frac{n\ell^2}{n+K} \mathbb{E} \left[\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid K, \mathbf{X}^K \right], \end{aligned}$$

car $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$. Donc, en utilisant l'inégalité de Jensen pour espérance conditionnelle puis l'inégalité du triangle,

$$\begin{aligned} & \tau \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \mathbb{E} \left[\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \right. \right. \\ & \quad \left. \left. \times \left(n(h(R^K, Y_1^K) - h(R^K, X_1^K))(Y_1^K - X_1^K) - \ell^2 h_y(R^K, X_1^K) \right) \mid K, \mathbf{X}^K \right] \right] \\ &= \tau \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \mathbb{E} \left[\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \right. \right. \\ & \quad \left. \left. \times \left(h_y(R^K, X_1^K) \left(\frac{n\ell^2}{n+K} - \ell^2 \right) + n h_{yy}(R^K, W)(Y_1^K - X_1^K)^3 \right) \mid K, \mathbf{X}^K \right] \right] \\ &\leq \tau \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \mathbb{E} \left[\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \right. \right. \\ & \quad \left. \left. \times \left| h_y(R^K, X_1^K) \left(\frac{n\ell^2}{n+K} - \ell^2 \right) + n h_{yy}(R^K, W)(Y_1^K - X_1^K)^3 \right| \mid K, \mathbf{X}^K \right] \right] \\ &\leq \tau \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \mathbb{E} \left[\left| h_y(R^K, X_1^K) \left(\frac{n\ell^2}{n+K} - \ell^2 \right) \right| \mid K, \mathbf{X}^K \right] \right] \\ & \quad + \tau \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \mathbb{E} \left[\left| n h_{yy}(R^K, W)(Y_1^K - X_1^K)^3 \right| \mid K, \mathbf{X}^K \right] \right], \end{aligned}$$

car

$$0 \leq \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \leq 1.$$

On démontre maintenant que chacun des deux derniers termes converge vers 0 lorsque $n \rightarrow \infty$. On débute par le premier :

$$\begin{aligned} & \tau \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \mathbb{E} \left[\left| h_y(R^K, X_1^K) \left(\frac{n\ell^2}{n+K} - \ell^2 \right) \right| \mid K, \mathbf{X}^K \right] \right] \\ & \leq \tau M \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \left| \frac{n\ell^2}{n+K} - \ell^2 \right| \right] \\ & = \tau M \ell^2 \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \left| \frac{K}{n+K} \right| \right] \leq \tau M \ell^2 \frac{\lfloor \sqrt{n} \log n \rfloor}{n+1} \mathbb{E} \left[\left| (\log f(X_1^K))' \right| \right] \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, car $|h_y| \leq M$, $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$, et $\mathbb{E}[|(\log f(X_1^K))'|] < \infty$ ($\mathbb{E}[((\log f(X_1^K))')^4] < \infty$ par hypothèse).

Pour ce qui est du deuxième terme,

$$\begin{aligned} & \tau \mathbb{E}[|(\log f(X_1^K))'| \mathbb{E}[|nh_{yy}(R^K, W)(Y_1^K - X_1^K)^3| \mid K, \mathbf{X}^K]] \\ & \leq \tau M \mathbb{E}[|(\log f(X_1^K))'| \mathbb{E}[|n(Y_1^K - X_1^K)^3| \mid K, \mathbf{X}^K]] \\ & = \tau M \mathbb{E} \left[|(\log f(X_1^K))'| \frac{2n\ell^3}{(n+K)^{3/2}} \sqrt{\frac{2}{\pi}} \right] \leq \tau M \frac{2n\ell^3}{(n+1)^{3/2}} \sqrt{\frac{2}{\pi}} \mathbb{E}[|(\log f(X_1^K))'|] \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, car $|h_{yy}| \leq M$, $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$, $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$, et $\mathbb{E}[|(\log f(X_1^K))'|] < \infty$ ($\mathbb{E}[((\log f(X_1^K))')^4] < \infty$ par hypothèse). ■

DÉMONSTRATION DU LEMME 3.6. Tout d'abord, en utilisant l'inégalité de Jensen pour espérance conditionnelle puis l'inégalité de Cauchy-Schwartz,

$$\begin{aligned} & \mathbb{E} \left[\left| \tau \ell^2 h_y(R^K, X_1^K) (\log f(X_1^K))' \mathbb{E} \left[\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \right] \right. \right. \\ & \quad \left. \left. - \tau \ell^2 h_y(R^K, X_1^K) (\log f(X_1^K))' \mathbb{E} \left[\exp \{W^K\} \mathbb{1} \{W^K < 0\} \mid R^K, \mathbf{X}^K \right] \right| \right] \\ & \leq \tau \ell^2 M \mathbb{E} \left[|(\log f(X_1^K))'| \right. \\ & \quad \left. \times \left| \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} - \exp \{W^K\} \mathbb{1} \{W^K < 0\} \right| \right] \\ & \leq \tau \ell^2 M \mathbb{E} \left[((\log f(X_1^K))')^2 \right]^{1/2} \mathbb{E} \left[\left(\exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \right. \right. \\ & \quad \left. \left. - \exp \{W^K\} \mathbb{1} \{W^K < 0\} \right)^2 \right]^{1/2}, \end{aligned}$$

car $|h_y| \leq M$. On sait que $\mathbb{E}[((\log f(X_1^K))')^2] < \infty$ ($\mathbb{E}[((\log f(X_1^K))')^4] < \infty$ par hypothèse). On démontre maintenant que la deuxième espérance converge vers 0 lorsque $n \rightarrow \infty$. Afin de simplifier l'écriture, on définit

$$S^K := \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)}.$$

On a

$$\begin{aligned} & \mathbb{E} \left[\left(\exp \{S^K\} \mathbb{1} \{S^K < 0\} - \exp \{W^K\} \mathbb{1} \{W^K < 0\} \right)^2 \right] \\ & = \mathbb{E} \left[\left(1 \wedge \exp \{S^K\} - \mathbb{1} \{S^K \geq 0\} - 1 \wedge \exp \{W^K\} + \mathbb{1} \{W^K \geq 0\} \right)^2 \right] \\ & = \mathbb{E} \left[\left(1 \wedge \exp \{S^K\} - 1 \wedge \exp \{W^K\} \right)^2 \right] + \mathbb{E} \left[\left(\mathbb{1} \{W^K \geq 0\} - \mathbb{1} \{S^K \geq 0\} \right)^2 \right] \\ & \quad + 2 \mathbb{E} \left[\left(1 \wedge \exp \{S^K\} - 1 \wedge \exp \{W^K\} \right) \left(\mathbb{1} \{W^K \geq 0\} - \mathbb{1} \{S^K \geq 0\} \right) \right]. \end{aligned}$$

On démontre que chacun de ces trois derniers termes converge vers 0 lorsque $n \rightarrow \infty$. On débute par le premier. Grâce à démonstration du lemme 3.3, on sait que

$$\mathbb{E} [|S^K - W^K|] \leq \frac{A}{n^{1/4}} \rightarrow 0,$$

lorsque $n \rightarrow \infty$, où A est une constante positive. Ainsi, on sait que $S^K - W^K$ converge vers 0 en probabilité, et donc,

$$\mathbb{E} \left[\left(1 \wedge \exp\{S^K\} - 1 \wedge \exp\{W^K\} \right)^2 \right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$, en utilisant la proposition 3.2 puisque la fonction $1 \wedge \exp\{x\}$ est bornée et uniformément continue (tel qu'indiqué par la proposition A.4 de Bédard (2006), $1 \wedge \exp\{x\}$ est Lipschitz continue).

Pour ce qui est du deuxième terme,

$$\begin{aligned} \mathbb{E} \left[\left(\mathbb{1}\{W^K \geq 0\} - \mathbb{1}\{S^K \geq 0\} \right)^2 \right] &= \mathbb{P}(W^K \geq 0, S^K < 0) + \mathbb{P}(W^K < 0, S^K \geq 0) \\ &= \mathbb{P}(W^K \geq 0, S^K < 0, |W^K - S^K| > n^{-1/4}) \\ &\quad + \mathbb{P}(W^K \geq 0, S^K < 0, |W^K - S^K| \leq n^{-1/4}) \\ &\quad + \mathbb{P}(W^K < 0, S^K \geq 0, |W^K - S^K| > n^{-1/4}) \\ &\quad + \mathbb{P}(W^K < 0, S^K \geq 0, |W^K - S^K| \leq n^{-1/4}) \\ &\leq 2\mathbb{P}(|W^K - S^K| > n^{-1/4}) \\ &\quad + \mathbb{P}(W^K \geq 0, S^K < 0, W^K - S^K \leq n^{-1/4}) \\ &\quad + \mathbb{P}(W^K < 0, S^K \geq 0, S^K - W^K \leq n^{-1/4}). \end{aligned}$$

En utilisant l'inégalité de Markov, on obtient que

$$\mathbb{P}(|W^K - S^K| > n^{-1/4}) \leq n^{1/4} \mathbb{E} [|S^K - W^K|] \leq A/n^{1/4} \rightarrow 0,$$

lorsque $n \rightarrow \infty$. Aussi,

$$\begin{aligned} \mathbb{P}(W^K \geq 0, S^K < 0, W^K - S^K \leq n^{-1/8}) &\leq \mathbb{P}(0 \leq W^K \leq n^{-1/4}), \\ \mathbb{P}(W^K < 0, S^K \geq 0, S^K - W^K \leq n^{-1/8}) &\leq \mathbb{P}(-n^{-1/4} \leq W^K < 0). \end{aligned}$$

On peut démontrer que

$$\begin{aligned} W^K | K, \mathbf{X}^K &= \sum_{i=2}^{n+K} (\log f(X_i^K))' (Y_i^K - X_i^K) - \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2 | K, \mathbf{X}^K \\ &\sim \mathcal{N} \left(-\frac{\ell^2}{2} \frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2, \frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2 \right). \end{aligned}$$

Alors,

$$\begin{aligned} \mathbb{P}(-n^{-1/4} \leq W^K \leq n^{-1/4}) &= \mathbb{E}[\mathbb{E}[\mathbb{1}\{-n^{-1/4} \leq W^K \leq n^{-1/4}\} \mid K, \mathbf{X}^K]] \\ &= \mathbb{E} \left[\Phi \left(\frac{n^{-1/4} + \frac{\ell^2}{2} \frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}{\sqrt{\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}} \right) \right. \\ &\quad \left. - \Phi \left(\frac{-n^{-1/4} + \frac{\ell^2}{2} \frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}{\sqrt{\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}} \right) \right]. \end{aligned}$$

On démontre que cette soustraction d'espérances converge vers 0 lorsque $n \rightarrow \infty$ en utilisant l'inégalité de Jensen (on fait entrer une valeur absolue à l'intérieur de l'espérance) puis la proposition 3.2 (Φ est bornée et uniformément continue). Plus précisément, on démontre que

$$\begin{aligned} &\frac{n^{-1/4} + \frac{\ell^2}{2} \frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}{\sqrt{\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}} - \frac{-n^{-1/4} + \frac{\ell^2}{2} \frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}{\sqrt{\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}} \\ &= \frac{2n^{-1/4}}{\sqrt{\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}} \rightarrow 0, \end{aligned}$$

en probabilité, lorsque $n \rightarrow \infty$. Grâce à la démonstration du lemme 3.4, on sait que $1/(n+K) \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2 \rightarrow \Upsilon$ en probabilité lorsque $n \rightarrow \infty$. Donc, en utilisant le théorème de Slutsky

$$\sqrt{\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2} \rightarrow \ell \sqrt{\Upsilon},$$

en probabilité lorsque $n \rightarrow \infty$. Alors, en utilisant encore une fois le théorème de Slutsky, on obtient

$$\frac{2n^{-1/4}}{\sqrt{\frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}} \rightarrow 0,$$

en probabilité lorsque $n \rightarrow \infty$.

Il ne reste qu'à démontrer que

$$\mathbb{E} \left[\left(\mathbb{1} \wedge \exp\{W^K\} - \mathbb{1} \wedge \exp\{S^K\} \right) \left(\mathbb{1}\{W^K \geq 0\} - \mathbb{1}\{S^K \geq 0\} \right) \right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$. On effectue ceci directement en utilisant l'inégalité de Jensen puis l'inégalité de Cauchy-Schwartz puisqu'il a été démontré que $\mathbb{E}[(\mathbb{1} \wedge \exp\{W^K\} - \mathbb{1} \wedge \exp\{S^K\})^2] \rightarrow 0$ et que $\mathbb{E}[(\mathbb{1}\{W^K \geq 0\} - \mathbb{1}\{S^K \geq 0\})^2] \rightarrow 0$ lorsque $n \rightarrow \infty$. ■

DÉMONSTRATION DU LEMME 3.7. Tout d'abord, en utilisant l'inégalité de Cauchy-Schwartz,

$$\mathbb{E} \left[\left| \tau \ell^2 h_y(R^K, X_1^K) (\log f(X_1^K))' \mathbb{E} \left[\exp\{W^K\} \mathbb{1}\{W^K < 0\} \mid R^K, \mathbf{X}^K \right] \right| \right]$$

$$\begin{aligned}
& \left. -\tau\ell^2\Phi\left(-\frac{\ell\sqrt{\Upsilon}}{2}\right)(\log f(X_1^K))'h_y(R^K, X_1^K)\right] \\
&= \mathbb{E}\left[\left|\tau\ell^2h_y(R^K, X_1^K)(\log f(X_1^K))'\right|\left|\mathbb{E}\left[\exp\{W^K\}\mathbb{1}\{W^K < 0\}\mid K, \mathbf{X}^K\right]-\Phi\left(-\frac{\ell\sqrt{\Upsilon}}{2}\right)\right|\right] \\
&\leq \tau\ell^2M\mathbb{E}\left[\left|(\log f(X_1^K))'\right|\left|\mathbb{E}\left[\exp\{W^K\}\mathbb{1}\{W^K < 0\}\mid K, \mathbf{X}^K\right]-\Phi\left(-\frac{\ell\sqrt{\Upsilon}}{2}\right)\right|\right] \\
&\leq \tau\ell^2M\mathbb{E}\left[\left((\log f(X_1^K))'\right)^2\right]^{1/2}\mathbb{E}\left[\left(\mathbb{E}\left[\exp\{W^K\}\mathbb{1}\{W^K < 0\}\mid K, \mathbf{X}^K\right]-\Phi\left(-\frac{\ell\sqrt{\Upsilon}}{2}\right)\right)^2\right]^{1/2},
\end{aligned}$$

car $R^K := (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$ et $|h_y| \leq M$. On sait que $\mathbb{E}[(\log f(X_1^K))']^2 < \infty$ ($\mathbb{E}[(\log f(X_1^K))']^4 < \infty$ par hypothèse). On démontre donc que la deuxième espérance converge vers 0 lorsque $n \rightarrow \infty$.

On peut démontrer que

$$\begin{aligned}
W^K \mid K, \mathbf{X}^K &= \sum_{i=2}^{n+K} (\log f(X_i^K))'(Y_i^K - X_i^K) - \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2 \mid K, \mathbf{X}^K \\
&\sim \mathcal{N}\left(-\frac{\ell^2}{2} \frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2, \frac{\ell^2}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2\right).
\end{aligned}$$

De plus, tel qu'indiqué dans la démonstration de la proposition A.5 de [Bédard \(2006\)](#), si $W \sim \mathcal{N}(\mu, \sigma^2)$, alors

$$\mathbb{E}[e^W \mathbb{1}\{W < 0\}] = \exp\{\mu + \sigma^2/2\} \Phi\left(-\sigma - \frac{\mu}{\sigma}\right).$$

Donc,

$$\mathbb{E}[\mathbb{1} \wedge \exp\{W^K\} \mathbb{1}\{W^K < 0\} \mid K, \mathbf{X}^K] = \Phi\left(-\frac{\ell}{2} \sqrt{\frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}\right).$$

Tel qu'expliqué dans la démonstration du lemme [3.4](#),

$$\mathbb{E}\left[\left(\Phi\left(-\frac{\ell}{2} \sqrt{\frac{1}{n+K} \sum_{i=2}^{n+K} ((\log f(X_i^K))')^2}\right) - \Phi\left(-\frac{\ell\sqrt{\Upsilon}}{2}\right)\right)^2\right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$. ■

3.4. COMPLÉMENT DE LA DÉMONSTRATION DU THÉORÈME [2.1](#) : VÉRIFICATION DES CONDITIONS DU RÉSULTAT (C) DU THÉORÈME [3.1](#)

Tel que mentionné à la section [2.6](#), le résultat (c) se résume à la convergence suivante : $\mathbb{E}[|\varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t))|] \rightarrow 0$. Dans cette section, nous expliquons en détail pourquoi nous pouvons faire cette affirmation. On rappelle tout d'abord le résultat (c) :

(c) For each $(h, g) \in B$ and $T > 0$, there exist $(\xi_n, \varphi_n) \in \hat{\mathcal{A}}_n$ such that

$$\sup_n \sup_{s \leq T} \mathbb{E}[|\xi_n(s)|] < \infty,$$

$$\sup_n \sup_{s \leq T} \mathbb{E}[|\varphi_n(s)|] < \infty,$$

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[(\xi_n(t) - h(\mathbf{X}_n(t))) \prod_{i=1}^k f_i(\mathbf{X}_n(t_i)) \right] = 0,$$

and

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[(\varphi_n(t) - g(\mathbf{X}_n(t))) \prod_{i=1}^k f_i(\mathbf{X}_n(t_i)) \right] = 0,$$

for all $k \geq 0, 0 \leq t_1 < t_2 < \dots < t_k \leq t \leq T$, and $f_1, \dots, f_k \in \mathcal{M}$.

La variable aléatoire $\varphi_n(t)$ représente le pseudo-générateur. Elle a été définie à la section 2.6.1. La variable aléatoire $\xi_n(t)$ est définie comme suit : $\xi_n(t) := h(\mathbf{Z}_{1,2}^n(t)) + \varphi_n(t^*)(t - t^*)$ où t^* est tel que $nt^* = \lfloor nt \rfloor$.

En utilisant l'inégalité de Jensen,

$$\begin{aligned} \left| \mathbb{E} \left[(\varphi_n(t) - g(X_n(t))) \prod_{i=1}^k f_i(X_n(t_i)) \right] \right| &= \left| \mathbb{E} \left[(\varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t))) \prod_{i=1}^k f_i(X_n(t_i)) \right] \right| \\ &\leq \mathbb{E} \left[|\varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t))| \left| \prod_{i=1}^k f_i(X_n(t_i)) \right| \right] \\ &\leq \kappa \mathbb{E} \left[|\varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t))| \right], \end{aligned}$$

où κ est une constante positive (les fonctions f_i sont bornées). Donc, si

$$\mathbb{E} \left[|\varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t))| \right] \rightarrow 0,$$

lorsque $n \rightarrow \infty$, alors

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[(\varphi_n(t) - g(X_n(t))) \prod_{i=1}^k f_i(X_n(t_i)) \right] = 0.$$

C'est pour cela que nous avons affirmé que le résultat (c) se résume à la convergence $\mathbb{E}[|\varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t))|] \rightarrow 0$. En effet, il s'agit de la pierre angulaire de la démonstration menant à la convergence faible $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\} \Rightarrow \{\mathbf{Z}(t), t \geq 0\}$ lorsque $n \rightarrow \infty$; c'est le résultat dont la démonstration demande le plus de travail. Les autres conditions du résultat (c) se vérifient assez facilement à partir de la démonstration de cette convergence. Le reste de cette section est dédié à la vérification détaillée des autres conditions du résultat (c).

Premièrement, en utilisant l'inégalité du triangle et l'inégalité de Jensen pour espérance conditionnelle,

$$\mathbb{E}[|\xi_n(t)|] \leq M + n(t - t^*) \mathbb{E} \left[\mathbb{E} \left[|h(\mathbf{Z}_{1,2}^n(t + 1/n)) - h(\mathbf{Z}_{1,2}^n(t))| \mid \mathcal{F}^{\mathbf{Z}^n}(t) \right] \right] \leq 3M,$$

car $|h| \leq M$ et $0 \leq n(t - t^*) < 1$. Donc, $\sup_n \sup_{s \leq T} \mathbb{E}[|\xi_n(s)|] < \infty$.

Ensuite, on vérifie que $\sup_n \sup_{s \leq T} \mathbb{E}[|\varphi_n(s)|] < \infty$. En utilisant l'inégalité du triangle,

$$|\varphi_n(t)| \leq |\varphi_{1,n}(t)| + |\varphi_{2,n}(t) + \varphi_{3,n}(t)|.$$

On analyse premièrement le deuxième terme à droite de l'inégalité. Il a été démontré à la section 2.6 que

$$\varphi_{2,n}(t) = \frac{n(1-\tau)}{A+1} \left(h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \frac{p(K+1)}{p(K)} \mathbb{1}(1 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1),$$

et que

$$\varphi_{3,n}(t) = \frac{n(1-\tau)}{A+1} \left(h(R^K - 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor).$$

En effectuant un développement en série de Taylor de h par rapport à $R^K + 1/\sqrt{n}$ autour de R^K , on obtient

$$h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) = \frac{1}{\sqrt{n}} h_x(R^K, X_1^K) + \frac{1}{2n} h_{xx}(W, X_1^K),$$

où W appartient à $(R^K, R^K + 1/\sqrt{n})$. Par ailleurs, en effectuant un développement en série de Taylor de h par rapport à $R^K - 1/\sqrt{n}$ autour de R^K , on obtient

$$h(R^K - 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) = -\frac{1}{\sqrt{n}} h_x(R^K, X_1^K) + \frac{1}{2n} h_{xx}(T, X_1^K),$$

où T appartient à $(R^K - 1/\sqrt{n}, R^K)$. Ainsi,

$$\begin{aligned} \varphi_{2,n}(t) + \varphi_{3,n}(t) &= \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) \\ &\quad + \mathbb{1}(K=1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \sqrt{n} \times \frac{p(K+1)}{p(K)} \\ &\quad - \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \sqrt{n} \\ &\quad + \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{2(A+1)} \left(h_{xx}(W, X_1^K) \frac{p(K+1)}{p(K)} + h_{xx}(T, X_1^K) \right) \\ &\quad + \mathbb{1}(K=1) \frac{1-\tau}{2(A+1)} h_{xx}(W, X_1^K) \frac{p(K+1)}{p(K)} \\ &\quad + \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{2(A+1)} h_{xx}(T, X_1^K). \end{aligned}$$

On analyse l'espérance de la valeur absolue de chacun de ces termes. En se basant sur la démonstration du lemme 2.2, on sait que

$$\sup_n \sup_{t \leq T} \mathbb{E} \left[\left| \mathbb{1}(K=1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \sqrt{n} \times \frac{p(K+1)}{p(K)} \right| \right] < \infty,$$

$$\sup_n \sup_{t \leq T} \mathbb{E} \left[\left| \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \sqrt{n} \right| \right] < \infty.$$

Pour ce qui est des termes liés à la deuxième dérivée, ils sont tous bornés. En effet, $|h_{xx}| \leq M$, $0 \leq \mathbb{1} \leq 1$, $0 \leq p(K+1)/p(K) \leq 2$, et τ et A sont des constantes. Donc, afin de démontrer que $\sup_n \sup_{t \leq T} \mathbb{E} \left[|\varphi_{2,n}(t) + \varphi_{3,n}(t)| \right] < \infty$, il ne reste qu'à analyser le premier terme. En considérant le cas où $\lfloor \sqrt{n} \log n \rfloor$ est impair,

$$\begin{aligned} & \mathbb{E} \left[\left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \sqrt{n} \left(\frac{p(K+1)}{p(K)} - 1 \right) \right| \right] \\ & \leq \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \sqrt{n} \left| \frac{p(K+1)}{p(K)} - 1 \right| \right| \right] \\ & = \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \sqrt{n} \left| \frac{p(K+1)}{p(K)} - 1 \right| \right| \right] \\ & \quad + \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \sqrt{n} \left| \frac{p(K+1)}{p(K)} - 1 \right| \right| \right], \end{aligned}$$

car $|h_x| \leq M$. On analyse chacun des termes séparément. Le premier correspond au cas où K est au mode ou à sa droite. Donc, $p(K+1)/p(K) = a_{K,n}$ et

$$\begin{aligned} & \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \sqrt{n} \left| \frac{p(K+1)}{p(K)} - 1 \right| \right| \right] \\ & = \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \sqrt{n} |a_{K,n} - 1| \right| \right] \\ & = \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left| \frac{K - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \right| \right| \right] \\ & \leq \frac{(1-\tau)M}{A+1} \mathbb{E} [|R^K|], \end{aligned}$$

car $a_{K,n} = 1 - b_{K,n}/\sqrt{n}$, $b_{K,n} = |K - \lfloor \sqrt{n} \log n \rfloor / 2| / \sqrt{n}$, $R^K = (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$, et $\mathbb{1} \in \{0, 1\}$. Ensuite, on étudie le cas où K est à gauche du mode. Donc, $p(K+1)/p(K) = a_{K,n}^{-1}$ et

$$\begin{aligned} & \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \sqrt{n} \left| \frac{p(K+1)}{p(K)} - 1 \right| \right| \right] \\ & = \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \sqrt{n} |a_{K,n}^{-1} - 1| \right| \right] \\ & = \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{\sqrt{n}}{\sqrt{n} - b_{K,n}} \right| \times b_{K,n} \right| \right] \\ & = \frac{(1-\tau)M}{A+1} \mathbb{E} \left[\left| \mathbb{1} \left(\frac{4 - \lfloor \sqrt{n} \log n \rfloor}{2\sqrt{n}} \leq R^K \leq \frac{-1}{2\sqrt{n}} \right) \left| \frac{\sqrt{n}}{\sqrt{n} - b_{K,n}} \right| \left| \frac{K - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \right| \right| \right] \\ & \leq \frac{\sqrt{n}(1-\tau)M}{(A+1)(\sqrt{n} - \log n)} \mathbb{E} [|R^K|] \leq \frac{4(1-\tau)M}{A+1} \mathbb{E} [|R^K|], \end{aligned}$$

car $a_{K,n} = 1 - b_{K,n}/\sqrt{n}$, $b_{K,n} = |K - \lfloor \sqrt{n} \log n \rfloor / 2| / \sqrt{n}$, $R^K = (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$, $\mathbb{1} \in \{0, 1\}$, et $b_{K,n} = |R^K| \leq \log n$ lorsque $(4 - \lfloor \sqrt{n} \log n \rfloor) / (2\sqrt{n}) \leq R^K \leq -1 / (2\sqrt{n})$. Donc, en utilisant la proposition 3.3, on sait que

$$\sup_n \sup_{t \leq T} \mathbb{E} \left[|\varphi_{2,n}(t) + \varphi_{3,n}(t)| \right] < \infty.$$

On démontre maintenant que $\sup_n \sup_{t \leq T} \mathbb{E} \left[|\varphi_{1,n}(t)| \right] < \infty$. En effectuant un développement en série de Taylor de la probabilité d'acceptation

$$\left(1 \wedge \prod_{i=1}^{n+K} \frac{f(Y_i^K)}{f(X_i^K)} \right) = \left(1 \wedge \exp \left\{ \sum_{i=1}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right),$$

par rapport à Y_1^K , et ceci autour du point X_1^K , on obtient

$$\begin{aligned} \varphi_{1,n}(t) &= \tau n \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K)) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right] \\ &\quad + \tau n \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K)) (Y_1^K - X_1^K) (\log f(\epsilon_1^K))' \right. \\ &\quad \left. \times \exp \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \right], \end{aligned}$$

où ϵ_1^K appartient à (X_1^K, Y_1^K) ou à (Y_1^K, X_1^K) . On analyse ces deux termes séparément. On débute par le premier. En effectuant un développement en série de Taylor de h par rapport à Y_1^K , et ceci autour du point X_1^K , on obtient

$$\begin{aligned} &\tau n \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K)) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid R^K, \mathbf{X}^K \right] \\ &= \tau n h_y(R^K, X_1^K) \mathbb{E} \left[(Y_1^K - X_1^K) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] \\ &\quad + \frac{\tau n}{2} \mathbb{E} \left[h_{yy}(R^K, W) (Y_1^K - X_1^K)^2 \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right], \end{aligned}$$

où W appartient à (X_1^K, Y_1^K) ou à (Y_1^K, X_1^K) , et car $R^K := (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$. En utilisant l'indépendance conditionnelle de Y_1^K, \dots, Y_{n+K}^K sachant (K, \mathbf{X}^K) , on obtient

$$\tau n h_y(R^K, X_1^K) \mathbb{E} \left[(Y_1^K - X_1^K) \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] = 0,$$

car $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2 / (n + K))$. Ensuite, en utilisant l'inégalité de Jensen pour espérance conditionnelle,

$$\mathbb{E} \left[\left| \frac{\tau n}{2} \mathbb{E} \left[h_{yy}(R^K, W) (Y_1^K - X_1^K)^2 \left(1 \wedge \exp \left\{ \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \right) \mid K, \mathbf{X}^K \right] \right| \right]$$

$$\leq \frac{\tau n M}{2} \mathbb{E} \left[\mathbb{E} \left[(Y_1^K - X_1^K)^2 \mid K, \mathbf{X}^K \right] \right] = \frac{\tau n M}{2} \mathbb{E} \left[\frac{\ell^2}{n+K} \right] \leq \frac{\tau n M \ell^2}{2(n+1)} \leq \frac{\tau M \ell^2}{2},$$

car $|h_{yy}| \leq M$, $0 \leq 1 \wedge \exp\{x\} \leq 1$, $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$, et $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$.

