

Stratégies de compromis pour l'estimation des paramètres de régression et pour la classification floue

THÈSE N° 5043 (2011)

PRÉSENTÉE LE 1^{ER} JUIN 2011

À LA FACULTÉ SCIENCES DE BASE

CHAIRE DE STATISTIQUE APPLIQUÉE

PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Nicolas FOURNIER

acceptée sur proposition du jury:

Prof. F. Eisenbrand, président du jury
Prof. S. Morgenthaler, directeur de thèse
Prof. F. Critchley, rapporteur
Prof. A. C. Davison, rapporteur
Prof. E. Ronchetti, rapporteur



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

A mes parents

La conclusion de cette thèse n'aurait pas pu avoir lieu sans l'aide de nombreuses personnes autour de moi, et je tiens à leur faire part de ma gratitude.

En premier lieu, je voudrais remercier le Professeur Stephan Morgenthaler, mon directeur de thèse, pour m'avoir permis de réaliser ce travail à ses côtés. Ses conseils très précieux, sa patience, sa générosité, ainsi que la confiance qu'il a placée en moi ont été essentiels pour mon développement scientifique et personnel, et ont fait de ces quatre années une expérience inoubliable.

Je tiens également à remercier le Professeur Frank Critchley (Open University, Angleterre), le Professeur Elvezio Ronchetti (Université de Genève, Suisse) et le Professeur Anthony C. Davison (EPFL, Suisse) pour avoir accepté de faire partie de mon jury. Merci également au Professeur Friedrich Eisenbrand pour avoir officié en tant que président.

Merci ensuite à mes deux géniales collègues de bureau successives, Sahar et Laurence. Leur constante bonne humeur a permis de créer une atmosphère décontractée et stimulante jour après jour, et leur disponibilité pour des questions de tout type m'aura enlevé plus d'une fois une épine du pied.

Merci à mes collègues de chaire, Jérôme, Ravi, Jean-Marc, Maya, Sonja et Anthea. J'ai beaucoup apprécié de travailler à leurs côtés, de même qu'avec Anne-Lise, toujours présente et de bonne humeur pour nous faciliter le côté administratif.

Enorme, énorme merci à Jacques, Jean-Benoît, Julian, Laurence et Stéphane pour les pauses café, pour les parties de cartes acharnées, pour les bières du mois et les sorties sportives, culturelles et gastronomiques. Avoir un tel groupe dynamique et joyeux à mes côtés a largement contribué à la réussite de ce projet. Je leur souhaite, ainsi qu'à tous les doctorants que j'ai pu côtoyer, le meilleur pour la suite de leur travail et de leur parcours.

Je suis également reconnaissant envers Maria-Pia et André pour avoir pris le temps de relire ce texte, et pour leurs nombreuses et très utiles remarques.

Merci enfin à mes parents, François et Geneviève, à mon frère, Antoine, et à mon amie, Isabelle, pour leur indéfectible soutien et leurs encouragements, non seulement durant mes années de thèse, mais également durant l'ensemble de mes études.

Lausanne, Mai 2011
Nicolas Fournier

Résumé

Le choix d'un modèle fait partie intégrante de tout problème d'inférence statistique. Plusieurs méthodes de sélection de modèle ont été développées afin de déterminer *le meilleur modèle* parmi une liste de candidats potentiels.

Une manière différente d'aborder la question est le compromis de modèles, et plus particulièrement son approche fréquentiste. Une estimation des paramètres d'intérêt est obtenue en construisant une moyenne pondérée des estimations de ces quantités sous chaque modèle candidat. Dans ce travail, nous développons des stratégies de compromis de type fréquentiste pour l'estimation des paramètres de régression, ainsi que pour la classification floue.

Dans le cas de la régression, nous construisons des compromis entre les estimateurs de Pitman associés à diverses distributions sous-jacentes pour les erreurs. Le poids associé à chaque modèle est égal à la vraisemblance de ce dernier, qui donne une mesure de la qualité de l'ajustement. Les propriétés asymptotiques des estimateurs de Pitman et de la vraisemblance nous permettent de définir une stratégie minimax pour le choix des distributions de compromis, faisant appel à une notion de distance entre distributions. Les performances de tels estimateurs sont par la suite comparées à celles d'autres estimateurs usuels et robustes.

Dans la seconde partie du travail, nous développons des stratégies de compromis pour la classification floue. Bien que cette méthode de classification se base sur des mélanges de distributions, nos stratégies de compromis ne sont pas faites directement sur les estimations des paramètres, mais sur les probabilités *a posteriori* d'appartenance aux classes. Deux types de compromis sont présentés, et les performances des règles de classification en résultant sont étudiées.

Mots-clé : compromis de modèles ; estimateurs de Pitman ; distance entre distributions ; classification floue.

Abstract

Model specification is an integral part of any statistical inference problem. Several model selection techniques have been developed in order to determine which model is *the best one* among a list of possible candidates.

Another way to deal with this question is the so-called model averaging, and in particular the frequentist approach. An estimation of the parameters of interest is obtained by constructing a weighted average of the estimates of these quantities under each candidate model. We develop compromise frequentist strategies for the estimation of regression parameters, as well as for the probabilistic clustering problem.

In the regression context, we construct compromise strategies based on the Pitman estimators associated with various underlying errors distributions. The weight given to each model is equal to its profile likelihood, which gives a measure of the goodness-of-fit. Asymptotic properties of both Pitman estimators and profile likelihood allow us to define a minimax strategy for choosing the distributions of the compromise, involving a notion of distance between distributions. Performances of such estimators are then compared to other usual and robust procedures.

In the second part of the thesis, we develop compromise strategies in the probabilistic clustering context. Although this clustering method is based on mixtures of distributions, our compromise strategies are not applied directly to the estimates of the parameters, but on the posterior probabilities of membership. Two types of compromise are presented, and the performances of resulting classification rules are investigated.

Keywords : model averaging; Pitman estimators; distance between distributions; probabilistic clustering.

Table des matières

Résumé	1
Abstract	3
1 Introduction	7
2 Sélection et compromis de modèles	11
2.1 Modèle statistique et estimation	11
2.2 Sélection de modèle	15
2.2.1 Le critère AIC	18
2.2.2 Le critère BIC	20
2.3 Compromis de modèles	21
3 Estimateurs de Pitman compromis	25
3.1 Estimateurs de Pitman	25
3.1.1 Estimateur de Pitman pour le paramètre de lieu	26
3.1.2 Estimateur de Pitman pour le paramètre d'échelle	33
3.1.3 Intervalles de confiance pour le paramètre de lieu	36
3.1.4 Remarques historiques	38
3.1.5 Comportement asymptotique des estimateurs de Pitman	39
3.2 Estimateurs de Pitman compromis	42
3.2.1 Estimateurs bi-optimaux	43
3.2.2 Estimateurs du maximum de vraisemblance compromis	46
3.2.3 Estimateur de Pitman compromis pour le paramètre de lieu	47
3.2.4 Estimateur de Pitman compromis pour le paramètre d'échelle	50
3.2.5 Choix des distributions de compromis	53
3.3 Simulations dans le cas de lois normales contaminées	54
3.4 Simulations dans le cas de lois t de Student	66
3.5 Généralisation à la régression linéaire	76
3.5.1 Estimateurs bi-optimaux pour la régression	78
3.5.2 Simulations dans le cas de lois t de Student	79

4	Stratégies de compromis pour la classification	97
4.1	Classification floue	98
4.1.1	L'algorithme EM	101
4.1.2	Classification floue robustifiée	104
4.2	Stratégie de compromis pour la classification floue	109
4.2.1	Simulations dans le cas de mélanges de lois elliptiques	115
4.2.2	Simulations dans le cas de mélanges de lois non-elliptiques	121
4.2.3	Simulations dans le cas de compromis sur la structure de covariance	123
4.2.4	Remarques concernant les résultats obtenus	127
5	Conclusion	131
A	Intégration numérique : méthode de Gauss-Legendre	135
B	Fonctions C₊₊ et R	139
	Bibliographie	150

Chapitre 1

Introduction

Les statistiques sont la science mathématique de la collecte, de l'analyse et de l'interprétation des données. Les statistiques sont utilisées dans un grand nombre de domaines aussi variés que la médecine, la psychologie, la finance, la physique ou encore l'écologie.

De manière générale, Fisher considérait que tout problème d'inférence statistique, c'est-à-dire la déduction de caractéristiques inconnues d'une population en se basant sur un échantillon limité issu de cette dernière, pouvait être décomposé en trois parties : le choix d'un modèle, l'ajustement de ce dernier, et l'évaluation de la qualité de l'ajustement. Durant la majorité du siècle passé, l'attention s'est énormément portée sur les deux derniers aspects. La théorie du maximum de vraisemblance notamment a permis l'application des statistiques aux nombreux domaines cités précédemment, et donne une méthode générale pour l'ajustement de modèles et l'analyse critique de la qualité de ce dernier.

En ce qui concerne le choix d'un modèle, on n'a montré que peu d'intérêt pour la question jusqu'au début des années 1970. En effet, les moyens techniques limités à l'époque suffisaient seulement à l'étude d'un seul modèle, et en prendre en considération plusieurs à la fois était dès lors difficile. C'est à la suite d'une série de papiers d'Akaike que des méthodes de sélection de modèle ont été développées et utilisées de manière croissante.

Le critère AIC, introduit par Akaike, est certainement une des méthodes de sélection de modèle les plus connues. Ce critère, intimement lié à la théorie de l'information introduite par Shannon au milieu du XX^e siècle, est basé sur l'estimation de la perte d'information qui se produit lorsque l'on utilise un modèle en particulier en lieu et place du modèle sous-jacent réel. Cette méthode fait de plus un compromis entre simplicité du modèle et qualité de l'ajustement, en pénalisant un grand nombre de paramètres.

Plusieurs autres critères de sélection de modèles ont été développés par la suite, et plusieurs d'entre eux ne sont que des modifications légères du critère AIC afin d'appli-

quer ce dernier à un problème particulier, chacun de ces critères possédant avantages et désavantages. Ces méthodes procèdent d'une manière similaire, en attribuant un score à chaque modèle candidat, et en sélectionnant le meilleur d'entre eux. Néanmoins, lorsque l'on choisit un modèle en particulier, et que l'on conduit l'inférence sur ce dernier uniquement, on ne tient pas compte d'une certaine incertitude quant à la sélection du modèle, et il s'ensuit souvent des intervalles de confiance trop reserrés ou des variances estimées trop faibles.

Une approche différente est alors le compromis de modèles, qui consiste à combiner plusieurs modèles candidats, en donnant à chacun un certain poids. Leamer introduit par exemple les compromis de modèles bayésiens, en construisant une distribution *a posteriori* pour le paramètre d'intérêt comme un mélange de distributions induites par chaque modèle candidat. Une approche fréquentiste est également possible, bien que nettement moins connue que le compromis bayésien. Dans ce cas, l'estimateur compromis est alors construit comme une moyenne pondérée des estimateurs résultant de chaque modèle candidat.

Dans ce travail, nous nous intéresserons à la construction de stratégies de compromis fréquentistes, dans le cas de l'estimation des paramètres du modèle simple de lieu et d'échelle, dans le cas de la régression linéaire simple et multiple, ainsi que dans le cas de la classification floue. Nous mettrons l'accent sur la partie probabiliste des modèles candidats en proposant un ensemble de modèles possibles possédant les mêmes paramètres, mais dont la distribution de probabilité sera différente. En construisant des compromis de la sorte, nous nous attendons à ce que les estimateurs en résultant se comportent de manière satisfaisante dans plusieurs situations. Cette idée correspond à une notion de robustesse.

Dans le chapitre 2, nous donnerons un aperçu plus détaillé de la notion de modèle statistique, ainsi que des méthodes d'estimation des paramètres, et de sélection de modèle, telles que les critères AIC et BIC. Nous présenterons également l'idée de compromis de modèles bayésiens et fréquentistes.

Dans le chapitre 3, nous construirons dans un premier temps des estimateurs compromis pour les paramètres du modèle de lieu et d'échelle, basés sur les estimateurs de Pitman. Nous commencerons par présenter en détail les estimateurs de Pitman. Puis nous définirons les estimateurs compromis, et nous nous baserons sur leurs comportements asymptotiques afin de sélectionner les modèles sur lesquels s'effectue le compromis, en utilisant une approche minimax.

Par la suite, nous évaluerons les performances de tels estimateurs au travers de simulations de Monte-Carlo, et ce pour plusieurs situations différentes. Nous comparerons les résultats avec des estimateurs bien connus des paramètres de lieu et d'échelles, y compris des estimateurs robustes comme le M-estimateur de Huber ou le Bisquare de Tukey. Nous utiliserons la généralisation des estimateurs de Pitman au cas de la régression linéaire simple et multiple afin d'y appliquer la construction d'estimateurs compromis.

Des simulations de Monte-Carlo seront également présentées, et nous comparerons les estimateurs compromis à d'autres estimateurs des paramètres de régression, dans le cas de designs fixes ou aléatoires.

Dans le chapitre 4, nous commencerons par présenter diverses méthodes de classification statistique, et plus particulièrement la méthode de classification floue, à laquelle nous appliquerons ensuite les stratégies de compromis. Nous donnerons ensuite les résultats de simulations de Monte-Carlo pour diverses situations et divers types de compromis.

Chapitre 2

Sélection et compromis de modèles

Un *modèle statistique* est une formalisation mathématique de relations entre différentes variables, habituellement présentée sous la forme d'équations mathématiques. Un modèle décrit comment une ou plusieurs variables aléatoires se comportent en fonction d'une ou plusieurs autres variables. Le modèle est dit statistique, car ces relations entre les différentes variables ne sont pas déterministes, mais stochastiques.

Plus précisément, à l'intérieur d'un modèle, les relations entre les variables sont exprimées à l'aide de *paramètres*. Ces derniers possèdent une interprétation utile, même lorsqu'ils sont associés à des quantités qui ne sont pas directement observables.

2.1 Modèle statistique et estimation

De manière générale, nous notons un modèle de la façon suivante :

$$Y = f(X_1, \dots, X_p \mid \boldsymbol{\theta}) + \varepsilon,$$

où Y est la variable dépendante, appelée également variable réponse, où X_1, \dots, X_p sont les variables indépendantes (ou encore explicatives), où $\boldsymbol{\theta}$ représente les paramètres à déterminer, et où ε est un terme d'erreur aléatoire, que l'on ne peut pas observer.

Exemple 2.1.1. Le modèle statistique certainement le plus connu est le modèle de *régression linéaire simple*, donné par

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où ε est distribué selon une loi normale centrée et de variance σ^2 , généralement inconnue. Dans ce modèle, on décrit la relation entre Y et X par une relation linéaire. β_0, β_1 , et σ s'il est inconnu, sont les paramètres à estimer.

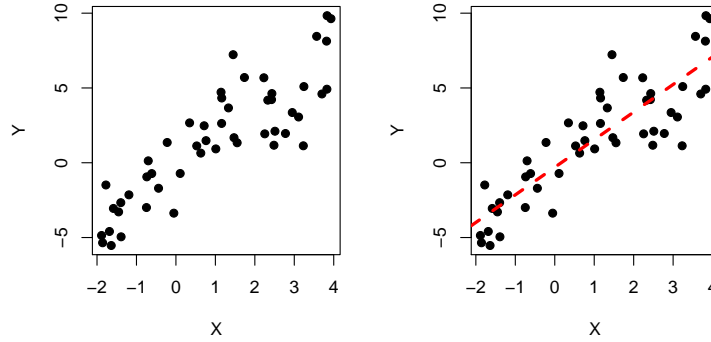


FIG. 2.1: Modèle de régression linéaire simple. On cherche la droite rouge représentant au mieux la relation entre X et Y .

Fisher (1922) décrit trois aspects successifs d'un problème général d'inférence statistique (voir également Mallows, 1998) :

1. spécification d'un modèle pour le problème en particulier ;
2. estimation des paramètres du modèle ;
3. estimation de la précision (qualité) de l'ajustement.

La deuxième et la troisième partie de cette démarche ont été très largement traitées durant une grande partie du siècle passé, et continuent de l'être aujourd'hui. Parmi les principales méthodes d'estimation des paramètres, citons la méthode des moindres carrés et la méthode du maximum de vraisemblance.

La méthode des moindres carrés, indépendamment développée par Gauss et Legendre au début du XIX^e siècle, est basée sur une notion de distance entre les observations y_i , $i = 1, \dots, n$, d'un échantillon et les valeurs prédites par le modèle. Les valeurs optimales des paramètres θ au sens des moindres carrés sont celles minimisant la quantité

$$\sum_{i=1}^n (y_i - f(x_{i,1}, \dots, x_{i,p} | \theta))^2 = \sum_{i=1}^n r_i^2(\theta),$$

où les $r_i(\theta)$ sont appelés les résidus du modèle. Ces derniers représentent l'écart entre les mesures y_i et les valeurs prédites par le modèle $f(x_{i,1}, \dots, x_{i,p} | \theta)$.

Exemple 2.1.2. Dans le modèle de régression linéaire simple, pour un échantillon (x_i, y_i) , $i = 1, \dots, n$, les estimateurs des moindres carrés de β_0 et β_1 sont donnés par

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

où $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ et $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. Le paramètre σ^2 est quant à lui estimé par la quantité

$$\hat{\sigma}^2 = \frac{\text{SCE}(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2.$$

La méthode des moindres carrés est le sujet de centaines d'ouvrages, et est utilisée aussi bien avec des modèles linéaires que non-linéaires, particulièrement lorsque les observations sont indépendantes, et lorsque les éléments stochastiques du modèles proviennent d'une loi normale.

La méthode du maximum de vraisemblance, développée par Fisher, est beaucoup plus générale que les moindres carrés. L'élément central de cette méthode est le modèle probabiliste régissant les observations, étant donné les paramètres : si nous avons

$$Y \sim G(y \mid \boldsymbol{\theta}),$$

où G est une distribution de probabilité, alors la vraisemblance pour un échantillon $y_i, i = 1, \dots, n$, est donnée par

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n g(y_i \mid \boldsymbol{\theta}).$$

La vraisemblance est fonction uniquement des paramètres $\boldsymbol{\theta}$, car tout le reste est connu (distribution G , taille de l'échantillon n , et observations y_i). La vraisemblance et le modèle probabiliste sont intimement liés, et chacun possède le rôle inverse de l'autre.

L'idée de Fisher est de choisir $\hat{\boldsymbol{\theta}}$ de telle sorte que la vraisemblance soit maximale :

$$\hat{\boldsymbol{\theta}}_{MV} = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}).$$

On choisit de cette manière la valeur des paramètres la plus probable, celle ayant le plus de chances d'avoir réellement engendré les données à disposition.

Exemple 2.1.3. Soit y_1, \dots, y_n des réalisations indépendantes d'une loi de Bernoulli de paramètre inconnu $p \in [0, 1]$. Le modèle probabiliste est donc $Y \sim \text{Bern}(p)$, et

$$g(y \mid p) = p^y (1-p)^{1-y} = \begin{cases} p, & \text{si } y = 1; \\ 1-p, & \text{si } y = 0. \end{cases}$$

La vraisemblance vaut alors

$$L(p) = \prod_{i=1}^n g(y_i \mid p) = p^{\sum_{i=1}^n y_i} (1-p)^{n - \sum_{i=1}^n y_i},$$

et l'estimateur du maximum de vraisemblance de p vaut

$$\hat{p}_{MV} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

La théorie du maximum de vraisemblance peut s'appliquer à un grand nombre de modèles et de situations. La méthode des moindres carrés peut être vue comme un cas particulier du maximum de vraisemblance, lorsque le modèle probabiliste est gaussien, comme c'est notamment le cas dans l'exemple de la régression linéaire. Les développements en informatique ont d'ailleurs permis la mise en oeuvre de la vraisemblance dans des problèmes de plus en plus complexes.

De plus, les méthodes de vraisemblance permettent l'estimation asymptotique optimale des paramètres. Il est également possible d'obtenir des intervalles de confiance et des tests d'hypothèses.

Ces méthodes d'estimation des paramètres ont été rejointes dans la seconde moitié du XX^e siècle par des méthodes dites robustes. Box (1953) a été le premier à utiliser le terme de *robustesse* en étudiant le comportement de certaines procédures statistiques usuelles lorsque les hypothèses sous-jacentes utilisées pour leur développement sont violées. Les statistiques robustes sont ainsi nées d'une critique des méthodes d'estimation classiques, à savoir leur incapacité à protéger leur utilisateur contre une mauvaise spécification du modèle. Un estimateur est ainsi dit robuste s'il est résistant à des petites déviations du modèle de base.

Exemple 2.1.4. Considérons un échantillon de n réalisations indépendantes d'une loi normale de moyenne $\mu \in \mathbb{R}$ et de variance 1 : $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$. Dans ce cas, l'estimateur optimal de μ est la moyenne arithmétique de l'échantillon :

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Imaginons à présent que la loi sous-jacente n'est plus une loi normale, mais une loi normale contaminée. Pour $\varepsilon \in [0, 1]$, les X_i sont alors distribués comme

$$X_i \sim \begin{cases} \mathcal{N}(\mu, 1), & \text{avec probabilité } 1 - \varepsilon; \\ \mathcal{N}(\mu, 10), & \text{avec probabilité } \varepsilon. \end{cases}$$

Quelles sont alors les conséquences d'un tel changement de modèle sur les performances de l'estimateur \bar{x} ? En calculant la variance de ce dernier sous le modèle de la loi normale contaminée, nous obtenons :

$$\text{Var}(\bar{x}) = \frac{1}{n} [(1 - \varepsilon) + 10\varepsilon].$$

La variance de \bar{x} augmente donc linéairement en fonction de la fraction de contamination ε , et même pour des petites contaminations, l'effet est assez important.

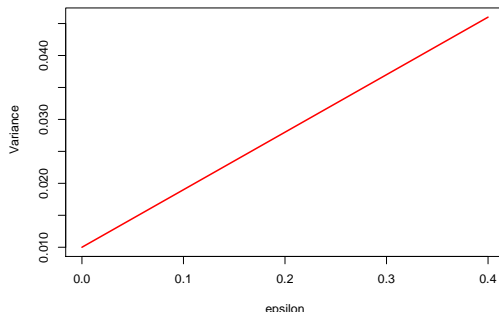


FIG. 2.2: Comportement de la variance de l'estimateur \bar{x} en fonction de la fraction de contamination ε .

Deux articles de Tukey (1960, 1962) vont poser les bases de l'estimation robuste, et Huber (1964) propose par la suite une généralisation de la méthode des moindres carrés et de la méthode du maximum de vraisemblance pour l'estimation du paramètre de lieu. Il définit ainsi une large classe d'estimateurs, appelés M-estimateurs, qui deviendra une méthode centrale dans l'estimation robuste de paramètres.

On notera également le développement d'outils permettant de *mesurer* la robustesse d'un estimateur, comme le point de rupture (*breakdown point* en anglais), introduit par Hampel (1971), ou encore la fonction d'influence, également introduite par Hampel (1974). Pour une vue d'ensemble des statistiques robustes, on pourra consulter Huber (1981), Hampel *et al.* (1986), ou encore Stigler (2010).

2.2 Sélection de modèle

Le problème de l'estimation des paramètres d'un modèle a donc été largement traité, au contraire peut-être de celui de la sélection du modèle en lui-même. Avant l'avènement des ordinateurs, il était déjà suffisamment difficile d'ajuster un modèle en particulier, et il n'était dès lors pas envisageable d'essayer d'appliquer plusieurs modèles différents au même jeu de données. Le problème du choix de modèle ne se posait alors pas.

La spécification d'un modèle peut être divisée en deux parties : la formulation d'une liste de modèles possibles, puis la sélection effective d'un (ou plusieurs) modèle pour l'inférence. Dans leur ouvrage traitant de la sélection de modèle, Burnham et Anderson (2002) mettent l'accent sur l'importance d'une réflexion poussée lorsqu'il s'agit de déterminer un ensemble de modèles possibles. Trop souvent, on se concentre sur l'analyse théorique et sur l'analyse des données, en oubliant quelque peu les raisons de l'étude.

Une fois que l'on a choisi un ensemble de candidats se pose la question *Quel est le meilleur modèle ?* Cette question doit tout d'abord être clarifiée par l'expérimentateur, dans le sens où ses buts et ses attentes doivent être précisés. Par la suite, on

essaiera de choisir un modèle proposant un certain équilibre entre adéquation aux données (*goodness-of-fit*) et simplicité (principe de parcimonie). Ce n'est qu'à partir des années 1970, suite à une série de papiers d'Akaike (1973, 1974, 1977), que des critères de sélection de modèle parmi une liste de candidats potentiels ont été développés et utilisés.

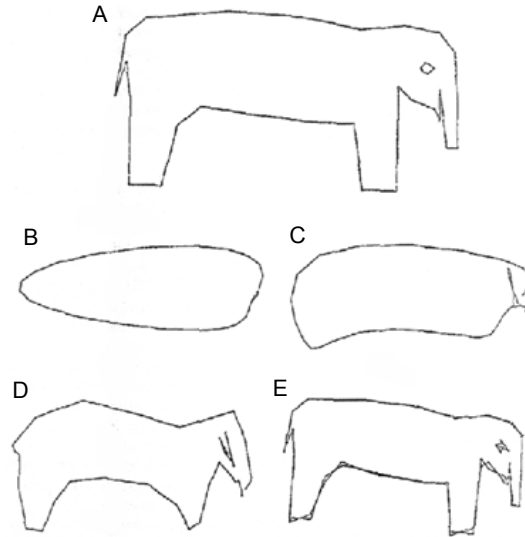


FIG. 2.3: Principe de parcimonie : *Combien faut-il de paramètres pour ajuster un éléphant ?* La question a été posée par Wel (1975). L'éléphant A est constitué de 36 points, et Wel a utilisé des séries de Fourier pour l'approcher. Il est arrivé à la conclusion que 30 termes étaient suffisants (E).

La méthode proposée par Akaike est liée au concept d'entropie. Elle permet d'obtenir une mesure de la perte d'information qui survient lorsque l'on utilise un modèle afin d'approcher la réalité. Cette méthode permet également de faire un compromis entre le biais (peu de paramètres) et la variance (trop de paramètres) dans la sélection d'un modèle, c'est-à-dire un compromis entre la précision et la simplicité. La théorie de l'information couvre de nombreux domaines (voir Cover et Thomas, 2006), et en particulier, la distance de Kullback-Leibler joue un rôle prépondérant dans la détermination de la méthode proposée par Akaike.

Définition 2.2.1. Soit F et G deux distributions de probabilité, et soit f et g leurs fonctions de densité respectives. La divergence (ou parfois appelée distance) de Kullback-Leibler est définie par :

$$D(G||F) = D(G(y)||F(y)) = \int \log \frac{g(y)}{f(y)} g(y) dy.$$

Introduite par Kullback et Leibler (1951) afin de donner une définition rigoureuse de l'information en rapport avec les statistiques suffisantes de Fisher, la divergence de

Kullback-Leibler représente la perte d'information lorsque F est utilisée au lieu de G . Strictement parlant, cette quantité n'est pas une distance, car elle n'est pas symétrique. En effet, en général nous avons $D(G||F) \neq D(F||G)$. C'est pourquoi Kullback préférerait le terme de *divergence*, ou encore *information de discrimination*.

Exemple 2.2.2. L'exemple suivant, tiré de Burnham et Anderson (2002), illustre la divergence de Kullback-Leibler entre G , une loi gamma de paramètres $\alpha = 4$ et $\beta = 4$, à différentes lois F_1, \dots, F_4 : une distribution de Weibull, lognormale, inverse gaussienne et une distribution F . Ces distributions sont représentées graphiquement dans la figure 2.4.

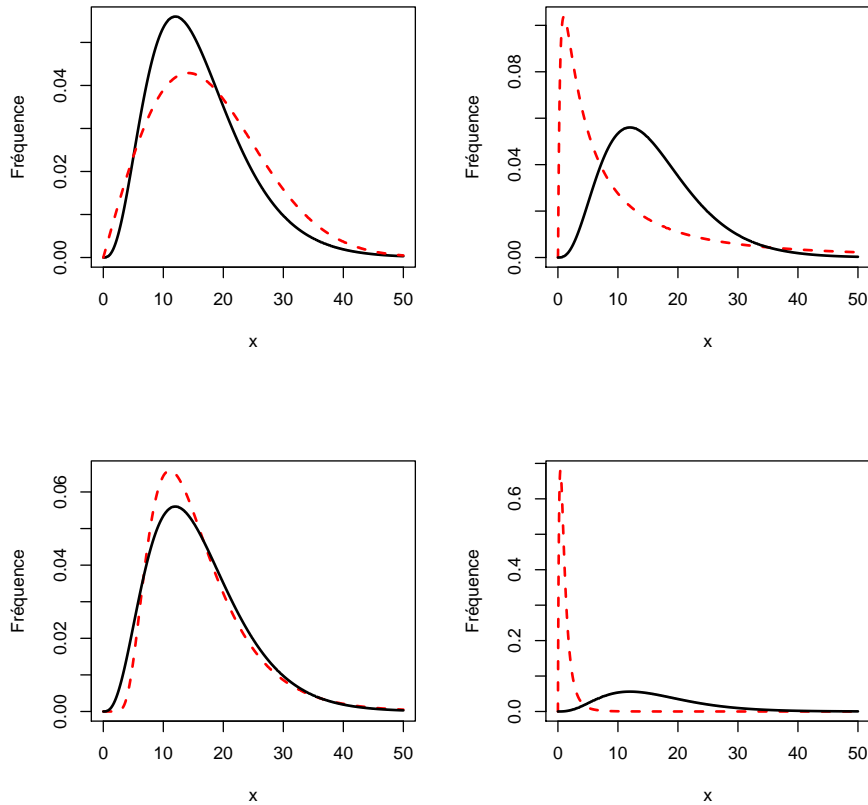


FIG. 2.4: Graphes de G (en noir) et de F_1, \dots, F_4 (en rouge).

Les paramètres des différentes distributions F_i sont ici connus et non pas estimés. Dans ce cas, la distribution de Weibull est la plus proche de G , suivie par la loi inverse gaussienne. La distribution F_4 est quant à elle très éloignée de G .

	Modèle	$D(G F_i)$
F_1	Weibull ($\alpha = 2, \beta = 20$)	0.04620
F_2	Lognormale ($\theta = 2, \sigma^2 = 2$)	0.67235
F_3	Inverse gaussienne ($\alpha = 16, \beta = 64$)	0.06008
F_4	Distribution F ($\alpha = 4, \beta = 10$)	5.74555

TAB. 2.1: Divergences de Kullback-Leibler entre G , une loi gamma de paramètres $\alpha = 4, \beta = 4$, et différentes distributions F_1, \dots, F_4 .

2.2.1 Le critère AIC

Akaike (1973) proposa l'utilisation de la divergence de Kullback-Leibler pour la sélection de modèles, à partir d'un ensemble de candidats potentiels. Néanmoins, cette distance ne peut être calculée sans connaître exactement le modèle réel sous-jacent ainsi que les paramètres de chaque modèle que l'on souhaite ajuster. Akaike contourne ce problème en déterminant une estimation de la divergence de Kullback-Leibler, basée sur la fonction de vraisemblance évaluée à son maximum.

Selon Akaike, afin de sélectionner un modèle sur la base de la divergence de Kullback-Leibler, la quantité critique pour un modèle $f(y | \theta)$ est la double espérance suivante :

$$\mathbb{E}_x \left[\mathbb{E}_y \left(\log(f(y | \hat{\theta}(x))) \right) \right],$$

où x et y sont des échantillons indépendants provenant de la même distribution, et où les deux espérances sont à calculer par rapport à la vraie distribution sous-jacente G .

Comme le montre Akaike, il ne suffit pas d'estimer cette double espérance par le logarithme de la vraisemblance évaluée à son maximum. En effet, ce dernier est biaisé vers le haut, et Akaike a montré que sous certaines conditions techniques (mais importantes), ce biais est approximativement égal au nombre de paramètres à estimer dans le modèle F .

Définition 2.2.3. Désignons par F un modèle, contenant les paramètres θ . Soit K le nombre de tels paramètres. Le critère AIC (pour *An Information Criterion*, ou encore *Akaike Information Criterion*), basé sur un échantillon $y_i, i = 1, \dots, n$, est alors défini par

$$\text{AIC}(F) = -2 \log L(\hat{\theta}_{MV}) + 2K,$$

où

$$L(\theta) = \prod_{i=1}^n f(y_i | \theta)$$

est la fonction de vraisemblance, et où

$$\hat{\boldsymbol{\theta}}_{MV} = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

est l'estimateur du maximum de vraisemblance des paramètres du modèle F .

Le facteur multiplicatif -2 est ajouté « pour des raisons historiques ». Le critère AIC est donc une estimation de la distance relative entre le modèle ajusté et la réalité. Il est formé de deux parties distinctes : une partie faisant appel à la vraisemblance observée, qui représente la qualité de l'ajustement aux données, et une seconde partie pénalisant le nombre de paramètres à estimer dans le modèle.

A partir d'un ensemble de modèles possibles, on calcule pour chacun la valeur du critère AIC, et le modèle présentant la valeur la plus petite est désigné comme étant le meilleur parmi les candidats. La valeur du critère en elle-même n'est pas importante, seule la valeur relative par rapport aux autres modèles importe.

Exemple 2.2.4. Considérons un ensemble de modèles candidats tel que la partie probabiliste de chaque modèle est une loi normale avec variance constante. C'est par exemple le cas dans les modèles de régression. Alors dans ce cas, le critère AIC peut être facilement calculé, et vaut

$$\text{AIC}(F) = n \log(\hat{\sigma}^2) + 2K,$$

où

$$\hat{\sigma}^2 = \frac{\text{SCE}(\boldsymbol{\theta})}{n}$$

est l'estimateur du maximum de vraisemblance de la variance σ^2 des erreurs. Notons que les termes au dénominateur de $\hat{\sigma}^2$ est bien n , la taille de l'échantillon à disposition, car il ne s'agit pas ici de l'estimateur des moindres carrés.

Plusieurs modifications du critère AIC ont été proposées par la suite. Mentionnons par exemple le critère AIC_c , où un terme de pénalisation supplémentaire valant $2K(K+1)/(n-K-1)$ est ajouté, afin de corriger le biais pour des petits échantillons (voir Hurvich et Tsai, 1989). Il a également été souvent reproché à Akaike que, dans les hypothèses de son travail, la vraie distribution sous-jacente G était incluse dans l'ensemble des modèles candidats. Akaike a pourtant confirmé que son estimation était non-biaisée asymptotiquement, et que le modèle réel pouvait ne pas faire partie des candidats. Takeuchi (1976) traite de cette question, et introduit le critère TIC (*Takeuchi Information Criterion*), où la correction du biais est ajustée dans le cas où le modèle sous-jacent est très éloigné des modèles candidats.

2.2.2 Le critère BIC

Très similaire au critère AIC, le critère BIC (*Bayesian Information Criterion*) a été introduit par Schwarz (1978) et Akaike (1977, 1978). Comme son nom l'indique, une composante bayésienne intervient dans son développement, mais ne se retrouve pas dans son utilisation. Lorsque plusieurs modèles sont possibles, une approche bayésienne consiste à sélectionner celui possédant la probabilité *a posteriori* la plus élevée. En posant une probabilité *a priori* égale pour chaque modèle candidat, et des hypothèses *a priori* très vagues sur la distribution des paramètres connaissant le modèle, on arrive alors au critère BIC, défini ci-dessous.

Définition 2.2.5. Désignons par F un modèle, contenant les paramètres θ . Soit K le nombre de tels paramètres. Le critère BIC (pour *Bayesian Information Criterion*), basé sur un échantillon y_i , $i = 1, \dots, n$, est alors défini par

$$\text{BIC}(F) = -2 \log L(\hat{\theta}_{MV}) + \log(n)K,$$

où

$$L(\theta) = \prod_{i=1}^n f(y_i | \theta)$$

est la fonction de vraisemblance, et où

$$\hat{\theta}_{MV} = \arg \max_{\theta} L(\theta)$$

est l'estimateur du maximum de vraisemblance des paramètres du modèle F .

Le critère BIC impose donc une pénalisation plus grande au nombre de paramètres du modèle, pour autant que la taille de l'échantillon soit plus grande que $n = 8$. Notons que cette quantité n'est pas une estimation de la divergence de Kullback-Leibler, au contraire du critère AIC.

Bien que très semblables, les critères AIC et BIC possèdent chacun des forces et des faiblesses. Par exemple, le critère BIC est consistant, dans le sens où la probabilité de sélectionner le vrai modèle, s'il fait partie des candidats, tend vers 1 lorsque la taille de l'échantillon tend vers l'infini. Au contraire, ce n'est pas le cas pour le critère AIC. Par contre, ce dernier est dit efficace, dans le sens où il sélectionnera un modèle tel que le rapport entre la perte moyenne sous ce modèle et la perte moyenne minimale converge vers 1 en probabilité, ceci dans le cas de la régression linéaire. Le critère BIC, lui, ne vérifie pas cette propriété. Pour plus de détails, voir Sin et White (1996) et Hurvich et Tsai (1995), et pour un ouvrage traitant de sélection de modèles en général, on pourra consulter McQuarrie et Tsai (1998).

2.3 Compromis de modèles

Les critères de sélection de modèle sont nombreux, et les critères AIC et BIC ne sont que les plus connus d'entre eux. Plusieurs modifications de ces méthodes ont été développées, plus particulièrement pour les appliquer à une discipline spécifique. Néanmoins, chacune de ces méthodes donnent un certain score à chaque modèle candidat, et sélectionne le meilleur d'entre eux. Parfois, un candidat est nettement plus performant que les autres, mais il se peut que plusieurs modèles donnent des résultats similaires et très proches. Dans ce cas, il peut être intéressant de tenir compte de tous ces modèles, et de ne pas les écarter au profit d'un seul candidat. On parle alors de compromis de modèles, *model averaging* en anglais.

Considérons un problème pour lequel deux modèles candidats F_1 et F_2 semblent bien ajustés aux données, mais donnent des estimations des paramètres très différentes. Un tel cas n'est pas rare, comme on peut le voir par exemple dans Regal et Hook (1991). Dès lors, baser l'inférence statistique sur un seul des deux modèles occulte une certaine incertitude quant à la sélection du modèle, et il en résulte souvent une sous-estimation de la variabilité et des intervalles de confiance trop étroits. Pour plus de détails concernant l'incertitude du modèle, on pourra consulter Draper (1995) et Chatfield (1995). Une solution consiste donc à tenir compte des deux modèles en même temps, et deux approches sont possibles.

Le compromis de modèles bayésien (*Bayesian model averaging*) est la première d'entre elles. Introduite par Leamer (1978), cette méthode consiste à construire une distribution *a posteriori* pour le paramètre d'intérêt qui serait un mélange entre la distribution *a posteriori* induite par chacun des modèles candidats. Le poids associé à chaque modèle est quant à lui donné par la probabilité *a posteriori* du candidat.

Soit $\mathcal{F} = \{F_1, \dots, F_m\}$ un ensemble de modèles candidats, soit θ_k les paramètres associés au modèle F_k , pour $k = 1, \dots, m$, et soit Δ le paramètre d'intérêt du problème en question. Alors la distribution *a posteriori* de Δ sachant les données D est définie par :

$$h(\Delta | D) = \sum_{k=1}^m h(\Delta | F_k, D)h(F_k | D),$$

où $h(F_k | D)$ est la probabilité *a posteriori* du modèle F_k , donnée par

$$h(F_k | D) = \frac{f(D | F_k)\pi(F_k)}{\sum_{l=1}^m f(D | F_l)\pi(F_l)},$$

avec

$$f(D | F_k) = \int f(D | \theta_k, F_k)f^*(\theta_k | F_k)d\theta_k$$

la vraisemblance intégrée du modèle F_k , où $\pi(F_k)$ est la probabilité *a priori* que le modèle F_k soit le vrai modèle (sachant que ce dernier se trouve dans \mathcal{F}), et où $f^*(\boldsymbol{\theta}_k | F_k)$ est la densité *a priori* des paramètres $\boldsymbol{\theta}_k$ dans le modèle F_k .

Plusieurs difficultés se posent lorsque l'on cherche à appliquer cette méthode :

- le nombre de modèles candidats peut être très grand, ce qui pose problème pour la sommation dans le mélange des distributions ;
- les intégrales nécessitent généralement une évaluation numérique ;
- le choix des probabilités *a priori* $\pi(F_k)$, $k = 1, \dots, m$, à assigner aux différents candidats est une question difficile ;
- choisir l'ensemble \mathcal{F} devient la question principale de la modélisation.

Hoeting *et al.* (1999) proposent diverses solutions, discutent de chacune de ces difficultés, et donnent des exemples concrets d'application à des problèmes aussi variés que la régression linéaire, la régression linéaire généralisée, ou les modèles à risque proportionnel et l'analyse de survie.

La seconde méthode de compromis de modèles est une approche fréquentiste (*frequentist model averaging*). Ici, le paramètre d'intérêt est estimé à l'aide de plusieurs modèles, et une moyenne pondérée est finalement formée. Le poids donné à chaque modèle peut être beaucoup plus général que dans le cas des compromis bayésiens.

La littérature traitant de cette manière de procéder est nettement moins dense que pour les compromis bayésiens. On trouve dans Rao et Tibshirani (1997) par exemple la description d'une méthode similaire au bootstrap, dans laquelle une observation est laissée de côté, et un poids est donné à chaque modèle selon la qualité de l'ajustement de ce dernier sur cette observation en particulier. Yang (2001) propose une méthode adaptative pour la régression linéaire, dans laquelle le jeu de données est divisé en deux. A nouveau, une partie est utilisée pour l'ajustement du modèle, et l'autre pour en juger la qualité, et ainsi déterminer des poids.

Hjort et Claeskens (2003, 2008) présentent un cadre général pour l'étude d'estimateurs compromis fréquentistes, et parviennent à déterminer le comportement asymptotique de tels estimateurs. Soit $\mathcal{F} = \{F_1, \dots, F_m\}$ un ensemble de modèles candidats, et soit $\boldsymbol{\theta}_k$ les paramètres associés au modèle F_k . L'estimateur compromis est alors défini par :

$$\hat{\boldsymbol{\theta}} = \sum_{k=1}^m w(F_k) \hat{\boldsymbol{\theta}}_k,$$

où $w(\cdot)$ est une fonction de poids non-négative et sommant à 1. Par exemple, on peut utiliser le critère AIC pour former des poids, en posant

$$w(F_k) = \frac{\exp\left(-\frac{1}{2}\text{AIC}(F_k)\right)}{\sum_{l=1}^m \exp\left(-\frac{1}{2}\text{AIC}(F_l)\right)},$$

où le modèle présentant l'AIC le plus faible reçoit le plus grand poids. On peut également former des poids semblables en se basant sur le critère BIC.

Hjort et Claeskens définissent le cadre suivant pour l'étude des estimateurs compromis. Soit $\mu = \mu(f)$ le paramètre d'intérêt, dépendant du modèle f . Les modèles candidats sont des augmentations d'un modèle de base

$$f(y \mid \boldsymbol{\theta}),$$

où $\boldsymbol{\theta} \in \mathbb{R}^p$, et sont de la forme

$$f(y \mid \boldsymbol{\theta}, \boldsymbol{\gamma}),$$

où $\boldsymbol{\gamma} \in \mathbb{R}^q$ est un vecteur de paramètres supplémentaires. De plus, il existe une valeur $\boldsymbol{\gamma}_0$ de $\boldsymbol{\gamma}$ telle que le modèle de base est équivalent au modèle augmenté avec ces paramètres, c'est-à-dire :

$$f(y \mid \boldsymbol{\theta}) = f(y \mid \boldsymbol{\theta}, \boldsymbol{\gamma}_0).$$

Les modèles candidats peuvent être décrits par S , un sous-ensemble de $\{1, \dots, q\}$, puisque certains γ_j seront égalés aux $\gamma_{j,0}$, et d'autres seront laissés libres. Il y a donc 2^q modèles possibles. Les estimateurs de la quantité d'intérêt μ peuvent alors être écrits comme

$$\hat{\mu}_S = \mu(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma}_{0,S^c}),$$

où $S \subset \{1, \dots, q\}$, et où S^c est le complémentaire de S . Finalement, Hjort et Claeskens étudient le comportement asymptotique des estimateurs compromis du type

$$\sum_{S \subset \{1, \dots, q\}} w(S) \hat{\mu}_S,$$

dans le cas où le vrai modèle sous-jacent est du type

$$g(y) = g_n(y) = f(y \mid \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n}).$$

Dans cette formulation, $\boldsymbol{\delta} \in \mathbb{R}^q$ représente des directions dans lesquelles le modèle réel s'éloigne du modèle candidat de base. On suppose donc que les données proviennent de $f(y \mid \boldsymbol{\theta}_0, \boldsymbol{\gamma})$, où $\boldsymbol{\gamma}$ n'est pas trop éloigné de $\boldsymbol{\gamma}_0$.

Dans ce qui suit, nous développerons des estimateurs compromis en suivant l'approche fréquentiste. Notre démarche sera néanmoins quelque peu différente de ce qui a été présenté ci-dessus, dans le sens où nous mettrons l'accent non pas sur les paramètres à inclure (ou non) dans un modèle, mais plutôt sur la partie probabiliste de ce dernier.

Dans le cas de la régression linéaire par exemple, nous ne nous intéresserons pas à faire des compromis entre des modèles comportant différentes covariables, mais plutôt

à combiner les résultats de modèles supposant diverses lois sous-jacentes pour les erreurs, la partie aléatoire du modèle. Cette approche va nous permettre d'utiliser les compromis de modèles dans une optique de robustesse, afin de construire un estimateur se comportant bien sous des petites modifications du modèle, et dans un nombre important de situations différentes.

Chapitre 3

Estimateurs de Pitman compromis

Dans un article inspiré par une idée de Fisher (1934), Pitman (1939) construit un estimateur équivariant pour le lieu, présentant un carré moyen de l'erreur minimal pour une distribution continue connue. La construction est basée sur un conditionnement sur une statistique ancillaire, et l'évaluation de plusieurs intégrales doubles est nécessaire au calcul de l'estimateur. Une importante bibliographie existe à propos des estimateurs de Pitman, et Lehmann et Casella (1998) présentent un excellent résumé, comprenant notamment les propriétés minimax de ces estimateurs, ainsi que leur admissibilité et leur efficacité asymptotique.

3.1 Estimateurs de Pitman

Considérons le modèle habituel pour le lieu $\mu \in \mathbb{R}$ et l'échelle $\sigma \in \mathbb{R}_+$

$$Y = \mu + \sigma E, \quad E \sim F,$$

où F est une distribution continue connue. Dans ce modèle, les données sont distribuées autour de la valeur μ (le lieu), selon la distribution F . Le paramètre d'échelle σ représente la dispersion des données autour de μ . Un problème classique de la statistique est l'estimation du paramètre de lieu sur la base d'un échantillon $\mathbf{y} = (y_1, \dots, y_n)$ de n réalisations indépendantes de Y .