On analyse maintenant le deuxième terme de $\varphi_{1,n}(t)$. En effectuant un développement en série de Taylor de h par rapport à Y_1^K , et ceci autour du point X_1^K ,

$$\begin{aligned} & \tau n \mathbb{E} \left[(h(R^K, Y_1^K) - h(R^K, X_1^K))(Y_1^K - X_1^K)(\log f(\epsilon_1^K))' \right. \\ & \quad \times \exp \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid R^K, \mathbf{X}^K \Big] \\ & = \tau n \mathbb{E} \left[h_y(R^K, W)(Y_1^K - X_1^K)^2 (\log f(\epsilon_1^K))' \right. \\ & \quad \times \exp \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid K, \mathbf{X}^K \Big], \end{aligned}$$

où W appartient à (X_1^K, Y_1^K) ou à (Y_1^K, X_1^K) ($R^K := (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$). Ensuite, en utilisant l'inégalité de Jensen pour espérance conditionnelle,

$$\begin{aligned} & \mathbb{E} \left[\left| \tau n \mathbb{E} \left[h_y(R^K, W)(Y_1^K - X_1^K)^2 (\log f(\epsilon_1^K))' \right. \right. \right. \\ & \quad \times \exp \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} \right\} \mathbb{1} \left\{ \log \frac{f(\epsilon_1^K)}{f(X_1^K)} + \sum_{i=2}^{n+K} \log \frac{f(Y_i^K)}{f(X_i^K)} < 0 \right\} \mid K, \mathbf{X}^K \Big] \Big| \Big] \\ & \leq \tau n M \mathbb{E} \left[\mathbb{E} \left[(Y_1^K - X_1^K)^2 |(\log f(\epsilon_1^K))'| \mid K, \mathbf{X}^K \right] \right], \end{aligned}$$

car $|h_y| \leq M$ et $0 \leq \exp\{x\} \mathbb{1}\{x < 0\} \leq 1$. Tel qu'expliqué à la démonstration du lemme 3.1,

$$|(\log f(\epsilon_1^K))'| \leq |(\log f(X_1^K))'| + M |Y_1^K - X_1^K|.$$

Donc, en utilisant l'inégalité du triangle,

$$\begin{aligned} \tau n M \mathbb{E} \left[\mathbb{E} \left[(Y_1^K - X_1^K)^2 |(\log f(\epsilon_1^K))'| \mid K, \mathbf{X}^K \right] \right] & \leq \tau n M \mathbb{E} \left[|(\log f(X_1^K))'| \mathbb{E} \left[(Y_1^K - X_1^K)^2 \mid K, \mathbf{X}^K \right] \right. \\ & \quad \left. + \tau n M^2 \mathbb{E} \left[\mathbb{E} \left[|Y_1^K - X_1^K|^3 \mid K, \mathbf{X}^K \right] \right] \right] \\ & = \tau n M \mathbb{E} \left[|(\log f(X_1^K))'| \frac{\ell^2}{n+K} \right] + \tau n M^2 \mathbb{E} \left[\frac{2\ell^3}{(n+K)^{3/2}} \sqrt{\frac{2}{\pi}} \right] \\ & \leq \tau M \ell^2 \mathbb{E} \left[|(\log f(X_1^K))'| \right] + \tau M^2 2\ell^3 \sqrt{\frac{2}{\pi}}, \end{aligned}$$

car $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n+K))$ et $K \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$. Ainsi,

$$\sup_n \sup_{t \leq T} \mathbb{E} \left[|\varphi_{1,n}(t)| \right] \leq \sup_n \sup_{t \leq T} \left\{ \frac{\tau M \ell^2}{2} + \tau M \ell^2 \mathbb{E} \left[|(\log f(X_1^K))'| \right] + \tau M^2 2\ell^3 \sqrt{\frac{2}{\pi}} \right\} < \infty,$$

car $\mathbb{E}[|(\log f(X_1^K))'|] < \infty$ ($\mathbb{E}[|((\log f(X_1^K))')^4|] < \infty$ par hypothèse et cette quantité ne dépend pas de n).

On démontre maintenant que $\lim_{n \rightarrow \infty} \mathbb{E}[(\xi_n(t) - h(X_n(t))) \prod_{i=1}^k f_i(X_n(t_i))] = 0$. En fait, on démontre plutôt que $\lim_{n \rightarrow \infty} \mathbb{E}[(\xi_n(t) - h(X_n(t)))] = 0$, ce qui est suffisant car les fonctions f_i sont bornées. On a que $\xi_n(t) - h(\mathbf{Z}_{1,2}^n(t)) = \varphi_n(t^*)(t - t^*)$ et donc,

$$\mathbb{E} \left[\left| \xi_n(t) - h(\mathbf{Z}_{1,2}^n(t)) \right| \right] = (t - t^*) \mathbb{E}[|\varphi_n(t^*)|] \leq (t - t^*) \sup_n \sup_{t \leq T} \mathbb{E} [|\varphi_n(t)|].$$

On sait que $\sup_n \sup_{t \leq T} \mathbb{E} [|\varphi_n(t)|] < \infty$ et que $(t - t^*) < 1/n \rightarrow 0$, donc

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\left| \xi_n(t) - h(\mathbf{Z}_{1,2}^n(t)) \right| \right] = 0.$$

Il ne reste qu'à démontrer que $\xi_n(t) - \int_0^t \varphi_n(s) ds$ est $\mathcal{F}^{\mathbf{Z}^n}(t)$ -martingale. On a que

$$\begin{aligned} \xi_n(t) - \int_0^t \varphi_n(s) ds &= h(\mathbf{Z}_{1,2}^n(t)) + \varphi_n(t^*)(t - t^*) - \left[\int_0^{t^*} \varphi_n(s) ds + \int_{t^*}^t \varphi_n(s) ds \right] \\ &= h(\mathbf{Z}_{1,2}^n(t)) - \int_0^{t^*} \varphi_n(s) ds, \end{aligned}$$

car $\varphi_n(s) = \varphi_n(t^*)$ pour tout $s \in [t^*, t)$ ($nt^* \leq nt < nt^* + 1 \Leftrightarrow t^* \leq t < t^* + 1/n$), ce qui implique que

$$\varphi_n(t^*)(t - t^*) = \int_{t^*}^t \varphi_n(s) ds.$$

En effet, $\{\varphi_n(t), t \geq 0\}$ est un processus à sauts dont l'intervalle entre les sauts est régulier et de $1/n$ (voir la définition de $\{\varphi_n(t), t \geq 0\}$ à la section 2.6.1). Alors,

$$\begin{aligned} &\mathbb{E} \left[\xi_n(t) - \int_0^t \varphi_n(u) du - \xi_n(s) + \int_0^s \varphi_n(u) du \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right] \\ &= \mathbb{E} \left[h(\mathbf{Z}_{1,2}^n(t)) - h(\mathbf{Z}_{1,2}^n(s)) - \int_{s^*}^{t^*} \varphi_n(u) du \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right]. \end{aligned}$$

L'objectif est de démontrer que cette espérance conditionnelle est égale à 0. En utilisant la définition de l'intégrale stochastique, où t_0^N, \dots, t_N^N forment une partition de $[s^*, t^*]$ telle que $\max(t_i^N - t_{i-1}^N) \rightarrow 0$ lorsque $N \rightarrow \infty$,

$$\begin{aligned} \int_{s^*}^{t^*} \varphi_n(u) du &= \lim_{N \rightarrow \infty} \sum_{i=0}^{N-1} \varphi_n(t_i^N)(t_{i+1}^N - t_i^N) \\ &= \varphi_n(s^*) \frac{1}{n} + \varphi_n(s^* + 1/n) \frac{1}{n} + \dots + \varphi_n(t^* - 1/n) \frac{1}{n} \\ &= \frac{1}{n} \sum_{i=ns^*}^{nt^*-1} \varphi_n(i/n), \end{aligned}$$

car le processus $\varphi_n(u) = \varphi_n(u^*)$ pour tout $u \in [u^*, u^* + 1/n)$, où $u^* := \lfloor ut \rfloor$. Alors,

$$\begin{aligned} \mathbb{E} \left[\int_{s^*}^{t^*} \varphi_n(u) du \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right] &= \mathbb{E} \left[\frac{1}{n} \sum_{i=ns^*}^{nt^*-1} \varphi_n(i/n) \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right] \\ &= \sum_{i=ns^*}^{nt^*-1} \mathbb{E} \left[\mathbb{E}[h(\mathbf{Z}_{1,2}^n((i+1)/n)) - h(\mathbf{Z}_{1,2}^n(i/n)) \mid \mathcal{F}^{\mathbf{Z}^n}(i/n)] \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right], \end{aligned}$$

en utilisant la définition de $\varphi_n(t)$. Ensuite, en utilisant le fait que $\mathcal{F}^{\mathbf{Z}^n}(s) \subseteq \mathcal{F}^{\mathbf{Z}^n}(i/n)$ pour tout $i \in \{ns^*, \dots, nt^* - 1\}$,

$$\begin{aligned} &\sum_{i=ns^*}^{nt^*-1} \mathbb{E} \left[\mathbb{E}[h(\mathbf{Z}_{1,2}^n((i+1)/n)) - h(\mathbf{Z}_{1,2}^n(i/n)) \mid \mathcal{F}^{\mathbf{Z}^n}(i/n)] \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right] \\ &= \mathbb{E} \left[\sum_{i=ns^*}^{nt^*-1} h(\mathbf{Z}_{1,2}^n((i+1)/n)) - h(\mathbf{Z}_{1,2}^n(i/n)) \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right] \\ &= \mathbb{E} \left[h(\mathbf{Z}_{1,2}^n(t^*)) - h(\mathbf{Z}_{1,2}^n(s^*)) \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right] = \mathbb{E} \left[h(\mathbf{Z}_{1,2}^n(t)) - h(\mathbf{Z}_{1,2}^n(s)) \mid \mathcal{F}^{\mathbf{Z}^n}(s) \right]. \end{aligned}$$

On a donc que $\xi_n(t) - \int_0^t \varphi_n(s) ds$ est $\mathcal{F}^{\mathbf{Z}^n}(t)$ -martingale et alors, $(\xi_n, \varphi_n) \in \hat{\mathcal{A}}_n$. Ceci termine la vérification des conditions du résultat (c) du théorème 3.1.

3.5. COROLLAIRE 8.6 DU CHAPITRE 4 DE [ETHIER ET KURTZ \(1986\)](#)

Tel que mentionné à la section 2.6, le théorème 3.1 permet d'établir la convergence faible des distributions à dimensions finies de $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$ vers celles de $\{\mathbf{Z}(t), t \geq 0\}$ lorsque $n \rightarrow \infty$. Des hypothèses supplémentaires garantissent la convergence faible $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\} \Rightarrow \{\mathbf{Z}(t), t \geq 0\}$ lorsque $n \rightarrow \infty$. Des conditions suffisantes sont énoncées dans le corollaire 8.6 tiré du chapitre 4 de [Ethier et Kurtz \(1986\)](#). Son énoncé est le suivant (il est à noter que tous les concepts et notations qui sont utilisés dans l'énoncé, mais qui ne sont pas encore définis, le seront ultérieurement) :

Corollaire 3.1 (8.6). *Suppose in Theorem 3.1 that $\{\mathbf{X}_n(t), t \geq 0\}$ and $\{\mathbf{X}(t), t \geq 0\}$ have sample paths in $D_E[0, \infty)$, and that there is an algebra $C_a \subset L$ that separates points. Suppose either that the compact containment condition ((7.9) of Chapter 3) holds for $\{\mathbf{X}_n(t), n = 1, 2, \dots\}$ or that C_a strongly separates points. If $(\{\xi_n(t), t \geq 0\}, \{\varphi_n(t), t \geq 0\})$ in condition (c) can be selected so that*

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\sup_{t \in \mathbb{Q} \cap [0, T]} |\xi_n(t) - h(\mathbf{X}_n(t))| \right] = 0,$$

and

$$\sup_n \mathbb{E}[\|\varphi_n\|_{p,T}] < \infty \text{ for some } p \in (1, \infty],$$

where $\|\varphi_n\|_{p,T} = [\int_0^T |\varphi_n(t)|^p dt]^{1/p}$ if $p < \infty$ and $\|\varphi_n\|_{\infty,T} = \text{ess sup}_{0 \leq t \leq T} |\varphi_n(t)|$, then $\{\mathbf{X}_n(t), t \geq 0\} \Rightarrow \{\mathbf{X}(t), t \geq 0\}$.

Afin d'utiliser le corollaire 3.1, les conditions mentionnées dans son énoncé doivent être vérifiées. La section 3.5.1 est dédiée à ceci. Notez que le résultat $\{\mathbf{X}_n(t), t \geq 0\} \Rightarrow \{\mathbf{X}(t), t \geq 0\}$ indique

que la convergence suivante est valide :

$$\lim_{n \rightarrow \infty} \mathbb{E}[h(\{\mathbf{X}_n(t), t \geq 0\})] = \mathbb{E}[h(\{\mathbf{X}(t), t \geq 0\})],$$

pour toute fonction h continue bornée.

3.5.1. Vérification des conditions du corollaire 3.1

Tout d'abord, la notation $D_E[0, \infty)$ représente l'espace des fonctions continues à droite $x : [0, \infty) \rightarrow E$ admettant une limite à gauche (les fonctions càdlàg) à valeurs dans E . Les processus $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$ et $\{\mathbf{Z}(t), t \geq 0\}$ ont des trajets dans $D_{\mathbb{R}^2}[0, \infty)$ car $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$ est un processus à sauts et $\{\mathbf{Z}(t), t \geq 0\}$ est un processus dont les trajets sont continus.

Ensuite, on vérifie que $C_a := C_c^\infty(\mathbb{R}^2) \subset L$ est une algèbre, où L est la fermeture du domaine de $B := \{(h, Gh) : h \in C_c^\infty(\mathbb{R}^2)\}$. Il faut donc vérifier que pour toutes fonctions $h, g \in C_c^\infty(\mathbb{R}^2)$, on a $hg \in C_c^\infty(\mathbb{R}^2)$, $h + g \in C_c^\infty(\mathbb{R}^2)$ et que pour tout $c \in \mathbb{R}$, $ch \in C_c^\infty(\mathbb{R}^2)$.

À la section 3.2.1, il a déjà été vérifié que pour toutes fonctions $h, g \in C_c^\infty(\mathbb{R}^2)$, on a $h + g \in C_c^\infty(\mathbb{R}^2)$ et que pour tout $c \in \mathbb{R}$, $ch \in C_c^\infty(\mathbb{R}^2)$. Ainsi, il ne reste qu'à vérifier que pour toutes fonctions $h, g \in C_c^\infty(\mathbb{R}^2)$, on a $hg \in C_c^\infty(\mathbb{R}^2)$.

Si h et g sont à support compact, il est clair que hg est à support compact. Aussi, la dérivée d'ordre quelconque de hg est un polynôme faisant intervenir les dérivées partielles de h et g . Donc, hg est infiniment dérivable. Alors, $C_c^\infty(\mathbb{R}^2)$ est une algèbre de fonctions sur \mathbb{R}^2 .

On définit maintenant la notion de séparation des points, puis on vérifie que $C_c^\infty(\mathbb{R}^2)$ sépare fortement les points.

Définition 3.3. *Un ensemble de fonctions $\mathcal{M} \subset \overline{C}(E)$ sépare les points si pour tous $x, y \in E$ où $x \neq y$, il existe une fonction $h \in \mathcal{M}$ telle que $h(x) \neq h(y)$. Par ailleurs, cet ensemble de fonctions sépare fortement les points si pour tout $x \in E$ et pour tout $\delta > 0$, il existe un ensemble fini $\{h_1, \dots, h_k\} \subset \mathcal{M}$ tel que*

$$\inf_{y: d(x,y) \geq \delta} \max_{1 \leq i \leq k} |h_i(y) - h_i(x)| > 0.$$

Pour tout $x \in \mathbb{R}^2$, il existe une fonction $h \in C_c^\infty(\mathbb{R}^2)$ qui atteint son maximum global (ou minimum global) en x et seulement en x , ce qui implique que pour tout $\delta > 0$,

$$\inf_{y: d(x,y) \geq \delta} |h(y) - h(x)| > 0.$$

Alors, $C_a := C_c^\infty(\mathbb{R}^2) \subset L$ est une algèbre qui sépare fortement les points. Le théorème 4.5 du chapitre 3 de [Ethier et Kurtz \(1986\)](#) indique que $C_c^\infty(\mathbb{R}^2)$ est conséquemment *convergence determining*. Ceci implique que, si $\mathbf{W}_n \in \mathbb{R}^2$ et $\mathbf{W} \in \mathbb{R}^2$ sont des variables aléatoires, et si

$$\lim_{n \rightarrow \infty} \mathbb{E}[h(\mathbf{W}_n)] = \mathbb{E}[h(\mathbf{W})],$$

pour tout $h \in C_c^\infty(\mathbb{R}^2)$, alors \mathbf{W}_n converge en distribution vers \mathbf{W} lorsque $n \rightarrow \infty$.

On démontre maintenant que $\lim_{n \rightarrow \infty} \mathbb{E} \left[\sup_{t \in \mathbb{Q} \cap [0, T]} |\xi_n(t) - h(X_n(t))| \right] = 0$. On a $\xi_n(t) - h(\mathbf{Z}_{1,2}^n(t)) = \varphi_n(t^*)(t - t^*)$ (car $\xi_n(t) := h(\mathbf{Z}_{1,2}^n(t)) + \varphi_n(t^*)(t - t^*)$) et

$$\begin{aligned} |\varphi_n(t^*)(t - t^*)| &= |n(t - t^*) \mathbb{E}[h(\mathbf{Z}_{1,2}^n(t^* + 1/n)) - h(\mathbf{Z}_{1,2}^n(t^*)) \mid \mathcal{F}^{\mathbf{Z}^n}(t^*)]| \\ &\leq \left| \mathbb{E}[h(\mathbf{Z}_{1,2}^n(t + 1/n)) - h(\mathbf{Z}_{1,2}^n(t)) \mid \mathcal{F}^{\mathbf{Z}^n}(t)] \right|, \end{aligned}$$

car $n(t - t^*) < 1$ ($nt^* = \lfloor nt \rfloor$). Tel qu'expliqué à la section 2.6.1,

$$\begin{aligned} &\mathbb{E}[h(\mathbf{Z}_{1,2}^n(t + 1/n)) - h(\mathbf{Z}_{1,2}^n(t)) \mid \mathcal{F}^{\mathbf{Z}^n}(t)] \\ &= \tau \mathbb{E} \left[\left(h(R^K, Y_1^K) - h(R^K, X_1^K) \right) \left(1 \wedge \frac{\prod_{i=1}^{n+K} f(Y_i^K)}{\prod_{i=1}^{n+K} f(X_i^K)} \right) \mid K, \mathbf{X}^K \right] \\ &+ \frac{1 - \tau}{A + 1} \left(h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \frac{p(K + 1)}{p(K)} \mathbb{1}(1 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \\ &+ \frac{1 - \tau}{A + 1} \left(h(R^K - 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor). \end{aligned}$$

On analyse chacun des termes séparément et la conclusion suivra par l'inégalité du triangle. On débute par le premier terme. En utilisant l'inégalité de Jensen pour espérance conditionnelle puis un développement en série de Taylor de la fonction h par rapport à Y_1^K , et ceci autour de X_1^K ,

$$\begin{aligned} &\left| \tau \mathbb{E} \left[\left(h(R^K, Y_1^K) - h(R^K, X_1^K) \right) \left(1 \wedge \frac{\prod_{i=1}^{n+K} f(Y_i^K)}{\prod_{i=1}^{n+K} f(X_i^K)} \right) \mid K, \mathbf{X}^K \right] \right| \\ &\leq \tau \mathbb{E} \left[\left| h(R^K, Y_1^K) - h(R^K, X_1^K) \right| \mid K, \mathbf{X}^K \right] \\ &= \tau \mathbb{E} \left[\left| h_y(R^K, W) \right| |Y_1^K - X_1^K| \mid K, \mathbf{X}^K \right] \\ &\leq \tau M \mathbb{E} \left[|Y_1^K - X_1^K| \mid K, \mathbf{X}^K \right] = \frac{\tau M \ell}{\sqrt{n}} \sqrt{\frac{2}{\pi}} \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, car $0 \leq 1 \wedge x \leq 1$ pour tout $x \geq 0$, $|h_y| \leq M$, et $Y_1^K \mid K, \mathbf{X}^K \sim \mathcal{N}(X_1^K, \ell^2/(n + K))$. La variable aléatoire W appartient à (X_1^K, Y_1^K) ou à (Y_1^K, X_1^K) .

Ensuite, en effectuant un développement en série de Taylor de la fonction h par rapport à $R^K + 1/\sqrt{n}$, et ceci autour de R^K ,

$$\begin{aligned} &\left| \frac{1 - \tau}{A + 1} \left(h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \frac{p(K + 1)}{p(K)} \mathbb{1}(1 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \right| \\ &\leq \frac{2(1 - \tau)}{A + 1} \left| h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right| \\ &= \frac{2(1 - \tau)}{\sqrt{n}(A + 1)} \left| h_x(W, X_1^K) \right| \leq \frac{2M(1 - \tau)}{\sqrt{n}(A + 1)} \rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$, car $\mathbb{1} \in \{0, 1\}$, $0 \leq p(K + 1)/p(K) \leq 2$ et $|h_x| \leq M$. La variable aléatoire W est entre R^K et $R^K + 1/\sqrt{n}$. De la même façon, on obtient

$$\left| \frac{1 - \tau}{A + 1} \left(h(R^K - 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \right| \leq \frac{M(1 - \tau)}{\sqrt{n}(A + 1)} \rightarrow 0.$$

Ainsi, $\lim_{n \rightarrow \infty} \mathbb{E}[\sup_{t \in \mathbb{Q} \cap [0, T]} |\xi_n(t) - h(X_n(t))|] = 0$.

On démontre maintenant que $\sup_n \mathbb{E}[\|\varphi_n\|_{2, T}] < \infty$. En utilisant l'inégalité de Jensen,

$$\mathbb{E} \left[\left(\int_0^T |\varphi_n(t)|^2 dt \right)^{1/2} \right] \leq \mathbb{E} \left[\int_0^T |\varphi_n(t)|^2 dt \right]^{1/2}.$$

En se basant sur les explications données à la section 3.4 en ce qui concerne l'intégrale de $\varphi_n(t)$,

$$\begin{aligned} \int_0^T |\varphi_n(t)|^2 dt &\leq \int_0^{T^*+1/n} |\varphi_n(t)|^2 dt = \lim_{N \rightarrow \infty} \sum_{i=0}^{N-1} |\varphi_n(t_i^N)|^2 (t_{i+1}^N - t_i^N) \\ &= \frac{1}{n} \sum_{i=0}^{nT^*} |\varphi_n(i/n)|^2, \end{aligned}$$

où T^* est telle que $\lfloor nT \rfloor = nT^*$ (et donc $nT^* \leq nT < nT^* + 1$) et t_0^N, \dots, t_N^N forment une partition de $[0, T^* + 1/n]$ telle que $\max(t_i^N - t_{i-1}^N) \rightarrow 0$ lorsque $N \rightarrow \infty$. Ainsi,

$$\mathbb{E} \left[\int_0^T |\varphi_n(t)|^2 dt \right] = \frac{1}{n} \sum_{i=0}^{nT^*} \mathbb{E}[|\varphi_n(i/n)|^2] = \frac{nT^* + 1}{n} \mathbb{E}[|\varphi_n(0)|^2],$$

car la distribution de $\varphi_n(i/n)$ est la même pour tout i . Donc, si $\sup_n \mathbb{E}[|\varphi_n(0)|^2] < \infty$, on saura que $\sup_n \mathbb{E}[\|\varphi_n\|_{2, T}] < \infty$.

À la section 3.4, il a aussi été démontré que

$$\begin{aligned} |\varphi_n(0)| &\leq \kappa + M\ell^2 |(\log f(X_1^K))'| \\ &\quad + \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{\sqrt{n}(1-\tau)M}{A+1} \left| \frac{p(K+1)}{p(K)} - 1 \right| \\ &\quad + \mathbb{1}(K=1) \frac{\sqrt{n}2M(1-\tau)}{A+1} + \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{\sqrt{n}(1-\tau)M}{A+1}, \end{aligned}$$

où κ est une constante positive. Cette dernière somme, mise au carré, est fonction du carré de chacun des termes puis de toutes les multiplications engendrées par deux termes. Afin de borner $\mathbb{E}[|\varphi_n(0)|^2]$, on peut analyser l'espérance du carré de chacun des termes et utiliser l'inégalité de Cauchy-Schwarz afin de borner les autres termes.

Premièrement, $M^2 \ell^4 \mathbb{E}[(\log f(X_1^K))']^2 < \infty$ ($\mathbb{E}[(\log f(X_1^K))']^4 < \infty$ par hypothèse et cette quantité ne dépend pas de n). Ensuite, on analyse les deux derniers termes. En utilisant la proposition 2.3,

$$\begin{aligned} \frac{n(2M(1-\tau))^2}{(A+1)^2} \mathbb{P}(K=1) &\rightarrow 0, \\ \frac{n((1-\tau)M)^2}{(A+1)^2} \mathbb{P}(K = \lfloor \sqrt{n} \log n \rfloor) &\rightarrow 0, \end{aligned}$$

lorsque $n \rightarrow \infty$. Finalement, en considérant le cas où $\lfloor \sqrt{n} \log n \rfloor$ est impair,

$$\left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) n \left(\frac{p(K+1)}{p(K)} - 1 \right)^2 \right]$$

$$\begin{aligned}
&= \left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) n \left(\frac{p(K+1)}{p(K)} - 1 \right)^2 \right] \\
&\quad + \left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) n \left(\frac{p(K+1)}{p(K)} - 1 \right)^2 \right],
\end{aligned}$$

On analyse chacun des termes séparément. Le premier correspond au cas où K est au mode ou à sa droite. Donc, $p(K+1)/p(K) = a_{K,n}$ et

$$\begin{aligned}
&\left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) n \left(\frac{p(K+1)}{p(K)} - 1 \right)^2 \right] \\
&= \left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) n (a_{K,n} - 1)^2 \right] \\
&= \left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left(\frac{K - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \right)^2 \right] \\
&\leq \left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} [(R^K)^2],
\end{aligned}$$

car $a_{K,n} = 1 - b_{K,n}/\sqrt{n}$, $b_{K,n} = |K - \lfloor \sqrt{n} \log n \rfloor / 2| / \sqrt{n}$, $R^K = (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$, et $\mathbb{1} \in \{0, 1\}$. Ensuite, on étudie le cas où K est à gauche du mode. Donc, $p(K+1)/p(K) = a_{K,n}^{-1}$ et

$$\begin{aligned}
&\left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) n \left(\frac{p(K+1)}{p(K)} - 1 \right)^2 \right] \\
&= \left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) n (a_{K,n}^{-1} - 1)^2 \right] \\
&= \left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left(\frac{\sqrt{n}}{\sqrt{n} - b_{K,n}} \right)^2 \times b_{K,n}^2 \right] \\
&= \left(\frac{(1-\tau)M}{A+1} \right)^2 \mathbb{E} \left[\mathbb{1} \left(\frac{4 - \lfloor \sqrt{n} \log n \rfloor}{2\sqrt{n}} \leq R^K \leq \frac{-1}{2\sqrt{n}} \right) \left(\frac{\sqrt{n}}{\sqrt{n} - b_{K,n}} \right)^2 \left(\frac{K - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \right)^2 \right] \\
&\leq \left(\frac{\sqrt{n}(1-\tau)M}{(A+1)(\sqrt{n} - \log n)} \right)^2 \mathbb{E} [(R^K)^2] \leq \left(\frac{4(1-\tau)M}{A+1} \right)^2 \mathbb{E} [(R^K)^2],
\end{aligned}$$

car $a_{K,n} = 1 - b_{K,n}/\sqrt{n}$, $b_{K,n} = |K - \lfloor \sqrt{n} \log n \rfloor / 2| / \sqrt{n}$, $R^K = (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$, $\mathbb{1} \in \{0, 1\}$, et $b_{K,n} = |R^K| \leq \log n$ lorsque $(4 - \lfloor \sqrt{n} \log n \rfloor) / (2\sqrt{n}) \leq R^K \leq -1 / (2\sqrt{n})$. Donc, en utilisant la proposition 3.3, on sait que

$$\sup_n \mathbb{E} [\|\varphi_n\|_{2,T}] < \infty.$$

Ainsi, les hypothèses supplémentaires garantissant la convergence faible $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\} \Rightarrow \{\mathbf{Z}(t), t \geq 0\}$ lorsque $n \rightarrow \infty$ sont vérifiées.

3.6. AUTRES RÉSULTATS UTILISÉS DANS LA DÉMONSTRATION DU THÉORÈME 2.1

Proposition 3.1. *Si f est une densité strictement positive telle que $f \in C^2(\mathbb{R})$ et $(\log f(x))'$ est une fonction Lipschitz continue, alors $\mathbb{E}[f''(X_1^K)/f(X_1^K)] = 0$ où $X_1^K \sim f$, et donc*

$$\mathbb{E}[(\log f(X_1^K))''] = \mathbb{E}\left[\frac{f''(X_1^K)}{f(X_1^K)} - \left(\frac{f'(X_1^K)}{f(X_1^K)}\right)^2\right] = -\mathbb{E}[(\log f(X_1^K))']^2.$$

DÉMONSTRATION DE LA PROPOSITION 3.1.

$$\begin{aligned} \mathbb{E}[f''(X_1^K)/f(X_1^K)] &= \int_{-\infty}^{\infty} f''(x)dx \\ &= \lim_{n \rightarrow \infty} \int_0^n f''(x)dx + \lim_{n \rightarrow \infty} \int_{-n}^0 f''(x)dx \\ &= \lim_{n \rightarrow \infty} f'(n) - f'(0) + f'(0) - \lim_{n \rightarrow \infty} f'(-n) \\ &= \lim_{n \rightarrow \infty} f'(n) - f'(-n) = 0, \end{aligned}$$

en utilisant le théorème fondamental du calcul à la troisième ligne et le fait que $f'(x) \rightarrow 0$ lorsque $x \rightarrow \pm\infty$ à la quatrième ligne (tel qu'indiqué par le lemme A.1 de [Bédard \(2006\)](#)). ■

Proposition 3.2. *Si X_n et Y_n , $n = 1, 2, \dots$, sont des suites de variables aléatoires telles que $X_n - Y_n \rightarrow 0$ en probabilité lorsque $n \rightarrow \infty$, alors, $\|g(X_n) - g(Y_n)\|_p \rightarrow 0$ lorsque $n \rightarrow \infty$, pour toute fonction g bornée et uniformément continue et pour tout $p \in (1, \infty)$, où $\|\cdot\|_p$ est la norme de l'espace L^p (c'est-à-dire $\|\cdot\|_p = \mathbb{E}[|\cdot|^p]^{1/p}$).*

DÉMONSTRATION DE LA PROPOSITION 3.2. Tout d'abord, pour $\delta > 0$,

$$\begin{aligned} \mathbb{E}[|g(X_n) - g(Y_n)|^p] &= \mathbb{E}[|g(X_n) - g(Y_n)|^p \mathbf{1}(|X_n - Y_n| \leq \delta)] \\ &\quad + \mathbb{E}[|g(X_n) - g(Y_n)|^p \mathbf{1}(|X_n - Y_n| > \delta)]. \end{aligned}$$

Par la continuité uniforme de g , pour tout $\epsilon > 0$, il existe un $\delta > 0$ tel que $\mathbb{E}[|g(X_n) - g(Y_n)|^p \mathbf{1}(|X_n - Y_n| \leq \delta)] \leq \epsilon \mathbb{E}[\mathbf{1}(|X_n - Y_n| \leq \delta)] \leq \epsilon$. De plus, pour n assez grand,

$$\mathbb{E}[|g(X_n) - g(Y_n)|^p \mathbf{1}(|X_n - Y_n| > \delta)] \leq (2M)^p \mathbb{P}(|X_n - Y_n| > \delta) \leq \epsilon,$$

par le fait qu'il existe une constante $M > 0$ telle que $|g| \leq M$ et par la convergence en probabilité. ■

Proposition 3.3. *La variable aléatoire $K \sim p$ (voir section 2.2 pour les hypothèses sur p) est telle que*

$$\sup_n \mathbb{E}\left[\left(\frac{K - \lfloor \sqrt{n} \log n \rfloor}{\sqrt{n}}\right)^2\right] < \infty.$$

DÉMONSTRATION DE LA PROPOSITION 3.3. On considère le cas où $\lfloor \sqrt{n} \log n \rfloor$ est pair. En utilisant l'équation (2.4),

$$\begin{aligned} \mathbb{E} \left[\left(\frac{K - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \right)^2 \right] &= \sum_{k=1}^{\lfloor \sqrt{n} \log n \rfloor} \left(\frac{k - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \right)^2 p(k) \\ &= \sum_{k=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \left(\frac{k}{\sqrt{n}} \right)^2 p \left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} \right) \prod_{i=1}^k \left(1 - \frac{i}{n} \right) + \left(\frac{1}{\sqrt{n}} \right)^2 p \left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} \right) \\ &\quad + \sum_{k=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \left(\frac{1+k}{\sqrt{n}} \right)^2 p \left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} \right) \prod_{i=1}^k \left(1 - \frac{i}{n} \right). \end{aligned}$$

Pour tout n fixe, cette quantité est finie. On analyse donc maintenant le comportement asymptotique. À la section 2.8, il a été démontré que $p(\lfloor \sqrt{n} \log n \rfloor / 2) \rightarrow 1 / \sqrt{2\pi}$ lorsque $n \rightarrow \infty$, et donc

$$\left(\frac{1}{\sqrt{n}} \right)^2 p \left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} \right) \rightarrow 0 \text{ lorsque } n \rightarrow \infty.$$

En utilisant des arguments semblables à ceux mentionnés à la section 2.8, on a

$$\begin{aligned} \sum_{k=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \left(\frac{1+k}{\sqrt{n}} \right)^2 p \left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} \right) \prod_{i=1}^k \left(1 - \frac{i}{n} \right) &\rightarrow \int_0^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx, \\ \sum_{k=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \left(\frac{k}{\sqrt{n}} \right)^2 p \left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} \right) \prod_{i=1}^k \left(1 - \frac{i}{n} \right) &\rightarrow \int_0^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx \\ &= \int_{-\infty}^0 x^2 \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx, \end{aligned}$$

lorsque $n \rightarrow \infty$. Ainsi,

$$\mathbb{E} \left[\left(\frac{K - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \right)^2 \right] \rightarrow \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx = 1 \text{ lorsque } n \rightarrow \infty.$$

■

3.7. DÉMONSTRATION DU COROLLAIRE 2.1

Tout d'abord,

$$\begin{aligned} \left| \mathbb{E} \left[1 \wedge \prod_{i=1}^{n+K} f(Y_i^K) / f(X_i^K) \right] - 2\Phi(-\ell \sqrt{\Upsilon} / 2) \right| &\leq \left| \mathbb{E} \left[1 \wedge \prod_{i=1}^{n+K} f(Y_i^K) / f(X_i^K) \right] - \mathbb{E}[1 \wedge \exp\{\psi^K\}] \right| \\ &\quad + \left| \mathbb{E}[1 \wedge \exp\{\psi^K\}] - 2\Phi(-\ell \sqrt{\Upsilon} / 2) \right|, \end{aligned}$$

où

$$\psi^K := \sum_{i=1}^{n+K} (\log f(X_i^K))' (Y_i^K - X_i^K) - \frac{\ell^2}{2(n+K)} ((\log f(X_i^K))')^2.$$

L'indice de temps a été omis puisque $(K(m), \mathbf{X}^K(m)) \sim \pi_n$ et $\mathbf{Y}^{K(m)}(m+1) \sim \mathcal{N}(\mathbf{X}^{K(m)}(m), (\ell^2/(n+K(m)))\mathcal{I}_{n+K(m)})$ pour tout $m \in \mathbb{N}$. En utilisant des arguments semblables à ceux de la démonstration du lemme 3.3, on peut démontrer que

$$\left| \mathbb{E} \left[1 \wedge \prod_{i=1}^{n+K} f(Y_i^K)/f(X_i^K) \right] - \mathbb{E}[1 \wedge \exp\{\psi^K\}] \right| \rightarrow 0,$$

lorsque $n \rightarrow \infty$. De plus, en utilisant des arguments semblables à ceux de la démonstration du lemme 3.4, on peut démontrer que

$$\left| \mathbb{E}[1 \wedge \exp\{\psi^K\}] - 2\Phi(-\ell \sqrt{\Upsilon}/2) \right| \rightarrow 0,$$

lorsque $n \rightarrow \infty$.