Parmi les estimateurs existants, la moyenne arithmétique de l'échantillon, donnée par $\bar{y} = n^{-1} \sum_{i=1}^n y_i$, est certainement la plus connue. Nous pouvons également mentionner les estimateurs du maximum de vraisemblance, la médiane, ainsi que les M-estimateurs de Huber. Pour plus de détails concernant l'estimation du paramètre de lieu, voir Lehmann et Casella (1998).

Définition 3.1.1. Un estimateur, ou une statistique, $T(\cdot)$ est dit(e) équivariant(e) pour le lieu et l'échelle s'il (elle) vérifie la propriété suivante :

$$T(s(\mathbf{y} + t\mathbf{1})) = s(T(\mathbf{y}) + t), \quad \forall s > 0, \forall t \in \mathbb{R},$$

où $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$.

Exemple 3.1.2. La moyenne arithmétique de l'échantillon $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ est une statistique équivariante pour le lieu et l'échelle.

La propriété d'équivariance pour le lieu et l'échelle semble être une propriété indispensable et naturelle pour tout estimateur de lieu. En effet, si l'on décale chacune des observations de la même valeur, on peut naturellement s'attendre à ce que l'estimateur du lieu soit également décalé. De même, si l'on change l'échelle des observations (par exemple un changement d'unités des mesures), l'estimateur du lieu devrait suivre.

Définition 3.1.3. Un estimateur, ou une statistique, $T(\cdot)$ est dit(e) équivariant(e) pour l'échelle et invariant(e) pour le lieu s'il (elle) vérifie la propriété suivante :

$$T(s(\mathbf{y} + t\mathbf{1})) = |s|T(\mathbf{y}), \quad \forall s > 0, \forall t \in \mathbb{R},$$

où $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$.

Exemple 3.1.4. L'estimateur classique de l'écart-type d'un échantillon

$$s = \left(\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right)^{\frac{1}{2}}$$

est un estimateur équivariant pour l'échelle.

3.1.1 Estimateur de Pitman pour le paramètre de lieu

Une idée développée par Pitman (1939) et basée sur un travail de Fisher (1934) permet l'estimation du paramètre de lieu pour une distribution continue F des erreurs dont la densité est connue. Cette méthode consiste à conditionner le raisonnement sur un sous-espace des observations possibles, en cherchant un estimateur équivariant pour le lieu et l'échelle. L'estimateur de Pitman minimise le carré moyen de l'erreur, et est déterminé de la façon suivante.

Soit l'ensemble $\mathcal{A}(\mathbf{y}) = \{\mathbf{z} = b(\mathbf{y} + a\mathbf{1}) \mid a \in \mathbb{R}, b > 0\}$. C'est sur ce demi-sous-espace linéaire de \mathbb{R}^n , l'ensemble des observations possibles, que le raisonnement de Pitman est basé.

Définition 3.1.5. Soit $a(\mathbf{y})$ une statistique équivariante pour le lieu et l'échelle, et soit $b(\mathbf{y})$ une statistique équivariante pour l'échelle. Le vecteur $\mathbf{c} = \mathbf{c}(\mathbf{y}) \in \mathbb{R}^n$ défini par

$$\mathbf{c} = \frac{1}{b(\mathbf{y})}(\mathbf{y} - a(\mathbf{y})\mathbf{1})$$

est appelé la configuration.

Exemple 3.1.6. Comme statistiques $a(\mathbf{y})$ et $b(\mathbf{y})$ on peut par exemple choisir

$$a(\mathbf{y}) = \bar{y} \quad \text{et} \quad b(\mathbf{y}) = \left(\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right)^{\frac{1}{2}},$$

ou encore

$$a(\mathbf{y}) = y_1 \quad \text{et} \quad b(\mathbf{y}) = |y_1 - y_n|,$$

si $y_1 \neq y_n$.

Remarquons que l'on a forcément $a(\mathbf{c}) = 0$ et $b(\mathbf{c}) = 1$. Ainsi, l'espace de toutes les configurations possibles est seulement de dimension $n - 2$. Cet élément de $\mathcal{A}(\mathbf{y})$ est choisi comme élément représentatif de cet ensemble. Il est alors possible d'exprimer tout élément de $\mathcal{A}(\mathbf{y})$ à l'aide de \mathbf{c} , de la manière suivante :

$$\mathcal{A}(\mathbf{y}) = \mathcal{A}(\mathbf{c}) = \{ \mathbf{z} = s(\mathbf{c} + t\mathbf{1}) \mid t \in \mathbb{R}, s > 0 \}.$$

Le point essentiel du raisonnement de Pitman tient dans la propriété suivante.

Proposition 3.1.7. *La statistique \mathbf{c} est une statistique ancillaire pour les paramètres μ et σ du modèle. C'est-à-dire, la distribution de \mathbf{c} est indépendante des paramètres μ et σ .*

Preuve Soit $\mathbf{x} = (x_1, \dots, x_n)$ un échantillon indépendant provenant de la distribution $F(y)$, et soit $y_1 = \mu + \sigma x_1, \dots, y_n = \mu + \sigma x_n$. L'échantillon $\mathbf{y} = (y_1, \dots, y_n)$ provient donc de la distribution inconnue $F((y - \mu)/\sigma)$. Néanmoins, grâce à l'équivariance des statistiques $a(\mathbf{y})$ et $b(\mathbf{y})$ utilisées dans la définition de la configuration \mathbf{c} , les deux échantillons \mathbf{x} et \mathbf{y} possèdent la même configuration. Ainsi, la distribution de \mathbf{c} est bien indépendante de μ et de σ . □

Considérons à présent le critère choisi par Pitman pour déterminer son estimateur optimal.

Définition 3.1.8. Pour un estimateur quelconque $T(\cdot)$ de μ , le carré moyen de l'erreur de $T(\cdot)$, sous la loi F des erreurs, et basé sur l'échantillon \mathbf{y} , est donné par

$$\text{CME}_F(T) = \mathbb{E}_F((T(\mathbf{y}) - \mu)^2).$$

L'idée de Pitman est donc de conditionner sur l'ensemble $\mathcal{A}(\mathbf{y})$, représenté par l'élément particulier \mathbf{c} , et ainsi de calculer le carré moyen de l'estimateur de la manière suivante :

$$\text{CME}_F(T) = \mathbb{E}_{F^*}(\mathbb{E}_F((T(\mathbf{y}) - \mu)^2 \mid \mathbf{c})),$$

où F^* représente la distribution de la configuration \mathbf{c} , induite par la loi F d'échantillonnage. Si l'on désire minimiser le carré moyen de l'erreur, il suffit en fait de minimiser le carré moyen de l'erreur conditionnel

$$\text{cCME}_F(T) = \mathbb{E}_F((T(\mathbf{y}) - \mu)^2 \mid \mathbf{c}).$$

En effet, $(T(\mathbf{y}) - \mu)^2 \geq 0$ et la distribution F^* étant indépendante de μ , l'estimateur minimisant CME_F sera celui minimisant cCME_F pour tout \mathbf{c} . Rappelons que l'échantillon \mathbf{y} peut être exprimé à l'aide de la configuration \mathbf{c} en posant $\mathbf{y} = s(\mathbf{c} + t\mathbf{1})$, pour un certain $t \in \mathbb{R}$ et un certain $s > 0$, et que nous cherchons un estimateur $T(\cdot)$ équivariant pour le lieu et l'échelle. Nous avons alors, en supposant sans perte de généralité que $\mu = 0$:

$$\begin{aligned} \text{cCME}_F(T) &= \mathbb{E}_F((T(\mathbf{y}) - \mu)^2 \mid \mathbf{c}) \\ &= \mathbb{E}_F((T(s(\mathbf{c} + t\mathbf{1})) - \mu)^2 \mid \mathbf{c}) \\ &= \mathbb{E}_F(s^2(T(\mathbf{c}) + t)^2 \mid \mathbf{c}) \\ &= T(\mathbf{c})^2 \mathbb{E}_F(s^2 \mid \mathbf{c}) + 2T(\mathbf{c}) \mathbb{E}_F(s^2 t \mid \mathbf{c}) + \mathbb{E}_F(s^2 t^2 \mid \mathbf{c}). \end{aligned}$$

En dérivant par rapport à $T(\mathbf{c})$ et en égalant à 0, on obtient immédiatement que

$$T_F(\mathbf{c}) = - \frac{\mathbb{E}_F(s^2 t \mid \mathbf{c})}{\mathbb{E}_F(s^2 \mid \mathbf{c})}.$$

L'estimateur de Pitman pour le paramètre de lieu est alors obtenu en utilisant une fois de plus l'équivariance de $T_F(\cdot)$, et est donné par

$$T_F(\mathbf{y}) = b(\mathbf{y})T_F(\mathbf{c}) + a(\mathbf{y}).$$

Notons également que le carré moyen de l'erreur conditionnel de l'estimateur de Pitman vaut

$$\text{cCME}_F(T_F) = \mathbb{E}_F(s^2 t \mid \mathbf{c})T_F(\mathbf{c}) + \mathbb{E}_F(s^2 t^2 \mid \mathbf{c}).$$

Dans l'expression de $T_F(\mathbf{c})$, ainsi que pour le calcul du carré moyen de l'erreur conditionnel, les espérances sont à calculer par rapport à la distribution conjointe conditionnelle de s et t connaissant la configuration \mathbf{c} . En effet, en réexprimant les éléments de $\mathcal{A}(\mathbf{y})$ à l'aide de \mathbf{c} , nous n'avons rien fait d'autre qu'un changement de variables, la distribution F se rapportant aux variables y_i .

Proposition 3.1.9. *La distribution conjointe conditionnelle de s et t sachant \mathbf{c} est proportionnelle à*

$$s^{n-1} \prod_{i=1}^n f(s(t + c_i)),$$

où f est la densité de la distribution F .

Preuve Il faut dans un premier temps calculer le déterminant de la matrice jacobienne de la transformation $y_i = s(t + c_i), i = 1, \dots, n$, donnée par

$$J = \begin{pmatrix} t + c_1 & s & s & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & s & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \\ t + c_{n-2} & s & 0 & 0 & & s \\ t + c_{n-1} & s & 0 & 0 & \cdots & 0 \\ t + c_n & s & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

En développant par rapport à la dernière ligne, nous avons

$$\det(J) = (-1)^{n-1}(t + c_n) \det(J_1) + (-1)^n s \det(J_2)$$

où

$$J_1 = \begin{pmatrix} s & s & 0 \\ \vdots & \ddots & \\ s & 0 & s \\ s & 0 & \cdots & 0 \end{pmatrix} \quad \text{et} \quad J_2 = \begin{pmatrix} t + c_1 & s & 0 \\ \vdots & \ddots & \\ t + c_{n-2} & 0 & s \\ t + c_{n-1} & 0 & \cdots & 0 \end{pmatrix}.$$

En développant à nouveau chacun des déterminants par rapport à la dernière ligne :

$$\det(J_1) = (-1)^n s \det(sI_{n-2}) = (-1)^n s^{n-1},$$

et

$$\det(J_2) = (-1)^n (t + c_{n-1}) \det(sI_{n-2}) = (-1)^n (t + c_{n-1}) s^{n-2},$$

où $I_n \in \mathbb{R}^{n \times n}$ est la matrice identité. Ainsi, $\det J$ est bien proportionnel à s^{n-1} , puisque les termes contenant la variable t s'annulent, du fait de leurs signes opposés. Enfin, puisque les observations y_1, \dots, y_n sont indépendantes, leur densité conjointe est le produit de leur densité. La densité conjointe conditionnelle de s et t sachant \mathbf{c} est alors bien de la forme

$$f(s, t \mid \mathbf{c}) = \frac{1}{K_F} s^{n-1} \prod_{i=1}^n f(s(t + c_i)),$$

où

$$K_F = \int_0^{+\infty} \int_{-\infty}^{+\infty} s^{n-1} \prod_{i=1}^n f(s(t + c_i)) dt ds$$

est une constante de normalisation. □

Remarque 3.1.10. Pitman a également démontré que f^* , la densité marginale de la configuration \mathbf{c} , était donnée par

$$f^*(\mathbf{c}) = k_{\mathbf{y}} \int_0^{+\infty} \int_{-\infty}^{+\infty} s^{n-1} \prod_{i=1}^n f(s(t + c_i)) dt ds,$$

où $k_{\mathbf{y}}$ est une constante ne dépendant que de l'échantillon \mathbf{y} , et non de f , la densité sous-jacente des erreurs.

L'évaluation des espérances conditionnelles nécessaires au calcul de l'estimateur n'est souvent pas possible sans passer par des méthodes d'intégration numériques. Néanmoins, dans le cas de la loi normale, il est possible de les calculer directement, comme le montre l'exemple suivant.

Exemple 3.1.11. Soit $F = \Phi$, la distribution normale centrée et réduite dont la densité est donnée par $\varphi(y) = (2\pi)^{-\frac{1}{2}} \exp(-y^2/2)$. Dans ce cas particulier, nous devons calculer des espérances du type

$$\begin{aligned} \mathbb{E}_F(g(s, t) \mid \mathbf{c}) &\propto \int_{-\infty}^{+\infty} \int_0^{+\infty} g(s, t) s^{n-1} \prod_{i=1}^n \exp(-s^2(t + c_i)^2/2) ds dt \\ &= \int_{-\infty}^{+\infty} \int_0^{+\infty} g(s, t) s^{n-1} \exp\left(\frac{-s^2}{2} \sum_{i=1}^n (t + c_i)^2\right) ds dt. \end{aligned}$$

Soit $\bar{c} = \frac{1}{n} \sum_{i=1}^n c_i$. Comme $\sum_{i=1}^n (c_i - \bar{c}) = 0$, nous avons que

$$\sum_{i=1}^n (t + c_i)^2 = \sum_{i=1}^n (t + \bar{c} + c_i - \bar{c})^2 = \sum_{i=1}^n (c_i - \bar{c})^2 + n(t + \bar{c})^2.$$

Le calcul des espérances est alors grandement facilité. Considérons pour commencer la constante de normalisation de la densité.

$$\begin{aligned} K_F &= \int_{-\infty}^{+\infty} \int_0^{+\infty} s^{n-1} \exp\left(\frac{-s^2}{2} \sum_{i=1}^n (t + c_i)^2\right) ds dt \\ &= \int_{-\infty}^{+\infty} \int_0^{+\infty} s^{n-1} \exp\left(\frac{-s^2}{2} \left[\sum_{i=1}^n (c_i - \bar{c})^2 + n(t + \bar{c})^2\right]\right) ds dt \\ &= \int_0^{+\infty} s^{n-1} \exp\left(\frac{-s^2}{2} \sum_{i=1}^n (c_i - \bar{c})^2\right) \left(\int_{-\infty}^{+\infty} \exp\left(-\frac{ns^2}{2}(t + \bar{c})^2\right) dt\right) ds \\ &= \left(\frac{2\pi}{n}\right)^{\frac{1}{2}} \int_0^{+\infty} s^{n-2} \exp\left(\frac{-s^2}{2} \sum_{i=1}^n (c_i - \bar{c})^2\right) ds, \end{aligned}$$

puisque l'on peut reconnaître à une constante près l'intégrale de la densité d'une loi normale de moyenne $-\bar{c}$ et de variance $\frac{1}{ns^2}$. En utilisant le même raisonnement, et une intégration par parties, nous avons :

$$\begin{aligned}
 K_F \mathbb{E}_F(s^2 \mid \bar{c}) &= \int_{-\infty}^{+\infty} \int_0^{+\infty} s^2 s^{n-1} \exp\left(\frac{-s^2}{2} \left[\sum_{i=1}^n (c_i - \bar{c})^2 + n(t + \bar{c})^2 \right]\right) ds dt \\
 &= \left(\frac{2\pi}{n}\right)^{\frac{1}{2}} \int_0^{+\infty} s^{n-1} s \exp\left(-\frac{s^2}{n} \sum_{i=1}^n (c_i - \bar{c})^2\right) ds \\
 &= -\left(\frac{2\pi}{n}\right)^{\frac{1}{2}} s^{n-1} \frac{1}{\sum_{i=1}^n (c_i - \bar{c})^2} \exp\left(-\frac{s^2}{2} \sum_{i=1}^n (c_i - \bar{c})^2\right) \Bigg|_{s=0}^{s=+\infty} \\
 &\quad + \frac{(n-1)}{\sum_{i=1}^n (c_i - \bar{c})^2} K_F \\
 &= \frac{(n-1)}{\sum_{i=1}^n (c_i - \bar{c})^2} K_F,
 \end{aligned}$$

et donc $\mathbb{E}_F(s^2 \mid \mathbf{c}) = (n-1) / \sum_{i=1}^n (c_i - \bar{c})^2$. De même, en faisant cette fois-ci apparaître le calcul de l'espérance d'une loi normale de moyenne $-\bar{c}$ et de variance $\frac{1}{ns^2}$, nous obtenons que $\mathbb{E}_F(s^2 t \mid \mathbf{c}) = (-\bar{c})(n-1) / \sum_{i=1}^n (c_i - \bar{c})^2$. Nous avons donc finalement

$$T_F(\mathbf{c}) = \bar{c},$$

et donc

$$\begin{aligned}
 T_F(\mathbf{y}) &= b(\mathbf{y})\bar{c} + a(\mathbf{y}) \\
 &= \bar{y},
 \end{aligned}$$

l'estimateur traditionnel du paramètre de lieu. De la même manière que précédemment, on peut montrer que $\mathbb{E}_F(s^2 t^2 \mid \mathbf{c}) = \bar{c}^2(n-1) / \sum_{i=1}^n (c_i - \bar{c})^2 + \frac{1}{n}$, et donc que le carré moyen de l'erreur conditionnel de $T_F(\mathbf{y})$ vaut $\frac{1}{n}$.

Dans l'exemple qui suit, nous présentons le calcul théorique de l'estimateur de Pitman dans le cas d'une loi normale contaminée, qui sera utilisée plus tard dans ce travail.

Exemple 3.1.12. Considérons la distribution normale contaminée. Cette loi représente la distribution d'une variable possédant une probabilité $\varepsilon \in [0, 0.5]$ de provenir d'une loi normale $\mathcal{N}(0, k^2)$, où $k \geq 1$, et une probabilité $1 - \varepsilon$ de provenir d'une loi normale centrée et réduite. La densité de cette loi est donnée par

$$f(y) = (1 - \varepsilon)\varphi(y) + \varepsilon k^{-1}\varphi(y/k),$$

où $\varphi(y)$ est la densité de la loi normale centrée et réduite. Afin de déterminer l'estimateur de Pitman pour le paramètre de lieu associé à cette distribution, nous allons

dans un premier temps conditionner notre raisonnement sur l'arrangement exact des erreurs, noté \mathbf{a} . C'est-à-dire, nous supposons que nous connaissons exactement si l'observation $i = 1, \dots, n$ provient de la partie contaminée de la distribution, ou de la partie standard. Une fois cette supposition faite, les calculs sont très semblables à ceux de l'exemple précédent. Soit

$$w_{\mathbf{a}}(i) = \begin{cases} 1, & \text{si la } i\text{-ème observation provient de } \mathcal{N}(0, 1); \\ k^{-2}, & \text{sinon.} \end{cases}$$

La densité conjointe de s et t , sachant la configuration \mathbf{c} et l'arrangement \mathbf{a} est alors de la forme

$$f(s, t \mid \mathbf{c}, \mathbf{a}) \propto s^{n-1} k^{-m} \exp\left(-\frac{1}{2} s^2 \sum_{i=1}^n w_{\mathbf{a}}(i) (t + c_i)^2\right),$$

où m est le nombre d'observations provenant de la partie contaminée de la loi. Soit maintenant $\bar{c}_{\mathbf{a}} = (\sum_i w_{\mathbf{a}}(i) c_i) / \sum_i w_{\mathbf{a}}(i) = (\sum_i w_{\mathbf{a}}(i) c_i) / (n - m + \frac{m}{k^2})$, et nous pouvons alors exprimer

$$\sum_{i=1}^n w_{\mathbf{a}}(i) (t + c_i)^2 = \sum_{i=1}^n w_{\mathbf{a}}(i) (c_i - \bar{c}_{\mathbf{a}})^2 + (t + \bar{c}_{\mathbf{a}})^2 \left(n - m + \frac{m}{k^2}\right).$$

La constante de normalisation de la densité conditionnelle vaut alors

$$\begin{aligned} K_F &= \int_0^{+\infty} \int_{-\infty}^{+\infty} s^{n-1} \frac{1}{k^m} \exp\left(-\frac{1}{2} s^2 \sum_{i=1}^n w_{\mathbf{a}}(i) (t + c_i)^2\right) dt ds \\ &= \left(\frac{2\pi}{n - m + \frac{m}{k^2}}\right)^{\frac{1}{2}} \frac{1}{k^m} \int_0^{+\infty} s^{n-2} \exp\left(-\frac{1}{2} s^2 \sum_{i=1}^n w_{\mathbf{a}}(i) (c_i - \bar{c}_{\mathbf{a}})^2\right) ds, \end{aligned}$$

et donc en intégrant par parties

$$\begin{aligned} \mathbb{E}_F(ts^2 \mid \mathbf{c}, \mathbf{a}) &= \frac{1}{K_F} \int_0^{+\infty} \int_{-\infty}^{+\infty} ts^2 s^{n-1} \frac{1}{k^m} \exp\left(-\frac{1}{2} s^2 \sum_{i=1}^n w_{\mathbf{a}}(i) (t + c_i)^2\right) dt ds \\ &= \frac{1}{K_F} \int_0^{+\infty} s^{n-1} s^2 \frac{1}{k^m} \exp\left(-\frac{1}{2} s^2 \sum_{i=1}^n w_{\mathbf{a}}(i) (c_i - \bar{c}_{\mathbf{a}})^2\right) ds \cdot \\ &\quad \left[\int_{-\infty}^{+\infty} t \exp\left(-\frac{1}{2} s^2 \left(n - m + \frac{m}{k^2}\right) (t + \bar{c}_{\mathbf{a}})^2\right) dt \right] \\ &= -\frac{1}{K_F} \left(\frac{2\pi}{n - m + \frac{m}{k^2}}\right)^{\frac{1}{2}} \frac{\bar{c}_{\mathbf{a}}}{k^m} \int_0^{+\infty} s^n \exp\left(-\frac{1}{2} s^2 \sum_{i=1}^n w_{\mathbf{a}}(i) (c_i - \bar{c}_{\mathbf{a}})^2\right) ds \\ &= -\bar{c}_{\mathbf{a}} \frac{n-1}{\sum_{i=1}^n w_{\mathbf{a}}(i) (c_i - \bar{c}_{\mathbf{a}})^2}. \end{aligned}$$

De manière tout à fait similaire :

$$\mathbb{E}_F(s^2 \mid \mathbf{c}, \mathbf{a}) = \frac{n-1}{\sum_{i=1}^n w_{\mathbf{a}}(i)(c_i - \bar{c}_{\mathbf{a}})^2}.$$

L'estimateur de Pitman sera alors donné par

$$\begin{aligned} T_F(\mathbf{c}) &= \frac{\sum_{\mathbf{a}} \mathbb{E}_F(ts^2 \mid \mathbf{c}, \mathbf{a}) \cdot \omega(\mathbf{a} \mid \mathbf{c})}{\sum_{\mathbf{a}} \mathbb{E}_F(s^2 \mid \mathbf{c}, \mathbf{a}) \cdot \omega(\mathbf{a} \mid \mathbf{c})} \\ &= \frac{\sum_{\mathbf{a}} \bar{c}_{\mathbf{a}} \omega(\mathbf{a} \mid \mathbf{c}) \mathbb{E}_F(s^2 \mid \mathbf{c}, \mathbf{a})}{\sum_{\mathbf{a}} \omega(\mathbf{a} \mid \mathbf{c}) \mathbb{E}_F(s^2 \mid \mathbf{c}, \mathbf{a})}, \end{aligned}$$

où $\omega(\mathbf{a} \mid \mathbf{c})$ est la probabilité d'observer l'arrangement \mathbf{a} , sachant la configuration \mathbf{c} , et où la somme parcourt les 2^n arrangements possibles. Nous avons

$$\begin{aligned} \omega(\mathbf{a} \mid \mathbf{c}) &= \int_{-\infty}^{+\infty} \int_0^{+\infty} \omega(\mathbf{a} \mid \mathbf{c}, s, t) ds dt \\ &= \varepsilon^m (1-\varepsilon)^{n-m} \int_{-\infty}^{+\infty} \int_0^{+\infty} s^{n-1} \frac{1}{k^m} \exp\left(-\frac{s^2}{2} \sum_{i=1}^n w_{\mathbf{a}}(i)(t+c_i)^2\right) ds dt \\ &= \left(\frac{2\pi}{n-m+m/k^2}\right)^{\frac{1}{2}} \frac{\varepsilon^m (1-\varepsilon)^{n-m}}{k^m} \underbrace{\int_0^{+\infty} s^{n-2} \exp\left(-\frac{s^2}{2} \sum_{i=1}^n w_{\mathbf{a}}(i)(c - \bar{c}_{\mathbf{a}})^2\right) ds}_{C}. \end{aligned}$$

Finalement, on montre par récurrence sur n que

$$C = \frac{1}{\left[\sum_{i=1}^n w_{\mathbf{a}}(i)(c_i - \bar{c}_{\mathbf{a}})^2\right]^{\frac{n-1}{2}}} \cdot \begin{cases} 2^{\frac{n-1}{2}} \left(\frac{n-1}{2}\right)!, & n \text{ impair}; \\ \frac{(2\pi)^{1/2} n!}{2^{\frac{n}{2}+1} \left(\frac{n}{2}\right)!}, & n \text{ pair}. \end{cases}$$

Ainsi, pour une loi normale contaminée, l'estimateur de Pitman pour le paramètre de lieu est une moyenne pondérée des configurations pondérées $\bar{c}_{\mathbf{a}} = (\sum_i w_{\mathbf{a}}(i)c_i) / \sum_i w_{\mathbf{a}}(i)$. En pratique, il n'est pas possible de parcourir les 2^n arrangements \mathbf{a} possibles, et des méthodes d'intégration numérique sont indispensables à la résolution de ce problème (voir Annexes A).

3.1.2 Estimateur de Pitman pour le paramètre d'échelle

Il est également possible de déterminer un estimateur du paramètre d'échelle selon la méthode de Pitman. Dans ce cas, la fonction de perte est souvent modifiée. En effet, la distribution des estimateurs du paramètre d'échelle est souvent asymétrique, et le carré moyen de l'erreur n'est dans ce cas pas révélateur de la qualité de l'estimateur. Ainsi, on cherche plutôt à minimiser $\mathbb{E}((\log S(\mathbf{y}) - \log \sigma)^2)$.

Supposons que l'on cherche l'estimateur optimal parmi les estimateurs équivariants pour l'échelle, c'est-à-dire satisfaisant la propriété

$$S(s(\mathbf{y} + t\mathbf{1})) = |s|S(\mathbf{y}), \quad \forall s > 0, \forall t \in \mathbb{R}.$$

En conditionnant comme dans le cas du paramètre de lieu sur la configuration \mathbf{c} , nous pouvons exprimer la perte moyenne comme

$$\begin{aligned} \mathbb{E}_F((\log S(\mathbf{y}) - \log \sigma)^2) &= \mathbb{E}_{F^*}(\mathbb{E}_F((\log S(\mathbf{y}) - \log \sigma)^2 | \mathbf{c})) \\ &= \mathbb{E}_{F^*}(\mathbb{E}_F((\log S(\mathbf{c}) + \log s)^2 | \mathbf{c})), \end{aligned}$$

où l'on a supposé sans perte de généralité que $\sigma = 1$. Comme auparavant le choix optimal de S est celui pour lequel l'espérance conditionnelle $\mathbb{E}_F((\log S(\mathbf{c}) + \log s)^2 | \mathbf{c})$ est minimale. En dérivant par rapport à S , on trouve immédiatement que

$$\log S_F(\mathbf{c}) = \mathbb{E}_F(-\log s | \mathbf{c}),$$

où de manière équivalente

$$S_F(\mathbf{c}) = \exp(\mathbb{E}_F(-\log s | \mathbf{c})),$$

et finalement

$$S_F(\mathbf{y}) = b(\mathbf{y})S_F(\mathbf{c}).$$

L'estimateur $S_F(\mathbf{y})$ est appelé l'estimateur de Pitman pour le paramètre d'échelle. Dans l'exemple ci-dessous, nous calculons cet estimateur dans le cas de la loi normale.

Exemple 3.1.13. Dans le cas où les erreurs proviennent d'une loi normale centrée et réduite, nous devons calculer dans un premier temps

$$\mathbb{E}_F(\log s | \mathbf{c}) \propto \int_{-\infty}^{+\infty} \int_0^{+\infty} \log s s^{n-1} \exp\left(\frac{-s^2}{2} \sum_{i=1}^n (t + c_i)^2\right) ds dt.$$

En suivant le même raisonnement que dans l'exemple 3.1.11, nous obtenons que

$$K_F \mathbb{E}_F(\log s | \mathbf{c}) = \left(\frac{2\pi}{n}\right)^{\frac{1}{2}} \int_0^{+\infty} \log s s^{n-1} \exp\left(\frac{-s^2}{2} \sum_{i=1}^n (c_i - \bar{c})^2\right) ds,$$

où K_F est la constante de normalisation de la densité conjointe conditionnelle de s et t sachant \mathbf{c} , donnée par

$$K_F = \left(\frac{2\pi}{n}\right)^{\frac{1}{2}} \int_0^{+\infty} s^{n-2} \exp\left(\frac{-s^2}{2} \sum_{i=1}^n (c_i - \bar{c})^2\right) ds.$$

Afin d'approximer ces intégrales, nous utilisons la fonction Gamma, définie par

$$\Gamma(x) = \int_0^{+\infty} u^{x-1} \exp(-u) du,$$

et sa dérivée

$$\Gamma'(x) = \int_0^{+\infty} \log u u^{x-1} \exp(-u) du.$$

En posant $\alpha = \sum_{i=1}^n (c_i - \bar{c})^2$, et en effectuant le changement de variables $u = \frac{\alpha s^2}{2}$, nous obtenons directement que

$$K_F \mathbb{E}_F(\log s \mid \mathbf{c}) = \frac{1}{4} \left(\frac{2}{\alpha} \right)^{\frac{n-1}{2}} \left[\log \left(\frac{2}{\alpha} \right) \Gamma \left(\frac{n-3}{2} \right) + \Gamma' \left(\frac{n-3}{2} \right) \right],$$

et

$$K_F = \frac{1}{2} \left(\frac{2}{\alpha} \right)^{\frac{n-1}{2}} \Gamma \left(\frac{n-3}{2} \right).$$

Ainsi, en utilisant le fait que $\Gamma'(n)/\Gamma(n) \approx \log n$ pour n suffisamment grand, nous obtenons finalement

$$\log S_F(\mathbf{c}) = \mathbb{E}_F(-\log s \mid \mathbf{c}) \approx -\frac{1}{2} \log \left(\frac{n-3}{\alpha} \right),$$

ou encore

$$S_F(\mathbf{c}) \approx \left(\frac{1}{n-3} \sum_{i=1}^n (c_i - \bar{c})^2 \right)^{\frac{1}{2}},$$

et

$$S_F(\mathbf{y}) \approx \left(\frac{1}{n-3} \sum_{i=1}^n (y_i - \bar{y})^2 \right)^{\frac{1}{2}}.$$

Remarque 3.1.14. Au premier abord, ce résultat peut sembler étonnant, du fait du facteur $(n-3)^{-1}$ en lieu et place du $(n-1)^{-1}$ présent dans l'estimateur habituel du paramètre d'échelle

$$s = \left(\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right)^{\frac{1}{2}}.$$

Rappelons que s est construit de telle sorte que

$$\mathbb{E}_F(s^2) = \sigma^2,$$

c'est-à-dire que le facteur $(n-1)^{-1}$ est choisi tel que s^2 soit non-biaisé. Or dans notre exemple, nous avons un critère totalement différent pour notre estimateur du paramètre

d'échelle, qui est de minimiser le carré moyen du logarithme du rapport entre l'estimateur et la vraie valeur du paramètre. C'est pourquoi on obtient alors un facteur différent.

3.1.3 Intervalles de confiance pour le paramètre de lieu

Cette idée de conditionnement et d'estimateurs équivariants peut également être appliquée à la détermination d'un intervalle de confiance pour le paramètre μ de notre modèle. Ne disposant pas de résultat général concernant la distribution de l'estimateur de Pitman, ou concernant sa variance, cette approche est donc très utile afin de pouvoir mener des tests d'hypothèses sur le paramètre de lieu.

Soit $L(\mathbf{y})$ et $U(\mathbf{y})$ deux fonctions équivariantes pour le lieu et l'échelle, c'est-à-dire telles que

$$L(a\mathbf{1} + b\mathbf{y}) = a + bL(\mathbf{y}), \quad \forall a \in \mathbb{R}, \forall b > 0.$$

Nous cherchons un intervalle $[L(\mathbf{y}), U(\mathbf{y})]$ qui couvre μ avec une probabilité donnée $1 - \alpha$, connaissant la configuration \mathbf{c} . Nous raisonnons donc à nouveau conditionnellement sur la configuration observée. En utilisant comme précédemment le système de coordonnées $y_i = s(t + c_i), t \in \mathbb{R}, s > 0$, nous cherchons donc à déterminer $L(\mathbf{y})$ et $U(\mathbf{y})$ de telle sorte que l'ensemble suivant ait une probabilité conditionnelle de $1 - \alpha$, pour $\alpha \in [0, 1]$:

$$\begin{aligned} \{(s, t) \mid \mu \in [L(\mathbf{y}), U(\mathbf{y})]\} &= \{(s, t) \mid s(t + L(\mathbf{c})) \leq \mu \leq s(t + U(\mathbf{c}))\} \\ &= \{(s, t) \mid -U(\mathbf{c}) \leq t \leq -L(\mathbf{c})\}, \end{aligned}$$

où l'on a supposé sans perte de généralité que $\mu = 0$ et où l'on a utilisé l'équivariance des bornes L et U .

La probabilité conditionnelle de se trouver dans cet ensemble est alors donnée par

$$\int_0^{+\infty} \int_{-U(\mathbf{c})}^{-L(\mathbf{c})} f(s, t \mid \mathbf{c}) dt ds.$$

Soit la fonction

$$CO_F(x \mid \mathbf{c}) = \int_0^{+\infty} \int_{-x}^{+\infty} f(s, t \mid \mathbf{c}) dt ds.$$

Cette fonction peut alors être utilisée afin de déterminer $L(\mathbf{c})$ et $U(\mathbf{c})$. Par exemple, on peut choisir

$$L(\mathbf{c}) = CO_F^{-1}(\alpha/2 \mid \mathbf{c}) \quad \text{et} \quad U(\mathbf{c}) = CO_F^{-1}(1 - \alpha/2 \mid \mathbf{c})$$

si l'on désire un intervalle bilatéral, ou encore

$$L(\mathbf{c}) = -\infty \quad \text{et} \quad U(\mathbf{c}) = CO_F^{-1}(1 - \alpha \mid \mathbf{c})$$

si l'on désire un intervalle unilatéral. Finalement, l'intervalle conditionnel cherché pour le paramètre μ est obtenu en utilisant une fois de plus l'équivariance de L et U en posant :

$$L(\mathbf{y}) = b(\mathbf{y})L(\mathbf{c}) + a(\mathbf{y}), \quad U(\mathbf{y}) = b(\mathbf{y})U(\mathbf{c}) + a(\mathbf{y}).$$

Comme pour les estimateurs de Pitman, il n'est souvent pas possible de calculer précisément la fonction CO_F , et le recours à des techniques d'intégration numériques et à l'interpolation de points est indispensable dans le cas général.

Exemple 3.1.15. Dans l'exemple 3.1.11, nous avons calculé l'estimateur de Pitman pour le paramètre de lieu dans le cas où les erreurs proviennent de la distribution normale Φ . Déterminons dans ce cas la forme de la fonction $CO_F(x \mid \mathbf{c})$. Soit d'abord la fonction

$$\begin{aligned} co_F(x \mid \mathbf{c}) &= \frac{d}{dx} CO_F(x \mid \mathbf{c}) \\ &\propto \int_0^{+\infty} f(s, -x \mid \mathbf{c}) ds, \end{aligned}$$

qui représente donc la densité conditionnelle marginale de t évaluée en $-x$.

Pour $f(s, t \mid \mathbf{c}) \propto s^{n-1} \exp\left(\frac{-s^2}{2} [\sum_{i=1}^n (c_i - \bar{c})^2 + n(t + \bar{c})^2]\right)$, nous avons que

$$CO_F(x \mid \mathbf{c}) \propto \int_{-x}^{+\infty} \left(n(t + \bar{c})^2 + \sum_{i=1}^n (c_i - \bar{c})^2 \right)^{-n/2} dt,$$

et donc

$$co_F(x \mid \mathbf{c}) \propto \left((x - \bar{c})^2 + \sum_{i=1}^n (c_i - \bar{c})^2 / n \right)^{-n/2}.$$

Comme $co_F(x \mid \mathbf{c})$ doit être une densité, nous reconnaissons alors la densité d'une loi t_ν de Student, où $\nu = n - 1$. En effet,

$$co_F(x \mid \mathbf{c}) \propto t_{n-1} \left(\frac{x - \bar{c}}{(\sum_{i=1}^n (c_i - \bar{c})^2 / n(n-1))^{1/2}} \right) \cdot \left(\sum_{i=1}^n (c_i - \bar{c})^2 / n(n-1) \right)^{-1/2},$$

puisque la densité d'une loi t_{n-1} est proportionnelle à $(1 + x^2/(n-1))^{-n/2}$. Nous pouvons ainsi conclure que l'intervalle de confiance conditionnel pour le paramètre de lieu dans le cas d'une loi normale est l'intervalle de Student habituel.

3.1.4 Remarques historiques

L'idée originale d'un conditionnement sur une statistique ancillaire est due à R. A. Fisher. Initialement, Fisher (1925) introduit la notion de statistique ancillaire afin de « récupérer » de l'information lorsqu'une statistique exhaustive n'existe pas. En effet, lorsque l'on peut déterminer une telle statistique, la variance de l'estimateur du maximum de vraisemblance est minimale, et atteint la borne de Cramer-Rao. Au contraire, lorsqu'il n'est pas possible de trouver une statistique exhaustive, la variance de l'estimateur dépasse cette borne. C'est en cherchant un compromis entre ces deux situations que Fisher eut l'idée d'introduire une statistique ancillaire. La méthode de Pitman présentée ci-dessus apparaît dans un autre travail de Fisher (1934). Pitman l'a reprise en la clarifiant et en l'appliquant à divers exemples.

Par cette idée, Fisher estimait avoir trouvé une manière d'obtenir une distribution *a posteriori* du paramètre d'intérêt, sans avoir à supposer une distribution *a priori*, au contraire des arguments bayésiens. En effet, en raisonnant conditionnellement sur une statistique ancillaire observée, et en appliquant un argument de *probabilité inverse*, Fisher (1930) obtint une distribution pour le paramètre inconnu, à l'aide des observations disponibles, qu'il nomma *distribution fiducielle* (*fiducial distribution* en anglais). Dans un travail ultérieur, Fisher (1956) insiste lourdement sur le fait que cet argument peut être utilisé sans aucune supposition *a priori* sur la distribution du paramètre.

Ci-dessus, nous avons présenté cette distribution fiducielle sous la forme

$$CO_F(x \mid \mathbf{c}) = \int_0^{+\infty} \int_{-x}^{+\infty} f(s, t \mid \mathbf{c}) dt ds$$

dans le cas du modèle $Y = \mu + \sigma E$. Ainsi, cette fonction peut être vue comme la distribution du paramètre μ , en adéquation avec les données observées, regroupées dans la statistique ancillaire \mathbf{c} . En effet, cette fonction peut par exemple être utilisée afin de déterminer des intervalles de confiance pour le paramètre. Mais il est également envisageable de définir un estimateur T de μ en utilisant cette loi, en choisissant par exemple T de telle sorte que

$$CO_F(T \mid \mathbf{c}) = \frac{1}{2}.$$

T pourrait donc être choisi comme la médiane de la loi.

Néanmoins, cette approche va être très critiquée, notamment suite à la polémique suivante. Fieller (1954) et Creasy (1954) présentent deux solutions différentes pour la construction d'un intervalle de confiance, basée sur l'argument de la distribution fiducielle, du rapport des moyennes de deux lois normales indépendantes de variances connues. La solution de Fieller est antérieure et a été approuvée par Fisher comme étant *la* solution fiducielle. Mais Creasy présente une autre solution, construite dans le même cadre que le problème de Behrens-Fisher, à savoir la distribution de la différence des deux moyennes de deux lois normales indépendantes avec variances inconnues (voir

Kim et Cohen, 1998). Ce problème montre donc que la distribution fiducielle d'un paramètre n'est pas unique.

3.1.5 Comportement asymptotique des estimateurs de Pitman

Dans cette section, nous démontrons que les estimateurs de Pitman sont asymptotiquement équivalents aux estimateurs du maximum de vraisemblance de μ et σ (voir Freedman, 1963, Hartigan, 1965 ou Morgenthaler, 1986). Pour ce faire, nous appliquons l'approximation de Laplace aux intégrales doubles utilisées pour la définition des estimateurs $T_F(\mathbf{y})$ et $S_F(\mathbf{y})$. Pour d'avantage de détails concernant la méthode de Laplace, voir de Bruijn (1981).

Définition 3.1.16. Considérons le modèle simple $Y = \mu + \sigma E$, où $E \sim F$, et soit $\mathbf{y} = (y_1, \dots, y_n)$ un échantillon indépendant de Y . La vraisemblance est donnée par

$$L(v, u | \mathbf{y}) = \left[\prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i - v}{u}\right) \right],$$

et la log-vraisemblance est

$$l(v, u | \mathbf{y}) = \log L(v, u | \mathbf{y}) = \sum_{i=1}^n \log \left(\frac{1}{u} f\left(\frac{y_i - v}{u}\right) \right).$$

On note également $\bar{l}(v, u | \mathbf{y}) = n^{-1}l(v, u | \mathbf{y})$.

Soit $\boldsymbol{\theta} = (\mu, \sigma)$. On note alors $\hat{\boldsymbol{\theta}}_F^n(\mathbf{y}) = \hat{\boldsymbol{\theta}}_F^n = (\hat{\mu}_F^n, \hat{\sigma}_F^n)$ l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ pour le modèle de lieu et d'échelle où les erreurs proviennent de la distribution F , et basé sur l'échantillon \mathbf{y} de taille n . Cet estimateur est défini par

$$\hat{\boldsymbol{\theta}}_F^n = \arg \max_{v \in \mathbb{R}, u > 0} l(v, u | \mathbf{y}).$$

Les deux conditions suivantes sont également nécessaires.

Définition 3.1.17. En reprenant les notations de la définition 3.1.16, notons également

$$\hat{m} = \bar{l}(\hat{\mu}_F^n, \hat{\sigma}_F^n | \mathbf{y})$$

le maximum de la log-vraisemblance, et

$$h(v, u | \mathbf{y}) = \bar{l}(v + \hat{\mu}_F^n, u + \hat{\sigma}_F^n | \mathbf{y}) - \hat{m}.$$

On dit que la distribution F satisfait la condition 1 si pour n suffisamment grand, la fonction h est strictement négative en dehors de n'importe quel voisinage de l'origine.

On dit également que la distribution F satisfait la condition 2 si pour tout voisinage de l'origine suffisamment petit, il doit être possible, pour n suffisamment grand, d'approximer h par son développement de Taylor d'ordre deux autour de l'origine, avec une erreur relative uniforme.

Proposition 3.1.18. *Soit $T_F^n(\mathbf{y})$ et $S_F^n(\mathbf{y})$ les estimateurs de Pitman pour les paramètres de lieu et d'échelle associés à un échantillon \mathbf{y} de taille n . Soit $\hat{\mu}_F^n$ et $\hat{\sigma}_F^n$ les estimateurs habituels du maximum de vraisemblance de μ et σ dans le modèle $Y = \mu + \sigma E$, où $E \sim F$. Supposons que F satisfait les conditions 1 et 2. Alors, pour $n \rightarrow +\infty$,*

$$T_F^n(\mathbf{y}) - \hat{\mu}_F^n \rightarrow 0 \quad \text{et} \quad S_F^n(\mathbf{y}) - \hat{\sigma}_F^n \rightarrow 0.$$

Preuve Rappelons que

$$T_F^n(\mathbf{y}) = b(\mathbf{y})T_F^n(\mathbf{c}) + a(\mathbf{y}),$$

où

$$T_F^n(\mathbf{c}) = -\frac{\mathbb{E}_F(s^2 t \mid \mathbf{c})}{\mathbb{E}_F(s^2 \mid \mathbf{c})},$$

et $\mathbf{c} = b(\mathbf{y})^{-1}(\mathbf{y} - a(\mathbf{y}))$. Notons pour simplifier $a = a(\mathbf{y})$ et $b = b(\mathbf{y})$ et considérons tout d'abord $\mathbb{E}_F(s^2 t \mid \mathbf{c})$. Nous avons

$$\begin{aligned} \mathbb{E}_F(s^2 t \mid \mathbf{c}) &\propto \int_0^{+\infty} \int_{-\infty}^{+\infty} s^2 t s^{n-1} \prod_{i=1}^n f(s(t + c_i)) dt ds \\ &= \int_0^{+\infty} \int_{-\infty}^{+\infty} s^2 t s^{n-1} \prod_{i=1}^n f\left(s\left(t + \frac{y_i - a}{b}\right)\right) dt ds. \end{aligned}$$

En effectuant le changement de variables $u = b/s$ et $v = a - tb$, nous obtenons

$$\begin{aligned} \mathbb{E}_F(s^2 t \mid \mathbf{c}) &\propto -b^n \int_0^{+\infty} \int_{-\infty}^{+\infty} v u^{-3} \prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i - v}{u}\right) dv du \\ &\quad + ab^n \int_0^{+\infty} \int_{-\infty}^{+\infty} u^{-3} \prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i - v}{u}\right) dv du. \end{aligned}$$

En appliquant le même raisonnement à $\mathbb{E}_F(s^2 \mid \mathbf{c})$, nous pouvons alors exprimer $T_F(\mathbf{c})$ comme

$$T_F^n(\mathbf{c}) = \frac{1}{b} \frac{\int_0^{+\infty} \int_{-\infty}^{+\infty} v u^{-3} \prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i - v}{u}\right) dv du}{\int_0^{+\infty} \int_{-\infty}^{+\infty} u^{-3} \prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i - v}{u}\right) dv du} - \frac{a}{b},$$

et donc

$$T_F^n(\mathbf{y}) = \frac{\int_0^{+\infty} \int_{-\infty}^{+\infty} vu^{-3} \prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i-v}{u}\right) dvdu}{\int_0^{+\infty} \int_{-\infty}^{+\infty} u^{-3} \prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i-v}{u}\right) dvdu}.$$

On reconnaît alors l'expression de la log-vraisemblance définie plus haut, et nous pouvons exprimer $T_F^n(\mathbf{y})$ comme

$$T_F^n(\mathbf{y}) = \frac{\int_0^{+\infty} \int_{-\infty}^{+\infty} vu^{-3} \exp(n\bar{l}(v, u | \mathbf{y})) dvdu}{\int_0^{+\infty} \int_{-\infty}^{+\infty} u^{-3} \exp(n\bar{l}(v, u | \mathbf{y})) dvdu}.$$

Afin d'approximer les intégrales du type $I(g) = \int_0^{+\infty} \int_{-\infty}^{+\infty} g(v, u) \exp(n\bar{l}(v, u | \mathbf{y})) dvdu$, nous appliquons à présent la méthode de Laplace. Grâce à un simple changement de variables, $I(g)$ peut alors être exprimée comme

$$I(g) = \exp(n\hat{m}) \int_{-\hat{\sigma}_F^n}^{+\infty} \int_{-\infty}^{+\infty} g(v + \hat{\mu}_F^n, u + \hat{\sigma}_F^n) \frac{1}{(u + \hat{\sigma}_F^n)^3} \exp(nh(v, u | \mathbf{y})) dvdu,$$

Le maximum de la fonction h est atteint à l'origine $(0, 0)$, et h est strictement négative partout ailleurs. La méthode de Laplace va nous permettre d'approximer $I(g)$, en raisonnant de la manière suivante :

1. Puisque F satisfait la condition 1, la fonction $\exp(nh(v, u | \mathbf{y}))$ devient si rapidement concentrée autour de l'origine que sa contribution à l'intégrale en dehors de n'importe quel voisinage de l'origine devient asymptotiquement négligeable. Ainsi, le domaine d'intégration peut être restreint à un voisinage infinitésimal de $(0, 0)$, et seules les propriétés infinitésimales des fonctions g et h sont asymptotiquement importantes.
2. Puisque F satisfait également la condition 2, nous pouvons remplacer h par son développement de Taylor et g par sa valeur prise en $(0, 0)$. Dès lors, le domaine d'intégration peut à nouveau être étendu à \mathbb{R}^2 tout entier, pour la même raison que nous avons pu le tronquer précédemment.