RÉFÉRENCES

- Bédard, M. 2006, «On the robustness of optimal scaling for random walk Metropolis algorithms», thèse de doctorat, University of Toronto.
- Ethier, S. N. et T. G. Kurtz. 1986, «Markov processes : Characterization and convergence», Wiley.
- Green, P. J. 1995, «Reversible jump Markov chain Monte Carlo computation and Bayesian model determination», *Biometrika*, vol. 82, n° 4, p. 711–732.
- Roberts, G. O., A. Gelman et W. R. Gilks. 1997, «Weak convergence and optimal scaling of random walk Metropolis algorithms», *Ann. Appl. Probab.*, vol. 7, n° 1, p. 110–120.

Chapitre 4

BAYESIAN ROBUSTNESS TO OUTLIERS IN LINEAR REGRESSION

Dans ce chapitre, un article soumis pour publication conjointement écrit avec mon co-directeur, Alain Desgagné, et ma directrice, Mylène Bédard, est présenté.

RÉSUMÉ

La robustesse complète est une propriété contribuant à la qualité d'un modèle statistique. Elle implique que l'impact des valeurs aberrantes, définies ici comme étant les observations qui ne sont pas en ligne avec la tendance générale, disparaît graduellement lorsque celles-ci s'éloignent de plus en plus de cette tendance. Ainsi, les conclusions obtenues sont cohérentes avec la majorité des observations. Les modèles non robustes peuvent mener à des résultats qui ne sont en ligne ni avec les valeurs aberrantes, ni avec les non aberrantes. Dans cet article, nous apportons la contribution suivante : nous généralisons des résultats de robustesse complète s'appliquant au modèle de régression linéaire simple passant par l'origine, au modèle de régression usuel. La stratégie pour atteindre la robustesse complète est simple. L'hypothèse traditionnelle de normalité sur le terme d'erreur est remplacée par une hypothèse de distribution à ailes extrêmement relevées. Les analyses sont donc effectuées selon la procédure habituelle et les méthodes typiques d'inférence statistique comme les tests d'hypothèses sont disponibles.

ABSTRACT

Whole robustness is an appealing attribute to look for in statistical models. It implies that the impact of outliers, defined here as the observations that are not in line with the general trend, gradually vanishes as they move further and further away from this trend. Consequently, the conclusions obtained are consistent with the bulk of the data. Nonrobust models may lead to inferences that are not in line with either the outliers or the nonoutliers. In this paper, we make the following contribution: we generalise existing whole robustness results for simple linear regression through the origin to the usual linear regression model. The strategy to attain whole robustness is simple:

replace the traditional normal assumption on the error term by a super heavy-tailed distribution assumption. Analyses are then conducted as usual, and typical methods of inference as statistical hypothesis testing are available.

MSC 2010 subject classifications: Primary 62J05; secondary 62F35.

Keywords: Bayesian inference; Built-in robustness; Maximum likelihood estimation; Super heavy-tailed distributions; Whole robustness.

4.1. INTRODUCTION

The linear regression model is commonly used to predict values of a dependent variable through auxiliary information (a set of observations from explanatory variables), or to quantify the strength of the relationship between the dependent variable and the explanatory variables. The underlying assumption in this model is that the explanatory variables are linearly related to the dependent variable, with an error term that traditionally is normally distributed. It is well understood that conflicting sources of information may contaminate the inference when normality is assumed. From a Bayesian perspective, these conflicting sources may represent the prior and outliers. We focus on the latter in this paper. A conflict therefore represents the fact that a group of observations produces a rather different inference than that arising from the bulk of the data and the prior. Under the normality assumption, this may lead to the following undesirable effect caused by the light normal tails: the posterior concentrates in an area that is not supported by any source of information in order to incorporate them all. In this context of robustness against outliers in linear regression, a contaminated inference corresponds to predictions that are not in line with either the outliers or the nonoutliers, and misleading quantification of the strength of the relationship between the dependent variable and the explanatory variables. We believe that the appropriate way to address the problem is to limit the impact of outliers in order to obtain conclusions that are consistent with the majority of the observations.

[Box and Tiao \(1968\)](#) were the first to propose a Bayesian solution for dealing with outliers in linear regression. They suggested to let the error term be distributed as a mixture of two normals: one component for the nonoutliers and the other one, with a larger variance, for the outliers. This approach has been generalised by [West \(1984\)](#) who modelled errors with heavy-tailed distributions constructed as scale mixtures of normals, which includes the Student distribution. [Peña et al. \(2009\)](#) introduced more recently a different robust Bayesian approach in which the weight of each observation decreases with the distance between this observation and the bulk of the data. The authors proved that the Kullback-Leibler divergence between the posterior arising from the nonoutliers only and the posterior arising from the sample containing outliers is bounded. This result essentially indicates that these two posterior densities can be different, but that their ratio is bounded.

The strategies of [Box and Tiao \(1968\)](#) and [West \(1984\)](#) are appealing by their simplicity. There is however a lack of theoretical results that validate the underlying ideas. In fact, heavy-tailed distributions as the Student only allow to attain partial robustness (see the results of [Andrade and O'Hagan \(2011\)](#) in a context of location-scale model), which means that outliers have, as they approach plus or minus infinity, a significant but limited impact on the inference. Consequently, we believe that there is still a need for simple strategies with evidences of their validity. In this paper, we aim at filling this gap through an approach as simple as those of [Box and Tiao \(1968\)](#) and [West \(1984\)](#), and theoretical results ensuring whole robustness. As these authors, we replace the traditional normal assumption on the error term by a distribution assumption that accommodates for the presence of outliers. The tails of the distribution are however heavier than those of heavy-tailed distributions; they are super heavy tails (e.g. log-Pareto or log-Student tails). We borrowed the idea from [Desgagné and Gagnon \(2016\)](#), in which simple linear regressions through the origin are considered and whole robustness is proved. Our work is thus aligned with the *theory of conflict resolution in Bayesian statistics*, as described by [O'Hagan and Pericchi \(2012\)](#) in their extensive literature review on that topic.

In the following, we first present the general model (without any specific distribution assumption on the error term) in Section [4.2.1](#). We then describe the super heavy-tailed distributions that we use to attain whole robustness in Section [4.2.2](#). They are log-regularly varying distributions. The linear regression model with an error term assumed to follow this type of distribution is characterised by its built-in robustness that resolves conflicts in a sensitive and automatic way. This shall be witnessed in Section [4.2.3](#) in which the robustness results are presented. The key result is the convergence of the posterior distribution (arising from the whole sample) towards the posterior arising from the nonoutliers only, when the outliers approach plus or minus infinity. This corresponds to whole robustness and indicates that users can conduct their analyses as usual, estimating parameters with posterior medians and Bayesian credible intervals for instance. The result follows from the pointwise convergence of the posterior density towards that arising from the nonoutliers only. The posterior density being proportional to the likelihood function if a prior proportional to 1 is used, we are able to establish the robustness of the maximum likelihood estimates. Our strategy is therefore also useful under the frequentist paradigm.

In Section [4.3](#), we illustrate the practical implications of the theoretical results presented in Section [4.2.3](#) through a numerical study. In particular, we show the relevance of our approach in the analysis of the relationship between the dependent variable and the explanatory variables. The relationship between two stock market indexes is studied. We also evaluate the accuracy of the estimates arising from the robust multiple linear regression model. Our robust model is compared with the nonrobust (with the normal assumption) and partially robust (with the Student distribution assumption) models. It is showed that our model performs as well as the nonrobust and partially robust ones in absence of outliers, in addition to being completely robust. It indicates that, by

simply changing the assumption on the error term, we obtain adequate estimates whether there are outliers or not.

4.2. RESOLUTION OF CONFLICTS IN LINEAR REGRESSION

4.2.1. Model

(i) Let $Y_1, \dots, Y_n \in \mathbb{R}$ be n random variables representing data points from the dependent variable and $\mathbf{x}_1^T := (1, x_{12}, \dots, x_{1p}), \dots, \mathbf{x}_n^T := (1, x_{n2}, \dots, x_{np})$ be n vectors of observations from the explanatory variables, where $p \in \{2, 3, \dots\}$, $n \geq p + 1$ and $x_{ij} \in \mathbb{R}$ are assumed to be known. We want to study the relationship between the dependent variable and the explanatory variables assuming that the following model is suitable:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n,$$

where the n random variables $\epsilon_1, \dots, \epsilon_n \in \mathbb{R}$ and the p -dimensional random variable $\boldsymbol{\beta} := (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ represent the errors and regression coefficients, respectively. These $n + 1$ random variables are conditionally independent given $\sigma > 0$, the scale parameter, with a conditional density for ϵ_i given by

$$\epsilon_i | \boldsymbol{\beta}, \sigma \stackrel{\mathcal{D}}{=} \epsilon_i | \sigma \stackrel{\mathcal{D}}{\sim} \frac{1}{\sigma} f\left(\frac{\epsilon_i}{\sigma}\right), \quad i = 1, \dots, n.$$

- (ii) We assume that f is a strictly positive continuous probability density function on \mathbb{R} that is symmetric with respect to the origin, and that is such that both tails of $|z|f(z)$ are monotonic (see Section 4.5 for the detailed definition of monotonicity), which implies that the tails of $f(z)$ are also monotonic. These assumptions on f imply that $f(z)$ and $|z|f(z)$ are bounded on the real line, and both converge to 0 as $|z| \rightarrow \infty$. Note that the function $f(z)$ can have parameters, e.g. a shape parameter; however, their value is assumed to be known.
- (iii) We assume that the joint prior density of $\boldsymbol{\beta}$ and σ , denoted $\pi(\boldsymbol{\beta}, \sigma)$, is bounded on $\sigma > 1$, and is such that $\pi(\boldsymbol{\beta}, \sigma)/(1/\sigma)$ is bounded on $0 < \sigma \leq 1$, for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Together, these assumptions are equivalent to: $\pi(\boldsymbol{\beta}, \sigma)/\max(1, 1/\sigma)$ is bounded on $\sigma > 0$, for all $\boldsymbol{\beta} \in \mathbb{R}^p$. A large variety of priors fit within this assumed structure; for instance, this is the case for all proper densities. In addition, non-informative priors such as $\pi(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$, the usual one for this type of random variables, and $\pi(\boldsymbol{\beta}, \sigma) \propto 1$ satisfy these assumptions.

Inference on the parameters $\boldsymbol{\beta}$ and σ helps us understanding the relationship between the dependent and the explanatory variables. For simplicity, we assume that all explanatory variables are continuous.

Given that the scale parameter of the distribution of the error term is σ , homoscedasticity is an underlying assumption of the model. When the classical normality is assumed, σ also represents the (conditional) standard deviation of each error ϵ_i . Note that if in addition $\pi(\boldsymbol{\beta}, \sigma) \propto 1$, the maximum a posteriori probability (MAP) estimator of $\boldsymbol{\beta}$ is the well-known least square estimator. The

MAP estimator corresponds to the maximum likelihood estimator when the prior is proportional to 1.

An important drawback of the classical normality assumption is that outliers have a significant impact on the estimation. In this paper, we study robustness of the estimation of $\boldsymbol{\beta}$ and σ . The objective is to find sufficient conditions to attain whole robustness, meaning that the impact of outliers gradually vanishes as they move further and further away from the trend emerging from the bulk of the data (this is the definition of outliers on which we focus in this paper). In other words, we want to attain robustness against observations with errors $\epsilon_i = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$ approaching $+\infty$ or $-\infty$, where $\boldsymbol{\beta}$ defines the probable hyperplanes for the bulk of the data. Our strategy to mathematically represent this situation is to let y_i approach $+\infty$ or $-\infty$ for the outliers and to consider the vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ as fixed. Studying this theoretical framework is indeed sufficient to attain in practise robustness against any type of outliers that fits our definition. Consider for instance an outlier with an extreme \mathbf{x} value, but a y value comparable to those in the bulk of the data. This outlier can be viewed as an observation with a fixed \mathbf{x} value, but an extremely low (or high) y value compared with the trend emerging from the bulk of the data.

We now set our mathematical framework for dealing with outliers. Among the n observations of Y_1, \dots, Y_n , denoted by $\mathbf{y}_n := (y_1, \dots, y_n)$, we assume that k of them form a group of nonoutlying observations, that we denote \mathbf{y}_k , while $\ell = n - k$ of them are considered as outliers. For $i = 1, \dots, n$, we define the binary functions k_i and ℓ_i as follows: if y_i is a nonoutlying value $k_i = 1$, and if it is an outlier $\ell_i = 1$. These functions take the value of 0 otherwise. Therefore, we have $k_i + \ell_i = 1$ for $i = 1, \dots, n$, with $\sum_{i=1}^n k_i = k$, and $\sum_{i=1}^n \ell_i = \ell$. We assume that each outlier converges towards $-\infty$ or $+\infty$ at its own specific rate, to the extent that the ratio of two outliers is bounded. More precisely, we assume that $y_i = a_i + b_i \omega$, for $i = 1, \dots, n$, where a_i and b_i are constants such that $a_i \in \mathbb{R}$ and

$$\text{(i)} \quad b_i = 0 \text{ if } k_i = 1, \quad \text{(ii)} \quad b_i \neq 0 \text{ if } \ell_i = 1,$$

and we let $\omega \rightarrow \infty$.

Let the joint posterior density of $\boldsymbol{\beta}$ and σ be denoted by $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)$ and the marginal density of (Y_1, \dots, Y_n) be denoted by $m(\mathbf{y}_n)$, where

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) = [m(\mathbf{y}_n)]^{-1} \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0.$$

Let the joint posterior density of $\boldsymbol{\beta}$ and σ arising from the nonoutlying observations only be denoted by $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)$ and the corresponding marginal density be denoted by $m(\mathbf{y}_k)$, where

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) = [m(\mathbf{y}_k)]^{-1} \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n \left[(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i}, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0.$$

Note that if the prior $\pi(\boldsymbol{\beta}, \sigma)$ is proportional to 1, the likelihood functions, given by the product terms in the posteriors above, can also be expressed as follows:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) = m(\mathbf{y}_n)\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) \quad \text{and} \quad \mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) = m(\mathbf{y}_k)\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k). \quad (4.1)$$

Proposition 4.1. *Consider the Bayesian context described in Section 4.2.1, and assume that $k > p + 1$. Then, the joint posterior densities $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)$ and $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)$ are proper. The condition $k \geq p + 1$ is sufficient if we assume that $\pi(\boldsymbol{\beta}, \sigma)/(1/\sigma)$ is bounded on $\sigma > 0$ (instead of $\pi(\boldsymbol{\beta}, \sigma)$ is bounded on $\sigma > 1$ and $\pi(\boldsymbol{\beta}, \sigma)/(1/\sigma)$ is bounded on $0 < \sigma \leq 1$).*

The proof of Proposition 4.1 is provided in Section 4.5.1. Note that, in fact, we only need $n > p + 1$ (or $n \geq p + 1$ if we assume that $\pi(\boldsymbol{\beta}, \sigma)/(1/\sigma)$ is bounded on $\sigma > 0$) to guarantee that $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)$ is proper. The condition $k > p + 1$ (or $k \geq p + 1$) is stronger.

4.2.2. Log-Regularly Varying Distributions

As mentioned in the introduction, our approach to attain robustness is to replace the traditional normal assumption on the error term by a log-regularly varying distribution assumption. In this section, we provide an overview of such distributions. For more information, we refer the reader to [Desgagné \(2013\)](#) and [Desgagné \(2015\)](#).

Definition 4.1 (Log-regularly varying distribution). *A random variable Z with a symmetric density $f(z)$ is said to have a log-regularly varying distribution with index $\rho \geq 1$ if $zf(z)$ is log-regularly varying at ∞ with this same index ρ (see Definition 4.2).*

Definition 4.2 (Log-regularly varying function). *We say that a measurable function g is log-regularly varying at ∞ with index $\rho \in \mathbb{R}$, denoted $g \in L_\rho(\infty)$, if*

$$\forall \epsilon > 0, \forall \tau \geq 1, \text{ there exists a constant } A(\epsilon, \tau) > 0 \text{ such that} \\ z \geq A(\epsilon, \tau) \text{ and } 1/\tau \leq \nu \leq \tau \Rightarrow |\nu^\rho g(z^\nu)/g(z) - 1| < \epsilon.$$

If $\rho = 0$, g is said to be log-slowly varying at ∞ .

Log-regularly varying functions form an interesting class of functions with useful properties for robustness. As indicated in Definition 4.2, they are such that $g \in L_\rho(\infty)$ if $g(z^\nu)/g(z)$ converges towards $\nu^{-\rho}$ uniformly in any set $\nu \in [1/\tau, \tau]$ (for any $\tau \geq 1$) as $z \rightarrow \infty$, where $\rho \in \mathbb{R}$. This implies that for any $\rho \in \mathbb{R}$, we have $g \in L_\rho(\infty)$ if and only if there exists a constant $A > 1$ and a function $s \in L_0(\infty)$ such that for $z \geq A$, g can be written as $g(z) = (\log z)^{-\rho} s(z)$. It means that a symmetric distribution with tails behaving like $(1/|z|)(\log |z|)^{-\rho}$ with $\rho \geq 1$ is a log-regularly varying distribution. Note that an example of such distributions is presented in Section 4.3.1.

4.2.3. Resolution of conflicts

We now present the main contribution of this paper, which is Theorem 4.1. This theorem contains our robustness results which are stated for models using auxiliary information, i.e. for models with $p \geq 2$. The case $p = 1$ corresponds to the location-scale model, and we refer the

reader to [Desgagné \(2015\)](#) for suitable assumptions under which robustness is attained in this specific situation.

Theorem 4.1. *Consider the model and context described in Section 4.2.1. If we assume that*

(i) *f is a log-regularly varying distribution (i.e. $zf(z) \in L_\rho(\infty)$ with $\rho \geq 1$),*

(ii) *$\ell \leq n/2 - (p - 1/2) \Leftrightarrow k \geq n/2 + (p - 1/2)$ (i.e. the difference between the number of nonoutliers and the number of outliers, $k - \ell$, is at least $2(p - 1/2)$),*

then, recalling that $y_i = a_i + b_i\omega$ with $b_i = 0$ for the nonoutliers and $b_i \neq 0$ for the outliers, we obtain the following results:

(a)

$$\lim_{\omega \rightarrow \infty} \frac{m(\mathbf{y}_n)}{\prod_{i=1}^n [f(y_i)]^{\ell_i}} = m(\mathbf{y}_k),$$

(b)

$$\lim_{\omega \rightarrow \infty} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) = \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k),$$

uniformly on $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$, for any $\lambda \geq 0$ and $\tau \geq 1$,

(c)

$$\lim_{\omega \rightarrow \infty} \int_0^\infty \int_{\mathbb{R}^p} |\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) - \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)| d\boldsymbol{\beta} d\sigma = 0,$$

(d) *as $\omega \rightarrow \infty$,*

$$\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n \xrightarrow{\mathcal{D}} \boldsymbol{\beta}, \sigma \mid \mathbf{y}_k,$$

and in particular

$$\beta_i \mid \mathbf{y}_n \xrightarrow{\mathcal{D}} \beta_i \mid \mathbf{y}_k, \quad i = 1, \dots, p, \quad \text{and} \quad \sigma \mid \mathbf{y}_n \xrightarrow{\mathcal{D}} \sigma \mid \mathbf{y}_k,$$

(e)

$$\lim_{\omega \rightarrow \infty} [m(\mathbf{y}_k)/m(\mathbf{y}_n)] \mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) = \mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k),$$

uniformly on $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$, for any $\lambda \geq 0$ and $\tau \geq 1$.

The proof of Theorem 4.1 is provided in Section 4.5.2.

Theorem 4.1 is of great practical use considering the simplicity of its two sufficient conditions. Condition (i) indicates that modelling must be performed using a density f with sufficiently heavy tails, and more precisely, using a log-regularly varying distribution (see Definition 4.1). Well-known heavy-tailed distributions, such as the Student distribution, do not satisfy this criterion. [Desgagné \(2015\)](#) introduced an appealing distribution family that belongs to the family of log-regularly varying distributions, and therefore, satisfies condition (i): the family of log-Pareto-tailed symmetric distributions. In our opinion, the most appealing log-Pareto-tailed symmetric distribution is a distribution called the log-Pareto-tailed standard normal distribution, with parameters $\alpha > 1$ and $\varphi > 1$. It exactly matches the standard normal on the interval $[-\alpha, \alpha]$, while having tails that behave like $(1/|z|)(\log |z|)^{-\varphi}$ (which is a log-Pareto behaviour). This distribution shall be described in more details in the numerical study in Section 4.3.

Condition (ii) indicates that there must be at most $\lfloor n/2 - (p - 1/2) \rfloor$ outliers, or equivalently, that there must be at least $\lceil n/2 + (p - 1/2) \rceil$ nonoutliers, where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the floor and ceiling functions, respectively. For instance, consider a sample of size $n = 15$ with one explanatory variable (the simple linear regression model is therefore used, i.e. $p = 2$). Robustness is attained provided that there are $\ell = 6$ outliers or less (which leaves $k = 9$ nonoutliers or more). The breakdown point, which is the proportion of outliers that an estimator can handle, is therefore $6/15 = 0.4$ when $n = 15$. More generally, the condition $\ell \leq n/2 - (p - 1/2)$ translates into a breakdown point of $\lfloor n/2 - (p - 1/2) \rfloor / n$. For fixed p , the breakdown point thus converges to 0.5 as $n \rightarrow \infty$, usually considered as the maximum desired value. Condition (ii) is generally satisfied in practise given that the number of outliers rarely reaches half the sample size (minus $p - 1/2$).

Another feature of Theorem 4.1 that contributes to its practical use is that the results are easy to interpret. In result (a), the asymptotic behaviour of the marginal $m(\mathbf{y}_n)$ is described. This result is used in Section 4.3.1 to assess robustness of Bayes factors for testing $H_0 : \beta_i = 0$ versus $H_0 : \beta_i \neq 0$ (when $i \geq 2$) through the convergence $m(\mathbf{y}_n)/m(\mathbf{y}_n | \beta_i = 0) \rightarrow m(\mathbf{y}_k)/m(\mathbf{y}_k | \beta_i = 0)$ as $\omega \rightarrow \infty$. Result (a) is in fact the centrepiece of Theorem 1; its demonstration requires considerable work, and then leads relatively easily to the other conclusions of the theorem. Result (b) indicates that the posterior density arising from the whole sample converges uniformly towards the posterior density arising from the nonoutliers only. The impact of outliers then gradually vanishes as they approach plus or minus infinity. Result (b) leads to result (c), stating that the posterior density arising from the whole sample converges in L^1 towards the posterior density arising from the nonoutlying observations only. This last result implies the following convergence: $\mathbb{P}(\boldsymbol{\beta}, \sigma \in E | \mathbf{y}_n) \rightarrow \mathbb{P}(\boldsymbol{\beta}, \sigma \in E | \mathbf{y}_k)$ as $\omega \rightarrow \infty$, uniformly for all sets $E \subset \mathbb{R}^p \times \mathbb{R}^+$. This result is slightly stronger than convergence in distribution (result (d)) which requires only pointwise convergence. The convergence of the posterior marginal distributions then follows. Result (d) is the most practical result; it indicates that any estimation of $\boldsymbol{\beta}$ and σ based on posterior quantiles (e.g. using posterior medians and Bayesian credible intervals) is robust to outliers. Result (e) follows from result (b) and indicates that for a given sample, the likelihood (up to a multiplicative constant that is independent of $\boldsymbol{\beta}$ and σ) converges uniformly to the likelihood arising from the nonoutliers only. Consequently, the maximum of $\mathcal{L}(\boldsymbol{\beta}, \sigma | \mathbf{y}_n)$ converges to the maximum of $\mathcal{L}(\boldsymbol{\beta}, \sigma | \mathbf{y}_k)$, and therefore, the maximum likelihood estimate also converges, as $\omega \rightarrow \infty$.

4.3. NUMERICAL STUDY

In this section, we aim at increasing the understanding of the practical implications of Theorem 4.1. This is first achieved, in Section 4.3.1, via an analysis of the relationship between two stock market indexes. Second, in Section 4.3.2, we study the behaviour of the posterior when we artificially move an observation towards plus or minus infinity. In other words, we visually represent the convergence of the posterior towards that arising from the nonoutliers only (result (d) of Theorem 4.1). In Section 4.3.3, a simulation study is conducted to evaluate the accuracy of the

estimates arising from our robust linear regression model with two explanatory variables. In all the analyses, we compare the results obtained from the robust model with those from the nonrobust (with the normal assumption) and partially robust (with the Student distribution assumption) models.

4.3.1. Analysis of the Relationship Between Two Stock Market Indexes

We know from Theorem 4.1 that the impact of outliers is asymptotically null if we assume a super heavy-tailed distribution on the error term of the linear regression model. In other words, this robust model excludes outliers in the limit. In this section, we show that this model turns out to effectively limit the impact of outliers in practise. In particular, we show how this translates into adequate conclusions, contrary to the nonrobust model. The data set presented in Figure 4.1 and Table 4. I is analysed for this purpose.

January 2011 daily returns of S&P 500 (y) and S&P/TSX (x)

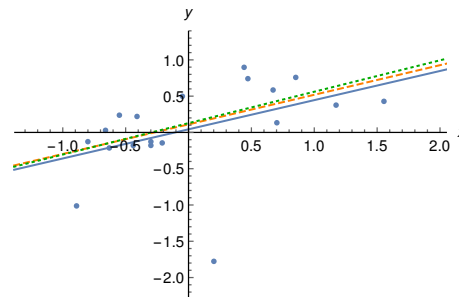


Figure 4.1. January 2011 daily returns of S&P 500 and S&P/TSX with posterior medians under the non-robust (blue solid line), partially robust (orange dashed line), and robust (green dotted line) models, respectively

y_i	-0.13	0.50	-0.21	-0.18	-0.14	0.37	0.90	-0.17	0.74	0.14
x_i	-0.30	-0.05	-0.63	-0.30	-0.20	1.18	0.44	-0.44	0.47	0.71
y_i	-1.01	-0.13	0.24	0.58	0.03	0.42	0.22	-1.79	0.77	
x_i	-0.89	-0.80	-0.55	0.67	-0.66	1.56	-0.41	0.20	0.85	

Table 4. I. Returns for day i of January 2011 for S&P 500 (y_i) and S&P/TSX (x_i) in percentages, $i = 1, \dots, 19$

The data are therefore analysed using a simple linear regression where the daily returns of the S&P 500 and S&P/TSX play the role of the dependent and explanatory variables, respectively. We set the prior to be non-informative, and more precisely, $\pi(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$. For each of the three models considered (robust, partially robust and nonrobust), we first compute posterior medians and highest posterior density (HPD) intervals for the parameters. For the partially robust model, the degrees of freedom of the heavy-tailed Student distribution are arbitrarily set to 10, and a known scale

parameter of 0.88 is added to this distribution in order to match the 2.5th and 97.5th percentiles with those of the standard normal. For the robust model, we assume that the error term has a log-Pareto-tailed standard normal distribution. Its density (depicted in Figure 4.2) is expressed as

$$f(x) = \begin{cases} (2\pi)^{-1/2} \exp(-x^2/2), & \text{if } |x| \leq \alpha \text{ (the standard normal part),} \\ (2\pi)^{-1/2} \exp(-\alpha^2/2)(\alpha/|x|)(\log \alpha / \log |x|)^\varphi, & \text{if } |x| > \alpha \text{ (the log-Pareto tails).} \end{cases} \quad (4.2)$$

We set α such that this distribution has also the same 2.5th and 97.5th percentiles as the standard normal, i.e. $\alpha = 1.96$. This implies that, according to the procedure described in Section 4 of Desgagné (2015), $\varphi = 4.08$ (this procedure ensures that f is a continuous probability density function). Therefore, all three error distributions studied in this section have 95% of their mass in the interval $[-1.96, 1.96]$.