On peut alors exprimer $I(g)$ comme

$$I(g) = \exp(n\hat{m}) g(\hat{\mu}_F^n, \hat{\sigma}_F^n) \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left(-\frac{n}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_F^n) \frac{-\partial^2 h(0, 0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_F^n)^T\right) d\boldsymbol{\theta} + O(n^{-1}).$$

On reconnaît alors la densité d'une loi normale bivariée, et il en résulte l'approximation suivante :

$$I(g) = \frac{2\pi g(\hat{\mu}_F^n, \hat{\sigma}_F^n) \exp(n\hat{m})}{n |\det(H)|^{\frac{1}{2}}} + O(n^{-1}),$$

où

$$H = \frac{-\partial^2 h(0, 0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

En appliquant ces approximations à l'expression de $T_F^n(\mathbf{y})$ ci-dessus, on obtient finalement

$$T_F^n(\mathbf{y}) = \frac{\hat{\mu}_F^n (\hat{\sigma}_F^n)^{-3}}{(\hat{\sigma}_F^n)^{-3}} + O(n^{-1}) = \hat{\mu}_F^n + O(n^{-1}).$$

Concernant le paramètre d'échelle, la démarche est la même. En effectuant le même changement de variables dans les intégrales définissant $S_F^n(\mathbf{c})$, on obtient

$$\log S_F^n(\mathbf{y}) = -\log b + \frac{\int_0^{+\infty} \int_{-\infty}^{+\infty} \log u \, u^{-1} \prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i - v}{u}\right) \, dv du}{\int_0^{+\infty} \int_{-\infty}^{+\infty} u^{-1} \prod_{i=1}^n \frac{1}{u} f\left(\frac{y_i - v}{u}\right) \, dv du},$$

et donc

$$S_F^n(\mathbf{y}) = \exp\left(\frac{\int_0^{+\infty} \int_{-\infty}^{+\infty} \log u \, u^{-1} \exp(n\bar{l}(v, u)) \, dv du}{\int_0^{+\infty} \int_{-\infty}^{+\infty} u^{-1} \exp(n\bar{l}(v, u)) \, dv du}\right).$$

En appliquant l'approximation de Laplace on obtient alors

$$S_F^n(\mathbf{y}) = \hat{\sigma}_F^n + O(n^{-1}).$$

□

Remarque 3.1.19. Easton (1985, 1986) a montré que les conditions 1 et 2 étaient vérifiées aussi bien par des distributions F à queues lourdes que par la loi normale, et ce pour une grande variété de distributions des erreurs.

Remarque 3.1.20. L'ordre de convergence de la méthode de Laplace est de $O(n^{-1})$. Ainsi, lorsque l'on effectue le rapport des deux intégrales définissant les estimateurs de Pitman, on obtient également un ordre de convergence de $O(n^{-1})$. Néanmoins, Tierney et Kadane (1986) ont montré qu'en utilisant un développement de Taylor d'ordre 3, il était possible d'obtenir un ordre de convergence de $O(n^{-2})$ pour un rapport de deux intégrales.

3.2 Estimateurs de Pitman compromis

Les estimateurs de Pitman ne peuvent être appliqués que lorsque la distribution des erreurs est connue. Dans le cas contraire, cette méthode n'est pas envisageable. Néanmoins, il est possible de construire des estimateurs optimaux (selon certains critères) simultanément pour plusieurs distributions sous-jacentes précisées à l'avance. L'idée que quelques distributions bien choisies peuvent résumer plusieurs aspects de toutes les distributions sous-jacentes possibles est au centre de la construction de tels estimateurs.

Pregibon et Tukey (1981), Tukey (1981, 1987) et Morgenthaler et Tukey (1991) ont envisagé plusieurs critères d'optimalité que l'on peut utiliser afin de construire des estimateurs, lorsque cette notion d'optimalité se rapporte à un ensemble de modèles possibles. Le résultat obtenu est alors un compromis entre les estimateurs de Pitman correspondant à chaque modèle, c'est-à-dire une moyenne pondérée des différents estimateurs de Pitman.

Easton (1991) a appliqué l'approximation de Laplace aux intégrales définissant les estimateurs de Pitman compromis, et en a obtenu les estimateurs du maximum de vraisemblance compromis. Ces estimateurs sont plus aisés à obtenir du point de vue calculatoire. De plus, les poids associés à chacune des distributions dépendent uniquement de la vraisemblance des résidus. Ces estimateurs donnent de meilleurs résultats que les M-estimateurs usuels, ou que les moyennes tronquées.

3.2.1 Estimateurs bi-optimaux

Définition 3.2.1. Soit $T = T(\mathbf{y})$ un estimateur équivariant du lieu, et soit $T_F = T_F(\mathbf{y})$ l'estimateur de Pitman du lieu, associé à la distribution F . L'efficacité de T , lorsque la vraie distribution sous-jacente est F , est définie par

$$\text{EFF}_F(T) = \frac{\text{Var}_F(T_F)}{\text{Var}_F(T)}.$$

Par efficacité, nous entendons donc l'efficacité dans les échantillons finis, et uniquement parmi les estimateurs équariants pour le lieu et l'échelle. Il ne s'agit donc pas de l'efficacité asymptotique, ou de l'efficacité par rapport à la borne inférieure de Cramer-Rao.

Définition 3.2.2. Soit $T = T(\mathbf{y})$ un estimateur équivariant du lieu, et soit $T_F = T_F(\mathbf{y})$ l'estimateur de Pitman du lieu, associé à la distribution F . L'excès relatif de variance (*relative excess variance* en anglais) de l'estimateur T , lorsque F est la distribution sous-jacente, est définie par

$$\text{REV}_F(T) = \frac{\text{Var}_F(T) - \text{Var}_F(T_F)}{\text{Var}_F(T_F)}.$$

Tant pour l'efficacité que pour l'excès relatif de variance, lorsque la distribution sous-jacente des erreurs est F , l'estimateur optimal est l'estimateur de Pitman T_F . Nous avons $\text{EFF}_F(T_F) = 1$ et $\text{REV}_F(T_F) = 0$.

Remarque 3.2.3. L'excès relatif de variance et l'efficacité sont deux mesures d'optimalité d'un estimateur totalement équivalentes, puisque l'on a

$$\text{REV}_F(T) = \frac{1}{\text{EFF}_F(T)} - 1.$$

Pour ce qui suit, il est néanmoins plus aisé de travailler avec l'excès relatif de variance.

Soient F_1 et F_2 deux distributions de lieu et d'échelle. Nous sommes intéressés par construire un estimateur T qui se comporte bien dans les deux cas. La figure 3.1 présente schématiquement les zones atteignables et inatteignables en termes d'excès relatifs de variance pour un estimateur T . L'intersection entre la limite de ces deux régions et la droite à 45 degrés représente un estimateur « bi-optimal » pour les deux distributions F_1 et F_2 envisagées, car il minimise simultanément les deux excès relatifs de variance.

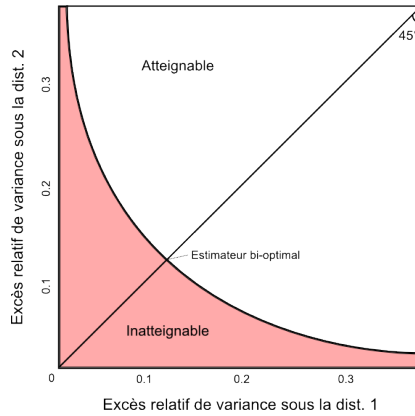


FIG. 3.1: Région atteignable pour les excès relatifs de variance de l'estimateur T

Formellement, l'estimateur bi-optimal T_{bo} est obtenu de la manière suivante : pour $\theta \in [0, 1]$, $T_{bo} = T_{bo}(\mathbf{y}, \theta)$ est tel que la quantité

$$\theta \text{REV}_{F_1}(T) + (1 - \theta) \text{REV}_{F_2}(T)$$

est minimale. En développant les variances des estimateurs comme les espérances des carrés moyens de l'erreur conditionnels, et en faisant apparaître f_1^* et f_2^* , les densités marginales de la configuration \mathbf{c} , on trouve finalement que l'estimateur bi-optimal T_{bo} est une combinaison convexe des deux estimateurs de Pitman associés aux distributions F_1 et F_2 .

Définition 3.2.4. Pour F_1 et F_2 deux distributions, et $\theta \in [0, 1]$, l'estimateur bi-optimal est défini par

$$T_{bo}(\mathbf{y}, \theta) = \frac{\theta \alpha_1(\mathbf{y}) T_{F_1}(\mathbf{y}) + (1 - \theta) \alpha_2(\mathbf{y}) T_{F_2}(\mathbf{y})}{\theta \alpha_1(\mathbf{y}) + (1 - \theta) \alpha_2(\mathbf{y})},$$

où

$$\alpha_i(\mathbf{y}) = \frac{\mathbb{E}_{F_i}(s^2 \mid \mathbf{c}) f_i^*(\mathbf{c})}{\text{Var}_{F_i}(T_{F_i}(\mathbf{y}))}, \quad i = 1, 2.$$

Ainsi, l'estimateur bi-optimal est une moyenne pondérée des estimateurs de Pitman T_{F_1} et T_{F_2} , dans laquelle les poids associés à chacun sont fonction à la fois de θ et de l'échantillon \mathbf{y} .

Deux autres approches sont possibles et donnent également des compromis entre deux estimateurs de Pitman.

Définition 3.2.5. Pour F_1 et F_2 deux distributions, l'estimateur bi-efficient minimise la quantité

$$\max_{i=1,2} \left\{ \frac{\mathbb{E}_{F_i}(s^2 \mid \mathbf{c})(T(\mathbf{y}) - T_{F_i}(\mathbf{y}))^2 f_i^*(\mathbf{c})}{\text{Var}_{F_i}(T_{F_i}(\mathbf{y}))} \right\},$$

et est donné par

$$T_{be} = T_{be}(\mathbf{y}) = \frac{\alpha_1(\mathbf{y})T_{F_1}(\mathbf{y}) + \alpha_2(\mathbf{y})T_{F_2}(\mathbf{y})}{\alpha_1(\mathbf{y}) + \alpha_2(\mathbf{y})},$$

où

$$\alpha_i(\mathbf{y}) = \left[\frac{\mathbb{E}_{F_i}(s^2 \mid \mathbf{c})f_i^*(\mathbf{c})}{\text{Var}_{F_i}(T_{F_i}(\mathbf{y}))} \right]^{\frac{1}{2}}, \quad i = 1, 2.$$

Définition 3.2.6. Pour F_1 et F_2 deux distributions, l'estimateur co-efficient minimise la quantité

$$\max_{i=1,2} \left\{ \frac{\mathbb{E}_{F_i}(s^2 \mid \mathbf{c})(T(\mathbf{y}) - T_{F_i}(\mathbf{y}))^2 f_i^*(\mathbf{c})}{\text{cCME}_{F_i}(T_{F_i}(\mathbf{y}))} \right\},$$

où $\text{cCME}_{F_i}(T)$ est le carré moyen de l'estimateur T conditionnel sur la configuration \mathbf{c} , et est donné par

$$T_{ce} = T_{ce}(\mathbf{y}) = \frac{\alpha_1(\mathbf{y})T_{F_1}(\mathbf{y}) + \alpha_2(\mathbf{y})T_{F_2}(\mathbf{y})}{\alpha_1(\mathbf{y}) + \alpha_2(\mathbf{y})},$$

où

$$\alpha_i(\mathbf{y}) = \left[\frac{\mathbb{E}_{F_i}(s^2 \mid \mathbf{c})f_i^*(\mathbf{c})}{\text{cCME}_{F_i}(T_{F_i}(\mathbf{y}))} \right]^{\frac{1}{2}}, \quad i = 1, 2.$$

L'estimateur co-efficient possède l'avantage que les variances des estimateurs de Pitman ne sont pas nécessaires pour son calcul, au contraire de l'estimateur bi-efficient. En effet, dans la plupart des cas, cette variance n'est pas directement calculable et doit être approximée par des simulations numériques.

3.2.2 Estimateurs du maximum de vraisemblance compromis

Le défaut principal des estimateurs de Pitman réside certainement dans le fait qu'il est souvent nécessaire de recourir à des méthodes d'intégration numérique afin de les déterminer. De même, pour les estimateurs compromis définis ci-dessus, dans lesquels il faut également approximer la variance des estimateurs de Pitman. Afin de contourner ce problème, Easton (1991) a appliqué la méthode d'approximation de Laplace aux intégrales définissant les estimateurs bi-optimaux, aussi bien pour les estimateurs de Pitman eux-mêmes que pour les différentes composantes des poids. La procédure est tout à fait similaire à celle présentée dans la preuve de la proposition 3.1.18. Easton obtient alors ce qu'il nomme les estimateurs du maximum de vraisemblance compromis, qui approximent les estimateurs bi-optimaux.

Exemple 3.2.7. Considérons l'estimateur bi-optimal $T_{bo}(\mathbf{y}, \theta)$. On a montré que T_{F_i} pouvait être approximé par $\hat{\mu}_{F_i}$, l'estimateur du maximum de vraisemblance de μ sous la distribution F_i . En reprenant les mêmes notations que dans la preuve de la proposition 3.1.18, les poids $\alpha_i(\mathbf{y})$ peuvent être approximatés par

$$\alpha_i(\mathbf{y}) \approx \frac{2\pi \exp(n\hat{m})}{n\hat{\sigma}_{F_i}^3 |\det(H_i)|^{\frac{1}{2}} \text{Var}_{F_i}(T_{F_i}(\mathbf{y}))}, \quad i = 1, 2,$$

où H_i est la matrice des dérivées deuxièmes de la log-vraisemblance associée à la distribution F_i . En simplifiant les constantes, on obtient alors que

$$T_{bo}(\mathbf{y}, \theta) \approx \frac{\theta \tilde{\alpha}_1(\mathbf{y}) \hat{\mu}_{F_1} + (1 - \theta) \tilde{\alpha}_2(\mathbf{y}) \hat{\mu}_{F_2}}{\theta \tilde{\alpha}_1(\mathbf{y}) + (1 - \theta) \tilde{\alpha}_2(\mathbf{y})},$$

où

$$\tilde{\alpha}_i(\mathbf{y}) \approx \frac{\exp(n\hat{m})}{\hat{\sigma}_{F_i}^3 |\det(H_i)|^{\frac{1}{2}} \text{Var}_{F_i}(T_{F_i}(\mathbf{y}))}, \quad i = 1, 2.$$

Les mêmes approximations peuvent être faites pour les estimateurs bi-efficient et co-efficient. Pour les estimateurs du maximum de vraisemblance compromis, l'avantage réside dans le fait que leur calcul, y compris l'évaluation des poids, dépend uniquement de la vraisemblance, plus aucune intégration n'étant nécessaire. En effet, nous pouvons également remplacer le terme $\text{Var}_{F_i}(T_{F_i}(\mathbf{y}))$ par le premier terme de la diagonale de H_i^{-1} , une estimation de la variance de l'estimateur du maximum de vraisemblance $\hat{\mu}_{F_i}$. Remarquons pour finir que ces constructions sont généralisables au cas de plus que deux densités.

3.2.3 Estimateur de Pitman compromis pour le paramètre de lieu

Lorsque la distribution des erreurs est inconnue, il est impossible d'utiliser immédiatement les estimateurs de Pitman. Néanmoins, l'idée d'un estimateur optimal simultanément pour plusieurs distributions est attrayante, et un compromis d'estimateurs est une solution envisageable.

Dans les constructions présentées ci-dessus, les distributions F_1 et F_2 sont souvent fixées dès le départ, comme étant la loi normale pour l'une, et la loi Slash pour la seconde, par exemple. Le fait de combiner une loi à queues courtes avec une loi à queues très lourdes doit permettre d'obtenir un comportement raisonnable dans la plupart des cas.

Dans ce qui suit, nous présentons la construction d'un estimateur basé sur le compromis d'estimateurs de Pitman, tout en simplifiant les poids associés aux distributions, et en insistant sur le choix des distributions de compromis.

Nous commençons par donner quelques notations et définitions qui seront utilisées pour le reste du travail.

Notation 3.2.8. Nous notons à présent G la distribution effective des erreurs dans le modèle $Y = \mu + \sigma E$. C'est-à-dire, dans ce modèle nous avons $E \sim G$. Cette distribution est dès à présent considérée comme inconnue.

Au lieu de supposer une certaine distribution pour G , nous allons plutôt supposer que cette distribution inconnue appartient à un ensemble de distributions, noté \mathcal{F} . Cet ensemble, choisi arbitrairement grand, reflète tous les modèles plausibles pour l'expérimentateur.

L'idée est alors de sélectionner quelques éléments $F_1, \dots, F_m \in \mathcal{F}$, et de construire un estimateur de Pitman compromis sur cette base.

Définition 3.2.9. L'estimateur de Pitman compromis pour le paramètre de lieu μ , basé sur les distributions $F_1, \dots, F_m \in \mathcal{F}$ est défini par

$$T(\mathbf{y}) = \frac{\sum_{k=1}^m w(F_k) T_{F_k}(\mathbf{y})}{\sum_{k=1}^m w(F_k)},$$

où $T_{F_k}(\mathbf{y})$ est l'estimateur de Pitman pour le paramètre de lieu associé à la distribution F_k , et où $w(F_k)$ est une fonction de poids non-négative.

Remarque 3.2.10. Nous avons ici supposé que les distributions de compromis contenues dans l'ensemble \mathcal{F} sont telles que le paramètre de lieu μ a la même signification pour chacune d'entre elles. Par exemple, nous pouvons supposer que ces distributions sont symétriques autour de 0.

Notre estimateur de Pitman compromis est donc une moyenne pondérée d'estimateurs de Pitman associés aux distributions F_1, \dots, F_m . Le choix de ces distributions reste encore à discuter, tout comme la fonction de poids, $w(F_k)$, qui devra représenter la qualité du choix de F_k comme modèle. Plus la distribution de compromis F_k sera bien adaptée aux données à disposition, plus elle recevra un poids élevé.

Définition 3.2.11. Pour F_k une distribution et \mathbf{y} un échantillon de taille n , la vraisemblance profil du modèle F_k est donnée par

$$L(F_k) = \prod_{i=1}^n \frac{1}{S_{F_k}(\mathbf{y})} f_k \left(\frac{y_i - T_{F_k}(\mathbf{y})}{S_{F_k}(\mathbf{y})} \right),$$

où $S_{F_k}(\mathbf{y})$ est l'estimateur de Pitman pour le paramètre d'échelle associé à la distribution F_k .

La vraisemblance profil du modèle F_k peut être vue comme la vraisemblance des résidus, après ajustement des estimateurs de Pitman associés à F_k pour les paramètres de lieu et d'échelle. Dès lors, cette vraisemblance est une mesure de crédibilité de F_k , étant données les observations. Plus la vraisemblance est grande, plus il semble raisonnable de supposer que F_k est proche de la vraie distribution G des erreurs.

Remarque 3.2.12. Il est également possible de voir $L(F)$ comme une approximation de $f^*(\mathbf{c})$, la densité marginale de la configuration \mathbf{c} , définie dans la remarque 3.1.10. En effet, en appliquant la méthode de Laplace à la double intégrale, en développant la fonction autour des estimateurs de Pitman (qui sont proches des estimateurs du maximum de vraisemblance), on obtient la vraisemblance profil du modèle F .

Ainsi, il nous semble approprié d'utiliser cette vraisemblance profil dans notre fonction de poids. Cette dernière pourra ainsi être une fonction croissante de la vraisemblance profil. Plus simplement, nous posons pour la suite

$$w(F_k) = L(F_k).$$

Ce choix très simple s'avèrera être crucial dans le comportement asymptotique de notre estimateur compromis, comme nous allons le voir par la suite. La question du choix des distributions de compromis sera également étudiée par le biais de ce comportement asymptotique.

Définition 3.2.13. Soit $D(G||F) = D(G(y)||F(y))$ la distance de Kullback-Leibler entre G et F . Nous définissons la distance de forme entre G et F par

$$sD(G||F) = \min_{s>0} D(G(y)||F(y/s)).$$

Cette distance est donc très similaire à la distance de Kullback-Leibler, mais nous permettons ici une remise à l'échelle de la distribution F , de sorte qu'elle soit le plus proche possible de G .

Remarque 3.2.14. Si nous ne désirons pas faire l'hypothèse que les distributions de compromis sont toutes symétriques autour de 0, il est nécessaire d'introduire dans la définition de la distance en forme une minimisation concernant également le paramètre de lieu.

A l'aide de cette définition, il nous est dès lors possible de déterminer le comportement asymptotique de l'estimateur de Pitman compromis pour le paramètre de lieu.