Comparison between the standard normal, Student and log-Pareto-tailed standard normal

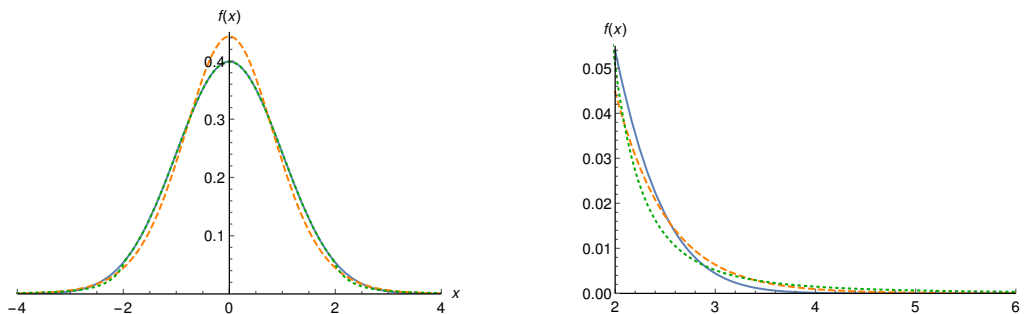


Figure 4.2. Densities of the standard normal (blue solid line), Student with 10 degrees of freedom and a known scale parameter of 0.88 (orange dashed line), and log-Pareto-tailed standard normal with $\alpha = 1.96$ and $\varphi = 4.08$ (green dotted line)

The parameter estimates are provided in Table 4. II. Contrary to the partially robust and robust models, the 95% HPD interval for β_2 of the nonrobust model includes 0. The posterior in this case puts therefore some mass around 0. We now perform Bayesian tests for hypotheses $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ through the Bayes factor $m(\mathbf{y}_n)/m(\mathbf{y}_n | H_0)$ (i.e. the marginal density of \mathbf{y}_n divided by the marginal density of \mathbf{y}_n under H_0 , which is the marginal density of \mathbf{y}_n under the simple location-scale model). The Bayes factors are 3.69, 9.23, and 23.01 for the nonrobust, partially robust and robust models, respectively. Under the robust model, it seems clear that there is enough evidence to conclude that the relationship between the variables is statistically significant. Under the nonrobust model, it is not that clear.

We now redo the analysis while excluding observation 18 (an outlier as can be seen in Figure 4.1) in order to show its impact. The results are presented in Table 4. III. The results are now relatively similar for all three models. In particular, the 95% HPD intervals for β_2 do not include 0. Also, the Bayes factors are 62.75, 51.16, and 56.37 for the nonrobust, partially robust, and robust models, respectively.

The impact of the outlier is much more significant on $\hat{\sigma}$ than on $\hat{\beta}$ (as can be seen by comparing the posterior medians in Table 4. II with those in Table 4. III). This is due to the position of $(x_{18}, y_{18}) = (0.20, -1.79)$, which is centred with respect to the x values in the data set (as seen in Figure 4.1). Its presence therefore results in a higher probability for a larger scaling. The impact of the outlier is however not only on the marginal posterior distribution of σ ; it propagates through the joint posterior of β and σ , increasing the posterior scaling of β . This is especially true for the nonrobust model (and this is what led to the differences in the test results presented above), but this is also true for the partially robust model. Our approach is appealing because it provides whole robustness for all parameters. In this example, whole robustness leads to an inference that best reflects the behaviour of the bulk of the data, compared with the inferences arising from the nonrobust and partially robust models.

Note that, as mentioned in Section 4.2.3, the Bayes factor is a robust measure when it is assumed that the error term has a super heavy-tailed distribution as the log-Pareto-tailed standard normal distribution. Indeed, result (a) of Theorem 4.1 states that the marginal $m(\mathbf{y}_n)$ behaves like $m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{\ell_i}$. Furthermore, the marginal $m(\mathbf{y}_n | H_0)$ behaves like $m(\mathbf{y}_k | H_0) \prod_{i=1}^n [f(y_i)]^{\ell_i}$, because when the assumptions of Theorem 4.1 are satisfied, those of Theorem 1 in Desgagné (2015), ensuring whole robustness for the location-scale model, are also satisfied. As a result, the Bayes factor $m(\mathbf{y}_n)/m(\mathbf{y}_n | H_0)$ behaves like $m(\mathbf{y}_k)/m(\mathbf{y}_k | H_0)$.

Assumptions on f	Estimates for		
	β_1	β_2	σ
Standard normal	0.04 (-0.25, 0.34)	0.40 (-0.02, 0.83)	0.62 (0.43, 0.88)
Student (10 d.f.)	0.11 (-0.14, 0.35)	0.41 (0.07, 0.76)	0.53 (0.33, 0.81)
Log-Pareto-tailed normal	0.13 (-0.09, 0.34)	0.43 (0.13, 0.72)	0.42 (0.26, 0.65)

Table 4. II. Posterior medians and 95% HPD intervals under the three models based on the analysis of the data set presented in Table 4. I

Assumptions on f	Estimates for		
	β_1	β_2	σ
Standard normal	0.15 (-0.03, 0.33)	0.44 (0.18, 0.69)	0.37 (0.25, 0.53)
Student (10 d.f.)	0.16 (-0.02, 0.34)	0.41 (0.16, 0.68)	0.39 (0.25, 0.58)
Log-Pareto-tailed normal	0.15 (-0.04, 0.33)	0.43 (0.17, 0.69)	0.37 (0.24, 0.54)

Table 4. III. Posterior medians and 95% HPD intervals under the three models based on the analysis of the data set presented in Table 4. I, but excluding observation 18

4.3.2. Convergence of the Posterior

In this section, we illustrate how the posterior behaves when we artificially move an observation towards plus or minus infinity, through the evolution of point estimates. More precisely, we use the same data set as in Section 4.3.1 (see Table 4. I and Figure 4.1), and we gradually move the observation identified as an outlier in that section (i.e. (x_{18}, y_{18})) according to the following procedure: y_{18} is gradually moved from the value -0.5 (a nonoutlier) to -6 (a large outlier), while x_{18} remains fixed. In Section 4.3.1, we have observed the impact of an outlier centred with respect to the x values in the data set. We therefore change the value of x_{18} to -1 to illustrate the impact of an outlier located at one of the endpoints of the x values. While we gradually move the observation, we estimate the parameters $\boldsymbol{\beta} := (\beta_1, \beta_2)^T$ and σ for each data set related to a different value of y_{18} using posterior medians with a prior proportional to $1/\sigma$. This whole process is performed under the three models (the nonrobust, partially robust, and wholly robust models with the same distribution assumptions as in Section 4.3.1). We now present the results in Figure 4.3, in which $|y_{18}| = -y_{18}$ is defined as the x -axis (instead of y_{18}) so that larger values correspond to larger distances to the bulk of the data.

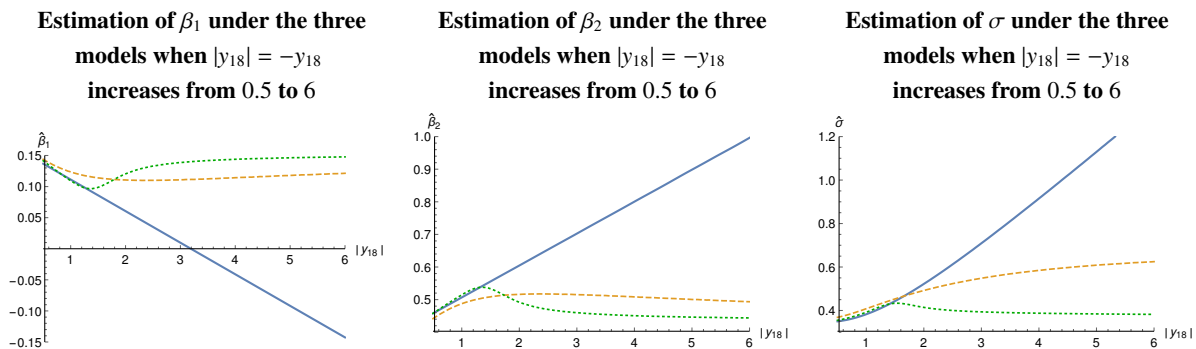


Figure 4.3. Estimation of $\boldsymbol{\beta}$ and σ when $|y_{18}| = -y_{18}$ increases from 0.5 to 6 under three different assumptions on f : standard normal density (blue solid line), Student density (orange dashed line) and log-Pareto-tailed standard normal density (green dotted line)

We first notice that, for the robust model, the impact of the point (x_{18}, y_{18}) grows until y_{18} reaches a threshold, and beyond this threshold, the impact vanishes as the observation approaches minus infinity. The threshold is around $y_{18} = -1.4$, and based on the data set with $y_{18} = -1.4$, we find $\hat{\beta}_1 = 0.10$, $\hat{\beta}_2 = 0.54$ and $\hat{\sigma} = 0.43$. The point estimates converge towards 0.16, 0.43 and 0.36 for β_1 , β_2 and σ , respectively, which are the point estimates when (x_{18}, y_{18}) is excluded from the sample. Whole robustness is, as expected, attained for both $\boldsymbol{\beta}$ and σ . Note that an increase in the value of the parameter α would result in an increase in the value of the threshold. Setting $\alpha = 1.96$ seems suitable for practical uses.

For the partially robust model, the estimation of $\boldsymbol{\beta}$ is robust as the impact of the outlier slowly decreases after a certain threshold. The impact on σ is limited, but does not decrease when the

outlying value moves away, reflecting partial robustness. The inference is clearly not robust when it is assumed that the error has a normal distribution, as the estimates of β_2 and σ grow, and that of β_1 diminishes.

4.3.3. Simulation Study

In this section, we evaluate through a simulation study the accuracy of the estimates arising from our robust linear regression model with two explanatory variables. The model is $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$ with $\boldsymbol{\beta} := (\beta_1, \beta_2, \beta_3)^T$ and $\epsilon_i \mid \sigma \stackrel{D}{\sim} (1/\sigma) f(\epsilon_i/\sigma)$, $i = 1, \dots, n$, where f is assumed to be a log-Pareto-tailed standard normal density with $\alpha = 1.96$ and $\varphi = 4.08$ in our robust model. It is compared with the same linear regression model, but where f is assumed to be a standard normal density in the nonrobust model, and a Student density with 10 degrees of freedom and a known scale parameter of 0.88 in the partially robust model. Note that the parameters of the distributions are the same as in Section 4.3.1.

For the simulation study, we set $n = 30$, $x_{1,2}, x_{2,2}, \dots, x_{30,2} = 1, 2, \dots, 30$ (which are the observations from the first explanatory variable), $x_{1,3}, x_{2,3}, \dots, x_{30,3} = 0^2, 1^2, \dots, 29^2$ (which are the observations from the second explanatory variable), and $\pi(\boldsymbol{\beta}, \sigma) \propto 1$. We simulate 50,000 data sets using values for $\boldsymbol{\beta}$ and σ respectively set to $(10, 1, -0.1)^T$ and 2 for each of the three scenarios that we now describe. In the first one, the errors are generated from a standard normal distribution; therefore, the probability to observe outliers is negligible. In the second scenario, the errors are generated from a mixture of two normals where the first component is a standard normal and the second has a mean of 10 and a variance of 1, with weights of 0.95 and 0.05, respectively. This last component can contaminate the data sets by generating extreme values. In the third and last scenario, the errors are also generated from a mixture of two normals, but the contamination is due to the second component's scaling. More precisely, the first component is again a standard normal distribution, but the second has a mean of 0 and a variance of 10^2 , with weights of 0.9 and 0.1, respectively.

For each simulated data set, we estimate $\boldsymbol{\beta}$ and σ using MAP estimation for the three models (recall that MAP estimation corresponds to maximum likelihood estimation when the prior is proportional to 1 as in this analysis). Within each simulation scenario, we evaluate the performance of each model using sample mean square errors (MSE), based on the true values $\beta_1 = 10, \beta_2 = 1, \beta_3 = -0.1$, and $\sigma = 2$; the performance of the estimation of $\boldsymbol{\beta}$ is evaluated through the sum of the MSE of the estimators of β_1, β_2 , and β_3 . The results are presented in Tables 4. IV and 4. V.

We first notice that when there is no outlier (the 100% $\mathcal{N}(0, 1)$ scenario), the accuracy of the estimates arising from the three models is approximately the same. This is not surprising given that the three densities studied are similar, especially on the interval $[-1.96, 1.96]$ in which 95% of their mass is located. They however differ in the thickness of their tails, a feature that plays a major role in robustness. The density with the slimmest tails, the normal, corresponds to the case where

the impact of outliers is the most significant. When assuming the Student distribution on the error term, the impact on $\hat{\beta}$ is limited, but the impact on $\hat{\sigma}$ is significant, reflecting partial robustness. Under the robust model, the impact on $(\hat{\beta}, \hat{\sigma})$ is, as expected, limited. Note that for all models, the impact on each of the $\hat{\beta}_i$ is in line with that on $\hat{\beta}$.

4.4. CONCLUSION

In this paper, we have proved whole robustness results for linear regression models, as the following key result: the convergence of the posterior distribution towards that arising from the nonoutliers only when the outliers approach plus or minus infinity (result (d), Theorem 4.1). These results hold under two simple and intuitive assumptions. First, assume that the error term follows a super heavy-tailed distribution to accommodate for the presence of outliers. Second, the number

Assumptions on f	Scenarios		
	100% $\mathcal{N}(0, 1)$	95% $\mathcal{N}(0, 1) + 5\%\mathcal{N}(10, 1)$	90% $\mathcal{N}(0, 1) + 10\%\mathcal{N}(0, 10^2)$
Standard normal	1.40	8.99	15.20
Student (10 d.f.)	1.42	3.04	3.61
Log-Pareto-tailed normal	1.43	1.83	2.43

Table 4. IV. Sum of the MSE of the estimators of β_1 , β_2 and β_3 under the three scenarios and the three assumptions of f

Assumptions on f	Scenarios		
	100% $\mathcal{N}(0, 1)$	95% $\mathcal{N}(0, 1) + 5\%\mathcal{N}(10, 1)$	90% $\mathcal{N}(0, 1) + 10\%\mathcal{N}(0, 10^2)$
Standard normal	0.07	9.06	23.80
Student (10 d.f.)	0.08	2.56	5.18
Log-Pareto-tailed normal	0.08	0.32	0.51

Table 4. V. MSE of the estimators of σ under the three scenarios and the three assumptions of f

of outliers must not exceed half the sample size minus $p - 1/2$. In other words, there must be at most $\lfloor n/2 - (p - 1/2) \rfloor$ outliers (or equivalently, that must be at least $\lceil n/2 + (p - 1/2) \rceil$ nonoutliers).

Although the whole robustness results are asymptotic, their practical relevance has been shown through numerical analyses in Section 4.3. The robust model used in these analyses results from the assumption that the error term follows the log-Pareto-tailed standard normal density provided in (4.2). This model has been compared with the nonrobust (with the normal assumption) and partially robust (with the Student distribution assumption) models. The conclusion is that our model performs as well as the nonrobust and partially robust models in absence of outliers, in addition to being completely robust. In fact, it performs *almost* as well in absence of outliers, as seen in Tables 4. IV and 4. V in Section 4.3.3. This is a rather small price to pay for whole

robustness. We therefore believe it is suitable to always assume that the error term has the density detailed in (4.2) to obtain adequate results, regardless of whether there are outliers or not, by computing estimates as usual from the posterior distribution.

4.5. PROOFS

The proof of Proposition 4.1 is given in Section 4.5.1 and the proof of Theorem 4.1 is given in Section 4.5.2. Prior to that, we define the constant $B > 0$ as follows:

$$B := \max \left\{ \sup_{z \in \mathbb{R}} f(z), \sup_{z \in \mathbb{R}} |z|f(z), \sup_{\beta \in \mathbb{R}^p, \sigma > 0} \pi(\beta, \sigma) / \max(1, 1/\sigma) \right\}.$$

The monotonicity of the tails of $f(z)$ and $|z|f(z)$ implies that there exists a constant $M > 0$ such that

$$|y| \geq |z| \geq M \Rightarrow f(y) \leq f(z) \text{ and } |y|f(y) \leq |z|f(z). \quad (4.3)$$

4.5.1. Proof of Proposition 4.1

To prove that $\pi(\beta, \sigma | \mathbf{y}_n)$ is proper (the proof for $\pi(\beta, \sigma | \mathbf{y}_k)$ is omitted because it is similar), it suffices to show that the marginal $m(\mathbf{y}_n)$ is finite. Recall that we require that $n > p + 1$. The reader will notice that only $n \geq p + 1$ is required if $\pi(\beta, \sigma)$ is bounded by B/σ for all $\sigma > 0$ (instead of $\pi(\beta, \sigma)$ is bounded by $B \max(1, 1/\sigma)$).

We first show that the function is integrable on the area where the ratio $1/\sigma$ is bounded. More precisely, we consider $\beta \in \mathbb{R}^p$ and $\delta M^{-1} \leq \sigma < \infty$, where δ is a positive constant that can be chosen as small as we want (upper bounds are provided in the proof). We next show that the function is integrable on the area where the ratio $1/\sigma$ approaches infinity, that is $0 < \sigma < \delta M^{-1}$. We have

$$\begin{aligned} & \int_{\delta M^{-1}}^{\infty} \int_{\mathbb{R}^p} \pi(\beta, \sigma) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \beta)/\sigma) d\beta d\sigma \\ & \stackrel{a}{\leq} B^{n-p+1} \int_{\delta M^{-1}}^{\infty} \max\left(1, \frac{1}{\sigma}\right) \frac{1}{\sigma^{n-p}} \int_{\mathbb{R}^p} \prod_{i=1}^p \frac{1}{\sigma} f\left(\frac{y_i - \mathbf{x}_i^T \beta}{\sigma}\right) d\beta d\sigma \\ & \stackrel{b}{\leq} \max\left(1, \frac{M}{\delta}\right) B^{n-p+1} \left| \det \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_p^T \end{pmatrix} \right|^{-1} \int_{\delta M^{-1}}^{\infty} \frac{1}{\sigma^{n-p}} d\sigma \int_{-\infty}^{\infty} f(u_1) du_1 \times \cdots \times \int_{-\infty}^{\infty} f(u_p) du_p \\ & \propto \int_{\delta M^{-1}}^{\infty} \frac{1}{\sigma^{n-p}} d\sigma \stackrel{c}{=} (M/\delta)^{n-p-1} / (n-p-1) < \infty. \end{aligned}$$

In step *a*, we use $\pi(\beta, \sigma) \leq B \max(1, 1/\sigma)$ and we bound each of $n - p$ densities f by B . In step *b*, we use the change of variables $u_i = (y_i - \mathbf{x}_i^T \beta)/\sigma$ for $i = 1, \dots, p$. The determinant is non-null because all explanatory variables are continuous. Indeed, consider the case $p = 2$ for instance (i.e. the simple linear regression); the determinant is different from 0 provided that $x_{12} \neq x_{22}$, and this

happens with probability 1. In step c , we use $n > p + 1$. Note that if, instead, we bound $\pi(\boldsymbol{\beta}, \sigma)$ by B/σ in step a , one can verify that the condition $n \geq p + 1$ is sufficient to bound above the integral.

We now show that the integral is finite on $\boldsymbol{\beta} \in \mathbb{R}^p$ and $0 < \sigma < \delta M^{-1}$. In this area, the ratio $(1/\sigma)$ approaches infinity. We have to carefully analyse the subareas where $y_i - \mathbf{x}_i^T \boldsymbol{\beta}$ is close to 0 in order to deal with the 0/0 form of the ratios $(y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma$. In order to achieve this, we split the domain of $\boldsymbol{\beta}$ as follows:

$$\begin{aligned} \mathbb{R}^p = & \left[\bigcap_{i_1=1}^n \mathcal{R}_{i_1}^c \right] \cup \left[\bigcup_{i_1=1}^n \left(\mathcal{R}_{i_1} \cap \left(\bigcap_{i_2=1(i_2 \neq i_1)}^n \mathcal{R}_{i_2}^c \right) \right) \right] \cup \left[\bigcup_{i_1, i_2=1(i_1 \neq i_2)}^n \left(\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \left(\bigcap_{i_3=1(i_3 \neq i_1, i_2)}^n \mathcal{R}_{i_3}^c \right) \right) \right] \\ & \cup \dots \cup \left[\bigcup_{i_1, i_2, \dots, i_p=1(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)}^n \left(\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p} \right) \right], \end{aligned} \quad (4.4)$$

where $\mathcal{R}_i := \{\boldsymbol{\beta} : |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| < \delta\}$, $i \in \{1, \dots, n\}$. The set \mathcal{R}_i represents the hyperplanes $y = \mathbf{x}_i^T \boldsymbol{\beta}$ characterised by the different values of $\boldsymbol{\beta}$ that satisfy $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| < \delta$. In other words, it represents the hyperplanes passing near the point (\mathbf{x}_i, y_i) , and more precisely, at a vertical distance of less than δ . The set $\bigcap_{i_1=1}^n \mathcal{R}_{i_1}^c$ is therefore comprised of the hyperplanes that are not passing close to any point. The set $\bigcup_{i_1=1}^n \left(\mathcal{R}_{i_1} \cap \left(\bigcap_{i_2=1(i_2 \neq i_1)}^n \mathcal{R}_{i_2}^c \right) \right)$ represents the hyperplanes passing near one (and only one) point. The set $\bigcup_{i_1, i_2=1(i_1 \neq i_2)}^n \left(\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \left(\bigcap_{i_3=1(i_3 \neq i_1, i_2)}^n \mathcal{R}_{i_3}^c \right) \right)$ represents the hyperplanes passing near two (and only two) points, and so on.

We choose δ small enough to ensure that $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p} \cap \mathcal{R}_{i_{p+1}} = \emptyset$ when i_1, \dots, i_{p+1} are all different. It is possible to do so because an hyperplane passes through no more than p points. This implies that

$$\begin{aligned} & \left[\bigcup_{i_1, i_2, \dots, i_p=1(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)}^n \left(\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p} \right) \right] \\ & = \left[\bigcup_{i_1, i_2, \dots, i_p=1(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)}^n \left(\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p} \cap \left(\bigcap_{i_{p+1}=1(i_{p+1} \neq i_1, i_2, \dots, i_p)}^n \mathcal{R}_{i_{p+1}}^c \right) \right) \right]. \end{aligned}$$

Note that all sets $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p}$ are nonempty when i_1, \dots, i_p are all different, because all explanatory variables are continuous (which implies that the $p \times p$ matrix with rows given by $\mathbf{x}_{i_1}^T, \dots, \mathbf{x}_{i_p}^T$ has a determinant different from 0). Note also that $\mathcal{R}_{i_1} \cap \left(\bigcap_{i_2=1(i_2 \neq i_1)}^n \mathcal{R}_{i_2}^c \right)$ is nonempty for all i_1 , $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \left(\bigcap_{i_3=1(i_3 \neq i_1, i_2)}^n \mathcal{R}_{i_3}^c \right)$ is nonempty for all i_1, i_2 such that $i_1 \neq i_2$, and so on. Finally note that the decomposition of \mathbb{R}^p given in (4.4) is comprised of $\sum_{i=0}^p \binom{n}{i}$ mutually exclusive sets given by $\bigcap_{i_1=1}^n \mathcal{R}_{i_1}^c$, $\mathcal{R}_{i_1} \cap \left(\bigcap_{i_2=1(i_2 \neq i_1)}^n \mathcal{R}_{i_2}^c \right)$, $i_1 = 1, \dots, n$, $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \left(\bigcap_{i_3=1(i_3 \neq i_1, i_2)}^n \mathcal{R}_{i_3}^c \right)$, $i_1, i_2 = 1, \dots, n$ with $i_1 \neq i_2$, and so on.

We now consider $0 < \sigma < \delta M^{-1}$ and $\boldsymbol{\beta}$ in one of the $\sum_{i=0}^p \binom{n}{i}$ mutually exclusive sets given in (4.4). As explained above, the difficulty lies in dealing with the hyperplanes $\boldsymbol{\beta}$ that are such that $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| < \delta$ for some points (\mathbf{x}_i, y_i) . The strategy is essentially to use the product of $(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)$ of these points to integrate over $\boldsymbol{\beta}$, and to bound the other terms of $m(\mathbf{y}_n)$. Therefore, if $\boldsymbol{\beta} \in \mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p}$, we consider the points $(\mathbf{x}_{i_1}, y_{i_1}), (\mathbf{x}_{i_2}, y_{i_2}), \dots, (\mathbf{x}_{i_p}, y_{i_p})$ to integrate over $\boldsymbol{\beta}$. If $\boldsymbol{\beta} \in \mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_{p-1}} \cap \left(\bigcap_{i_p=1(i_p \neq i_1, \dots, i_{p-1})}^n \mathcal{R}_{i_p}^c \right)$, we consider the points $(\mathbf{x}_{i_1}, y_{i_1}), (\mathbf{x}_{i_2}, y_{i_2}), \dots, (\mathbf{x}_{i_{p-1}}, y_{i_{p-1}})$,

and any other point $(\mathbf{x}_{i_p}, y_{i_p})$ to integrate over $\boldsymbol{\beta}$, and so on. We have

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) &\stackrel{a}{\leq} (B/\sigma) \max(\delta M^{-1}, 1) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\
&\propto (1/\sigma) \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \prod_{i \notin \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\
&\stackrel{b}{\leq} (1/\sigma) [(1/\sigma) f(\delta/\sigma)]^{n-p} \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\
&\stackrel{c}{\leq} [B/\delta]^{n-p-1} (1/\sigma^2) f(\delta/\sigma) \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma).
\end{aligned}$$

In step *a*, we use $\pi(\boldsymbol{\beta}, \sigma) \leq B \max(1, 1/\sigma) = (B/\sigma) \max(\sigma, 1) \leq (B/\sigma) \max(\delta M^{-1}, 1)$. In step *b*, for all $i \notin \{i_1, \dots, i_p\}$ we use $f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \leq f(\delta/\sigma)$ by the monotonicity of the tails of f because $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma \geq \delta/\sigma \geq \delta \delta^{-1} M = M$. In step *c*, we bound $n - p - 1$ terms $(1/\sigma) f(\delta/\sigma)$ by B/δ .

Finally, we bound the integral of $(1/\sigma^2) f(\delta/\sigma) \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)$ by

$$\begin{aligned}
&\int_0^\infty (1/\sigma^2) f(\delta/\sigma) \int_{\mathbb{R}^p} \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) d\boldsymbol{\beta} d\sigma \\
&\stackrel{a}{=} \left| \det \begin{pmatrix} \mathbf{x}_{i_1}^T \\ \vdots \\ \mathbf{x}_{i_p}^T \end{pmatrix} \right|^{-1} \int_0^\infty (1/\sigma^2) f(\delta/\sigma) d\sigma \stackrel{b}{=} \left| \det \begin{pmatrix} \mathbf{x}_{i_1}^T \\ \vdots \\ \mathbf{x}_{i_p}^T \end{pmatrix} \right|^{-1} \int_0^\infty f(\sigma') d\sigma' < \infty.
\end{aligned}$$

In step *a*, we use the same change of variables as above: $u_j = (y_{i_j} - \mathbf{x}_{i_j}^T \boldsymbol{\beta})/\sigma$ for $j = 1, \dots, p$. In step *b*, we use the change of variable $\sigma' = \delta/\sigma$.

4.5.2. Proof of Theorem 4.1

Consider the model and the context described in Section 4.2.1. We assume that $\ell \leq n/2 - (p - 1/2) \Leftrightarrow k \geq n/2 + (p - 1/2) \Leftrightarrow k \geq \ell + 2p - 1$. In addition, we assume that $\ell \geq 1$, i.e. that there is at least one outlier, otherwise the proof would be trivial. Two propositions and a lemma that are used in the proof of Theorem 4.1 are first given, and the proofs of results (a) to (e) follow. The proofs of these propositions and this lemma can be found in [Desgagné \(2015\)](#).

Proposition 4.2 (Dominance). *If $s \in L_0(\infty)$ and $g \in L_\rho(\infty)$, then for all $\delta > 0$, there exists a constant $A(\delta) > 1$ such that $z \geq A(\delta)$ implies that*

$$(\log z)^{-\delta} < s(z) < (\log z)^\delta \quad \text{and} \quad (\log z)^{-\rho-\delta} < g(z) < (\log z)^{-\rho+\delta}.$$

Proposition 4.3 (Location-scale transformation). *If $zf(z) \in L_\rho(\infty)$, then we have*

$$(1/\sigma) f((z - \mu)/\sigma) / f(z) \rightarrow 1 \text{ as } z \rightarrow \infty,$$

uniformly on $(\mu, \sigma) \in [-\lambda, \lambda] \times [1/\tau, \tau]$, for any $\lambda \geq 0$ and $\tau \geq 1$.

Lemma 4.1. For all $\lambda \geq 0$, $\forall \tau \geq 1$, there exists a constant $D(\lambda, \tau) \geq 1$ such that $z \in \mathbb{R}$ and $(\mu, \sigma) \in [-\lambda, \lambda] \times [1/\tau, \tau]$ implies that

$$1/D(\lambda, \tau) \leq (1/\sigma)f((z - \mu)/\sigma)/f(z) \leq D(\lambda, \tau).$$

Note that Lemma 4.1 is a corollary of Proposition 4.3.

PROOF OF RESULT (A). We first observe that

$$\begin{aligned} \frac{m(\mathbf{y}_n)}{m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{\ell_i}} &= \frac{m(\mathbf{y}_n)}{m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{\ell_i}} \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_n) d\sigma d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \frac{\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{k_i + \ell_i}}{m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{\ell_i}} d\sigma d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} d\sigma d\boldsymbol{\beta}. \end{aligned}$$

We show that the last integral converges towards 1 as $\omega \rightarrow \infty$ to prove result (a). If we use Lebesgue's dominated convergence theorem to interchange the limit $\omega \rightarrow \infty$ and the integral, we have

$$\begin{aligned} &\lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} d\sigma d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \lim_{\omega \rightarrow \infty} \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} d\sigma d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) d\sigma d\boldsymbol{\beta} = 1, \end{aligned}$$

using Proposition 4.3 in the second equality, since $\mathbf{x}_1, \dots, \mathbf{x}_n$ are fixed, and then Proposition 4.1. Note that the condition of Proposition 4.1 is satisfied because $k \geq l + 2p - 1 \Rightarrow k \geq p + 2$ (because $\ell \geq 1$ and $p \geq 2$). Note also that pointwise convergence is sufficient, for any value of $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma > 0$, once the limit is inside the integral. However, in order to use Lebesgue's dominated convergence theorem, we need to prove that the integrand is bounded by an integrable function of $\boldsymbol{\beta}$ and σ that does not depend on ω , for any value of $\omega \geq y$, where y is a constant. The constant y can be chosen as large as we want, and minimum values for y will be given throughout the proof. In order to bound the integrand, we divide the domain of integration into two areas: $1 \leq \sigma < \infty$ and $0 < \sigma < 1$. Again, we want to separately analyse the area where the ratio $1/\sigma$ approaches infinity.

We assumed that y_i can be written as $y_i = a_i + b_i \omega$, where $\omega \rightarrow \infty$, and a_i and b_i are constants such that $a_i \in \mathbb{R}$ and $b_i \neq 0$ if $\ell_i = 1$ (if the observation is an outlier). Therefore, the ranking of the elements in the set $\{|y_i| : \ell_i = 1\}$ is primarily determined by the values $|b_1|, \dots, |b_n|$, and we can choose the constant y larger than a certain threshold to ensure that this ranking remains unchanged

for all $\omega \geq y$. Without loss of generality, we assume for convenience that

$$\omega = \min_{\{i: \ell_i=1\}} |y_i| \quad \text{and consequently} \quad \min_{\{i: \ell_i=1\}} |b_i| = 1.$$

We now bound above the integrand on the first area.

Area 1: Consider $1 \leq \sigma < \infty$ and assume without loss of generality that y_1, \dots, y_p are p nonoutliers (therefore $k_1 = \dots = k_p = 1$). We have

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} &\propto \frac{\pi(\boldsymbol{\beta}, \sigma)}{\sigma^n} \prod_{i=1}^n \frac{f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{[f(y_i)]^{\ell_i}} \\ &\stackrel{a}{\leq} \frac{B}{\sigma^n} \prod_{i=1}^n \frac{D(|a_i|, 1)f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{[f(y_i)]^{\ell_i}} \\ &\stackrel{b}{\leq} \frac{1}{[f(\omega)]^\ell} \frac{B}{\sigma^n} \prod_{i=1}^n D(|a_i|, 1)f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) [|b_i|D(|a_i|, |b_i|)]^{\ell_i} \\ &\propto \frac{1}{[f(\omega)]^\ell} \frac{1}{\sigma^n} \prod_{i=1}^n f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\ &\stackrel{c}{=} \frac{1}{[f(\omega)]^\ell} \frac{1}{\sigma^n} \prod_{i=1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} [f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)]^{\ell_i} \\ &\stackrel{d}{=} \frac{\prod_{i=1}^p (1/\sigma)f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)}{\sigma^{k-p-1/2}} \left[\frac{\omega/\sigma}{\omega f(\omega)} \right]^\ell \frac{1}{\sigma^{1/2}} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} [f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)]^{\ell_i}. \end{aligned}$$

In step *a*, we use $y_i = a_i + b_i\omega$ and Lemma 4.1 to obtain

$$f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) = f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma + a_i/\sigma) \leq D(|a_i|, 1)f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma),$$

because $|a_i/\sigma| \leq |a_i|$ for all i . We also use $\pi(\boldsymbol{\beta}, \sigma) \leq B \max(1, 1/\sigma) = B$. In step *b*, we use again Lemma 4.1 to obtain $f(\omega)/f(y_i) = f((y_i - a_i)/b_i)/f(y_i) \leq |b_i|D(|a_i|, |b_i|)$. In step *c*, we set $b_i = 0$ if $k_i = 1$ and we use the symmetry of f to obtain $f(-\mathbf{x}_i^T \boldsymbol{\beta}/\sigma) = f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)$. In step *d*, we use the assumption $k_1 = \dots = k_p = 1$.

Now it suffices to demonstrate that

$$\left[\frac{\omega/\sigma}{\omega f(\omega)} \right]^\ell \frac{1}{\sigma^{1/2}} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} [f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)]^{\ell_i} \quad (4.5)$$

is bounded by a constant that does not depend on $\omega, \boldsymbol{\beta}$ and σ since $(1/\sigma)^{k-p-1/2} \prod_{i=1}^p (1/\sigma)f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)$ is an integrable function on area 1. Indeed, since $k > p + 1$, we have

$$\int_1^\infty (1/\sigma)^{k-p-1/2} \int_{\mathbb{R}^p} \prod_{i=1}^p (1/\sigma)f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma) d\boldsymbol{\beta} d\sigma = \left| \det \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_p^T \end{pmatrix} \right|^{-1} \int_1^\infty \frac{1}{\sigma^{k-p-1/2}} d\sigma < \infty,$$

using the following change of variables: $u_i = \mathbf{x}_i^T \boldsymbol{\beta} / \sigma$, $i = 1, \dots, p$. The determinant is different from 0 because all the explanatory variables are continuous. Note that if instead, in step a above, we bound $\pi(\boldsymbol{\beta}, \sigma)$ by B/σ , one can verify that the condition $k \geq p + 1$ is sufficient to bound above the integral.

In order to bound the function in (4.5), we split area 1 into three parts: $1 \leq \sigma < \omega^{1/2}$, $\omega^{1/2} \leq \sigma < \omega/(\gamma M)$ and $\omega/(\gamma M) \leq \sigma < \infty$, where M is defined in (4.3) and γ is a positive constant that can be chosen as large as we want (lower bounds are provided in the proof). Note that this split is well defined if $y > \max(1, (\gamma M)^2)$ since $\omega \geq y$.

First, consider $\omega/(\gamma M) \leq \sigma < \infty$. We have

$$\begin{aligned} & \left[\frac{\omega/\sigma}{\omega f(\omega)} \right]^\ell \frac{1}{\sigma^{1/2}} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta} / \sigma)]^{k_i} [f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta}) / \sigma)]^{\ell_i} \stackrel{a}{\leq} \frac{B^{n-p}}{\sigma^{1/2}} \left[\frac{\omega/\sigma}{\omega f(\omega)} \right]^\ell \\ & \stackrel{b}{\leq} B^{n-p} (\gamma M)^{\ell+1/2} \frac{(1/\omega)^{1/2}}{[\omega f(\omega)]^\ell} \stackrel{c}{\leq} B^{n-p} (\gamma M)^{\ell+1/2} \frac{(1/\omega)^{1/2}}{(\log \omega)^{-(\rho+1)\ell}} \\ & \stackrel{d}{\leq} B^{n-p} (\gamma M)^{\ell+1/2} [2(\rho+1)\ell/e]^{(\rho+1)\ell} < \infty. \end{aligned}$$

In step a , we use $f \leq B$. In step b , we use $\omega/\sigma \leq \gamma M$ and $1/\sigma \leq \gamma M/\omega$. In step c , we use $\omega f(\omega) > (\log \omega)^{-\rho-1}$ if $\omega \geq y \geq A(1)$, where $A(1)$ comes from Proposition 4.2. For step d , it is purely algebraic to show that the maximum of $(\log \omega)^\xi / \omega^{1/2}$ is $(2\xi/e)^\xi$ for $\omega > 1$ and $\xi > 0$, where $\xi = (\rho+1)\ell$ in our situation.

Now, consider the two other parts combined (we will split them in the next step), that is $1 \leq \sigma \leq \omega/(\gamma M)$. We have

$$\begin{aligned} & \left[\frac{\omega/\sigma}{\omega f(\omega)} \right]^\ell \frac{1}{\sigma^{1/2}} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta} / \sigma)]^{k_i} [f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta}) / \sigma)]^{\ell_i} \\ & = \frac{1}{\sigma^{1/2}} \left[\frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^\ell \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta} / \sigma)]^{k_i} \left[\frac{f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta}) / \sigma)}{f(\omega/\sigma)} \right]^{\ell_i} \\ & \stackrel{a}{\leq} \frac{1}{\sigma^{1/2}} \left[\frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^\ell B^{k-p} [D(0, \gamma)\gamma]^\ell. \end{aligned}$$

In step a , we use

$$\prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta} / \sigma)]^{k_i} \left[\frac{f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta}) / \sigma)}{f(\omega/\sigma)} \right]^{\ell_i} \leq B^{k-p} [D(0, \gamma)\gamma]^\ell. \quad (4.6)$$

The proof of this inequality is substantial. Therefore, to ease the reading, it is deferred after the demonstration that the remaining term, i.e.

$$\frac{1}{\sigma^{1/2}} \left[\frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^\ell,$$

is bounded.

We first consider $\omega^{1/2} \leq \sigma \leq \omega/(\gamma M)$. We have

$$\frac{1}{\sigma^{1/2}} \left[\frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^\ell \stackrel{a}{\leq} B^\ell \frac{(1/\omega)^{1/4}}{[\omega f(\omega)]^\ell} \stackrel{b}{\leq} B^\ell \frac{(1/\omega)^{1/4}}{(\log \omega)^{-(\rho+1)\ell}} \stackrel{c}{\leq} B^\ell [4(\rho+1)\ell/e]^{(\rho+1)\ell} < \infty.$$

In step *a*, we use $(\omega/\sigma)f(\omega/\sigma) \leq B$ and $(1/\sigma)^{1/2} \leq (1/\omega)^{1/4}$. In step *b*, we use $\omega f(\omega) > (\log \omega)^{-\rho-1}$ if $\omega \geq y \geq A(1)$, where $A(1)$ comes from Proposition 4.2. In step *c*, it is purely algebraic to show that the maximum of $(\log \omega)^\xi/\omega^{1/4}$ is $(4\xi/e)^\xi$ for $\omega > 1$ and $\xi > 0$, where $\xi = (\rho+1)\ell$ in our situation.

We now consider $1 \leq \sigma \leq \omega^{1/2}$. We have

$$\frac{1}{\sigma^{1/2}} \left[\frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^\ell \stackrel{a}{\leq} \left[\frac{\omega^{1/2} f(\omega^{1/2})}{\omega f(\omega)} \right]^\ell \stackrel{b}{\leq} 2^{(\rho+1)\ell} < \infty.$$

In step *a*, we use $1/\sigma \leq 1$ and $(\omega/\sigma)f(\omega/\sigma) \leq \omega^{1/2} f(\omega^{1/2})$ by the monotonicity of the tails of $|z|f(z)$ since $\omega/\sigma \geq \omega^{1/2} \geq y^{1/2} \geq M$ if $y \geq M^2$. In step *b*, we use $\omega^{1/2} f(\omega^{1/2})/(\omega f(\omega)) \leq 2(1/2)^{-\rho} = 2^{\rho+1}$ if $\omega \geq y \geq A(1, 2)$, where $A(1, 2)$ comes from the definition of log-regularly varying functions (see Definition 4.2).

Finally, we prove the inequality in (4.6). Recall that we assume that $k \geq n/2 + (p-1/2) \Leftrightarrow k \geq \ell + 2p - 1$. We know that y_1, \dots, y_p are p nonoutliers (therefore $k_1 = \dots = k_p = 1$). We also know that the number of remaining nonoutliers (the nonoutliers among observations $p+1$ to n) is greater than or equal to p because $k - p \geq \ell + p - 1 \geq p$ (because we assume that $\ell \geq 1$).

In order to prove the result, we split the domain of β as follows:

$$\begin{aligned} \mathbb{R}^p = & [\cap_i \mathcal{O}_i^c] \cup \left[\cup_i \left(\mathcal{O}_i \cap \left(\cap_{i_1} \mathcal{F}_{i_1}^c \right) \right) \right] \cup \left[\cup_{i, i_1} \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \left(\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c \right) \right) \right] \\ & \cup \dots \cup \left[\cup_{i, i_1, \dots, i_{p-1}} (i_j \neq i_s, \forall i_j, i_s \text{ s.t. } j \neq s) \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap \left(\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c \right) \right) \right] \\ & \cup \left[\cup_{i, i_1, \dots, i_p} (i_j \neq i_s, \forall i_j, i_s \text{ s.t. } j \neq s) \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} \right) \right], \end{aligned}$$

where

$$\mathcal{O}_i := \{\beta : |b_i \omega - \mathbf{x}_i^T \beta| < \omega/2\}, \forall i \in \mathcal{I}_O, \quad (4.7)$$

$$\mathcal{F}_i := \{\beta : |\mathbf{x}_i^T \beta| < \omega/\gamma\}, \forall i \in \mathcal{I}_\mathcal{F}, \quad (4.8)$$

$\mathcal{I}_O := \{i : i \in \{p+1, \dots, n\} \text{ and } \ell_i = 1\}$ and $\mathcal{I}_\mathcal{F} := \{i : i \in \{p+1, \dots, n\} \text{ and } k_i = 1\}$ are the sets of indexes of outliers and remaining fixed observations (nonoutliers), respectively.

The set \mathcal{O}_i represents the hyperplanes $y = \mathbf{x}_i^T \beta$ characterised by the different values of β that satisfy $|b_i \omega - \mathbf{x}_i^T \beta| < \omega/2$. In other words, it represents the hyperplanes that pass at a vertical distance of less than $\omega/2$ of the point $(\mathbf{x}_i, b_i \omega)$, which is considered as an outlier since $\omega \rightarrow \infty$ (recall that $b_i \omega = y_i - a_i$). Analogously, the set \mathcal{F}_i represents the hyperplanes that pass at a vertical distance of less than ω/γ of the point $(\mathbf{x}_i, 0)$, which is considered as a nonoutlier. Therefore, the set $\cap_i \mathcal{O}_i^c$ represents the hyperplanes that pass at a vertical distance of at least $\omega/2$ of all the

points $(\mathbf{x}_i, b_i\omega)$ (all the outliers). The set $\cup_i(\mathcal{O}_i \cap (\cap_{i_1} \mathcal{F}_{i_1}^c))$ represents the hyperplanes that pass at a vertical distance of less than $\omega/2$ of at least one point $(\mathbf{x}_i, b_i\omega)$ (an outlier), but at a vertical distance of at least ω/γ of all the points $(\mathbf{x}_i, 0)$ (all the nonoutliers). For each $i_1 \in \mathcal{I}_{\mathcal{F}}$, the set $\cup_i(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap (\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c))$ represents the hyperplanes that pass at a vertical distance of less than $\omega/2$ of at least one point $(\mathbf{x}_i, b_i\omega)$ (an outlier), at a vertical distance of less than ω/γ of the point $(\mathbf{x}_{i_1}, 0)$ (a nonoutlier), but at a vertical distance of at least ω/γ of all the other nonoutliers, and so on.

Now, we claim that $\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} = \emptyset$ for all i, i_1, \dots, i_p with $i_j \neq i_s, \forall i_j, i_s$ such that $j \neq s$. To prove this, we use the fact that \mathbf{x}_i (a vector of size p) can be expressed as a linear combination of $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_p}$. This is true because all explanatory variables are continuous, and therefore, linearly independent with probability 1. As a result, considering that $\boldsymbol{\beta} \in \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p}$ and $\mathbf{x}_i = \sum_{s=1}^p a_s \mathbf{x}_{i_s}$ for some $a_1, \dots, a_p \in \mathbb{R}$, we have

$$|b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta}| = \left| b_i\omega - \left(\sum_{s=1}^p a_s \mathbf{x}_{i_s} \right)^T \boldsymbol{\beta} \right| \stackrel{a}{\geq} |b_i\omega| - \left| \sum_{s=1}^p a_s \mathbf{x}_{i_s}^T \boldsymbol{\beta} \right| \stackrel{b}{\geq} \omega - \frac{\omega}{\gamma} \sum_{s=1}^p |a_s| \stackrel{c}{\geq} \omega - \frac{\omega}{2}.$$

In step *a*, we use the reverse triangle inequality. In step *b*, we use that $|b_i| \geq 1$ and $|\sum_{s=1}^p a_s \mathbf{x}_{i_s}^T \boldsymbol{\beta}| \leq \sum_{s=1}^p |a_s| \|\mathbf{x}_{i_s}^T \boldsymbol{\beta}\| \leq \sum_{s=1}^p |a_s| \omega/\gamma$ because $\boldsymbol{\beta} \in \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p}$, which means that $|\mathbf{x}_i^T \boldsymbol{\beta}| < \omega/\gamma$ for all $i \in \{i_1, \dots, i_p\}$. In step *c*, we define the constant γ such that $\gamma \geq 2 \sum_{s=1}^p |a_s|$ (we choose γ such that it satisfies this inequality for any combination of i and i_1, \dots, i_p). Therefore, we have that $\boldsymbol{\beta} \notin \mathcal{O}_i$. This proves that $\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} = \emptyset$ for all i, i_1, \dots, i_p with $i_j \neq i_s, \forall i_j, i_s$ such that $j \neq s$. This in turn implies that

$$\begin{aligned} \mathbb{R}^p = & [\cap_i \mathcal{O}_i^c] \cup \left[\cup_i (\mathcal{O}_i \cap (\cap_{i_1} \mathcal{F}_{i_1}^c)) \right] \cup \left[\cup_{i, i_1} (\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap (\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c)) \right] \\ & \cup \dots \cup \left[\cup_{i, i_1, \dots, i_{p-1}} (\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap (\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c)) \right]. \end{aligned}$$

This decomposition of \mathbb{R}^p is comprised of $1 + \sum_{i=0}^{p-1} \binom{k-p}{i}$ mutually exclusive sets given by $\cap_i \mathcal{O}_i^c$, $\cup_i(\mathcal{O}_i \cap (\cap_{i_1} \mathcal{F}_{i_1}^c))$, $\cup_i(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap (\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c))$ for $i_1 \in \mathcal{I}_{\mathcal{F}}$, and so on.

We are now ready to bound the function on the left-hand side in (4.6). We first show that the function is bounded on $\boldsymbol{\beta} \in \cap_i \mathcal{O}_i^c$ and $1 \leq \sigma \leq \omega/(\gamma M)$. For all $i \in \mathcal{I}_{\mathcal{O}}$, we have

$$\frac{f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega/\sigma)} \leq \frac{f(\omega/(2\sigma))}{f(\omega/\sigma)} \leq 2D(0, 2) \leq D(0, \gamma)\gamma,$$

using the monotonicity of f because $|b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma \geq \omega/(2\sigma) \geq \gamma M/2 \geq M$ (we choose $\gamma \geq 2$), and then Lemma 4.1. Therefore, on $\boldsymbol{\beta} \in \cap_i \mathcal{O}_i^c$ and $1 \leq \sigma \leq \omega/(\gamma M)$,

$$\prod_{i=p+1}^n \left[f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma) \right]^{k_i} \left[\frac{f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega/\sigma)} \right]^{\ell_i} \leq B^{k-p} [D(0, \gamma)\gamma]^\ell,$$

using $f \leq B$.

Now, we consider the area defined by: $1 \leq \sigma \leq \omega/(\gamma M)$ and $\boldsymbol{\beta}$ belongs to one of the $\sum_{i=0}^{p-1} \binom{k-p}{i}$ mutually exclusive sets $\cup_i(\mathcal{O}_i \cap (\cap_{i_1} \mathcal{F}_{i_1}^c))$, $\cup_i(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap (\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c))$ for $i_1 \in \mathcal{I}_{\mathcal{F}}$, etc. We have

$$\begin{aligned} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} \left[\frac{f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega/\sigma)} \right]^{\ell_i} &\stackrel{a}{\leq} B^\ell \prod_{i=p+1}^n \frac{[f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i}}{[f(\omega/\sigma)]^{\ell_i}} \stackrel{b}{\leq} B^{\ell+(k-p)-\ell} [D(0, \gamma)\gamma]^\ell \\ &= B^{k-p} [D(0, \gamma)\gamma]^\ell. \end{aligned}$$

In step *a*, we use $f \leq B$ for all $i \in \mathcal{I}_{\mathcal{O}}$. In step *b*, we use the fact that in any of the sets in which $\boldsymbol{\beta}$ can belong, there are at least ℓ nonoutlying points $(\mathbf{x}_i, 0)$ such that $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/\gamma$. Indeed, the case in which there are the least nonoutliers such that $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/\gamma$ corresponds to $\boldsymbol{\beta} \in \cup_i(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap (\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c))$. In this case there are $p-1$ nonoutliers such that $|\mathbf{x}_i^T \boldsymbol{\beta}| < \omega/\gamma$ (observations i_1 to i_{p-1}), which leaves $k-p-(p-1) \geq \ell$ nonoutliers such that $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/\gamma$ (i.e. that there are $k-p-(p-1) \geq \ell$ sets in the intersection $\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c$). Recall that we assume that $k \geq \ell + 2p - 1$, that y_1, \dots, y_p are p nonoutliers, and that the product above only involves observations $p+1$ to n . Therefore, for ℓ nonoutliers such that $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/\gamma$, we use

$$f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)/f(\omega/\sigma) \leq f(\omega/(\gamma\sigma))/f(\omega/\sigma) \leq D(0, \gamma)\gamma,$$

by the monotonicity of f because $|\mathbf{x}_i^T \boldsymbol{\beta}|/\sigma \geq \omega/(\gamma\sigma) \geq M$, and then Lemma 4.1. For the remaining $k-p-\ell$ nonoutlying points, we use $f \leq B$. Note that this argument justifies the need of the assumption $k \geq \ell + 2p - 1$.

Area 2: Consider $0 < \sigma < 1$. We actually need to show that

$$\lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} d\sigma d\boldsymbol{\beta} = \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) d\sigma d\boldsymbol{\beta}.$$

For area 2, we proceed in a slightly different manner than for area 1. We begin by dividing the first integral above into two parts as follows:

$$\begin{aligned} &\lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} d\sigma d\boldsymbol{\beta} \\ &= \lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} \mathbb{1}_{\cap_i \mathcal{O}_i^c}(\boldsymbol{\beta}) d\sigma d\boldsymbol{\beta} \\ &\quad + \lim_{\omega \rightarrow \infty} \int_{\cup_i \mathcal{O}_i} \int_0^1 \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} d\sigma d\boldsymbol{\beta}, \end{aligned}$$

where

$$\mathcal{O}_i := \{\boldsymbol{\beta} : |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| < \omega/2\}, \forall i \in \mathcal{I}_{\mathcal{O}},$$

with $\mathcal{I}_{\mathcal{O}} := \{i : i \in \{1, \dots, n\} \text{ and } \ell_i = 1\}$. Note that the definition of \mathcal{O}_i is very similar as that of the set defined in (4.7) (this is why we use the same notation); its interpretation is also very similar. We

show that the first part above is equal to the integral $\int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) d\sigma d\boldsymbol{\beta}$ and that the second part is equal to 0.

For the first part, we again use Lebesgue's dominated convergence theorem in order to interchange the limit $\omega \rightarrow \infty$ and the integral. We have

$$\begin{aligned} & \lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} \mathbb{1}_{\cap_i \mathcal{O}_i^c}(\boldsymbol{\beta}) d\sigma d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \lim_{\omega \rightarrow \infty} \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} \mathbb{1}_{\cap_i \mathcal{O}_i^c}(\boldsymbol{\beta}) d\sigma d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \times 1 \times \mathbb{1}_{\mathbb{R}^p}(\boldsymbol{\beta}) d\sigma d\boldsymbol{\beta} = \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) d\sigma d\boldsymbol{\beta}, \end{aligned}$$

using Proposition 4.3 in the second equality since $\mathbf{x}_1, \dots, \mathbf{x}_n$ are fixed, and $\lim_{\omega \rightarrow \infty} \mathbb{1}_{\cap_i \mathcal{O}_i^c}(\boldsymbol{\beta}) = \mathbb{1}_{\mathbb{R}^p}(\boldsymbol{\beta}) = 1 \Leftrightarrow \lim_{\omega \rightarrow \infty} \mathbb{1}_{\cup_i \mathcal{O}_i}(\boldsymbol{\beta}) = 0$. Indeed, if $\ell_i = 1$ and $b_i > 0$ (which implies that $y_i > 0$), $\boldsymbol{\beta} \in \mathcal{O}_i$ implies that $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| < \omega/2 \leq y_i/2$, which in turn implies that $y_i/2 < \mathbf{x}_i^T \boldsymbol{\beta} < 3y_i/2$, and in the limit, no $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfies this (we have the same conclusion if $b_i < 0$). Note that pointwise convergence is sufficient, for any value of $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma > 0$, once the limit is inside the integral. We now demonstrate that the integrand is bounded, for any value of $\omega \geq y$, by an integrable function of $\boldsymbol{\beta}$ and σ that does not depend on ω .

Consider $\boldsymbol{\beta} \in \cap_i \mathcal{O}_i^c$, that is $\{\boldsymbol{\beta} : |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/2 \text{ for all } i \in \mathcal{I}_O\}$, and $0 < \sigma < 1$. Note that the integrand is equal to 0 if $\boldsymbol{\beta} \notin \cap_i \mathcal{O}_i^c$. For all $i \in \mathcal{I}_O$, we have

$$(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \leq f(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \leq f(\omega/2) \leq 2|b_i|D(|a_i|, 2|b_i|)f(y_i),$$

by the monotonicity of the tails of $|z|f(z)$ and then the monotonicity of the tails of $f(z)$, because $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma \geq |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/2 \geq y/2 \geq M$, if we choose $y \geq 2M$. Lemma 4.1 is used in the last inequality with $\omega = (y_i - a_i)/b_i$. Therefore,

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} \mathbb{1}_{\cap_i \mathcal{O}_i^c}(\boldsymbol{\beta}) \leq \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n [2|b_i|D(|a_i|, 2|b_i|)]^{\ell_i},$$

which is an integrable function.

We now prove that

$$\lim_{\omega \rightarrow \infty} \int_{\cup_i \mathcal{O}_i} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} d\sigma d\boldsymbol{\beta} = 0.$$

We first bound above the integrand and then we prove that the integral of the upper bound converges towards 0 as $\omega \rightarrow \infty$. In the same manner as in the proof of the inequality in (4.6), we split the domain of $\boldsymbol{\beta}$ as follows:

$$\cup_i \mathcal{O}_i = \left[\cup_i \left(\mathcal{O}_i \cap \left(\cap_{i_1} \mathcal{F}_{i_1}^c \right) \right) \right] \cup \left[\cup_{i, i_1} \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \left(\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c \right) \right) \right]$$

$$\cup \dots \cup \left[\cup_{i,i_1,\dots,i_{p-1}(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)} \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap \left(\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c \right) \right) \right] \\ \cup \left[\cup_{i,i_1,\dots,i_p(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)} \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} \right) \right],$$

where

$$\mathcal{F}_i := \{\boldsymbol{\beta} : |\mathbf{x}_i^T \boldsymbol{\beta}| < \omega/\gamma\}, \forall i \in \mathcal{I}_{\mathcal{F}},$$

and $\mathcal{I}_{\mathcal{F}} := \{i : i \in \{1, \dots, n\} \text{ and } k_i = 1\}$. The definition of \mathcal{F}_i is the same as that of the set defined in (4.8). For an interpretation of this set and of the sets involve in the decomposition of $\cup_i \mathcal{O}_i$, see the proof of (4.6). Given that $|y_i| \geq \omega$ for all $i \in \mathcal{O}_i$, we can use the same mathematical arguments as in the proof of (4.6) to show that $\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} = \emptyset$ for all i, i_1, \dots, i_p with $i_j \neq i_s, \forall i_j \neq i_s$ such that $j \neq s$. Therefore,

$$\cup_i \mathcal{O}_i = \left[\cup_i \left(\mathcal{O}_i \cap \left(\cap_{i_1} \mathcal{F}_{i_1}^c \right) \right) \right] \cup \left[\cup_{i,i_1} \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \left(\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c \right) \right) \right] \\ \cup \dots \cup \left[\cup_{i,i_1,\dots,i_{p-1}(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)} \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap \left(\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c \right) \right) \right].$$

This decomposition of $\cup_i \mathcal{O}_i$ is comprised of $\sum_{i=0}^{p-1} \binom{k}{i}$ mutually exclusive sets given by $\cup_i (\mathcal{O}_i \cap (\cap_{i_1} \mathcal{F}_{i_1}^c))$, $\cup_i (\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap (\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c))$ for $i_1 \in \mathcal{I}_{\mathcal{F}}$, and so on. We now consider the area defined by: $0 < \sigma < 1$ and $\boldsymbol{\beta}$ belongs to one of these $\sum_{i=0}^{p-1} \binom{k}{i}$ mutually exclusive sets. We have

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} \\ \stackrel{a}{\leq} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[\frac{|b_i| D(|a_i|, |b_i|) (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega)} \right]^{\ell_i} \\ \propto \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n \left[(1/\sigma) f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i} \left[\frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega)} \right]^{\ell_i} \\ \stackrel{b}{\leq} (B/\sigma) [2\gamma D(0, 2\gamma) (1/\sigma) f(\omega/\sigma)]^{\ell+1} \prod_{i=1(i \neq i_p, \dots, i_{\ell+p})}^n \left[(1/\sigma) f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i} \\ \times \left[\frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega)} \right]^{\ell_i} \\ \propto (1/\sigma) [(1/\sigma) f(\omega/\sigma)]^{\ell+1} \prod_{i=1(i \neq i_p, \dots, i_{\ell+p})}^n \left[(1/\sigma) f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i} \\ \times \left[\frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega)} \right]^{\ell_i} \\ \stackrel{c}{\leq} (1/\sigma) (1/\sigma) f(\omega/\sigma) \prod_{i=1(i \neq i_p, \dots, i_{\ell+p})}^n \left[(1/\sigma) f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i} \left[(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{\ell_i}$$

$$\stackrel{d}{\leq} (B/\omega)(1/\sigma) \prod_{i=1(i \neq i_p, \dots, i_{\ell+p})}^n (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma).$$

In step *a*, we use Lemma 4.1 to obtain $f(\omega)/f(y_i) = f((y_i - a_i)/b_i)/f(y_i) \leq |b_i|D(|a_i|, |b_i|)$ for all $i \in \mathcal{I}_O$. In step *b*, we use $\pi(\boldsymbol{\beta}, \sigma) \leq B \max(1, 1/\sigma) = B/\sigma$. We also use that in any of the sets in which $\boldsymbol{\beta}$ can belong, there are at least $\ell+1$ nonoutlying points such that $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/\gamma$ (corresponding to $\boldsymbol{\beta} \in \mathcal{F}_i^c$ for at least $\ell+1$ nonoutlying points). Indeed, the case in which there are the least nonoutliers such that $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/\gamma$ corresponds to $\boldsymbol{\beta} \in \cup_i (\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap (\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c))$. In this case there are $p-1$ nonoutliers such that $|\mathbf{x}_i^T \boldsymbol{\beta}| < \omega/\gamma$ (say observations i_1 to i_{p-1}), which leaves $k - (p-1)$ nonoutliers such that $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/\gamma$ (i.e. that there are $k - (p-1)$ sets in the intersection $\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c$), and we know that $k - (p-1) \geq \ell + p > \ell + 1$ because we assume that $k \geq \ell + 2p - 1$ and we only consider the models with $p \geq 2$. This implies that there exists a set of $\ell + 1$ indices, say $\{i_p, \dots, i_{\ell+p}\} \subset \mathcal{I}_{\mathcal{F}}$, such that for all $i \in \{i_p, \dots, i_{\ell+p}\}$,

$$f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \leq f(\omega/(2\gamma\sigma)) \leq 2\gamma D(0, 2\gamma) f(\omega/\sigma),$$

using the monotonicity of the tails of f in the first inequality because, if we define the constant $a_{(k)} := \max_{i \in \{1, \dots, k\}} |a_i|$ with $\omega \geq y \geq (2\gamma)a_{(k)}$, we have $|a_i - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma \geq (|\mathbf{x}_i^T \boldsymbol{\beta}| - |a_i|)/\sigma \geq (\omega/\gamma - a_{(k)})/\sigma \geq \omega/(2\gamma\sigma) \geq \omega/(2\gamma) \geq y/(2\gamma) \geq M$ if we choose $y \geq 2\gamma M$. In the second inequality, we use Lemma 4.1 (as mentioned in the proof of (4.6), we choose $\gamma \geq 2$). In step *c* above, we use the monotonicity of the tails of $|z|f(z)$ to obtain $(\omega/\sigma)f(\omega/\sigma) \leq \omega f(\omega)$ for ℓ terms, because $\omega/\sigma \geq \omega \geq y \geq M$ if we choose $y \geq M$. In step *d*, we use $(1/\sigma)f(\omega/\sigma) \leq B/\omega$.