Proposition 3.2.15. *Soit $T(\mathbf{y})$ l'estimateur de Pitman compromis pour le paramètre de lieu, basé sur F_1, \dots, F_m , où la fonction de poids est donnée par $w(F_k) = L(F_k)$. Alors, lorsque $n \rightarrow \infty$,*

$$T(\mathbf{y}) \rightarrow T_{F_k}(\mathbf{y}),$$

où $k \in \{1, \dots, m\}$ est tel que $sD(G||F_k)$ est minimale, ou encore

$$F_k = \arg \min_{F \in \{F_1, \dots, F_m\}} sD(G||F).$$

Preuve Il convient donc d'étudier le comportement asymptotique de la fonction de poids, et donc plus particulièrement de la vraisemblance profil $L(F_k)$. Soit

$$l(F_k) = \log L(F_k) = \sum_{i=1}^n \log \left[\frac{1}{S_{F_k}(\mathbf{y})} f_k \left(\frac{y_i - T_{F_k}(\mathbf{y})}{S_{F_k}(\mathbf{y})} \right) \right],$$

et rappelons que les estimateurs de Pitman $T_{F_k}(\mathbf{y})$ et $S_{F_k}(\mathbf{y})$ sont proches des estimateurs du maximum de vraisemblance $\hat{\mu}_{F_k}$ et $\hat{\sigma}_{F_k}$, et convergent vers μ_{F_k} et σ_{F_k} . Rappelons encore que G est la vraie distribution sous-jacente des erreurs. Ainsi :

$$\begin{aligned} \frac{1}{n} l(F_k) &\rightarrow \int \log f_k \left(\frac{y - \mu_{F_k}}{\sigma_{F_k}} \right) \frac{1}{\sigma_{F_k}} dG \left(\frac{y - \mu}{\sigma} \right) \\ &= \int \log g(v) dG(v) - D \left(G \left(\frac{y - \mu}{\sigma} \right) \middle| \middle| F_k \left(\frac{y - \mu_{F_k}}{\sigma_{F_k}} \right) \right), \end{aligned}$$

lorsque $n \rightarrow \infty$. En supposant que le paramètre μ a la même signification pour les distributions F_k et G (par exemple que les deux distributions sont symétriques autour de l'origine), nous pouvons alors poser $\mu = \hat{\mu}_{F_k} = 0$ sans perte de généralité. Avec un changement de variables, et comme $l(F_k)$ atteint son maximum pour $\hat{\sigma}_{F_k}$, nous avons

$$\frac{1}{n} l(F_k) \rightarrow c(G) - D(G(y)||F_k(y/\sigma^*(F_k))),$$

lorsque $n \rightarrow \infty$, où $c(G)$ est une constante ne dépendant que de la distribution sous-jacente G , et où $\sigma^*(F_k)$ est un paramètre d'échelle, dépendant de la distribution F_k , et tel que $D(G(y)||F(y/\sigma^*(F_k)))$ est minimale. Comme nous n'avons supposé aucune signification particulière pour le paramètre d'échelle σ dans notre modèle, une telle remise à l'échelle est nécessaire et introduit donc la notion de distance en forme. Ainsi,

$$\frac{1}{n}l(F_k) \rightarrow c(G) - sD(G||F_k),$$

lorsque $n \rightarrow \infty$. Pour n grand, on a

$$L(F_k) \sim \exp(nc(G) - nsD(G||F_k)),$$

où le signe \sim signifie ici que les deux quantités sont asymptotiquement équivalentes.

Comme $c(G)$ est la même constante pour chaque distribution de compromis $F_k, k = 1, \dots, m$, et comme $\exp(-nsD(G||F_k)) \rightarrow 0 \forall k = 1, \dots, m$, lorsque $n \rightarrow \infty$, il en découle que

$$\frac{w(F_k)}{\sum_{j=1}^m w(F_j)} \rightarrow \begin{cases} 1, & \text{si } sD(G||F_k) \leq sD(G||F_j), \forall j = 1, \dots, m; \\ 0, & \text{sinon,} \end{cases}$$

lorsque $n \rightarrow \infty$. Ainsi, l'estimateur de Pitman compromis pour le paramètre de lieu converge vers l'estimateur de Pitman associé à la distribution de compromis F_k telle que F_k est la plus proche de G au sens de la distance en forme. □

3.2.4 Estimateur de Pitman compromis pour le paramètre d'échelle

Afin d'appliquer la même démarche à la construction d'un estimateur du paramètre d'échelle σ , basé sur les estimateurs de Pitman, nous devons poser une condition supplémentaire sur les distributions de compromis contenues dans l'ensemble \mathcal{F} . En effet, comme nous l'avons déjà vu dans la preuve de la proposition 3.2.15, le paramètre d'échelle σ ne possède pas forcément la même signification pour chacune des distributions sous-jacentes prises en considération. Au contraire du paramètre de lieu, qui représente le centre de symétrie de chaque distribution, la dispersion de ces dernières est différente. Si l'on ne pose aucune condition supplémentaire sur les distributions de compromis, une estimation de σ comme nous l'avons construite pour μ n'aura alors aucun sens.

Afin de résoudre ce problème, il est nécessaire de standardiser les distributions de l'ensemble \mathcal{F} . Par exemple, nous pouvons exiger que chaque distribution de compromis vérifie la propriété

$$\text{DIQ}(F_j) = 1, \quad \forall F_j \in \mathcal{F},$$

où $\text{DIQ}(F) = q_F(0.75) - q_F(0.25)$ représente la distance interquartile de F . En imposant cette condition, σ possède alors la même signification pour chacune des distributions

sous-jacentes envisagées. Plus précisément, σ représente la distance interquartile de l'échantillon observé, puisque

$$\text{DIQ}(F(\cdot/\sigma)) = \sigma \text{DIQ}(F) = \sigma.$$

Nous pouvons alors définir l'estimateur de Pitman compromis pour le paramètre d'échelle de la même manière que précédemment.

Définition 3.2.16. L'estimateur de Pitman compromis pour le paramètre d'échelle σ , basé sur les distributions $F_1, \dots, F_m \in \mathcal{F}$ telles que $\text{DIQ}(F_j) = 1, \forall F_j \in \mathcal{F}$, est défini par

$$S(\mathbf{y}) = \frac{\sum_{k=1}^m w(F_k) S_{F_k}(\mathbf{y})}{\sum_{k=1}^m w(F_k)},$$

où $S_{F_k}(\mathbf{y})$ est l'estimateur de Pitman pour le paramètre de lieu associé à la distribution F_k , et où $w(F_k)$ est une fonction de poids non-négative.

Remarque 3.2.17. Pour le reste de ce chapitre, nous supposons que les distributions de compromis sont standardisées de la manière présentée ci-dessus, sauf mention explicite du contraire.

Cette condition supplémentaire nous permet de formuler immédiatement le corollaire à la proposition 3.2.15 suivant.

Corollaire 3.2.18. Soit $T(\mathbf{y})$ et $S(\mathbf{y})$ les estimateurs de Pitman compromis pour les paramètres de lieu et d'échelle, basés sur F_1, \dots, F_m , où la fonction de poids est donnée par $w(F_k) = L(F_k)$. Alors, lorsque $n \rightarrow \infty$,

$$T(\mathbf{y}) \rightarrow T_{F_k}(\mathbf{y}), \quad S(\mathbf{y}) \rightarrow S_{F_k}(\mathbf{y}),$$

où $k \in \{1, \dots, m\}$ est tel que $D(G||F_k)$ est minimale, ou encore

$$F_k = \arg \min_{F \in \{F_1, \dots, F_m\}} D(G||F).$$

Preuve La preuve est similaire à celle de la proposition 3.2.15, mais cette fois-ci, nous pouvons poser sans perte de généralité que

$$\sigma = \hat{\sigma}_{F_k} = 1,$$

puisque le paramètre d'échelle a la même signification pour toutes les distributions. Ainsi, la log-vraisemblance devient

$$\frac{1}{n} l(F_k) \rightarrow c(G) - D(G||F_k)$$

lorsque $n \rightarrow \infty$, où $c(G) = \int \log g(v) dG(v)$ ne dépend que de la distribution sous-jacente des erreurs. Ainsi,

$$\frac{w(F_k)}{\sum_{j=1}^m w(F_j)} \rightarrow \begin{cases} 1, & \text{si } D(G||F_k) \leq D(G||F_j), \forall j = 1, \dots, m; \\ 0, & \text{sinon,} \end{cases}$$

lorsque $n \rightarrow \infty$. Ainsi, les estimateurs de Pitman compromis pour les paramètres de lieu et d'échelle convergent vers les estimateurs de Pitman associés à la distribution de compromis F_k telle que F_k est la plus proche de G au sens de la distance de Kullback-Leibler. □

Exemple 3.2.19. Dans cet exemple, nous allons étudier le comportement des estimateurs de Pitman compromis basés sur deux distributions, afin d'illustrer le corollaire précédent. Soit $F_1 = \Phi$, la loi distribution normale, et soit $F_2 = t_3$, la distribution de Student avec 3 degrés de liberté. Nous générons des erreurs suivant également deux distributions : soit $G_1 = t_2$ et $G_2 = t_8$. Nous avons alors que

$$D(G_1||F_1) = 0.5229, \quad D(G_1||F_2) = 0.2592,$$

$$D(G_2||F_1) = 0.1998, \quad D(G_2||F_2) = 0.3879.$$

Ainsi, si la vraie distribution des erreurs est G_1 , la distribution t_3 est la plus proche au sens de la distance de Kullback-Leibler. Au contraire, si G_2 est la vraie distribution, c'est Φ qui s'en approche le plus.

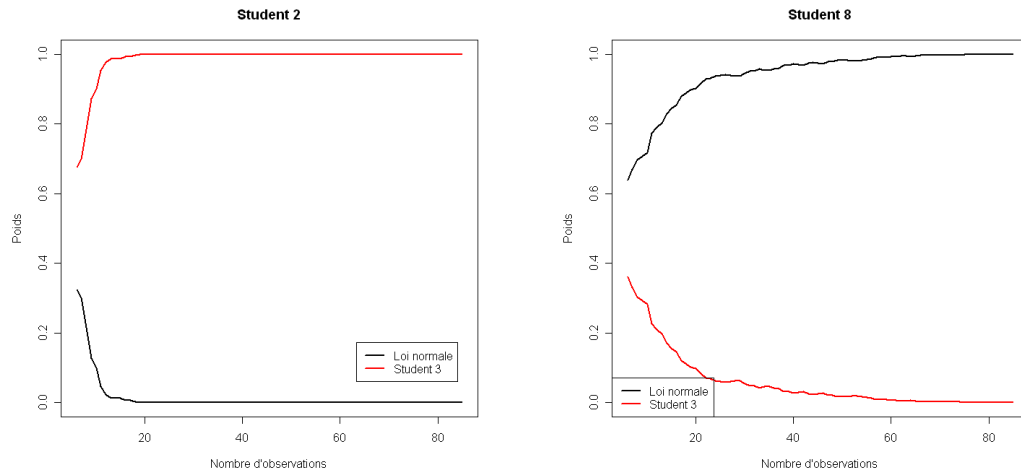


FIG. 3.2: A gauche : $G = G_1 = t_2$. A droite : $G = G_2 = t_8$.

La figure 3.2 montre l'évolution des poids accordés aux estimateurs de Pitman associés à F_1 et F_2 en fonction du nombre d'observations disponibles. Lorsque $G = G_1 = t_2$,

l'estimateur de Pitman compromis devient relativement rapidement l'estimateur de Pitman associé à la distribution t_3 , puisque c'est elle qui est la plus proche de t_2 . Au contraire, lorsque $G = G_2 = t_8$, l'estimateur de Pitman compromis devient celui associé à la loi normale Φ . Cet exemple illustre que ce type d'estimateurs compromis semble approprié pour des échantillons de taille relativement petite déjà.

3.2.5 Choix des distributions de compromis

Un aspect qui reste à traiter dans la construction des estimateurs de Pitman compromis est le choix des distributions de compromis $F_1, \dots, F_m \in \mathcal{F}$. Nous allons nous baser sur le comportement asymptotique de ces estimateurs afin de proposer une manière de faire ce choix.

Nous avons vu que les estimateurs de Pitman compromis convergent vers les estimateurs de Pitman associés à la distribution de compromis la plus proche de la vraie distribution des erreurs, au sens de la distance de Kullback-Leibler. Un choix intéressant serait un nombre raisonnable de distributions qui « couvrent » relativement bien l'ensemble des distributions possibles \mathcal{F} . Nous voudrions donc choisir nos distributions de telle manière que G n'est « pas trop éloignée » de l'une d'entre elles. Ceci nous amène à une approche de type minimax, qui peut être résumée comme suit : choisir $F_1, \dots, F_m \in \mathcal{F}$ de telle sorte que

$$\bigvee_{G \in \mathcal{F}} (D(G||F_1) \wedge \dots \wedge D(G||F_m))$$

est minimale, où \wedge , respectivement \vee , est l'opérateur minimum, respectivement l'opérateur maximum. Ainsi, pour chaque distribution possible G , notre choix contiendra une distribution qui se trouve relativement proche de cette dernière.

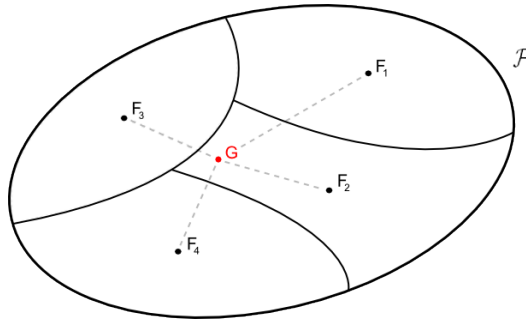


FIG. 3.3: Schématiquement, il s'agit de placer les distributions F_1, \dots, F_4 de sorte que G se situe « à proximité », quelle que soit sa position dans \mathcal{F} .

3.3 Simulations dans le cas de lois normales contaminées

Dans cette section, nous présentons les résultats de simulations de Monte-Carlo pour des estimateurs de Pitman compromis basés sur des lois normales contaminées. Afin d'alléger les notations, nous utiliserons dès à présent le terme de *stratégie* afin de désigner un estimateur compromis. Nous allons nous intéresser tout d'abord au choix des distributions, puis à la variance asymptotique des stratégies, et enfin à leur performance sur des petits échantillons.

Choix des distributions

Notation 3.3.1. Nous notons $F(y; \varepsilon, k)$ la loi normale contaminée avec proportion de contamination $\varepsilon \in [0, 0.5]$ et écart-type de contamination $k \geq 1$. Rappelons que la densité de cette loi est donnée par

$$f(y; \varepsilon, k) = (1 - \varepsilon)\varphi(y) + \varepsilon k^{-1}\varphi(y/k),$$

où $\varphi(y) = (2\pi)^{-1/2} \exp(-y^2/2)$ est la densité de la loi normale centrée et réduite.

Considérons l'ensemble suivant des distributions sous-jacentes possibles pour les erreurs du modèle de lieu et d'échelle :

$$\mathcal{F} = \{F(y; \varepsilon, k) \mid \varepsilon \in [0, 0.35], k \in [1, 10]\}.$$

Cet ensemble contient des distributions normales contaminées relativement plausibles, dans le sens qu'il nous semble improbable qu'un jeu de données contienne plus de 35% de données contaminées, et possédant de plus une variance 100 fois supérieure aux données standard. Notons que l'ensemble \mathcal{F} contient également Φ , la loi normale centrée et réduite. Afin de simplifier quelque peu notre analyse, nous allons travailler sur une discrétisation \mathcal{F}' de \mathcal{F} , en apposant une grille sur cet ensemble. Il est en effet nécessaire de limiter les distributions possibles pour résoudre le problème de choix des distributions de compromis de manière raisonnable. Dès lors, nous prendrons en compte uniquement les distributions normales contaminées avec

$$\varepsilon \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35\}$$

et

$$k \in \{0, 1, 2, \dots, 10\}.$$

Nous présentons maintenant dans la table 3.1 les solutions du problème de minimisation consistant à choisir $F_1, \dots, F_m \in \mathcal{F}'$ de telle sorte que

$$\bigvee_{G \in \mathcal{F}'} (D(G||F_1) \wedge \dots \wedge D(G||F_m))$$

soit minimale, pour différentes valeurs de m , et la stratégie correspondante est notée S_m . De plus, dans certains cas, nous exigeons que Φ fasse partie de la stratégie de

compromis. Dans ce cas, la stratégie est notée $S_{m\Phi}$ et contient donc $m+1$ distributions.

Stratégie	m	Paramètres	Dist. max.
S_1	1	$\varepsilon_1 = 0.15, k_1 = 7$	0.1010
S_2	2	$\varepsilon_1 = 0.10, k_1 = 6$ $\varepsilon_2 = 0.25, k_2 = 7$	0.0576
S_3	3	$\varepsilon_1 = 0.05, k_1 = 7$ $\varepsilon_2 = 0.20, k_2 = 4$ $\varepsilon_3 = 0.25, k_3 = 8$	0.0367
$S_{1\Phi}$	2	$\varepsilon_1 = 0.20, k_1 = 7$	0.0781
$S_{2\Phi}$	3	$\varepsilon_1 = 0.15, k_1 = 6$ $\varepsilon_2 = 0.20, k_2 = 8$	0.0542
$S_{3\Phi}$	4	$\varepsilon_1 = 0.05, k_1 = 7$ $\varepsilon_2 = 0.20, k_2 = 4$ $\varepsilon_3 = 0.25, k_3 = 8$	0.0367

TAB. 3.1: Stratégies minimax pour différentes valeurs de m . La dernière colonne contient la distance maximale d'une distribution $G \in \mathcal{F}'$ possible à une des distributions de la stratégie.

La stratégie S_1 contient une seule distribution, plus précisément celle avec $\varepsilon = 0.15$ et $k = 7$. Cette distribution peut être considérée comme étant « au centre » de \mathcal{F}' au sens de la distance de Kullback-Leibler. Si nous forçons Φ à faire partie de la stratégie, et que nous y ajoutons une seconde distribution, la fraction de contamination augmente légèrement ($\varepsilon = 0.2$) mais k reste inchangé (stratégie $S_{1\Phi}$). Les autres stratégies sont construites de manière relativement naturelle, en combinant une distribution proche de Φ avec un modèle alternatif possédant des queues plus lourdes, comme par exemple S_2 .

Nous remarquons que le fait d'ajouter une troisième distribution de compromis est moins important que d'en ajouter une seconde, en termes de distance maximale à une distribution G possible. En effet, le passage de S_1 à S_2 entraîne une diminution de la distance maximale de près de 50%, tandis que la diminution due au passage de S_2 à S_3 est moindre. De même, ajouter la loi normale à deux autres distributions (stratégie $S_{2\Phi}$) est pratiquement inutile au vu de la distance maximale. Ainsi, pour ce qui suit, nous ne prendrons en considération que les stratégies comportant au maximum deux distributions de compromis, c'est-à-dire les stratégies $S_1, S_{1\Phi}$ et S_2 .

Variance asymptotique

Soit $\text{Var}_G(S)$ la variance asymptotique de l'estimateur associé à la stratégie S lorsque G est la distribution des erreurs. Pour de grands échantillons, nous avons vu que l'estimateur de Pitman compromis convergeait vers l'estimateur du maximum de vraisemblance

associé à la distribution F_j la plus proche de G . Or, la variance asymptotique de l'estimateur du maximum de vraisemblance est donnée par

$$\text{Var}_G(F_j) = \left(\int \psi_{F_j}(u)^2 dG(u) \right) \left(\int \psi'_{F_j}(u) dG(u) \right)^{-2},$$

où

$$\psi_{F_j}(u) = -\frac{d}{du} \log f_j(u), \text{ et } \psi'_{F_j}(u) = \frac{d}{du} \psi_{F_j}(u).$$

La quantité

$$\frac{\text{Var}_G(S_i)}{\text{Var}_G(S_j)}$$

représente ainsi l'efficacité asymptotique de la stratégie S_j , en comparaison avec la stratégie S_i , si G est la distribution des erreurs. Cette quantité peut être vue comme le gain (ou la perte) relatif en variance asymptotique lorsque l'on utilise la stratégie S_j en lieu et place de S_i . Comme G est inconnue, nous comparons alors deux stratégies de compromis en regardant l'efficacité asymptotique maximale et minimale :

$$\max_{G \in \mathcal{F}'} \frac{\text{Var}_G(S_i)}{\text{Var}_G(S_j)} \text{ et } \min_{G \in \mathcal{F}'} \frac{\text{Var}_G(S_i)}{\text{Var}_G(S_j)}.$$

Comparons à présent les stratégies S_1 et $S_{1\Phi}$ sur la base des quantités ci-dessus : nous avons

$$\max_{G \in \mathcal{F}'} \frac{\text{Var}_G(S_1)}{\text{Var}_G(S_{1\Phi})} = 1.012 \text{ et } \min_{G \in \mathcal{F}'} \frac{\text{Var}_G(S_1)}{\text{Var}_G(S_{1\Phi})} = 0.81.$$

Le gain maximal d'efficacité en utilisant $S_{1\Phi}$ au lieu de S_1 est d'à peine 1% et se produit lorsque $G = \Phi$. Au contraire, la perte maximale est de l'ordre de 20% et se produit lorsque $G = F(\cdot; \epsilon = 0.05, k = 3)$, c'est-à-dire pour une distribution relativement proche de Φ . Ceci montre qu'une légère déviation autour de Φ va avoir un grand impact sur l'estimateur associé à la stratégie $S_{1\Phi}$. En ce sens, nous pouvons dire que la stratégie $S_{1\Phi}$ est nettement moins performante que S_1 qui, elle, est moins sensible aux déviations du modèle.

Comparons les stratégies $S_{1\Phi}$ et S_2 :

$$\max_{G \in \mathcal{F}'} \frac{\text{Var}_G(S_{1\Phi})}{\text{Var}_G(S_2)} = 1.223 \text{ et } \min_{G \in \mathcal{F}'} \frac{\text{Var}_G(S_{1\Phi})}{\text{Var}_G(S_2)} = 0.986.$$

A nouveau, le pire cas pour la stratégie S_2 , en comparaison avec $S_{1\Phi}$, se produit lorsque $G = \Phi$, mais la perte d'efficacité relative est d'à peine plus de 1%. Au contraire, le gain maximal est de plus de 20% et n'est pas négligeable. Ceci montre que si une stratégie de compromis à deux distributions doit être utilisée, laisser libre choix pour les deux distributions donne de meilleurs résultats que de forcer Φ à faire partie de la stratégie. Le caractère « extrême » de Φ est trop pénalisant.

Finalement, la comparaison de S_1 et S_2 montre que les deux estimateurs résultants sont quasiment asymptotiquement équivalents, puisque la perte ou le gain maximal d'efficacité ne dépassent pas les 2%.

Echantillons finis

Nous présentons maintenant dans les tableaux suivants les résultats de simulations de Monte-Carlo pour les estimateurs de Pitman compromis de lieu. Nous avons généré 1000 échantillons de taille 5, 10 et 20, et ce sous différentes distributions G . Pour chaque échantillon, nous avons déterminé les estimateurs associés aux stratégies $S_1, S_{1\Phi}$ et S_2 , ainsi que des estimateurs usuels du lieu : moyenne arithmétique, médiane, M-estimateur de Huber ($c = 1.345$) et Bisquare de Tukey ($c = 4.685$).

Remarque 3.3.2. Concernant les M-estimateurs de Huber et de Tukey, les constantes ont été choisies de telle sorte que l'efficacité asymptotique relative des estimateurs (par rapport à l'estimateur du maximum de vraisemblance) est de 95%. Par ailleurs, l'estimation de l'échelle est donnée par le MAD, la médiane des déviations absolues à la médiane.

Pour chaque situation, nous donnons l'efficacité relative de chaque estimateur, c'est-à-dire le rapport entre la variance de ce dernier et la variance minimale parmi tous les estimateurs. Pour chaque distribution sous-jacente G considérée, le meilleur estimateur parmi ceux pris en considération aura ainsi une efficacité relative de 1 (100%).