The integral of $(B/\omega)(1/\sigma) \prod_{i=1(i \neq i_p, \dots, i_{\ell+p})}^n (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)$ is bounded by

$$(B/\omega) \int_{\mathbb{R}^p} \int_0^\infty (1/\sigma) \prod_{i=1(i \neq i_p, \dots, i_{\ell+p})}^n (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) d\sigma d\boldsymbol{\beta} = (B/\omega)m(\mathbf{y}_{\mathcal{I}_R}),$$

where $m(\mathbf{y}_{\mathcal{I}_R})$ is the marginal density arising from a prior proportional to $1/\sigma$ and $n - (\ell + 1) = k - 1$ observations (\mathbf{x}_i, y_i) , $i \in \mathcal{I}_R := \{1, \dots, n\} \setminus \{i_p, \dots, i_{\ell+p}\}$. In order to prove that $(B/\omega)m(\mathbf{y}_{\mathcal{I}_R}) \rightarrow 0$ as $\omega \rightarrow \infty$, it suffices to prove that $m(\mathbf{y}_{\mathcal{I}_R})$ is bounded by a constant that does not depend on ω , because $1/\omega \rightarrow 0$. In the proof of Proposition 4.1, we proved that a marginal, as $m(\mathbf{y}_{\mathcal{I}_R})$, is bounded by a constant that does not depend on ω if the number of observations (which is $k-1$ in our case) is greater than or equal to $p+1$ if the prior divided by $1/\sigma$ is bounded (which is the case for $m(\mathbf{y}_{\mathcal{I}_R})$). Because we assume that $k \geq \ell + 2p - 1$ and $\ell \geq 1$ (the proof for the case $\ell = 0$ is trivial), and because we only consider the models with $p \geq 2$, $m(\mathbf{y}_{\mathcal{I}_R})$ is the marginal of $k-1 \geq \ell + 2p - 2 \geq p+1$ observations. As a result,

$$(B/\omega) \int_{\mathbb{R}^p} \int_0^\infty (1/\sigma) \prod_{i=1(i \neq i_p, \dots, i_{\ell+p})}^n (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) d\sigma d\boldsymbol{\beta} \rightarrow 0 \text{ as } \omega \rightarrow \infty.$$

We therefore have that

$$\int_{\cup_i \mathcal{O}_i} \int_0^1 \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} d\sigma d\boldsymbol{\beta} \rightarrow 0 \text{ as } \omega \rightarrow \infty.$$

■

PROOF OF RESULT (B). Consider $(\boldsymbol{\beta}, \sigma)$ such that $\pi(\boldsymbol{\beta}, \sigma) > 0$ (the proof for the case $(\boldsymbol{\beta}, \sigma)$ such that $\pi(\boldsymbol{\beta}, \sigma) = 0$ is trivial). We have, as $\omega \rightarrow \infty$,

$$\begin{aligned} \frac{\pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_n)}{\pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k)} &= \frac{m(\mathbf{y}_k)}{m(\mathbf{y}_n)} \times \frac{\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n [(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)]^{k_i}} \\ &= \frac{m(\mathbf{y}_k)}{m(\mathbf{y}_n)} \prod_{i=1}^n \left[(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{\ell_i} \\ &= \frac{m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{\ell_i}}{m(\mathbf{y}_n)} \prod_{i=1}^n \left[\frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{\ell_i} \rightarrow 1. \end{aligned}$$

The first ratio in the last equality does not depend on $\boldsymbol{\beta}$ and σ and converges towards 1 as $\omega \rightarrow \infty$ using result (a). Also, the product converges towards 1 uniformly in any set $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$ using Proposition 4.3 since $\mathbf{x}_1, \dots, \mathbf{x}_n$ are fixed. Furthermore, since f and $\pi(\boldsymbol{\beta}, \sigma)/\max(1, 1/\sigma)$ are bounded, $\pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k)$ is also bounded on any set $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$. Then, we have

$$\left| \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_n) - \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \right| = \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \left| \frac{\pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_n)}{\pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k)} - 1 \right| \rightarrow 0 \text{ as } \omega \rightarrow \infty.$$

■

PROOF OF RESULTS (C) AND (D). Using Proposition 4.1, we know that $\pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k)$ and $\pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_n)$ are proper. Moreover, using result (b), we have the pointwise convergence $\pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_n) \rightarrow \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k)$ as $\omega \rightarrow \infty$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma > 0$, as a result of the uniform convergence. Then, the conditions of Scheffé's theorem (see Scheffé (1947)) are satisfied and we obtain the convergence in L^1 given by result (c) as well as the following result:

$$\lim_{\omega \rightarrow \infty} \int_E \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_n) d\boldsymbol{\beta} d\sigma = \int_E \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) d\boldsymbol{\beta} d\sigma,$$

uniformly for all sets $E \subset \mathbb{R}^p \times \mathbb{R}^+$. Result (d) follows directly.

■

PROOF OF RESULT (E). Using (4.1), result (e) follows directly from result (b).

■

REFERENCES

Andrade, J. A. A. and A. O'Hagan. 2011, Bayesian robustness modelling of location and scale parameters, *Scand. J. Stat.*, vol. 38, n° 4, p. 691–711.

- Box, G. E. P. and G. C. Tiao. 1968, A Bayesian approach to some outlier problems, *Biometrika*, vol. 55, n° 1, p. 119–129.
- Desgagné, A. 2013, Full robustness in Bayesian modelling of a scale parameter, *Bayesian Anal.*, vol. 8, n° 1, p. 187–220.
- Desgagné, A. 2015, Robustness to outliers in location–scale parameter model using log-regularly varying distributions, *Ann. Statist.*, vol. 43, n° 4, p. 1568–1595.
- Desgagné, A. and P. Gagnon. 2016, Bayesian robustness to outliers in linear regression and ratio estimation, Submitted for publication.
- O’Hagan, A. and L. Pericchi. 2012, Bayesian heavy-tailed models and conflict resolution: A review, *Braz. J. Probab. Stat.*, vol. 26, n° 4, p. 372–401.
- Peña, D., R. Zamar and G. Yan. 2009, Bayesian likelihood robustness in linear models, *J. Statist. Plann. Inference*, vol. 139, n° 7, p. 2196–2207.
- Scheffé, H. 1947, A useful convergence theorem for probability distributions, *Ann. Math. Statist.*, p. 434–438.
- West, M. 1984, Outlier models and prior distributions in Bayesian linear regression, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 46, n° 3, p. 431–439.

Chapitre 5

AN EFFICIENT BAYESIAN ROBUST PRINCIPAL COMPONENT REGRESSION

Dans ce chapitre, un article soumis pour publication conjointement écrit avec ma directrice, Mylène Bédard, et mon co-directeur, Alain Desgagné, est présenté.

RÉSUMÉ

La régression sur composantes principales est un modèle de régression linéaire dans lequel les régresseurs sont des composantes principales. Ce type de modélisation est particulièrement utile dans des contextes de prédiction où le nombre de variables explicatives est élevé. Surprenamment, peu d'ouvrages traitant d'approches bayésiennes peuvent être trouvés dans la littérature. Dans cet article, nous tentons de combler certaines lacunes par l'intermédiaire de la contribution pratique suivante : nous introduisons une approche bayésienne avec toutes les instructions nécessaires à son utilisation. L'approche possède deux caractéristiques que nous croyons importantes. Premièrement, elle fait intervenir de façon efficace les composantes principales appropriées dans le processus de prédiction. Ceci est effectué en deux étapes. La première est la sélection de modèles ; la deuxième est le *model averaging*, permettant d'incorporer l'information provenant des modèles importants. Les probabilités *a posteriori* sont requises pour ces étapes. Une procédure menant à un algorithme à sauts réversibles efficace est fournie à cet effet. La seconde caractéristique de notre approche est la robustesse complète, ce qui implique que l'impact des valeurs aberrantes disparaît lorsque celles-ci s'éloignent à plus ou moins l'infini. Les conclusions obtenues sont conséquemment en ligne avec la majorité des observations.

ABSTRACT

Principal component regression is a linear regression model with principal components as regressors. This type of modelling is particularly useful for prediction in settings with high-dimensional covariates. Surprisingly, the existing literature treating of Bayesian approaches is relatively sparse.

In this paper, we aim at filling some gaps through the following practical contribution: we introduce a Bayesian approach with detailed guidelines for a straightforward implementation. The approach features two characteristics that we believe are important. First, it effectively involves the relevant principal components in the prediction process. This is achieved in two steps. The first one is model selection; the second one is to average out the predictions obtained from the selected models according to model averaging mechanisms, allowing to account for model uncertainty. The model posterior probabilities are required for model selection and model averaging. For this purpose, we include a procedure leading to an efficient reversible jump algorithm. The second characteristic of our approach is whole robustness, meaning that the impact of outliers on inference gradually vanishes as they approach plus or minus infinity. The conclusions obtained are consequently consistent with the majority of observations (the bulk of the data).

MSC 2010 subject classifications: Primary 62J05; secondary 62F35.

Keywords: Dimension reduction; Linear regression; Outliers; Principal component analysis; Reversible jump algorithm; Whole robustness.

5.1. INTRODUCTION

Principal component regression (PCR) corresponds to the usual linear regression model in which the covariates arise from a principal component analysis (PCA) of some explanatory variables. PCA is commonly used to reduce the dimensionality of data sets through two steps: first, the transformation of the “original” variables into principal components; second, the selection of the first d components, where d is smaller than the number of “original” variables. PCR is thus especially useful in situations where the number of explanatory variables is larger than the number of observations, or simply when one wishes to deal with smaller and therefore more stable models. The price to pay is that we do not obtain the typical inference on the “original” explanatory variables (as the identification of which variables have a statistically significant relationship with the dependent variable). PCR is, as a result, mostly used for prediction. Still, it can be useful for elucidating underlying structure in these “original” explanatory variables, as explained and shown in the very interesting paper of [West \(2003\)](#). We focus on prediction in this paper, but we believe that our contributions can also be beneficial for elucidating underlying structure.

Traditionally, linear regression analysis assumes normality of the errors. We explain in [Section 5.2.1](#) that this has the advantage that estimates are easily computed. A well-known problem however arises: inference may be contaminated by outliers. This problem is due to the slimness of the normal tails, which causes a shift in the posterior density to incorporate all data. It may find itself concentrated between the outliers and the bulk of the data, in an area that is not supported by any source of information. This translates into predictions that are not in line with either the nonoutliers or the outliers. Following [Box and Tiao \(1968\)](#) and [West \(1984\)](#), we propose to model the errors in a different way, and more precisely, to replace the traditional normal assumption by

an assumption that accommodates for the presence of outliers. This strategy is presented in Section 5.2.2. We instead assume that the error term has a super heavy-tailed distribution that allows attaining whole robustness, as stated in Chapter 4. Consequently, the predictions obtained reflect the behaviour of the bulk of the data.

Another challenge of PCR is the determination of which components to incorporate in the predictions. As a first step, we recommend to set d such that a predetermined percentage of the total variation is accounted for (provided that it is possible to do the estimation), as done in West (2003). In our analyses, we aim at reaching 90%. This step does not involve a study of the relationship between the variable of interest and the components. West (2003) proposes, as a second step, to use a product of Student distributions centered at 0 as prior for the regression coefficients, seeking to attract towards 0 the coefficients that are not significantly far from this value. This supports the effort of dimension reduction. We believe this approach, although interesting, may contaminate the inference as the prior may act as “conflicting information”, in the sense that it may be in contradiction with the information carried by the data. The prior may have similar impact as outliers on the inference. We propose instead to identify the relevant components via model selection. More precisely, we introduce a random variable K representing the model indicator and compute its posterior probabilities. Given that the components carrying significant information for prediction are usually the first ones, we consider d models, where Model $K = k$ is comprised of the first k components, $k \in \{1, \dots, d\}$. This means that K also represents the number of components in a given model. This strategy aims at simplifying the computations. After having identified the relevant models, we propose to account for model uncertainty using model averaging (see, e.g., Raftery et al. (1997) and Hoeting et al. (1999)). More precisely, we propose to identify models with posterior probabilities larger than 0.01, and to average over predictions arising from these models by weighting them according to the normalised probabilities. Model selection and model averaging allow to further reduce the dimensionality, while effectively considering the relevant components in predictions.

The need of computing the model posterior probabilities, combined with the increased complexity of the posterior due to the new assumption on the error term, point towards the use of the reversible jump algorithm. Indeed, it is a useful Markov chain Monte Carlo method introduced by Green (1995) that allows switches between subspaces of differing dimensionality. It is thus a useful tool, not only for parameter estimation, but also for model selection; the method is described Section 5.2.3. The implementation of such samplers requires the specification of some functions, and their efficiency relies heavily on their design (i.e. on how the functions are specified). In Section 5.2.4, we provide a detailed procedure to implement an efficient reversible jump algorithm.

The applicability and performance of our PCR approach is illustrated in Sections 5.3 and 5.4, containing respectively a simulation study and a real data analysis. The former is used to illustrate the performance of our approach under ideal conditions. The attained level of performance is used as a reference in the latter to establish that our method is suitable in real life situations.

5.2. PRINCIPAL COMPONENT REGRESSION

Consider that a PCA has been performed on $p \in \{1, 2, \dots\}$ explanatory variables with n observations each, from which the first $d \in \{1, \dots, p\}$ components are retained in order to be able to estimate the models (i.e. d is such that $n \geq d + 1$, see Chapter 4 for conditions that guarantee a proper posterior distribution). Let the associated design matrix be denoted by

$$\mathbf{X} := \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix},$$

where $x_{11} = \dots = x_{n1} = 1$, and $x_{ij} \in \mathbb{R}$ for $i = 1, \dots, n$ and $j = 2, \dots, d$. For simplicity, we consider the vector $(x_{11}, \dots, x_{n1})^T = (1, \dots, 1)^T$ as a component, but in fact, the principal components are the vectors $(x_{12}, \dots, x_{n2})^T, \dots, (x_{1d}, \dots, x_{nd})^T$. These principal components are such that $\sum_{i=1}^n x_{ij}x_{is} = 0$ for all $j, s \in \{2, \dots, d\}$ with $j \neq s$ (they are pairwise orthogonal vectors), and $(1/n) \sum_{i=1}^n x_{ij} = 0$ for all $j \in \{2, \dots, d\}$ (their mean is 0). They are ordered, as usual, by their variability; the first one has the largest sample variance, the second one has the second largest sample variance, and so on.

The main goal is to study the relationship between a dependent variable, represented by the random variables $Y_1, \dots, Y_n \in \mathbb{R}$, and the covariates (the components) in order to predict values for the dependent variable. We start from the premise that the following models are suitable:

$$Y_i = (\mathbf{x}_i^K)^T \boldsymbol{\beta}^K + \epsilon_i^K, \quad i = 1, \dots, n,$$

where $K \in \{1, \dots, d\}$ is the model indicator, which also represents the number of components included in the model (i.e. Model $K = k$ is the model with the first k components), $(\mathbf{x}_i^K)^T := (x_{i1}, \dots, x_{iK})$, $\epsilon_1^K, \dots, \epsilon_n^K \in \mathbb{R}$ are n random variables that represent the error term of Model K , and $\boldsymbol{\beta}^K := (\beta_1^K, \dots, \beta_K^K)^T \in \mathbb{R}^K$ is a vector comprised of K random variables that represent the regression coefficients of Model K . The scale of the error term of Model K is represented by the random variable $\sigma^K > 0$. Throughout the paper, the superscript indicates which model is used. Just keep in mind that, for $K = k$, we simply deal with a linear regression model with k covariates; in our specific case, these k covariates are the first k components.

From this perspective of statistical modelling, we can achieve the main goal (study the relationship between the dependent variable and the covariates for prediction purpose) through the identification of the relevant principal components (i.e. determination of the “best” models) and estimation of the parameters. Under the Bayesian paradigm, the model selection is performed considering K as a random variable. In this paper, we choose to perform model selection and parameter estimation simultaneously through the computation of $\pi(k, \sigma^k, \boldsymbol{\beta}^k \mid \mathbf{y})$, the joint posterior distribution of $(K, \sigma^K, \boldsymbol{\beta}^K)$ given $\mathbf{y} := (y_1, \dots, y_n)$.

A prior structure has to be determined to proceed with the computations. It is set to be fully non-informative, letting the data completely drive the inference. More precisely, we set the prior of K , denoted $\pi(k)$, equal to $1/d$ for all k in $\{1, \dots, d\}$, the prior of σ^K , denoted $\pi(\sigma^K)$, proportional

to $1/\sigma^k$ for $\sigma^k > 0$, and the conditional prior of $\boldsymbol{\beta}^k \mid K$, denoted $\pi(\boldsymbol{\beta}^k \mid k)$, proportional to 1 for $\boldsymbol{\beta}^k \in \mathbb{R}^k$, which are the usual non-informative priors for these types of random variables. We obviously assume the appropriate prior independence to obtain the following prior on $(K, \sigma^K, \boldsymbol{\beta}^K)$: $\pi(k, \sigma^k, \boldsymbol{\beta}^k) = \pi(k) \times \pi(\sigma^k) \times \pi(\boldsymbol{\beta}^k \mid k) \propto 1/\sigma^k$ for $k \in \{1, \dots, d\}$, $\sigma^k > 0$, and $\boldsymbol{\beta}^k \in \mathbb{R}^k$. It might seem that the so-called Lindley paradox (Lindley (1957) and Jeffreys (1967)) may arise when using such prior structure. Casella et al. (2009) explain that this is due to the improper reference prior on the parameters $\pi(\sigma^k, \boldsymbol{\beta}^k \mid k) = c/\sigma^k$, and propose a solution. The constant c can indeed be arbitrarily fixed, and it can even be different for the different models. We are however constant in our choices and use 1 for all models. We agree that this choice is arbitrary, but no approach is completely objective, including that of Casella et al. (2009). We believe further theoretical investigations are needed to verify the validity of using such prior structure. That said, the model selection results presented in the numerical analyses are in line with those arising from the Bayesian information criterion (BIC, see Schwarz (1978)), which suggests that the paradox does not arise. Note that the approach proposed in this paper is still valid if informative priors are used, such as those in Raftery et al. (1997). If users are looking for simpler models, they can set the prior on K to penalise for the number of parameters in the same vein as the BIC. For instance, they can set $\pi(k) \propto 1/n^{k/2}$ for all $k \in \{1, \dots, d\}$. A power different from $k/2$ would result in a more (less) aggressive penalty if it is increased (decreased).

As is typically done in Bayesian linear regression, we assume that $\epsilon_1^K, \dots, \epsilon_n^K$ and $\beta_1^K, \dots, \beta_K^K$ are $n + K$ conditionally independent random variables given (K, σ^K) , with a conditional density for ϵ_i^K given by

$$\epsilon_i^K \mid K, \sigma^K, \boldsymbol{\beta}^K \stackrel{D}{=} \epsilon_i^K \mid K, \sigma^K \stackrel{D}{\sim} (1/\sigma^K) f(\epsilon_i^K / \sigma^K), \quad i = 1, \dots, n.$$

5.2.1. First Situation: No Outliers

If one is sure that there is no outlier in the data, the most convenient way to model the errors is probably to assume their normality, i.e. $f = \mathcal{N}(0, 1)$. Indeed, considering this assumption and using the structure of the principal components, estimates and model posterior probabilities have closed-form expressions, as indicated in Proposition 5.1.

Proposition 5.1. *Assuming that $f = \mathcal{N}(0, 1)$, the posterior is given by*

$$\pi(k, \sigma^k, \boldsymbol{\beta}^k \mid \mathbf{y}) = \pi(k \mid \mathbf{y}) \pi(\sigma^k \mid k, \mathbf{y}) \prod_{j=1}^k \pi(\beta_j^k \mid k, \sigma^k, \mathbf{y}), \quad k \in \{1, \dots, d\}, \sigma^k > 0, \boldsymbol{\beta}^k \in \mathbb{R}^k,$$

where

$$\pi(k \mid \mathbf{y}) \propto \frac{\pi(k) \Gamma((n-k)/2) \pi^{k/2} \left(\mathbf{1}(k=1) + \mathbf{1}(k \geq 2) \prod_{j=2}^k \sqrt{\sum_{i=1}^n x_{ij}^2} \right)^{-1}}{\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 - \mathbf{1}(k \geq 2) \sum_{j=2}^k \frac{(\sum_{i=1}^n x_{ij} y_i)^2}{\sum_{i=1}^n x_{ij}^2} \right)^{\frac{n-k}{2}}}, \quad (5.1)$$

$$\pi(\sigma^k | k, \mathbf{y}) = \frac{2^{1-\frac{n-k}{2}} \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 - \mathbb{1}(k \geq 2) \sum_{j=2}^k \frac{(\sum_{i=1}^n x_{ij}y_i)^2}{\sum_{i=1}^n x_{ij}^2} \right)^{\frac{n-k}{2}}}{\Gamma((n-k)/2)(\sigma^k)^{n-k+1}} \times \exp \left\{ -\frac{1}{2(\sigma^k)^2} \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 - \mathbb{1}(k \geq 2) \sum_{j=2}^k \frac{(\sum_{i=1}^n x_{ij}y_i)^2}{\sum_{i=1}^n x_{ij}^2} \right) \right\},$$

$\beta_1^K | K, \sigma^K, \mathbf{y} \sim \mathcal{N}(\bar{y}, (\sigma^K)^2/n)$, and $\beta_j^K | K, \sigma^K, \mathbf{y} \sim \mathcal{N}(\sum_{i=1}^n x_{ij}y_i / \sum_{i=1}^n x_{ij}^2, (\sigma^K)^2 / \sum_{i=1}^n x_{ij}^2)$ for $j = 2, \dots, K$ (if $K \geq 2$), with $\bar{y} := (1/n) \sum_{i=1}^n y_i$. Note that the normalisation constant of $\pi(k | \mathbf{y})$ is the sum over k of the expression on the right-hand side in (5.1).

PROOF. See Section 5.6. ■

Proposition 5.1 indicates that users can easily determine which Model k^* has the highest posterior probability and estimate its parameters. Indeed, they can evaluate $\pi(k | \mathbf{y})$ (up to a constant) for all $k \in \{1, \dots, d\}$ in order to find the maximum, and given Model k^* , the parameters can be estimated using posterior means $\hat{\beta}_1^{k^*} = \bar{y}$ and $\hat{\beta}_j^{k^*} = \sum_{i=1}^n x_{ij}y_i / \sum_{i=1}^n x_{ij}^2$ for $j = 2, \dots, k^*$ (if $k^* \geq 2$). In our analyses, we use Bayesian model averaging (see Raftery et al. (1997)) to predict values for the dependent variable given sets of observations of the covariates. When normality is assumed, we can therefore use $\mathbb{E}[Y_{n+1} | \mathbf{y}] = \sum_k \pi(k | \mathbf{y}) \sum_{j=1}^k x_{n+1,j} \mathbb{E}[\beta_j^k | k, \mathbf{y}] = \sum_k \pi(k | \mathbf{y}) (\mathbf{x}_{n+1}^k)^T \hat{\boldsymbol{\beta}}^k$, where $\hat{\boldsymbol{\beta}}^k := (\hat{\beta}_1^k, \dots, \hat{\beta}_k^k)^T$. Note that, in this context of normality assumption, $(\sigma^k)^2 | K, \mathbf{y}$ has an inverse-gamma distribution with shape and scale parameters given by $(n-k)/2$ and $(\sum_{i=1}^n y_i^2 - n\bar{y}^2 - \mathbb{1}(k \geq 2) \sum_{j=2}^k (\sum_{i=1}^n x_{ij}y_i)^2 / \sum_{i=1}^n x_{ij}^2) / 2$, respectively.

As explained in Section 5.1, an important drawback of the classical normal assumption on the error term is that outliers have a significant impact on the estimation. This issue is addressed in Section 5.2.2.

5.2.2. Second Situation: Possible Presence of Outliers

The proposed solution to limit the impact of outliers is simple: replace the traditional normal assumption on the error term by a super heavy-tailed distribution assumption, aiming at accommodating for the presence of outliers. As explained in Chapter 4, this is the strategy to attain whole robustness, meaning that the impact of outliers gradually vanishes as they approach plus or minus infinity. Whole robustness ensures that the resulting inference is consistent with the majority of observations (the bulk of the data). The super heavy-tailed distribution that we use is the log-Pareto-tailed standard normal distribution, with parameters $\alpha > 1$ and $\psi > 1$. This is a distribution introduced by Desgagné (2015) that is expressed as

$$f(x) = \begin{cases} (2\pi)^{-1/2} \exp(-x^2/2), & \text{if } |x| \leq \alpha, \\ (2\pi)^{-1/2} \exp(-\alpha^2/2)(\alpha/|x|)(\log \alpha / \log |x|)^\psi, & \text{if } |x| > \alpha. \end{cases} \quad (5.2)$$

It exactly matches the standard normal on the interval $[-\alpha, \alpha]$, while having tails that behave like $(1/|z|)(\log |z|)^{-\psi}$ (which is a log-Pareto behaviour). Assuming that the error term follows this distribution and provided that there are at most $\lfloor n/2 - (d - 1/2) \rfloor$ outliers ($\lfloor \cdot \rfloor$ is the floor function), Theorem 1 in Chapter 4 indicates that the posterior distribution of $(\sigma^K, \boldsymbol{\beta}^K)$ converges towards the posterior of $(\sigma^K, \boldsymbol{\beta}^K)$ arising from the nonoutliers only, when the outliers approach plus or minus infinity, for any given model (i.e. for any given K). Whole robustness is thus attained for any given model. Consider for instance a sample of size $n = 20$ and that the first 7 components are retained for the analysis (i.e. $d = 7$). The convergence holds for any model if there is at most three outliers. The result ensures that, for any given model, any estimation of σ^K and $\boldsymbol{\beta}^K$ based on posterior quantiles (e.g. using posterior medians and Bayesian credible intervals) is robust to outliers. Given that this result is valid for any given K , we conjecture that it is also valid for the whole joint posterior of $(K, \sigma^K, \boldsymbol{\beta}^K)$; this is empirically verified in Sections 5.3 and 5.4. In the analyses we set $\alpha = 1.96$, which implies that $\psi = 4.08$, according to the procedure described in Section 4 of Desgagné (2015) (this procedure ensures that f is a continuous probability density function (PDF)). As explained in Chapter 4, setting $\alpha = 1.96$ seems suitable for practical purposes.

5.2.3. Reversible Jump Algorithm

The price to pay for robustness is an increase in the complexity of the posterior. We consequently need a numerical approximation method for the computation of integrals with respect to this posterior. The method commonly used in contexts of model selection and parameter estimation within the Bayesian paradigm is the reversible jump algorithm, a Markov chain Monte Carlo method introduced by Green (1995). The implementation of this sampler requires the specification of some functions, a step typically driven by the structure of the posterior. In the current context, we rely on Proposition 5.1 for specifying these functions. As illustrated in Chapter 4, if there is no outlier, the posterior arising from an error term with density f as in (5.2) is similar to that arising from the normality assumption, for any given model. Furthermore, in presence of outliers, the posterior is similar to that arising from the normality assumption, but based on the nonoutliers only (i.e. excluding the outliers), again for any given model. Therefore, whether there are outliers or not, the posterior should have a structure similar to that expressed in Proposition 5.1. In other words, the posterior should reflect some kind of conditional independence between the regression coefficients, given K and σ^K . We consider this feature to design the reversible jump algorithm. In particular, we borrow ideas from Chapter 2, in which an efficient reversible jump algorithm is built to sample from distributions reflecting a specific type of conditional independence (the sampler that we use is described in detail later in this section).

The reversible jump algorithm is the preferred method because it allows switches between subspaces of differing dimensionality, and therefore, to select models and estimate parameters from a single output. One iteration of this sampler is essentially as follows: a type of movement is first randomly chosen, then a candidate for the next state of the Markov chain (that depends

on the type of movement) is proposed. The candidate is accepted with a specific probability; if it is rejected, the chain remains at the same state. We consider three types of movements: first, updating of the parameters (for a given model); second, switching from Model K to Model $K + 1$ (this requires that a regression coefficient be added); third, switching from Model K to Model $K - 1$ (this requires that a regression coefficient be withdrawn). The probability mass function (PMF) used to randomly select the type of movement is the following:

$$g(j) := \begin{cases} \tau, & \text{if } j = 1, \\ (1 - \tau)/2, & \text{if } j = 2, 3, \end{cases} \quad (5.3)$$

where $0 < \tau < 1$ is a constant. Therefore, at each iteration, an update of the parameters is attempted with probability τ , and a switch to either Model $K + 1$ or Model $K - 1$ is attempted with probability $(1 - \tau)/2$ in each case.

Updating the parameters of Model K is achieved here by using a $(K + 1)$ -dimensional proposal distribution centered around the current value of the parameter (σ^K, β^K) and scaled according to $\ell / \sqrt{K + 1}$, where ℓ is a positive constant. We assume that each of the $K + 1$ candidates is generated independently from the others, according to the one-dimensional strictly positive PDF $\varphi_i, i = 1, \dots, K + 1$. Although the chosen PDF φ_i usually is the normal density, using the PDF in (5.2) induces larger candidate steps, and therefore, results in a better exploration of the state space. This claim has been empirically verified, and we thus rely on this updating strategy in the analyses in Sections 5.3 and 5.4. Note that one can easily simulate from (5.2) using the inverse transformation method.

A major issue with the design of the reversible jump algorithm is related to the fact that there may be a great difference between the “good” values for the parameters under Model K and those under Model $K + 1$. As witnessed from the posterior density detailed in Proposition 5.1, this should not be a concern in the case where there is no outlier. Furthermore, when the same data points are diagnosed as outliers for both Model K and Model $K + 1$, this should not be a concern either, as explained previously. When observations are outliers with respect to Model $K + 1$ but not to Model K however, there will be a difference between the “good” values for the parameters of these models (because of the robustness provided by the super heavy-tailed distribution assumption). Therefore, when switching from Model K to Model $K + 1$ for instance, the parameters that were already in Model K need to be moved to a position that is, combined with a candidate value for the new parameter, appropriate under Model $K + 1$. Otherwise, this model switching will be less likely to be accepted, and if it is, it will possibly require some (maybe a lot of) iterations before the chain reaches the high probability area. This may result in inaccurate estimates. Existing research has focused on this issue (e.g. Brooks et al. (2003), Al-Awadhi et al. (2004), Hastie (2005), and Karagiannis and Andrieu (2013)).

Our strategy is simple and easy to implement. First, add up the vector \mathbf{c}^{K+1} to the current value of the parameters of Model K to end up in a suitable area under Model $K + 1$. Then, generate a

value u^{K+1} for the new parameter β_{K+1}^{K+1} from a strictly positive PDF q_{K+1} . The candidates for the parameters under Model $K + 1$ are thus $(\sigma^{K+1}, \beta^{K+1}) = ((\sigma^K, \beta^K) + \mathbf{c}^{K+1}, u^{K+1})$, where (σ^K, β^K) are the current values of the parameters under Model K . The rationale behind this is the following. Suppose, for instance, that β_1^1 (the intercept in Model 1) takes values around 0.5, but β_1^2 (the intercept in Model 2) takes values around 1, while β_2^2 takes values around 2 with a scale of 1. Setting $\mathbf{c}^2 := (0, 0.5)^T$ and q_2 equal to the distribution given in (5.2) with location and scale parameters of 2 and 1, respectively, should result in a vector $((\sigma^1, \beta^1) + \mathbf{c}^2, u^2)$ that is in the high probability area under Model 2. Note that, in order not to obtain negative values for σ^K , we always set the first component of the vectors \mathbf{c}^i to 0.

We now provide a pseudo-code for the reversible jump algorithm used to sample from $\pi(k, \sigma^k, \beta^k \mid \mathbf{y})$ when assuming that f is given in (5.2). We thereafter explain how to specify the inputs required to implement it.

1. Initialisation: set $(K, \sigma^K, \beta^K)(0)$. *Remark:* the number in parentheses beside a vector (or a scalar) indicates the iteration index of the vector (or of the scalar). For instance, $(K, \sigma^K, \beta^K)(m)$, where $m \in \mathbb{N}$, implies that the iteration index of any component in $(K, \sigma^K, \beta^K)(m)$ is m .

Iteration $m + 1$.

2. Generate $u \sim \mathcal{U}(0, 1)$.

- 3.(a) If $u \leq \tau$, attempt an update of the parameters. Generate a candidate $\mathbf{w}^{K(m)} := (w_1, \dots, w_{K(m)+1})$, where $w_1 \sim \varphi_1(\cdot \mid K(m), \sigma^{K(m)}, \ell)$, and $w_i \sim \varphi_i(\cdot \mid K(m), \beta_{i-1}^{K(m)}, \ell)$ for $i = 2, \dots, K(m) + 1$. Generate $u_a \sim \mathcal{U}(0, 1)$. If

$$u_a \leq \left(1 \wedge \frac{(1/w_1)f(\mathbf{y} \mid K(m), \mathbf{w}^{K(m)})}{(1/\sigma^{K(m)})f(\mathbf{y} \mid (K, \sigma^K, \beta^K)(m))} \right),$$

where

$$f(\mathbf{y} \mid k, \sigma^k, \beta^k) := \prod_{i=1}^n \frac{1}{\sigma^k} f\left(\frac{y_i - (\mathbf{x}_i^k)^T \beta^k}{\sigma^k}\right),$$

set $(K, \sigma^K, \beta^K)(m + 1) = (K(m), \mathbf{w}^{K(m)})$.

- 3.(b) If $\tau < u \leq \tau + (1 - \tau)/2$, attempt adding a parameter to switch from Model $K(m)$ to Model $K(m) + 1$. Generate $u^{K(m)+1} \sim q_{K(m)+1}$ and generate $u_a \sim \mathcal{U}(0, 1)$. If

$$u_a \leq \left(1 \wedge \frac{\pi(K(m) + 1)f(\mathbf{y} \mid K(m) + 1, (\sigma^K, \beta^K)(m) + \mathbf{c}^{K(m)+1}, u^{K(m)+1})}{\pi(K(m))f(\mathbf{y} \mid (K, \sigma^K, \beta^K)(m))q_{K(m)+1}(u^{K(m)+1})} \right),$$

set $(K, \sigma^K, \beta^K)(m + 1) = (K(m) + 1, (\sigma^K, \beta^K)(m) + \mathbf{c}^{K(m)+1}, u^{K(m)+1})$.

- 3.(c) If $u > \tau + (1 - \tau)/2$, attempt withdrawing the last parameter to switch from Model $K(m)$ to Model $K(m) - 1$. Generate $u_a \sim \mathcal{U}(0, 1)$. If

$$u_a \leq \left(1 \wedge \frac{\pi(K(m) - 1)f(\mathbf{y} \mid K(m) - 1, (\sigma^K, \beta^{K^-})(m) - \mathbf{c}^{K(m)})q_{K(m)}(\beta_K^K(m))}{\pi(K(m))f(\mathbf{y} \mid (K, \sigma^K, \beta^K)(m))} \right),$$

set $(K, \sigma^K, \boldsymbol{\beta}^K)(m+1) = (K(m)-1, (\sigma^K, \boldsymbol{\beta}^{K-})(m) - \mathbf{c}^{K(m)})$, where $(\sigma^K, \boldsymbol{\beta}^{K-})(m)$ is $(\sigma^K, \boldsymbol{\beta}^K)(m)$ without the last component (more precisely, $(\sigma^K, \boldsymbol{\beta}^{K-})(m) := (\sigma^K, \beta_1^K, \dots, \beta_{K-1}^K)(m)$).

4. In case of rejection, set $(K, \sigma^K, \boldsymbol{\beta}^K)(m+1) = (K, \sigma^K, \boldsymbol{\beta}^K)(m)$.
5. Go to Step 2.

It is easily verified that the resulting stochastic process $\{(K, \sigma^K, \boldsymbol{\beta}^K)(m), m \in \mathbb{N}\}$ is a $\pi(k, \sigma^k, \boldsymbol{\beta}^k | \mathbf{y})$ -irreducible and aperiodic Markov chain. Furthermore, it satisfies the reversibility condition with respect to the posterior, as stated in the following proposition. Therefore, it is an ergodic Markov chain, which guarantees that the Law of Large Numbers holds.

Proposition 5.2. *Consider the reversible jump algorithm described above. The Markov chain $\{(K, \sigma^K, \boldsymbol{\beta}^K)(m), m \in \mathbb{N}\}$ satisfies the reversibility condition with respect to the posterior $\pi(k, \sigma^k, \boldsymbol{\beta}^k | \mathbf{y})$ arising from the assumption that f is given in (5.2).*

PROOF. See Section 5.6. ■

5.2.4. Optimal Implementation

In order to implement the reversible jump algorithm described above, we have to specify the PDFs q_i , the constants τ and ℓ , and the vectors \mathbf{c}^i . In the following paragraphs, we explain how we achieve this task.

In Chapter 2, a simple structure for the posterior is considered in order to obtain theoretical results that lead to an optimal design of the reversible jump algorithm. The structure is the following. The parameters are conditionally independent and identically distributed for any given model. The model indicator K indicates the number of parameters that are added to the simplest model (the one with the fewest parameters). Finally, when switching from Model K to Model $K+1$ (i.e. when adding a parameter), the distributions of the parameters that were in Model K do not change. The authors find asymptotically optimal (as the number of parameters approaches infinity) values for τ and ℓ . They conjecture that their results are valid (to some extent) when the parameters are conditionally independent, but not identically distributed, for any given model. They also provide guidelines to suitably design the PDFs q_i . Their setting is somewhat similar to ours, because, as explained above, the posterior should reflect some kind of conditional independence between the regression coefficients, given K and σ^K . Also, when switching from Model K to Model $K+1$, if there is no outlier or if the same data points are diagnosed as outliers for both models, there should not be a great difference between the “good” values for the parameters under Model K and those under Model $K+1$. We therefore use their results to design the reversible jump algorithm (at least as a starting point).

In the setting of Chapter 2, the asymptotically optimal value for τ depends on the PDFs q_i . But for moderate values of d , selecting any value between 0.2 and 0.6 seems almost optimal. We recommend to set $\tau = 0.6$ if d is rather small (because there are not so many models to visit), as in our analyses in Sections 5.3 and 5.4. Selecting larger values for τ leaves more time between model

switchings for the chain to explore the state space of the parameters. We ran several reversible jump algorithms with different values for τ to verify if 0.6 actually is a good choice. The optimal values are close to 0.6 for the data sets analysed in Sections 5.3 and 5.4.

If the parameters $(\sigma^K, \boldsymbol{\beta}^K)$ were independent and identically distributed for any given model, the asymptotically optimal value for ℓ would correspond to an acceptance rate of candidates \mathbf{w}^K of approximately 0.234. Note that this rate have to be computed by considering only iterations in which there has been an attempt at updating the parameters. If the parameters are nearly independent but not identically distributed, the asymptotically optimal value may correspond to an acceptance rate smaller than 0.234, as explained in Bédard (2007). If in addition σ^K has a great impact on the distribution of $\boldsymbol{\beta}^K$, the asymptotically optimal value may correspond to a further reduced acceptance rate, as explained in Bédard (2015). Considering this, combined to the fact that d can be rather small, we recommend to perform trial runs to identify the optimal value for ℓ . We use the 0.234 rule to initiate the process. In our analyses in Sections 5.3 and 5.4, the optimal values for ℓ correspond to acceptance rates relatively close to 0.234.

We propose to specify the PDFs q_i and the vectors \mathbf{c}^i through trial runs too. Specifying these functions and vectors requires information about the location of all regression coefficients for all models and about the scaling of β_K^K for Model K , $K \geq 2$. In order to gather this information and to identify the optimal value for ℓ , we propose a naive but simple strategy for the trial runs that can be executed automatically: run a random walk Metropolis (RWM) algorithm for each model. The RWM algorithm is the reversible jump algorithm described above in which $\tau = 1$ (it is an algorithm in which only updates of the parameters are proposed). Here is the procedure that we recommend for the trial runs:

For each $k \in \{1, \dots, d\}$:

1. Tune the value of ℓ so that the acceptance rate of candidates \mathbf{w}^k is approximately 0.234. Record the corresponding value ℓ^k .
2. Select a sequence of values for ℓ around ℓ^k : $(\ell_1, \dots, \ell_{j_0} := \ell^k, \dots, \ell_L)$, where L is a positive integer.
3. For each ℓ_j : run a RWM algorithm with: $(\sigma^k(0))^2 \sim \text{Inv-}\Gamma$ with shape and scale parameters given by $(n - k)/2$ and $(\sum_{i=1}^n y_i^2 - n\bar{y}^2 - \mathbb{1}(k \geq 2) \sum_{j=2}^k (\sum_{i=1}^n x_{ij}y_i)^2 / \sum_{i=1}^n x_{ij}^2)/2$, respectively, $\beta_1^k(0) \sim \mathcal{N}(\bar{y}, (\sigma^k(0))^2/n)$, and $\beta_j^k(0) \sim \mathcal{N}(\sum_{i=1}^n x_{ij}y_i / \sum_{i=1}^n x_{ij}^2, (\sigma^k(0))^2 / \sum_{i=1}^n x_{ij}^2)$, $j = 2, \dots, k$ (if $k \geq 2$).
4. For each ℓ_j : estimate the location of each of the parameters β_i^k using the median (that we denote $m_{i,j}^k$) of $\{\beta_i^k(m), m \in \{B+1, \dots, T\}\}$, for $i = 1, \dots, k$, where B is the length of the burn-in period and T is the number of iterations. If $k \geq 2$, estimate the scaling of β_k^k using the interquartile range (IQR) divided by 1.349 (that we denote $s_{k,j}^k$) of $\{\beta_k^k(m), m \in \{B+1, \dots, T\}\}$ (this is robust measure of the scaling). Measure the efficiency of the algorithm with respect to ℓ_j using the integrated autocorrelation time (IAT) of $\{\sigma^k(m), m \in \{B+1, \dots, T\}\}$. Recall

that all parameters move at the same time when there is an update of the parameters, which indicates that this efficiency measure is appropriate. Record the value ℓ_{opt}^k corresponding to the smallest IAT.

5. Change the sequence of values for ℓ if the smallest IAT corresponds to the lower or upper bound of the current range.
6. Compute the average of $\{m_{i,1}^k, \dots, m_{i,L}^k\}$ (that we denote m_i^k) for $i = 1, \dots, k$, and the average of $\{s_{k,1}^k, \dots, s_{k,L}^k\}$ (that we denote s_k^k).

First notice that these trial runs can be performed in parallel for computational efficiency. The whole process can be executed automatically, in the sense that users only have to retrieve the output at the end. From this output, users may first verify if $m_1^1 = m_1^2 = \dots = m_1^d = \bar{y}$ (which indicates that the intercept estimate is the same for all models and that it is \bar{y}), and $m_j^1 = m_j^{j+1} = \dots = m_j^d = \sum_{i=1}^n x_{ij}y_i / \sum_{i=1}^n x_{ij}^2$ (which indicates that the j th regression coefficient estimate is the same for all models and that it is $\sum_{i=1}^n x_{ij}y_i / \sum_{i=1}^n x_{ij}^2$) for $j = 2, \dots, d$. If it is the case, it indicates that there should not be any outliers because the estimates are the same as when assuming normality (see Proposition 5.1). Consequently, users may assume normality of the errors and proceed as in Section 5.2.1 for the computation of the predictions. Otherwise, set q_j equal to the distribution given in (5.2) with location and scale parameters given by m_j^j and s_j^j , respectively, for $j = 2, \dots, d$. Set $\mathbf{c}^j = (0, m_1^j - m_1^{j-1}, \dots, m_{j-1}^j - m_{j-1}^{j-1})^T$, for $j = 2, \dots, d$. Set ℓ equal to the median of $\{\ell_{\text{opt}}^1, \dots, \ell_{\text{opt}}^d\}$.

The only remaining inputs required to implement the reversible jump algorithm are initial values for the model indicator and the parameters. We recommend to generate $K(0) \sim \mathcal{U}\{1, \dots, d\}$, $(\sigma^K(0))^2 \sim \text{Inv-}\Gamma$ with shape and scale parameters given by $(n - K(0))/2$ and $(\sum_{i=1}^n y_i^2 - n\bar{y}^2 - \mathbb{1}(K(0) \geq 2) \sum_{j=2}^{K(0)} (\sum_{i=1}^n x_{ij}y_i)^2 / \sum_{i=1}^n x_{ij}^2) / 2$, respectively, $\beta_1^K(0) \sim \mathcal{N}(m_1^{K(0)}, (s_1^{K(0)})^2)$, and $\beta_j^K(0) \sim \mathcal{N}(m_j^{K(0)}, (s_j^{K(0)})^2)$, $j = 2, \dots, K(0)$ (if $K(0) \geq 2$). In the analyses in Sections 5.3 and 5.4, we use sequences of length $L = 10$ for ℓ , $T = 100,000$ iterations and a burn-in period of length $B = 10,000$ for the trial runs. When running the reversible jump, we use 1,000,000 iterations and a burn-in period of length 100,000.

5.3. SIMULATION STUDY

Through this simulation study, we empirically verify the conjecture stated in Section 5.2.2 about the convergence of the posterior of $(K, \sigma^K, \boldsymbol{\beta}^K)$ towards that arising from the nonoutliers only when the outliers approach plus or minus infinity. A framework that includes the presence of large outliers is thus set up. This can be considered as ideal conditions for the application of our method (given that this simulates the limiting situation). This nevertheless also allows to illustrate the performance of our approach under these conditions (recall that in the next section, its performance is evaluated when applied to a real data set). To help evaluate the performance, we compare the results with those arising from the normality assumption. The impact of model selection and model averaging is not analysed. We believe that the beneficial effect of these statistical techniques is well

understood, contrary to that of using super heavy-tailed distributions in linear regression analyses. Indeed, this strategy has been recently introduced in [Desgagné and Gagnon \(2016\)](#), in which the special case of simple linear regressions through the origin is considered. Then, its validity when applied the usual linear regression model was verified in Chapter 4. It is the first time that this strategy is considered in a context of model selection.

We begin by generating $n = 20$ observations from explanatory variables. We shall work with 24 such variables which, by construction, are expected to be represented by 4 principal components. We define the vectors $\mathbf{u}_1 := (1, \dots, 20)^T$, $\mathbf{u}_2 := (0^2, 1^2, \dots, 19^2)^T$, $\mathbf{u}_3 := (\log(1), \dots, \log(20))^T$, and $\mathbf{u}_4 := (\exp(1), \dots, \exp(20))^T$. From these vectors, four orthogonal, centered, and standardised vectors $\mathbf{e}_1, \dots, \mathbf{e}_4$ are obtained through the Gram-Schmidt method. The vectors $\mathbf{e}_1, \dots, \mathbf{e}_4$ represent the observations from the first four explanatory variables. The remaining 20 explanatory variables are obtained from $\mathbf{e}_1, \dots, \mathbf{e}_4$, introducing correlation between the explanatory variables. More precisely, for each of the vectors \mathbf{e}_i , we generate five vectors from a 20-dimensional normal distribution with mean \mathbf{e}_i and diagonal covariance matrix, where each of the elements on the diagonal is arbitrarily set to 0.1^2 .

From these variables, we first generate observations from the dependent variable using the linear regression model with a normal error term. The intercept and the second regression coefficient are arbitrarily set to 10 and 1, respectively, and each of the following regression coefficients are equal to minus the half of the previous one (therefore, $-1/2$, $1/4$, and so on). The scale parameter of the error term is arbitrarily set to 1. We now perform the statistical analysis of the resulting data set. We shall then add a large outlier to evaluate the impact on the estimation (and therefore, on the predictions) under both the normality and the super heavy-tailed distribution assumptions. We apply the PCA on the initial data set and select d upon examining the variation explained by each of the principal components (see Figure 5.1 and Table 5. I).

As expected, everything suggests that we should retain 4 principal components, and therefore, select $d = 5$. We now compute the posterior probabilities of the models and the parameter estimates. The results under the super heavy-tailed distribution assumption are presented in Table 5. II.

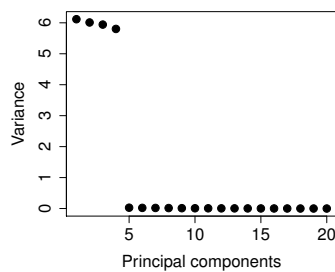


Figure 5.1. Sample variance for each of the principal components

Princ. comp.	Cum. expl. var. (in %)
1	25.48
2	50.53
3	75.29
4	99.47

Table 5. I. Cumulative explained variations for the four first principal components

We notice that the parameter estimates do not vary significantly from one model to another. As explained in Section 5.2.2, this is expected when there is no outlier or when the same data points are diagnosed as outliers for all models. The results under normality assumption are essentially the same, which suggests the absence of outliers. In particular, the posterior probabilities of Models 1 to 5 are 0.02, 0.02, 0.04, 0.49, and 0.42, respectively. The estimates of the first to the fifth regression coefficients are 10.14, -0.18 , 0.21, 0.28, and 0.15, respectively. Therefore, predictions shall be essentially the same under both assumptions, provided that we use $\hat{y}_{n+1} = \sum_k \pi(k | \mathbf{y})(\mathbf{x}_{n+1}^k)^T \hat{\boldsymbol{\beta}}^k$ in both cases, along with posterior medians under the super heavy-tailed distribution assumption.

We now add an outlier to the data set. In order to preserve the properties on the columns in the design matrix \mathbf{X} containing the principal components, we set $\mathbf{x}_{21}^d = (1, 0, \dots, 0)^T$, ensuring that the conclusions of Proposition 5.1 still hold. Indeed, the vectors comprised of the observations from the covariates are still pairwise orthogonal and their mean is still 0. This prevents us from doing another PCA, and therefore, facilitates the comparison of results. We set $y_{21} = 30$, a large outlier compared with the trend emerging from the bulk of the data (this can correspond to the estimated intercept from the original data set for most models, i.e. $\hat{\beta}_1^K = 10.14$). With or without the outlier, the estimates for $\beta_i^K, i = 2, \dots, 5$, under normality do not vary; recall that they are given by $\hat{\beta}_i^K = \sum_{i=1}^n x_{ij}y_i / \sum_{i=1}^n x_{ij}^2, i = 2, \dots, 5$. The estimates for β_1^K and σ^K are however different: while $\hat{\beta}_1^K = 11.09$ for all models, we find for instance that the estimate for the scale parameter of the error term in Model 5 is 5.04 (compared with 0.95 based on the original data set). The impact of the outlier thus mainly translates into higher probabilities on large scale parameter values for the error term. This is due to the position of the outlier (the impact of an outlier with a different position is shown in Section 5.4). Its impact is however not only on the marginal posterior distributions of σ^K and β_1^K . It propagates through the whole joint posterior of K, σ^K and $\boldsymbol{\beta}^K$, increasing the posterior scaling of the entire vector $\boldsymbol{\beta}^K$ and modifying the posterior probabilities of the models. The latter are 0.12, 0.14, 0.18, 0.25 and 0.32, respectively for Models 1 to 5.

The results under the super heavy-tailed distribution assumption are now presented in Table 5. III. Whole robustness seems attained. Indeed, the posterior probabilities of the models and the parameter estimates are essentially the same as those based on the original data set. As a result, predictions made under the super heavy-tailed distribution assumption reflect the behaviour of the bulk of the data, contrary to those made under normality. In other words, new observations from the dependent variable in line with the bulk of the data are more accurately predicted under the robust approach.

5.4. REAL DATA ANALYSIS: PREDICTION OF RETURNS FOR THE S&P 500

In Section 5.3, we showed that our strategy allows to obtain adequate inference in presence of large outliers, corresponding to an ideal condition for the application of our method. In this section, we show how our strategy performs when applied to a real data set containing outliers.

Again, the results obtained from our method are compared with those arising from the normality assumption.

Models	Posterior prob.	Posterior medians of				
		β_1	β_2	β_3	β_4	β_5
1	0.02	10.15	—	—	—	—
2	0.02	10.14	-0.19	—	—	—
3	0.06	10.23	-0.20	0.25	—	—
4	0.49	10.14	-0.19	0.21	0.28	—
5	0.41	10.14	-0.18	0.22	0.28	0.15

Table 5. II. Posterior probabilities of models and parameter estimates under the super heavy-tailed distribution assumption based on the data set without outliers

Models	Posterior prob.	Posterior medians of				
		β_1	β_2	β_3	β_4	β_5
1	0.03	10.16	—	—	—	—
2	0.03	10.15	-0.18	—	—	—
3	0.06	10.23	-0.19	0.24	—	—
4	0.49	10.14	-0.19	0.21	0.29	—
5	0.39	10.14	-0.18	0.22	0.29	0.15

Table 5. III. Posterior probabilities of models and parameter estimates under the super heavy-tailed distribution assumption based on the data set with the outlier

The context is the following: we model the January 2011 daily returns of the S&P 500 by exploiting their potential linear relationship with some financial assets and indicators, aiming at predicting the February 2011 daily returns of this stock index. The detailed list of explanatory variables is provided in Section 5.7. There are 18 explanatory variables in total, and obviously, their observations on day i are used to predict the return of the S&P 500 on day $i + 1$. For each of the explanatory variables and for the dependent variable, there are $n = 19$ observations that can be used to estimate the models, and the linear regression model with all explanatory variables has 20 parameters (18 regression coefficients for the variables, the intercept, and the scale parameter). The PCA should be a beneficial procedure given that financial assets and indicators are likely to carry redundant information.

As in Section 5.3, we start by selecting d . We decide to retain 10 principal components which account for approximately 92% of the total variation (see Figure 5.2 and Table 5. IV), and therefore, to set $d = 11$.

We now present the results under the super heavy-tailed distribution assumption in Table 5. V. The posterior probabilities of models other than 2 to 6 are all less than 0.01. We notice from Table 5. V that outliers seem present with respect to either Models 2 and 3 or Models 4, 5 and 6 (because parameter estimates are significantly different). To help us in this outlier investigation,

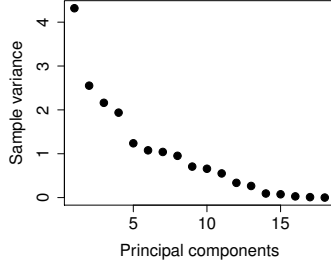


Figure 5.2. Sample variance for each of the principal components

Princ. comp.	Cum. expl. var. (in %)
8	84.87
9	88.80
10	92.45
11	95.51

Table 5. IV. Cumulative explained variations for the first eighth to eleventh components

we now present the results under the normality assumption. The probabilities of Models 2 to 6 are 0.92, 0.07, 0.01, 0.00, and 0.00 (to two decimal places), respectively, and the estimates of the first to sixth regression coefficients are 0.06, -0.23 , 0.03, 0.08, 0.06, and -0.09 , respectively. This suggests that there are outliers with respect to Models 4, 5, and 6 (because the parameter estimates under the normality assumption are different from those under the super heavy-tailed distribution assumption for these models). It is noteworthy that Model 4 is an unlikely model under the normality assumption, but a relatively likely model under the super heavy-tailed assumption. This leads, along with the discrepancies in the parameter estimates, to different predictions. In particular, the mean absolute deviation between the predicted and actual returns in percentage for February 2011 is 0.62% under the normality assumption, whereas it is 0.57% under the super heavy-tailed distribution assumption. Under both assumptions, we use the same expression to predict, i.e. $\hat{y}_{n+1} = \sum_k \pi(k | \mathbf{y})(\mathbf{x}_{n+1}^k)^T \hat{\boldsymbol{\beta}}^k$, but the estimates are posterior medians for the super heavy-tailed distribution assumption. It may be of interest to only predict whether the financial asset (in this case, the S&P 500) will go up or down the day after. Using the sign of our predicted returns to answer this question, we are respectively correct 10 and 12 times out of 19, under the normality and the super heavy-tailed assumptions.

We now identify the outliers and redo the statistical analysis excluding these observations to evaluate their impact. For the identification step, we compute the errors, i.e. $y_i - (\mathbf{x}_i^k)^T \hat{\boldsymbol{\beta}}^k$, $i = 1, \dots, 19$, of Model 4 (the most probable model among those for which there are outliers) under the super heavy-tailed assumption (see Figure 5.3). Note that the computation of these errors is based on the data used to estimate the models, i.e. the January 2011 daily returns.

The second observation seems to be not in line with the trend emerging from the bulk of the data. This is also true under Models 5 and 6. The results excluding this observation under the super heavy-tailed distribution assumption are presented in Table 5. VI. Again, the probabilities of models other than 2 to 6 are all less than 0.01. The results under the normality assumption are now essentially the same. Note that the conclusions of Proposition 5.1 do not hold any more. Indeed, without the second observation, the properties on the columns in the design matrix are not

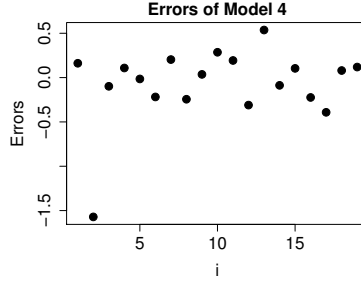


Figure 5.3. Errors of Model 4 under the super heavy-tailed assumption computed using posterior medians

preserved. The reversible jump algorithm has therefore been used, and the strategy explained in Section 5.2.2 to design it has been applied.

The second observation has a significant impact on the posterior distribution under the normality assumption. The robust alternative is able to limit the impact (as can be seen by comparing the results in Table 5. V with those in Table 5. VI). In fact, this example illustrates another feature of our method. It has the ability of reflecting the level of uncertainty about the fact that an observation really is an outlier or not. This phenomenon was discussed in Chapter 4. In this example, the method seems to diagnose the second observation as an outlier (which explains why it limits its impact), but not as a “clear” outlier, as it does in Section 5.3 with the twenty-first observation. It is a question of whether or not the observation is far enough from the bulk of the data, and more precisely, far enough from the probable hyperplanes for the bulk of the data. Note that the differences between the predictions made under the two different assumptions, based on the original data set, are due to the presence of second observation. Indeed, without this observation they are the same. In particular, the mean absolute deviation between the predicted and actual returns in percentage is now 0.50% (it is 0.57% for the robust approach with the outlier), and the number of correct predictions of whether the S&P 500 will go up or down is 12 (it also 12 for the robust approach with the outlier).

Models	Posterior prob.	Posterior medians of					
		β_1	β_2	β_3	β_4	β_5	β_6
2	0.61	0.08	-0.22	—	—	—	—
3	0.12	0.08	-0.22	0.04	—	—	—
4	0.20	0.13	-0.16	0.05	0.14	—	—
5	0.05	0.13	-0.17	0.05	0.14	0.06	—
6	0.01	0.12	-0.17	0.05	0.14	0.06	-0.06

Table 5. V. Posterior probabilities of models and parameter estimates under the super heavy-tailed distribution assumption based on the returns in %

Models	Posterior prob.	Posterior medians of					
		β_1	β_2	β_3	β_4	β_5	β_6
2	0.16	0.13	-0.17	—	—	—	—
3	0.04	0.13	-0.16	0.05	—	—	—
4	0.61	0.15	-0.15	0.06	0.16	—	—
5	0.16	0.15	-0.15	0.06	0.16	0.06	—
6	0.03	0.15	-0.15	0.06	0.16	0.06	-0.04

Table 5. VI. Posterior probabilities of models and parameter estimates under the super heavy-tailed distribution assumption based on the returns in %, without the second observation

5.5. CONCLUSION AND FURTHER REMARKS

In this paper, we have introduced a Bayesian PCR approach that addresses the two following issues: first, how to identify the components that should be involved in the prediction process (and how to involve them); second, how to limit the impact of outliers so that inferences are not contaminated. All the required guidelines for a straightforward implementation of our method have been provided. In particular, a detailed procedure to implement an efficient reversible jump algorithm has been included, allowing to obtain the required model posterior probabilities and parameter estimates. The relevance of our approach has been shown via a simulation study and a real data analysis in Sections 5.3 and 5.4, respectively.

The method used to reduce dimensionality in our paper is the traditional PCA. Recall that this method transforms the “original” explanatory variables into principal components. This transformation is based on the sample correlation matrix of the observations from the explanatory variables. This process can be contaminated if there are outliers among these observations. The approach to attain robustness presented in this paper can thus be viewed as a first step towards fully robust PCR. Further research is therefore needed. We believe it would be particularly useful to develop a robust procedure to reduce dimensionality that also induces sparsity to deal with the case of interest $p \gg n$.

REFERENCES

- Al-Awadhi, F., M. Hurn and C. Jennison. 2004, Improving the acceptance rate of reversible jump MCMC proposals, *Statist. Probab. Lett.*, vol. 69, n° 2, p. 189–198.
- Bédard, M. 2007, Weak convergence of Metropolis algorithms for non-i.i.d. target distributions, *Ann. Appl. Probab.*, vol. 17, p. 1222–1244.
- Bédard, M. 2015, On the optimal scaling problem of Metropolis algorithms for hierarchical target distributions, Preprint.
- Box, G. E. P. and G. C. Tiao. 1968, A Bayesian approach to some outlier problems, *Biometrika*, vol. 55, n° 1, p. 119–129.

- Brooks, S. P., P. Giudici and G. O. Roberts. 2003, Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 65, n° 1, p. 3–39.
- Casella, G., F. J. Giròn, M. L. Martínez and E. Moreno. 2009, Consistency of Bayesian procedures for variable selection, *Ann. Statist.*, vol. 37, n° 3, p. 1207–1228.
- Desgagné, A. 2015, Robustness to outliers in location–scale parameter model using log-regularly varying distributions, *Ann. Statist.*, vol. 43, n° 4, p. 1568–1595.
- Desgagné, A. and P. Gagnon. 2016, Bayesian robustness to outliers in linear regression and ratio estimation, Submitted for publication.
- Green, P. J. 1995, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, vol. 82, n° 4, p. 711–732.
- Hastie, D. 2005, Towards automatic reversible jump markov chain monte carlo, Ph.D. Thesis, University of Bristol.
- Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky. 1999, Bayesian model averaging: A tutorial, *Statist. Sci.*, p. 382–401.
- Jeffreys, H. 1967, Theory of probability, Oxford Univ. Press, London.
- Karagiannis, G. and C. Andrieu. 2013, Annealed importance sampling reversible jump MCMC algorithms, *J. Comp. Graph. Stat.*, vol. 22, n° 3, p. 623–648.
- Lindley, D. V. 1957, A statistical paradox, *Biometrika*, vol. 44, n° 1, p. 187–192.
- Raftery, A. E., D. Madigan and J. A. Hoeting. 1997, Bayesian model averaging for linear regression models, *J. Amer. Statist. Assoc.*, vol. 92, n° 437, p. 179–191.
- Schwarz, G. 1978, Estimating the dimension of a model, *Ann. Statist.*, vol. 6, n° 2, p. 461–464.
- West, M. 1984, Outlier models and prior distributions in Bayesian linear regression, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 46, n° 3, p. 431–439.
- West, M. 2003, Bayesian factor regression models in the “large p, small n” paradigm, in *Bayesian Statistics 7*, Oxford Univ. Press, London, p. 723–732.