Estimateur	Φ	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.15$ $k = 7$	$\varepsilon = 0.25$ $k = 7$	$\varepsilon = 0.3$ $k = 9$	t_6	t_2	Slash
S_1	69.7	75.1	100	100	100	82.3	97.6	36.1
$S_{1\Phi}$	92.7	96.5	76.6	58.5	61.3	99.8	85.4	20.1
S_2	74.1	79.8	98.7	95.5	97.5	86.5	100	31.4
Moyenne	100	100	29.5	26.3	28.2	96.3	25.6	0.0
Médiane	69.5	75.6	95.2	92.5	96.5	80.5	93.1	100
Huber	93.1	97.7	67.8	49.7	52.0	100	77.4	37.8
Bisquare	89.3	92.8	83.1	66.3	77.4	96.5	95.7	75.8

TAB. 3.2: Efficacité relative (en %) des estimateurs de lieu considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 5$.

La table 3.2 présente les résultats pour des échantillons de taille $n = 5$. Lorsque $G = \Phi$, il est clair que la moyenne arithmétique possède la variance la plus faible. De même, parmi les 3 estimateurs compromis, celui de la stratégie $S_{1\Phi}$ est clairement le meilleur, puisque la stratégie contient la distribution Φ . Néanmoins, son efficacité relative n'est que de 93%, en raison de la seconde distribution utilisée pour sa construction et de

la taille relativement petite de l'échantillon. Les estimateurs compromis S_1 et S_2 sont similaires, avec un petit avantage pour S_2 , car ce dernier contient la distribution avec $\varepsilon = 0.1$ et $k = 6$, qui est plus proche de Φ que la distribution utilisée dans S_1 .

Pour des distributions à queues légères, comme la loi normale contaminée avec $\varepsilon = 0.05$ et $k = 2$, ou la distribution t de Student avec 6 degrés de liberté, $S_{1\Phi}$ est à nouveau le meilleur des trois estimateurs compromis, et présente même de meilleurs résultats que la moyenne arithmétique dans le dernier cas. Comme auparavant, S_2 se comporte mieux que S_1 , et ce pour les mêmes raisons que précédemment.

Lorsque la distribution sous-jacente possède des queues modérées, comme c'est le cas pour $F(\cdot; 0.15, 7)$, $F(\cdot; 0.25, 7)$, $F(\cdot; 0.3, 9)$ ou pour t_2 , les estimateurs compromis S_1 et S_2 présentent les meilleures performances parmi tous les autres estimateurs. Ceci était attendu au moins pour le cas d'une distribution sous-jacente normale contaminée, puisque les deux estimateurs sont construits sur des compromis de ce type de distribution. Nous nous attendions également à ce que S_2 domine S_1 dans le cas particulier de $G = F(\cdot; 0.25, 7)$, puisque cette loi fait partie de S_2 . Cela ne semble pas être le cas, vraisemblablement à cause du petit nombre d'observations. Notons également que S_1 et S_2 présentent de bons résultats même avec une distribution sous-jacente totalement différente comme la loi t_2 de Student. Finalement, nous remarquons que l'efficacité de $S_{1\Phi}$ chute significativement. Cet estimateur semble être très pénalisé par la présence de Φ dans le compromis, tout du moins pour les échantillons de petite taille.

Lorsque G est la distribution Slash, une loi à queues très lourdes, les trois estimateurs compromis semblent inefficaces. Ceci peut s'expliquer en étudiant la fonction Ψ associée à une distribution normale contaminée. Cette fonction représente l'influence d'une contamination (infinitésimale) des données, et est une indication de la robustesse de l'estimateur du maximum de vraisemblance.

Dans le cas d'une loi normale contaminée, la figure 3.4 montre que la fonction Ψ est faite de deux parties distinctes : une première partie linéaire se rapportant à la distribution Φ du mélange, puis une seconde partie linéaire, avec une pente plus faible, se rapportant à la contamination. Cette seconde partie est toutefois croissante, au contraire du M-estimateur de Huber par exemple. Dès lors, l'estimateur n'est pas résistant dans le cas d'une loi à queues très lourdes, comme la loi Slash. Les estimateurs de Pitman compromis basés sur des lois normales contaminées ne peuvent pas nous protéger contre ce genre de distribution, du fait de leur nature gaussienne sous-jacente.

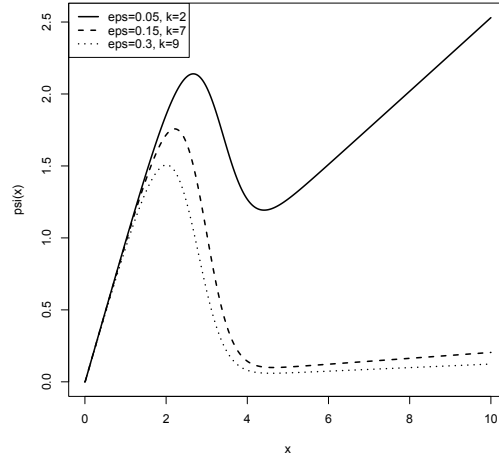


FIG. 3.4: Fonction Ψ pour des lois normales contaminées.

Les tables 3.3 et 3.4 présentent les résultats des simulations pour des échantillons de taille $n = 10$ et $n = 20$. Lorsque $G = \Phi$, la performance des estimateurs S_1 et S_2 augmente légèrement avec n , avec toujours un avantage pour S_2 , comme discuté précédemment. $S_{1\Phi}$ est toujours le meilleur estimateur compromis en cas de distribution à queues légères, mais S_1 et S_2 se comportent très bien. Dans le cas de la distribution t_6 , c'est même S_2 qui est le meilleur des trois pour des échantillons de taille $n = 20$.

Estimateur	Φ	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.15$ $k = 7$	$\varepsilon = 0.25$ $k = 7$	$\varepsilon = 0.3$ $k = 9$	t_6	t_2	Slash
S_1	76.8	80.2	99.7	99.3	94.5	86.5	100	69.3
$S_{1\Phi}$	93.9	97.7	91.4	76.6	52.6	97.5	87.8	60.3
S_2	80.6	83.5	100	92.3	81.2	89.3	98.9	69.3
Moyenne	100	100	21.2	21.1	12.6	92.8	20.9	0.4
Médiane	73.1	74.0	86.9	100	100	83.9	94.2	100
Huber	93.9	99.8	78.8	55.2	35.3	100	85.1	57.1
Bisquare	90.5	95.2	92.0	72.8	54.7	94.4	89.4	80.1

TAB. 3.3: Efficacité relative (en %) des estimateurs de lieu considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 10$.

Estimateur	Φ	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.15$ $k = 7$	$\varepsilon = 0.25$ $k = 7$	$\varepsilon = 0.3$ $k = 9$	t_6	t_2	Slash
S_1	88.1	94.2	100	99.2	95.4	95.4	100	30.5
$S_{1\Phi}$	97.9	99.5	95.0	94.7	81.7	96.3	89.5	30.3
S_2	90.3	95.8	99.7	100	100	96.6	99.1	27.1
Moyenne	100	98.6	18.1	15.6	10.3	85.5	22.0	0.0
Médiane	70.0	72.6	69.8	82.4	85.0	84.3	93.6	99.5
Huber	95.2	100	75.8	62.9	48.0	100	88.1	67.8
Bisquare	92.8	98.0	91.6	85.5	72.7	97.5	95.0	100

TAB. 3.4: Efficacité relative (en %) des estimateurs de lieu considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 20$.

Lorsque $n = 10$, S_1 est le meilleur choix en cas de distribution à queues modérées, et est le plus efficace dans la plupart des cas. Lorsque $n = 20$, S_2 devient meilleur, sauf si $G = F(\cdot; 0.15, 7)$, car cette distribution est en fait l'unique distribution utilisée dans S_1 , qui est alors le meilleur choix. Nous remarquons de plus que lorsque G est l'une des distributions utilisées dans S_2 , ce dernier est au moins aussi efficace que S_1 pour des échantillons de taille supérieure à 10.

Finalement, nous observons que les performances de l'estimateur $S_{1\Phi}$ augmentent dans le cas d'une distribution sous-jacente à queues modérées, lorsque la taille de l'échantillon augmente. Afin d'expliquer ce phénomène, rappelons qu'une seconde distribution est utilisée dans ce compromis, plus précisément la loi $F(\cdot; 0.2, 7)$, qui possède des queues relativement lourdes. Les résultats ci-dessus nous indiquent donc que le poids associé à la distribution Φ dans le compromis $S_{1\Phi}$ devient faible relativement rapidement. Par exemple, l'efficacité relative de $S_{1\Phi}$ dans le cas où $G = t_2$ est de 87.2% lorsque $n = 10$ et de 89.5% lorsque $n = 20$.

En ce qui concerne le cas de la loi Slash, même si les performances des estimateurs compromis lorsque la taille de l'échantillon vaut $n = 10$ sont surprenants, il semble raisonnable de penser que ces estimateurs possèdent une variance infinie lorsque la loi des erreurs a également une variance infinie, du fait de la nature gaussienne sous-jacente de ces estimateurs. En effet, comme nous avons pu le voir sur la figure 3.4, la fonction Ψ associée à l'estimateur du maximum de vraisemblance pour une loi normale contaminée est linéaire. De plus la densité de la loi Slash est donnée par

$$f(y) = \begin{cases} y^{-2} (((2\pi)^{-1/2} - \varphi(y))), & y \neq 0; \\ \frac{1}{2\sqrt{2\pi}}, & y = 0. \end{cases}$$

et donc l'intégrale $\int_{-\infty}^{+\infty} \Psi^2(y)f(y)dy$ diverge.

3.3. Simulations dans le cas de lois normales contaminées

Estimateur	Φ	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.15$ $k = 7$	$\varepsilon = 0.25$ $k = 7$	$\varepsilon = 0.3$ $k = 9$	t_6	t_2	Slash
S_1	2.4	2.9	0.6	3.8	5.9	2.4	0.0	6.4
$S_{1\Phi}$	1.5	1.6	2.6	5.4	3.6	1.2	2.9	4.9
S_2	2.4	2.8	0.0	5.1	7.2	2.2	0.6	6.0
Moyenne	0.0	0.1	1.3	1.6	0.7	2.3	8.1	0.2
Médiane	2.4	2.7	2.2	3.6	0.0	2.7	2.2	0.0
Huber	1.4	1.6	3.2	3.9	2.2	0.0	3.2	3.3
Bisquare	1.9	2.3	2.8	4.9	4.1	1.3	2.8	4.1

TAB. 3.5: Ecart-type (en %) de l'efficacité relative de chaque estimateur, estimé à l'aide de la méthode du *jackknife*. Taille de l'échantillon : $n = 10$.

La table 3.5 donne pour chaque estimateur une estimation de l'écart-type de son efficacité relative, obtenue à l'aide de la méthode du *jackknife* (voir Shao et Tu, 1995). Le comportement de cet écart-type est comparable pour $n = 5$ et $n = 20$, ce qui nous indique que les variances de chaque estimateur évoluent de manière similaire lorsque la taille de l'échantillon change.

	$n = 5$	$n = 10$	$n = 20$
S_1	69.7	76.8	88.1
$S_{1\Phi}$	58.5	52.6	81.7
S_2	74.1	80.6	90.3
Moyenne	25.6	12.6	10.3
Médiane	69.5	73.1	69.8
Huber	49.7	35.3	48.0
Bisquare	66.3	54.7	72.7

TAB. 3.6: Efficacité relative minimale (en %) des estimateurs de lieu considérés parmi toutes les distributions sous-jacentes, exceptée la distribution Slash.

La table 3.6 nous donne une manière d'apprécier la qualité globale de chaque estimateur, en présentant l'efficacité relative minimale de chacun, lorsque toutes les distributions sous-jacentes sont prises en considération, excepté la loi Slash. Cette table nous donne donc une approche minimax, puisque l'on regarde la perte maximale d'efficacité de chaque estimateur. $S_{1\Phi}$ devrait être évité pour des échantillons relativement petits, mais est aussi performant que les deux autres estimateurs compromis lorsque $n = 20$. S_1 et S_2 sont pratiquement équivalents. S_2 nous protège plus efficacement lorsque la distribution sous-jacente possède des queues légères, mais S_1 possède l'avantage de sa simplicité, puisqu'il est nécessaire de déterminer les estimateurs de Pitman associés à une seule distribution.

En conclusion, nous pouvons dire que l'estimateur compromis dans lequel nous avons fixé une distribution, en l'occurrence Φ , se comporte moins bien que les estimateurs compromis dans lesquels les distributions sont libres. Lorsque l'on ajoute une distribution, de S_1 à S_2 , le gain est modéré, et il est raisonnable d'imaginer que l'ajout d'une troisième distribution soit inutile, ou ne vaille pas le coût calculatoire. S_1 représente en un certain sens l'approche robuste traditionnelle : on recherche un estimateur se comportant bien dans la plupart des cas, sans faire de compromis.

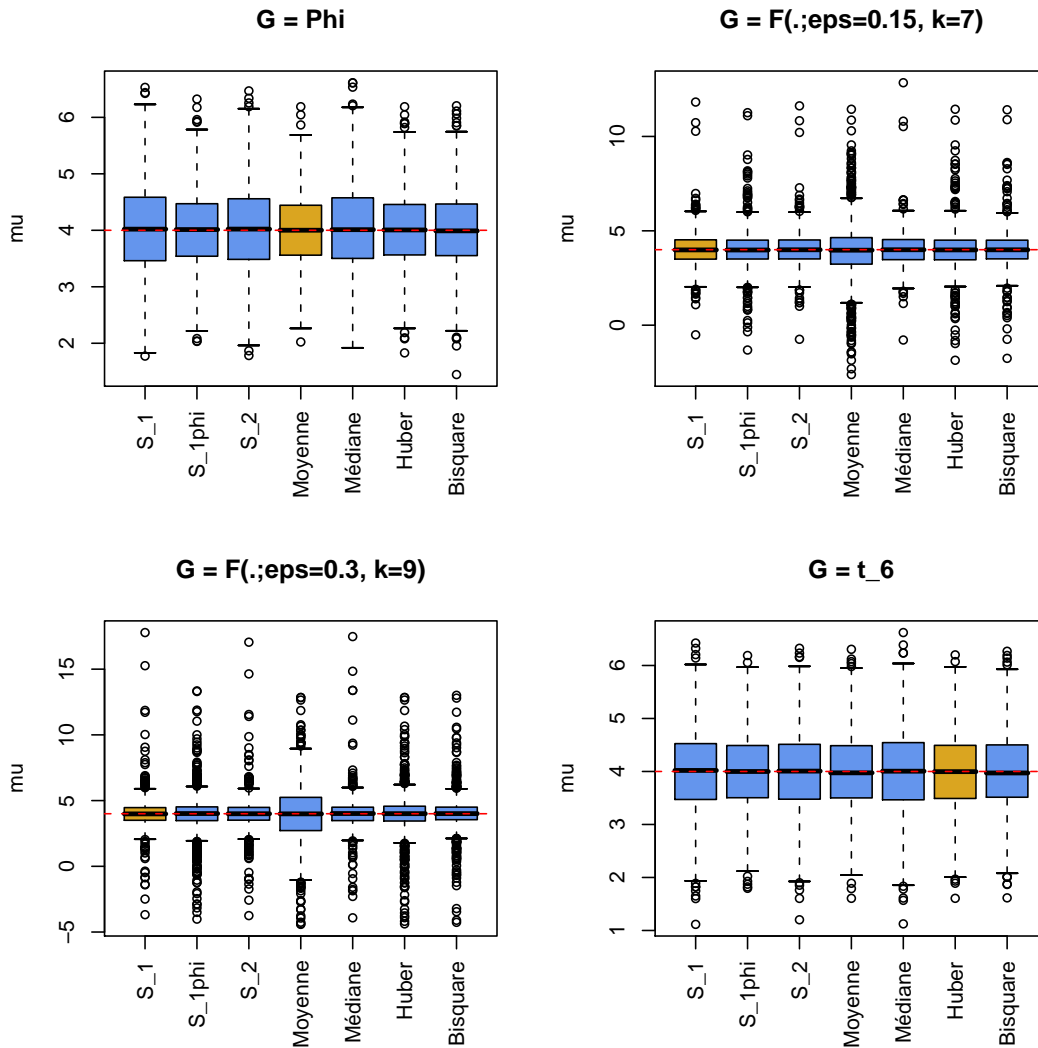


FIG. 3.5: Boxplots des résultats des simulations pour différentes loi sous-jacentes. En jaune, l'estimateur possédant l'efficacité relative maximale. La ligne rouge représente la vraie valeur du paramètre de lieu μ . Taille de l'échantillon : $n = 5$.

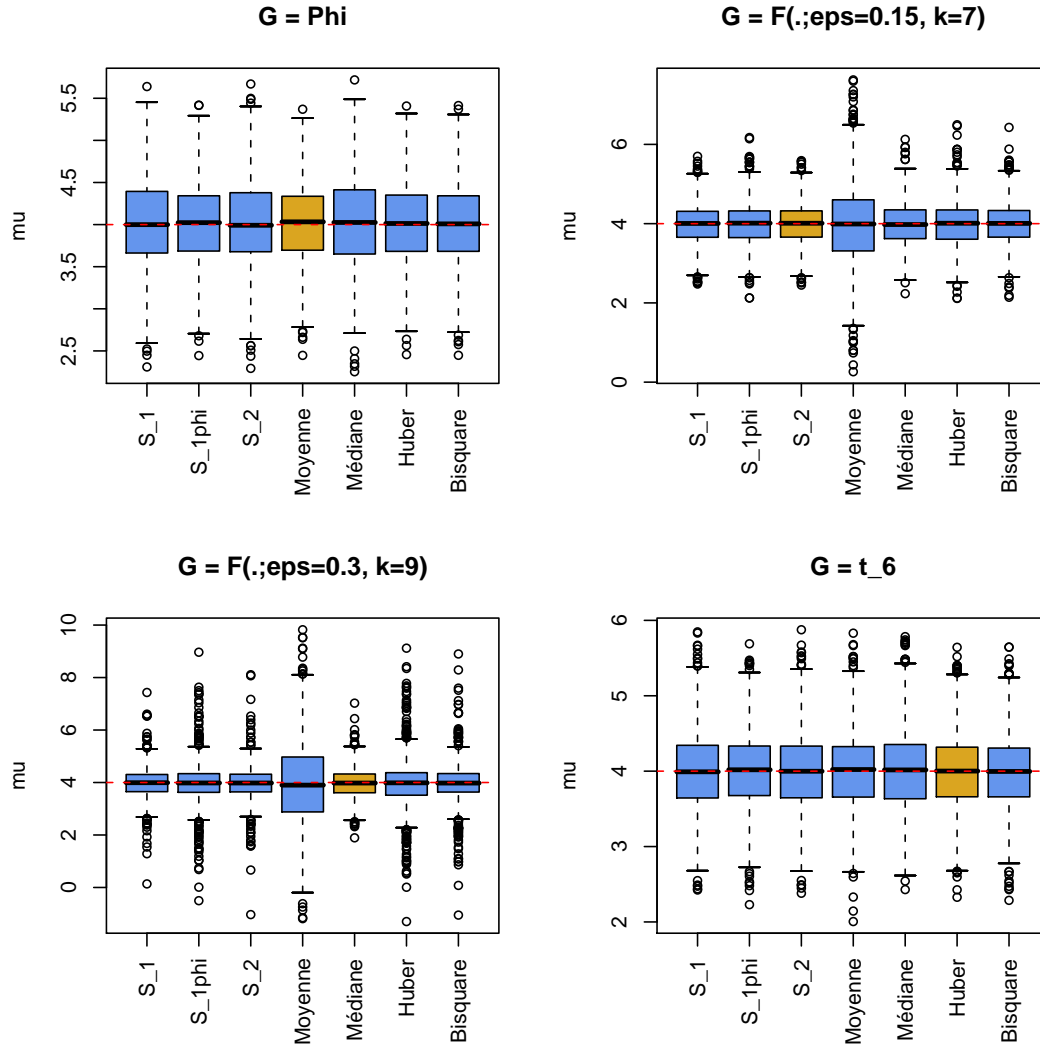


FIG. 3.6: Boxplots des résultats des simulations pour différentes loi sous-jacentes. En jaune, l'estimateur possédant l'efficacité relative maximale. La ligne rouge représente la vraie valeur du paramètre de lieu μ . Taille de l'échantillon : $n = 10$.

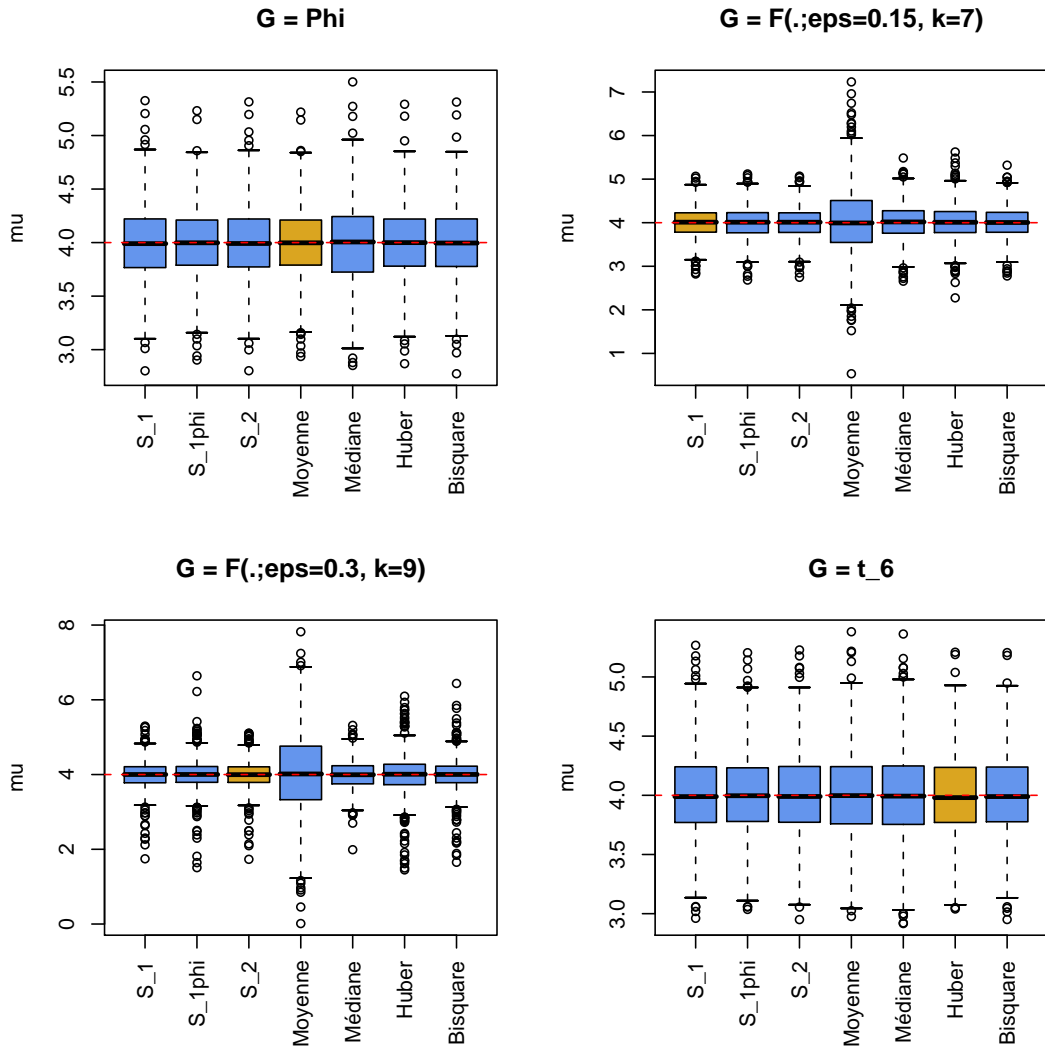


FIG. 3.7: Boxplots des résultats des simulations pour différentes loi sous-jacentes. En jaune, l'estimateur possédant l'efficacité relative maximale. La ligne rouge représente la vraie valeur du paramètre de lieu μ . Taille de l'échantillon : $n = 20$.