5.6. PROOFS

PROOF OF PROPOSITION 5.1. The proof is essentially a computation using that $f = \mathcal{N}(0, 1)$ and the structure of the principal components. First,

$$\pi(k, (\sigma^k, \boldsymbol{\beta}^k) \mid \mathbf{y}) \propto f(\mathbf{y} \mid k, (\sigma^k, \boldsymbol{\beta}^k))\pi(k)\pi(\sigma^k)\pi(\boldsymbol{\beta}^k \mid k) \propto f(\mathbf{y} \mid k, (\sigma^k, \boldsymbol{\beta}^k))(1/\sigma^k)\pi(k).$$

The likelihood function for a given model is

$$\begin{aligned} f(\mathbf{y} \mid k, (\sigma^k, \boldsymbol{\beta}^k)) &= \prod_{i=1}^n \frac{1}{\sigma^k \sqrt{2\pi}} \exp \left\{ -\frac{1}{2(\sigma^k)^2} (y_i - (\mathbf{x}_i^k)^T \boldsymbol{\beta}^k)^2 \right\} \\ &= \frac{1}{(\sigma^k)^n (2\pi)^{n/2}} \exp \left\{ -\frac{1}{2(\sigma^k)^2} \sum_{i=1}^n (y_i - (\mathbf{x}_i^k)^T \boldsymbol{\beta}^k)^2 \right\}. \end{aligned}$$

We now analyse the sum in the exponential:

$$\begin{aligned}\sum_{i=1}^n (y_i - (\mathbf{x}_i^k)^T \boldsymbol{\beta}^k)^2 &= \sum_{i=1}^n y_i^2 - 2 \sum_{i=1}^n y_i \sum_{j=1}^k x_{ij} \beta_j^k + \sum_{i=1}^n \left(\sum_{j=1}^k x_{ij} \beta_j^k \right)^2 \\ &= \sum_{i=1}^n y_i^2 - 2 \sum_{j=1}^k \beta_j^k \sum_{i=1}^n y_i x_{ij} + \sum_{i=1}^n \left(\sum_{j=1}^k x_{ij} \beta_j^k \right)^2.\end{aligned}$$

We also have

$$\sum_{i=1}^n \left(\sum_{j=1}^k x_{ij} \beta_j^k \right)^2 = \sum_{i=1}^n \left(\sum_{j=1}^k x_{ij}^2 (\beta_j^k)^2 + \sum_{j,s=1(j \neq s)}^k x_{ij} \beta_j^k x_{is} \beta_s^k \right) = \sum_{j=1}^k (\beta_j^k)^2 \sum_{i=1}^n x_{ij}^2,$$

using $\sum_{i=1}^n x_{ij} x_{is} = 0$ for all $j, s \in \{2, \dots, d\}$ with $j \neq s$, $x_{11} = \dots = x_{n1} = 1$, and $(1/n) \sum_{i=1}^n x_{ij} = 0$ for all $j \in \{2, \dots, d\}$. Consequently,

$$\begin{aligned}\sum_{i=1}^n (y_i - (\mathbf{x}_i^k)^T \boldsymbol{\beta}^k)^2 &= \sum_{i=1}^n y_i^2 - 2 \sum_{j=1}^k \beta_j^k \sum_{i=1}^n y_i x_{ij} + \sum_{j=1}^k (\beta_j^k)^2 \sum_{i=1}^n x_{ij}^2 \\ &= \sum_{i=1}^n y_i^2 - 2\beta_1^k \sum_{i=1}^n y_i - \mathbb{1}(k \geq 2) 2 \sum_{j=2}^k \beta_j^k \sum_{i=1}^n y_i x_{ij} + n(\beta_1^k)^2 + \mathbb{1}(k \geq 2) \sum_{j=2}^k (\beta_j^k)^2 \sum_{i=1}^n x_{ij}^2,\end{aligned}$$

using again $x_{11} = \dots = x_{n1} = 1$. We have

$$n(\beta_1^k)^2 - 2\beta_1^k \sum_{i=1}^n y_i = n \left(\beta_1^k - \frac{1}{n} \sum_{i=1}^n y_i \right)^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2.$$

We also have

$$\begin{aligned}\mathbb{1}(k \geq 2) \left(\sum_{j=2}^k (\beta_j^k)^2 \sum_{i=1}^n x_{ij}^2 - 2 \sum_{j=2}^k \beta_j^k \sum_{i=1}^n y_i x_{ij} \right) &= \mathbb{1}(k \geq 2) \sum_{j=2}^k \sum_{i=1}^n x_{ij}^2 \left((\beta_j^k)^2 - 2\beta_j^k \frac{\sum_{i=1}^n y_i x_{ij}}{\sum_{i=1}^n x_{ij}^2} \right) \\ &= \mathbb{1}(k \geq 2) \sum_{j=2}^k \sum_{i=1}^n x_{ij}^2 \left(\beta_j^k - \frac{\sum_{i=1}^n x_{ij} y_i}{\sum_{i=1}^n x_{ij}^2} \right)^2 - \mathbb{1}(k \geq 2) \sum_{j=2}^k \sum_{i=1}^n x_{ij}^2 \left(\frac{\sum_{i=1}^n x_{ij} y_i}{\sum_{i=1}^n x_{ij}^2} \right)^2.\end{aligned}$$

Putting this together leads to:

$$\begin{aligned}\pi(k, (\sigma^k, \boldsymbol{\beta}^k) \mid \mathbf{y}) &\propto \pi(k) \frac{1}{(\sigma^k)^{n-k+1}} \exp \left\{ -\frac{1}{2(\sigma^k)^2} \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 - \mathbb{1}(k \geq 2) \sum_{j=2}^k \frac{(\sum_{i=1}^n x_{ij} y_i)^2}{\sum_{i=1}^n x_{ij}^2} \right) \right\} \\ &\quad \times \frac{1}{\sigma^k \sqrt{2\pi}} \exp \left\{ -\frac{n}{2(\sigma^k)^2} (\beta_1^k - \bar{y})^2 \right\} \\ &\quad \times \left(\mathbb{1}(k = 1) + \mathbb{1}(k \geq 2) \prod_{j=2}^k \frac{1}{\sigma^k \sqrt{2\pi}} \exp \left\{ -\frac{\sum_{i=1}^n x_{ij}^2}{2(\sigma^k)^2} \left(\beta_j^k - \frac{\sum_{i=1}^n x_{ij} y_i}{\sum_{i=1}^n x_{ij}^2} \right)^2 \right\} \right).\end{aligned}$$

Multiplying and dividing by the appropriate terms leads to the result. ■

PROOF OF PROPOSITION 5.2. As explained in Green (1995), it suffices to separately verify that the probability to go from a set A to a set B is equal to the probability to go from B to A when updating the parameters and when switching models, for accepted movements and for all appropriate A, B .

When updating the parameters, the probability to go from a set A to a set B is given by

$$\int_A \pi(k, (\sigma^k, \boldsymbol{\beta}^k) | \mathbf{y}) g(1) \int_B \prod_{i=1}^{1+k} \varphi_i(w_i | k, (\sigma^k, \boldsymbol{\beta}^k)_i, \ell) \left(1 \wedge \frac{w_1^{-1} f(\mathbf{y} | k, \mathbf{w}^k)}{(\sigma^k)^{-1} f(\mathbf{y} | k, (\sigma^k, \boldsymbol{\beta}^k))} \right) d\mathbf{w}^k d(\sigma^k, \boldsymbol{\beta}^k),$$

and, using Fubini's theorem, this probability yields

$$\int_B \pi(k, \mathbf{w}^k | \mathbf{y}) g(1) \int_A \prod_{i=1}^{1+k} \varphi_i((\sigma^k, \boldsymbol{\beta}^k)_i | k, w_i, \ell) \left(1 \wedge \frac{(\sigma^k)^{-1} f(\mathbf{y} | k, (\sigma^k, \boldsymbol{\beta}^k))}{w_1^{-1} f(\mathbf{y} | k, \mathbf{w}^k)} \right) d(\sigma^k, \boldsymbol{\beta}^k) d\mathbf{w}^k,$$

which is the probability to go from B to A . Note that this is valid for all $k \in \{1, \dots, d\}$.

The probability to switch from Model $k \in \{1, \dots, d-1\}$, where the parameters are in the set A , to Model $k+1$, where the parameters are in the set $A' \times B$ (the set A' is a modified version of A to account for the addition of \mathbf{c}^{k+1}), is given by

$$\int_A \pi(k, (\sigma^k, \boldsymbol{\beta}^k) | \mathbf{y}) g(2) \int_B q_{k+1}(u^{k+1}) \times \left(1 \wedge \frac{\pi(k+1) f(\mathbf{y} | k+1, ((\sigma^k, \boldsymbol{\beta}^k) + \mathbf{c}^{k+1}), u^{k+1}))}{\pi(k) f(\mathbf{y} | k, (\sigma^k, \boldsymbol{\beta}^k)) q_{k+1}(u^{k+1})} \right) du^{k+1} d(\sigma^k, \boldsymbol{\beta}^k).$$

After the change of variables $(\sigma^{k+1}, \boldsymbol{\beta}^{k+1}) = ((\sigma^k, \boldsymbol{\beta}^k) + \mathbf{c}^{k+1}, u^{k+1})$, we have

$$\int_{A' \times B} \pi(k, ((\sigma^{k+1}, (\boldsymbol{\beta}^{k+1})^-) - \mathbf{c}^{k+1}) | \mathbf{y}) g(2) q_{k+1}(\boldsymbol{\beta}_{k+1}^{k+1}) \times \left(1 \wedge \frac{\pi(k+1) f(\mathbf{y} | k+1, (\sigma^{k+1}, \boldsymbol{\beta}^{k+1}))}{\pi(k) f(\mathbf{y} | k, ((\sigma^{k+1}, (\boldsymbol{\beta}^{k+1})^-) - \mathbf{c}^{k+1})) q_{k+1}(\boldsymbol{\beta}_{k+1}^{k+1})} \right) d(\sigma^{k+1}, \boldsymbol{\beta}^{k+1}).$$

This last probability is equal to

$$\int_{A' \times B} \pi(k+1, (\sigma^{k+1}, \boldsymbol{\beta}^{k+1}) | \mathbf{y}) g(3) \times \left(1 \wedge \frac{\pi(k) f(\mathbf{y} | k, ((\sigma^{k+1}, (\boldsymbol{\beta}^{k+1})^-) - \mathbf{c}^{k+1})) q_{k+1}(\boldsymbol{\beta}_{k+1}^{k+1})}{\pi(k+1) f(\mathbf{y} | k+1, (\sigma^{k+1}, \boldsymbol{\beta}^{k+1}))} \right) d(\sigma^{k+1}, \boldsymbol{\beta}^{k+1}),$$

which is the probability to switch from Model $k+1$, where the parameters are in the set $A' \times B$, to Model k , where the parameters are in the set A .

Therefore, the Markov chain $\{(K(m), (\sigma^{K(m)}, \boldsymbol{\beta}^{K(m)})(m)), m \in \mathbb{N}\}$ satisfies the reversibility condition with respect to the posterior. ■

5.7. APPENDIX

Name	Ticker symbol
Artis Real Estate Investment Trust	AX-UN.TO
Asanko Gold Inc.	AKG.TO
Bonterra Energy Corp.	BNE.TO
Canadian Imperial Bank Of Commerce	CM.TO
CI Financial Corp.	CIX.TO
Celestica Inc. Subordinate Voting Shares	CLS.TO
DHX Media Ltd.	DHX-B.TO
Dominion Diamond Corporation	DDC.TO
Gildan Activewear Inc.	GIL.TO
Husky Energy Inc.	HSE.TO
iPath Bloomberg Sugar Subindex	SGG
iShares MSCI Japan	EWJ
iShares 20+ Year Treasury Bond	TLT
Laurentian Bank of Canada	LB.TO
Parkland Fuel Corporation	PKI.TO
United States Oil Fund LP	USO
Vermilion Energy Inc.	VET.TO
Volume of the S&P 500	N/A

Table 5. VII. Names of the companies, funds, and indicators used in Section 5.4 with their ticker symbol (if available)

CONCLUSION

Dans cette thèse, nous avons proposé des stratégies simples pour résoudre des problèmes concrets. D'une part, au chapitre 2 nous avons présenté une stratégie menant à la spécification des fonctions nécessaires à l'implémentation de l'algorithme à sauts réversibles dans un contexte spécifique, ce qui facilite l'utilisation de cette méthode MCMC dans cette situation. Cette stratégie découle d'une optimisation de l'efficacité (mesurée par la capacité d'exploration de l'espace d'états) du processus stochastique limite. Il s'agit de la limite en terme de convergence faible des processus stochastiques engendrés par l'algorithme lorsque la dimension des modèles tend vers l'infini, soit le premier résultat de ce type pour cette méthode MCMC. D'autre part, au chapitre 4 une stratégie permettant l'atteinte de la robustesse complète dans un contexte d'estimation des paramètres du modèle de régression linéaire a été présentée. Cette robustesse complète se manifeste par la convergence de la distribution *a posteriori* des paramètres (basée sur l'échantillon complet) vers celle excluant les valeurs aberrantes (autrement dit, la distribution *a posteriori* basée sur les valeurs non aberrantes seulement) lorsque celles-ci s'éloignent à plus ou moins l'infini. L'impact des valeurs aberrantes est alors limité (voire éliminé), et ainsi, les conclusions obtenues sont en accord avec la majorité des observations. Au chapitre 5, ces stratégies ont été combinées pour permettre l'introduction d'une approche bayésienne de régression robuste sur composantes principales où les composantes importantes sont identifiées à l'aide d'un processus de sélection de modèles, ce dernier basé sur un algorithme à sauts réversibles efficace.

En plus de ces stratégies, de possibles avenues de recherche ont été proposées. Il s'agit premièrement de la généralisation des résultats présentés au chapitre 2 à des distributions *a posteriori* plus complexes, ce qui permettrait de guider les utilisateurs dans l'implémentation de l'algorithme à sauts réversibles dans un éventail plus large de situations. Deuxièmement, il est question de l'atteinte de la robustesse complète dans un contexte de sélection de variables pour des modèles de régression linéaire. Au chapitre 5, il a été vérifié empiriquement que la stratégie présentée au chapitre 4 pouvait être appropriée dans des contextes différents de celui initialement établi. Il serait intéressant d'obtenir des résultats théoriques qui viendraient confirmer la validité de la stratégie dans le contexte de sélection de modèles. Ceci permettrait par le fait même de fournir des conditions sous lesquelles la robustesse est garantie dans ce contexte. Finalement, il a été expliqué au

chapitre 5 qu'une procédure de réduction de dimension robuste est nécessaire afin d'obtenir une régression sur composantes principales complètement robuste aux valeurs aberrantes. Suffisamment de recherche permettrait d'introduire, par exemple, une approche de régression sur composantes principales où la robustesse complète est garantie pour toutes les étapes du processus de prédiction (sous certaines conditions qu'il serait possible d'identifier). Idéalement, cette méthodologie serait combinée à une procédure qui engendrerait un algorithme à sauts réversibles optimal (avec les résultats théoriques validant l'atteinte de l'efficacité optimale dans cette situation) et qui pourrait être exécutée automatiquement. Maintenant, voyons ce que l'avenir nous réserve...

APPENDICE : FONCTIONS R

Nous fournissons maintenant quelques fonctions R ayant été utilisées pour produire certaines analyses numériques présentées dans la thèse. Nous débutons par les fonctions servant à l'étude de simulation de la section 2.4.2. Ensuite, nous présentons les fonctions permettant le calcul des probabilités *a posteriori* des modèles et l'estimation des paramètres au chapitre 5.

```
fonction_type_prop<-function(A,tau){ ### fonction utilisée pour déterminer
  ### quel est le type de mouvement tenté lors d'une itération donnée

  u<-runif(1)

  if(u<=tau){prop=1} # on propose de tenter une mise à jour des
    # paramètres
  else{

    if(u>tau && u<=tau+(1-tau)*A/(A+1)){prop=2} # on propose de
    # tenter un ajout de paramètre
    else{prop=3} # on propose de tenter un retrait de paramètre

  }

  return(prop)
}

fonction_mcmc<-function(nb_iter=100000,burnin=0,val_ini,n,A,ell=2.38,tau){
  ### algorithme à sauts réversibles
  # le burn-in est égal à 0 par défaut car souvent démarré à stationnarité
  # ell=2.38 car valeur optimale lorsque f=N(0,1)

  vect_K<-matrix(ncol=1,nrow=nb_iter+1)
  vect_X_1<-matrix(ncol=1,nrow=nb_iter+1)
  # seulement les valeurs de K et X_1 sont gardées
```

```

vect_K[1]<-val_ini[1]
vect_X_1[1]<-val_ini[2]

### pair ou impair? pour distinguer dans quel cas on se trouve
# pour la distribution p
k_max=floor(sqrt(n)*log(n))
if(k_max %% 2==0){pair=T} else{pair=F}

# on définit un vecteur qui contient l'état actuel de la chaîne
vect_param<-matrix(ncol=1,nrow=n+k_max+1)
# premiere position <= K, les autres <= les paramètres
# des NA pour le reste
# pour tout de suite, l'état actuel est val_ini
vect_param<-val_ini

# on compte le nombre de fois qu'on propose une m-à-j
# des paramètres et le nombre de fois que c'est accepté
# pour comparer avec 0.234
compte_1<-0
nb_accept_1<-0

for(i in 2:(nb_iter+1)){

    ### On choisit le type de mouvement

    k<-vect_param[1]

    type<-fonction_type_prop(A,tau)

    if(type==1){ ### On met à jour les paramètres

        ### on veut simuler d'une normale de dimension
        ### n+k, centrée à l'état actuel

        esp<-vect_param[2:(n+k+1)]

        ### on simule le candidat y

        y<-rnorm(n+k,mean=esp,sd=(2.38 / sqrt(n+k)))

        ### on calcule le log de la prob d'accepter

        log_prob<--0.5*sum(y^2-esp^2)
    }
}

```

```

compte_1<-compte_1+1

if(log(runif(1))<=log_prob){
# on accepte le candidat

    vect_param[1,2:(n+k+1)]<-y

    nb_accept_1<-nb_accept_1+1

}
#else{

# rien à faire parce qu'on reste au même état

#}

}

if(type==2){ ### on ajoute un paramètre

# le cas impair
if(pair==F){

    if(k>=(k_max+1)/2 && k<k_max){
# on est au mode de K ou à droite,
# mais pas au max, si max, on refuse
# voir (2.2) pour la prob d'acceptation

        log_num<-log(1-(k-k_max)/2)/n

    }

    if(k<(k_max+1)/2){
# on est à gauche du mode

        log_num<--log(1-(k_max/2-k)/n)

    }

}

# le cas pair
if(pair==T){

    if(k==k_max/2){log_num<-0}
# p(k+1)/p(k)=1

```

```

    if(k>=k_max/2+1 && k<k_max){
      # on est au mode ou à droite

      log_num<-log(1-(k-k_max/2)/n)

    }
    if(k<k_max/2){
      # on est à gauche du mode

      log_num<--log(1-(k_max/2-k)/n)

    }

  }

log_denom<-log(A)

if(k<k_max){

  if(log(runif(1))<=(log_num - log_denom)){
    ### on accepte le candidat

    vect_param[1,1]<-k+1
    # on simule sa valeur
    vect_param[1,(n+k+2)]<-rnorm(1,mean=0,sd=1)

  }

}

#else{

# rien à faire parce qu'on reste au même état

#}

}

if(type==3){ ### on élimine un paramètre

  if(k>1){ ### on accepte le candidat si k>1
    # voir section 2.2 pour explication

    vect_param[1,(n+k+1)]<-NA
  }
}

```

```

                                vect_param[1,1]<-k-1

                                }
                                #else{

                                # rien à faire parce qu'on reste au même état

                                #}

                                }

                                # on enregistre les valeurs de K et de X_1
                                vect_K[i]<-vect_param[1]
                                vect_X_1[i]<-vect_param[2]

                                }

                                return(list(vect_K=vect_K,vect_X_1=vect_X_1,tx_acc_1=nb_accept_1/compte_1)

                                }

```

Nous présentons maintenant les fonctions servant au calcul des probabilités *a posteriori* des modèles et à l'estimation des paramètres au chapitre 5.

```

# on définit la valeur des paramètres de la log-Pareto-tailed normal
a<-1.96
v<-1+(2/sqrt(2*pi))*exp(-(a^2)/2)*a*log(a)/(1-.95) # environ 4.08
# on définit la densité de cette distribution
f<-function(x,sc,a,v){

    if(abs(x)<=(sc*a)){

        return(1/(sqrt(2*pi)*sc)*exp(-x^2/(2*sc^2)))

    }
    else{

        return( 1/sqrt(2*pi)*exp(-a^2/2)*(a/abs(x))*(log(a)/log(abs(x)/sc))^v)

    }

}

```

```

# fonction pour simuler d'une log-Pareto-tailed normal
simul_log_pareto<-function(a,v){

  u<-runif(1)

  if(u>=pnorm(-a) && u<=pnorm(a)){
    # on est dans la partie normale

    return(qnorm(u))

  }
  else{

    if(u<pnorm(-a)){

      return(-exp((sqrt(2*pi)*exp(a^2/2)*(v-1)*u/(a*(log(a))^v))^(1/(1-v))))

    }
    else{

      return(exp((sqrt(2*pi)*exp(a^2/2)*(v-1)*u/(a*(log(a))^v))^(1/(1-v))))

    }

  }

}

# log-vraisemblance sous la log-Pareto-tailed normal
logL<-function(param,y,x,a,v){
# param[1] <= sigma
# le reste, les coefficients de la régression

  return(sum(log(apply(as.matrix(y-x%*%param[2:length(param)],)
                      ,1,f,sc=param[1],a=a,v=v))))

}

# log-vraisemblance sous la normal
# utile lorsque les propriétés sur les colonnes de X ne sont plus respectées
# comme dans l'analyse à la section 5.4
logLnorm<-function(param,y,x){

  return(-length(y)*log(param[1])-(1/(2*param[1]^2))

```

```

        *sum((y-x%*%param[2:length(param),])^2))
    }

fonction_type_prop<-function(tau){ ### fonction utilisée pour déterminer
    ### quel est le type de mouvement tenté lors d'une itération donnée

    u<-runif(1)

    if(u<=tau){prop=1} # on propose de tenter une mise à jour des
        # paramètres

    else{

        if(u<=tau+(1-tau)/2){prop=2} # on propose de tenter un
            # ajout de paramètre
        else{prop=3} # on propose de tenter un retrait de paramètre

    }

    return(prop)

}

### algorithme à sauts réversibles
fonction_mcmc<-function(nb_iter, val_ini, d, tau, ell, y, x, a, v, vect_loc,
                        vect_sca, matrix_adjust){
    # vect_loc <= vecteurs contenant les positions des coefficients
    # lorsqu'on les ajoute, notées  $m_j^j$  dans le chapitre 5
    # la première position ne sera pas utilisée car on n'ajoute
    # jamais le premier coefficient (il est toujours présent)
    # vect_sca <= même chose, mais pour les ``scaling''
    # notés  $s_j^j$  dans le chapitre 5
    # matrix_adjust <= matrice dont les colonnes sont les vecteurs
    # notés  $c^2, c^3, \dots, c^d$  dans le chapitre 5 pour déplacement
    # des paramètres présents quand on change de modèle

    # on compte le nombre de fois où l'on tente une m-à-j et le nombre
    # de fois que ça fonctionne
    nb_accept_1<-0
    compte_1<-0

    n<-length(y)

```

```

# on enregistre toutes les valeurs de toutes les v.a.
matrix_rv<-matrix(ncol=2+d,nrow=nb_iter+1)

matrix_rv[1,]<-val_ini

for(i in 2:(nb_iter+1)){

  # quel est le modèle actuel?
  k<-matrix_rv[(i-1),1]

  ### On choisit le type de mouvement

  type<-fonction_type_prop(tau)

  if(type==1){ ### On met à jour les paramètres

    compte_1<-compte_1+1

    ### on veut simuler le candidat

    # position actuelle
    esp<-as.matrix(matrix_rv[(i-1),2:(k+2)])

    # candidat
    w<-matrix(nrow=k+1,ncol=1) ### log-Pareto

    for(j in 1:(k+1)){

      w[j]<-esp[j]+(ell / sqrt(1+k))*simul_log_pareto(a,v)

    }

    # si on souhaite simuler les candidats d'une normale
    # on l'utilise sous l'hypothèse de normalité
    #w<-as.matrix(rnorm(k+1,mean=esp,sd=(ell / sqrt(1+k))))

    ### on calcule le log de la prob d'accepter

    if(w[1]>0){
      # on vérifie que le candidat pour sigma > 0

      # log of numerator
      log_num<--log(w[1])+logL(w,y,as.matrix(x[,1:k]),a,v)
      # changer logL par la ligne dessus sous la normalité

```



```

#logLnorm(w, y, as.matrix(x[,1:k]))

#log of denom
log_denom<--log(esp[1])
                +logL(esp, y, as.matrix(x[,1:k]), a, v)
# +logLnorm(esp, y, as.matrix(x[,1:k]))

if(log(runif(1))<=log_num-log_denom){
### on accepte le candidat

        matrix_rv[i,1]<-k
        matrix_rv[i,2:(k+2)]<-w

        nb_accept_1<-nb_accept_1+1

}
else{

        matrix_rv[i,1]<-k
        matrix_rv[i,2:(k+2)]<-matrix_rv[(i-1),2:(k+2)]

}

}
else{

        matrix_rv[i,1]<-k
        matrix_rv[i,2:(k+2)]<-matrix_rv[(i-1),2:(k+2)]

}

}

if(type==2){ ### on ajoute un paramètre

        if(k<d){ # si on peut en ajouter un

                # on simule le candidat pour le nouveau paramètre
                U<-vect_loc[k+1]+vect_sca[k+1]*simul_log_pareto(a, v)
                # utiliser la ligne ci-dessous sous la normalité
                #U<-rnorm(1,mean=vect_loc[k+1],sd=vect_sca[k+1])

                # log of numerator
                log_num<-logL(as.matrix(c(matrix_rv[(i-1),2:(k+2)]

```

```

+c(0,matrix_adjust[1:k,k]),U),y,x[,1:(k+1)],a,v)
# utiliser la ligne ci-dessous sous la normalité
#logLnorm(as.matrix(c(matrix_rv[(i-1),2:(k+2)]
+c(0,matrix_adjust[1:k,k]),U),y,x[,1:(k+1)])

#log of denom
log_denom<-logL(as.matrix(matrix_rv[(i-1),2:(k+2)]),y,
                as.matrix(x[,1:k]),a,v)
                +log(f(U-vect_loc[k+1],vect_sca[k+1],a,v))
# utiliser la ligne ci-dessous sous la normalité
#logLnorm(as.matrix(matrix_rv[(i-1),2:(k+2)]),y,
                as.matrix(x[,1:k]))
                +dnorm(U,mean=vect_loc[k+1],sd=vect_sca[k+1],log=T)

if(log(runif(1))<=log_num-log_denom){
### on accepte le candidat

    matrix_rv[i,1]<-k+1
    matrix_rv[i,2:(k+2)]<-matrix_rv[(i-1),2:(k+2)]
                                +c(0,matrix_adjust[1:k,k])
    matrix_rv[i,(k+3)]<-U

}
else{

    matrix_rv[i,1]<-k
    matrix_rv[i,2:(k+2)]<-matrix_rv[(i-1),2:(k+2)]

}

}
else{

    matrix_rv[i,1]<-k
    matrix_rv[i,2:(k+2)]<-matrix_rv[(i-1),2:(k+2)]

}

}

if(type==3){ ### on élimine un paramètre

    if(k>1){ # si on peut en éliminer un

```

```

sca<-vect_sca[k]

# log of numerator
log_num<-logL(as.matrix(matrix_rv[(i-1),2:(k+1)]
  -c(0,matrix_adjust[1:(k-1),k-1])),
  y,as.matrix(x[,1:(k-1)]),a,v)
+log(f(matrix_rv[(i-1),k+2]-vect_loc[k],sca,a,v))
# utiliser la ligne ci-dessous sous la normalité
#logLnorm(as.matrix(matrix_rv[(i-1),2:(k+1)]
  -c(0,matrix_adjust[1:(k-1),k-1])),
  y,as.matrix(x[,1:(k-1)]))+dnorm(matrix_rv[(i-1),k+2],
  mean=vect_loc[k],sd=sca,log=T)

#log of denom
log_denom<-logL(as.matrix(matrix_rv[(i-1),2:(k+2)]),y,
  x[,1:k],a,v)
# utiliser la ligne ci-dessous sous la normalité
#logLnorm(as.matrix(matrix_rv[(i-1),2:(k+2)]),y,x[,1:k])

if(log(runif(1))<=log_num-log_denom){
### on accepte le candidat

matrix_rv[i,1]<-k-1
matrix_rv[i,2:(k+1)]<-matrix_rv[(i-1),2:(k+1)]
  -c(0,matrix_adjust[1:(k-1),k-1])

}
else{

  matrix_rv[i,1]<-k
  matrix_rv[i,2:(k+2)]<-matrix_rv[(i-1),2:(k+2)]

}

}
else{

  matrix_rv[i,1]<-k
  matrix_rv[i,2:(k+2)]<-matrix_rv[(i-1),2:(k+2)]

}

}
}

```

```

    }

    return(list(matrix_rv=matrix_rv[2:(nb_iter+1)],,
              tx_acc_1=nb_accept_1/compte_1))

}

# fonction pour calculer les prob des modèles
cal_post_prob_mcmc<-function(result_k,d){
# result_k <= vecteur contenant les réalisations de K

# on définit le dénominateur pour les prob
div<-length(result_k)

vect_prob_K_post<-matrix(ncol=1,nrow=d)
for(i in 1:d){

# on compte le nombre de réalisations pour chacune des valeurs
# et on divise par le nombre total
vect_prob_K_post[i]<-length(which(result_k==i))/div

}

return(vect_prob_K_post)

}

```

Ce code peut être utilisé afin d'effectuer les *trial runs* servant à estimer la position et l'échelle des paramètres pour chacun des modèles en posant $\tau = 1$, comme mentionné au chapitre 5. Il peut ainsi servir à l'estimation des paramètres pour un modèle donné dans un contexte plus général de robustesse en régression linéaire (lorsque les covariables ne sont pas nécessairement des composantes principales), comme celui étudié au chapitre 4. En fait, nous avons utilisé la stratégie de design de l'algorithme à sauts réversibles expliquée à la section 5.2.4 afin de calculer les cotes de Bayes (qui est un rapport de deux probabilités *a posteriori* de deux modèles différents) et d'estimer les paramètres à la section 4.3.1.