Les tables 3.7, 3.8 et 3.9 présentent les résultats de simulations de Monte-Carlo pour le paramètre d'échelle, similaires à celles réalisées pour le paramètre de lieu. Les estimateurs pris en considération sont les trois estimateurs de Pitman compromis S_1 , $S_{1\phi}$ et S_2 , l'estimateur usuel de l'écart-type $s = [(n-1)^{-1} \sum_{i=1}^n (y_i - \bar{y})^2]^{\frac{1}{2}}$, la déviation absolue médiane (*median absolute deviation*, *MAD* en anglais) donnée par la médiane des écarts absolus à la médiane, et la distance interquartile empirique de l'échantillon (DIQ).

Notons que s et MAD n'estiment pas la même quantité que les estimateurs de Pitman compromis ou DIQ, qui eux estiment la distance interquartile, c'est-à-dire le paramètre

d'échelle σ , puisque les distributions ont été standardisées. Dès lors, nous présentons les efficacités relatives du log de ces estimateurs, puisqu'ils ne diffèrent que d'une constante multiplicative. s et MAD seront biaisés, mais nous ne nous intéressons qu'à leur variance.

Les résultats pour le paramètre d'échelle sont très similaires à ceux du paramètre de lieu. Lorsque la distribution sous-jacente est la loi normale, ou une loi à queues légères, l'estimateur usuel s est à chaque fois le meilleur choix, et ce également lorsque la taille de l'échantillon augmente. Les estimateurs compromis présentent tous dans ce cas une efficacité relativement décevante par rapport aux résultats du paramètre de lieu. S_1 et S_2 se comportent de manière très similaire, quelle que soit la taille de l'échantillon. Lorsque $n = 10$, respectivement $n = 20$, nous remarquons que l'estimateur compromis $S_{1\Phi}$ présente une efficacité relative de seulement 67.1%, respectivement 78.4%.

Lorsque la distribution sous-jacente possède des queues modérées, les estimateurs compromis se comportent tous relativement bien, en particulier S_2 . On note toutefois une grande perte d'efficacité pour S_1 et $S_{1\Phi}$ lorsque $G = F(\cdot; 0.3, 9)$ pour $n = 20$.

Notons finalement que lorsque G est la loi Slash, comme dans le cas du paramètre de lieu, les estimateurs compromis voient leur efficacité relative chuter de manière importante, du fait de leur nature gaussienne sous-jacente.

Estimateur	Φ	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.15$ $k = 7$	$\varepsilon = 0.25$ $k = 7$	$\varepsilon = 0.3$ $k = 9$	t_6	t_2	Slash
S_1	70.4	71.4	100	100	99.3	80.4	97.6	91.2
$S_{1\Phi}$	81.8	83.1	94.8	93.8	93.9	90.5	100	90.1
S_2	72.4	72.4	98.9	100	100	82.1	96.7	88.3
s	100	100	62.3	70.4	81.0	100	75.7	54.3
MAD	34.2	36.3	71.7	72.0	66.8	42.9	55.8	100
DIQ	37.6	35.5	70.2	68.8	60.2	43.2	55.5	97.2

TAB. 3.7: Efficacité relative (en %) du log des estimateurs d'échelle considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 5$.

Estimateur	Φ	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.15$ $k = 7$	$\varepsilon = 0.25$ $k = 7$	$\varepsilon = 0.3$ $k = 9$	t_6	t_2	Slash
S_1	51.9	59.3	100	94.0	91.4	69.9	96.7	54.2
$S_{1\Phi}$	67.1	76.4	95.0	84.8	80.5	85.4	100	54.4
S_2	49.5	56.0	97.5	100	100	65.8	93.6	49.9
s	100	100	32.8	54.5	64.3	100	53.9	23.4
MAD	39.2	43.8	73.2	76.0	78.6	54.5	83.6	100
DIQ	40.0	44.1	69.8	66.8	61.7	57.8	86.1	92.8

TAB. 3.8: Efficacité relative (en %) du log des estimateurs d'échelle considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 10$.

Estimateur	Φ	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.15$ $k = 7$	$\varepsilon = 0.25$ $k = 7$	$\varepsilon = 0.3$ $k = 9$	t_6	t_2	Slash
S_1	56.7	73.4	100	89.0	77.0	82.5	87.8	20.5
$S_{1\Phi}$	78.4	91.1	96.6	85.9	74.6	94.2	85.3	20.2
S_2	53.9	68.5	96.1	100	100	75.5	80.6	18.8
s	100	100	23.9	37.1	56.0	100	34.6	9.2
MAD	36.7	50.4	63.4	62.1	69.0	61.9	96.9	100
DIQ	37.3	53.1	63.0	54.9	55.4	67.0	100	93.4

TAB. 3.9: Efficacité relative (en %) du log des estimateurs d'échelle considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 20$.

3.4 Simulations dans le cas de lois t de Student

Dans cette section, nous présentons les résultats de simulations de Monte-Carlo pour des estimateurs de Pitman compromis basés sur des lois t de Student. Comme précédemment, nous utiliserons le terme de *stratégie* afin de désigner un estimateur compromis. Nous allons nous intéresser tout d'abord au choix des distributions, puis à la variance asymptotique des stratégies, et enfin à leur performance sur des petits échantillons.

Choix des distributions

Considérons l'ensemble des distributions sous-jacentes possibles suivant :

$$\mathcal{F} = \{t_\nu \mid \nu = 1, 2, \dots, 20\} \cup \{\Phi\}.$$

Cet ensemble contient donc les distributions t de Student avec différents degrés de liberté, ainsi que la loi normale Φ . Nous avons choisi un degré de liberté maximal de 20, car la distribution t_{20} est très semblable à Φ .

Au contraire du cas des lois normales contaminées présenté ci-dessus, nous n'avons pas besoin ici de discrétiser l'ensemble \mathcal{F} , puisque nous pouvons facilement dénombrer les distributions qui en font partie.

Les résultats de la minimisation de

$$\bigvee_{G \in \mathcal{F}} (D(G||F_1) \wedge \dots \wedge D(G||F_m))$$

pour différentes valeurs de m sont présentés dans la table 3.10.

Stratégie	m	Paramètres	Distance max.
S_1	1	$\nu_1 = 2$	0.0804
S_2	2	$\nu_1 = 2$ $\nu_2 = 5$	0.0287
S_3	3	$\nu_1 = 1$ $\nu_2 = 3$ $\nu_3 = 8$	0.0107
$S_{1\Phi}$	2	$\nu_1 = 2$	0.0373
$S_{2\Phi}$	3	$\nu_1 = 1$ $\nu_2 = 3$	0.0193
$S_{3\Phi}$	4	$\nu_1 = 1$ $\nu_2 = 2$ $\nu_3 = 5$	0.0096

TAB. 3.10: Stratégies minimax pour différentes valeurs de m . La dernière colonne contient la distance maximale d'une distribution $G \in \mathcal{F}$ possible à une des distributions de la stratégie.

La stratégie S_1 ne contient que la distribution t_2 , qui peut être considérée comme étant « au centre » de \mathcal{F} par rapport à la distance de Kullback-Leibler. En rajoutant la distribution t_5 , comme dans S_2 , nous observons que la distance maximale à toute distribution G est divisée par un facteur quatre, tandis que l'ajout d'une troisième distribution divise cette distance par un facteur deux seulement. A nouveau, nous ne considérerons que des compromis comportant au maximum deux distributions.

Comme Φ peut être vue comme une loi t de Student avec un nombre infini de degrés de liberté, il n'est pas étonnant de voir des distributions à queues relativement lourdes associées à Φ dans les stratégies $S_{1\Phi}$, $S_{2\Phi}$ et $S_{3\Phi}$, du fait du caractère « extrême » de la loi normale dans notre ensemble \mathcal{F} . Cela indique également qu'une loi t avec 8 ou 10 degrés de liberté n'est pas si éloignée de Φ , au sens de la distance de Kullback-Leibler.

Variance asymptotique

Nous comparons à présent les stratégies $S_1, S_{1\Phi}$ et S_2 sur la base de leur variance asymptotique. Si l'on utilise $S_{1\Phi}$ en lieu et place de S_1 ,

$$\max_{G \in \mathcal{F}} \frac{\text{Var}_G(S_1)}{\text{Var}_G(S_{1\Phi})} = 1.112 \text{ et } \min_{G \in \mathcal{F}} \frac{\text{Var}_G(S_1)}{\text{Var}_G(S_{1\Phi})} = 0.921.$$

Le gain maximal d'efficacité est de l'ordre de 11% et se produit lorsque $G = \Phi$, ce qui n'est pas une surprise, étant donné que cette distribution fait partie du compromis $S_{1\Phi}$. Au contraire, la perte maximale est de l'ordre de 8% seulement et se produit lorsque $G = t_7$, une distribution que l'on peut considérer « au milieu » de t_2 et Φ .

Comparons à présent $S_{1\Phi}$ et S_2 :

$$\max_{G \in \mathcal{F}} \frac{\text{Var}_G(S_{1\Phi})}{\text{Var}_G(S_2)} = 1.117 \text{ et } \min_{G \in \mathcal{F}} \frac{\text{Var}_G(S_{1\Phi})}{\text{Var}_G(S_2)} = 0.951.$$

A nouveau, la perte maximale d'efficacité intervient lorsque $G = \Phi$, mais n'est que de 5% dans ce cas. De même, le gain maximal est du même ordre que précédemment. Comme dans le cas des lois normales contaminées, il semble préférable, si l'on désire un compromis entre deux distributions, de ne pas forcer Φ à faire partie du compromis, mais plutôt de laisser libre choix pour les deux distributions. A nouveau, le caractère extrême de Φ est pénalisant.

Finalement, la comparaison de S_1 et S_2 nous montre que les deux estimateurs sont relativement similaires, avec un léger avantage pour S_2 lorsque $G = \Phi$, comme on pouvait s'y attendre.

Echantillons finis

Nous présentons à présent les résultats de simulations de Monte-Carlo pour le paramètre de lieu. Comme dans le cas des lois normales contaminées, nous avons généré 1000 échantillons de taille 5, 10 et 20, et ce sous différentes distributions G . Pour chaque échantillon, nous avons déterminé les estimateurs associés aux stratégies $S_1, S_{1\Phi}$ et S_2 , ainsi que des estimateurs usuels du lieu : moyenne arithmétique, médiane, M-estimateur de Huber ($c = 1.345$) et Bisquare de Tukey ($c = 4.685$). Nous présentons dans chaque cas l'efficacité relative de chaque estimateur.

La table 3.11 présente les résultats pour des échantillons de taille $n = 5$, qui s'avèrent être très similaires à ceux obtenus dans le cas des lois normales contaminées. Lorsque $G = \Phi$, $S_{1\Phi}$ est à nouveau le meilleur des trois estimateurs compromis, puisqu'il contient Φ . Néanmoins, la perte d'efficacité relative de S_2 n'est que de 17%, contre 26% dans le cas des lois normales contaminées.

Lorsque la distribution sous-jacente possède des queues légères, comme les lois t_{10} et t_6 , ou $F(\cdot; 0.05, 2)$, les trois estimateurs compromis se comportent bien, avec une légère

Estimateur	Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$	$\varepsilon = 0.25$	Slash
						$k = 2$	$k = 7$	
S_1	74.2	84.3	90.9	94.2	100	85.0	97.9	93.1
$S_{1\Phi}$	90.5	99.1	100	99.9	85.8	99.0	63.0	61.7
S_2	82.9	92.9	97.9	100	95.8	93.5	78.3	78.1
Moyenne	100	100	91.4	69.4	25.8	98.6	27.5	0.2
Médiane	69.1	77.5	80.5	85.0	96.4	74.9	100	100
Huber	90.7	98.7	99.8	99.6	78.8	100	51.6	42.9
Bisquare	85.4	93.9	98.7	98.6	91.9	97.7	68.4	80.3

TAB. 3.11: Efficacité relative (en %) des estimateurs de lieu considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 5$.

préférence pour $S_{1\Phi}$. On remarque également une perte d'efficacité relative moindre par rapport aux estimateurs compromis basés sur des lois normales contaminées.

Dans le cas d'une distribution à queues relativement lourdes, comme les lois t_3 et t_2 , les estimateurs compromis S_1 et S_2 sont de très bons choix, et $S_{1\Phi}$ présente également de bons résultats. Ce dernier semble moins pénalisé par la présence de la distribution Φ que dans le cas des lois normales contaminées. Les résultats pour le cas $G = F(\cdot; 0.25, 7)$ sont quelque peu surprenants en ce qui concerne $S_{1\Phi}$ et S_2 . Dans ce cas, la médiane se comporte très bien, et S_1 est le meilleur des trois estimateurs compromis.

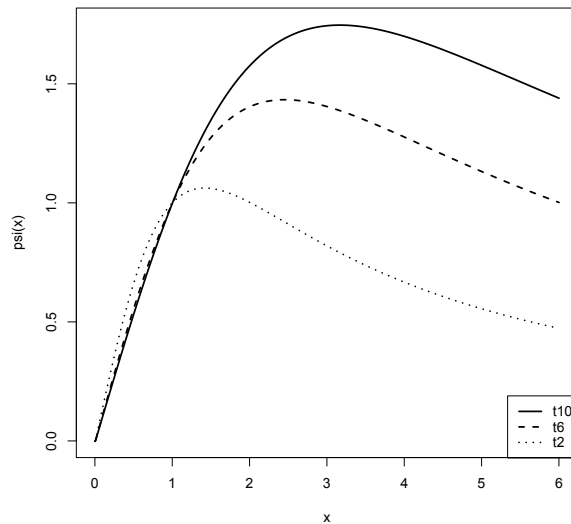


FIG. 3.8: Fonction Ψ pour des lois t de Student.

Finalement, dans le cas où G est la loi Slash, nous remarquons que les estimateurs compromis se comportent nettement mieux lorsqu'ils sont construits avec des distributions t qu'avec des lois normales contaminées. Néanmoins, pour $n = 5$, la perte d'efficacité relative est importante pour $S_{1\phi}$ et S_2 . Comme précédemment, la figure 3.8 montre le comportement de la fonction Ψ pour une loi t avec différents degrés de liberté. Nous remarquons que dans ce cas, la fonction Ψ possède une seconde partie décroissante, ce qui permet aux estimateurs basés sur ces distributions de se comporter relativement bien en cas de distributions à queues très lourdes comme la loi Slash.

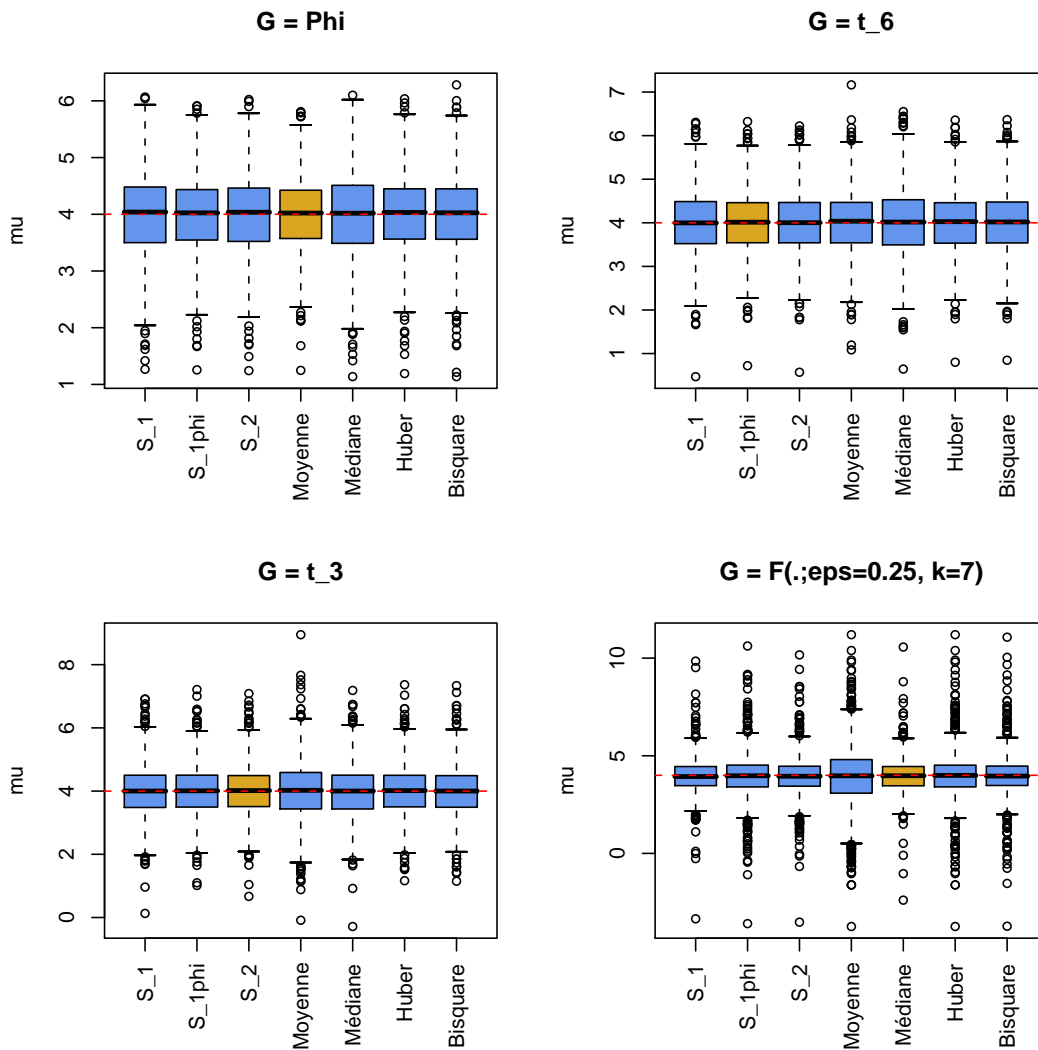


FIG. 3.9: Boxplots des résultats des simulations pour différentes loi sous-jacentes. En jaune, l'estimateur possédant l'efficacité relative maximale. La ligne rouge représente la vraie valeur du paramètre de lieu μ . Taille de l'échantillon : $n = 5$.

3.4. Simulations dans le cas de lois t de Student

Les tables 3.12 et 3.13 présentent les résultats pour des échantillons de taille $n = 10$ et $n = 20$. Lorsque $G = \Phi$, $S_{1\Phi}$ est le meilleur des trois estimateurs compromis, comme on pouvait s'y attendre. S_2 se comporte relativement bien et présente une perte d'efficacité relative moins importante que S_1 , cette dernière stratégie ne comprenant que la loi t_2 .

Estimateur	Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.25$ $k = 7$	Slash
S_1	74.1	91.1	92.6	98.9	100	81.8	100	50.2
$S_{1\Phi}$	92.7	100	99.7	95.4	85.8	99.6	67.1	45.3
S_2	83.9	97.5	98.7	100	94.9	92.0	82.8	47.8
Moyenne	100	95.5	89.0	58.1	18.5	98.1	18.6	0.0
Médiane	68.7	84.5	84.8	93.7	93.4	74.0	91.8	100
Huber	92.6	99.8	100	93.3	75.3	100	50.1	55.9
Bisquare	87.7	94.6	96.9	93.2	83.6	96.0	67.9	83.9

TAB. 3.12: Efficacité relative (en %) des estimateurs de lieu considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 10$.

Dans le cas d'une distribution sous-jacente à queues modérées, les trois estimateurs compromis sont très similaires, et $S_{1\Phi}$ ne semble pas trop pénalisé par la distribution Φ . Comme le montre l'efficacité relative de S_1 , la loi t_2 , utilisée comme seconde distribution dans $S_{1\Phi}$, donne de très bons résultats de manière générale. L'ajout d'une seconde distribution libre, la loi t_5 dans S_2 , semble moins significative que dans le cas des lois normales contaminées.

Lorsque la distribution sous-jacente est la loi normale contaminée $F(\cdot; 0.25, 7)$, nous n'observons plus de perte importante d'efficacité pour les estimateurs compromis, sauf pour $S_{1\Phi}$ lorsque $n = 10$.

Estimateur	Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.25$ $k = 7$	Slash
S_1	72.7	87.8	91.8	98.3	100	85.2	100	78.3
$S_{1\Phi}$	94.0	99.0	99.0	95.9	91.6	98.9	88.3	44.7
S_2	84.4	96.8	98.4	100	97.0	95.1	93.5	62.0
Moyenne	100	94.9	89.6	53.2	19.4	96.9	16.8	0.0
Médiane	63.4	76.2	80.9	86.5	88.9	73.2	83.1	100
Huber	92.9	100	100	93.0	81.4	100	64.2	35.9
Bisquare	89.3	97.3	98.3	96.7	86.1	98.4	79.5	77.8

TAB. 3.13: Efficacité relative (en %) des estimateurs de lieu considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 20$.

Finalement, lorsque $n = 20$, les trois estimateurs compromis présentent une meilleure efficacité relative que l'estimateur de Huber, si la loi sous-jacente est la loi Slash, bien que la médiane reste de loin le meilleur choix dans ce cas.

Estimateur	Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.25$ $k = 7$	Slash
S_1	2.2	1.8	2.0	1.1	0.0	1.9	0.0	33.1
$S_{1\Phi}$	1.5	0.1	0.9	1.2	2.0	0.8	2.2	25.3
S_2	2.0	1.1	1.3	0.0	1.1	1.4	1.4	29.1
Moyenne	0.0	1.8	2.2	3.4	3.5	1.9	1.3	0.1
Médiane	2.4	2.1	2.1	1.9	1.8	2.2	2.4	0.0
Huber	1.4	0.9	0.0	1.7	2.6	0.0	2.6	5.85
Bisquare	2.1	1.2	1.1	1.5	2.1	1.1	2.8	5.18

TAB. 3.14: Ecart-type (en %) de l'efficacité relative de chaque estimateur, estimé à l'aide de la méthode du *jackknife*. Taille de l'échantillon : $n = 10$.

La table 3.14 montre l'écart-type estimé de l'efficacité relative pour chaque estimateur. Nous remarquons que les écarts-types pour les estimateurs compromis sont légèrement plus faibles que dans le cas de stratégies basées sur des lois normales contaminées, comme présenté auparavant. Par contre, lorsque la loi sous-jacente est la loi Slash, ce n'est pas le cas. A nouveau, lorsque la taille de l'échantillon varie, ces écarts-types se comportent de manière similaire.

	$n = 5$	$n = 10$	$n = 20$
S_1	74.2	74.1	72.7
$S_{1\Phi}$	63.0	67.1	88.3
S_2	78.3	82.8	84.4
Moyenne	25.8	18.5	16.8
Médiane	69.1	68.7	63.4
Huber	51.6	50.1	64.2
Bisquare	68.4	67.9	79.5

TAB. 3.15: Efficacité relative minimale (en %) des estimateurs de lieu considérés parmi toutes les distributions sous-jacentes, exceptée la distribution Slash.

La table 3.15 montre la même approche minimax que dans le cas des lois normales contaminées. De ce point de vue, nous remarquons que S_2 est supérieur à S_1 , tandis que $S_{1\Phi}$ devrait être évité pour des petits échantillons. Pour $n = 20$ par contre, ce dernier est même le meilleur estimateur parmi ceux considérés. De manière générale, nous pouvons dire également que S_2 est un excellent choix, quelle que soit la taille de l'échantillon.

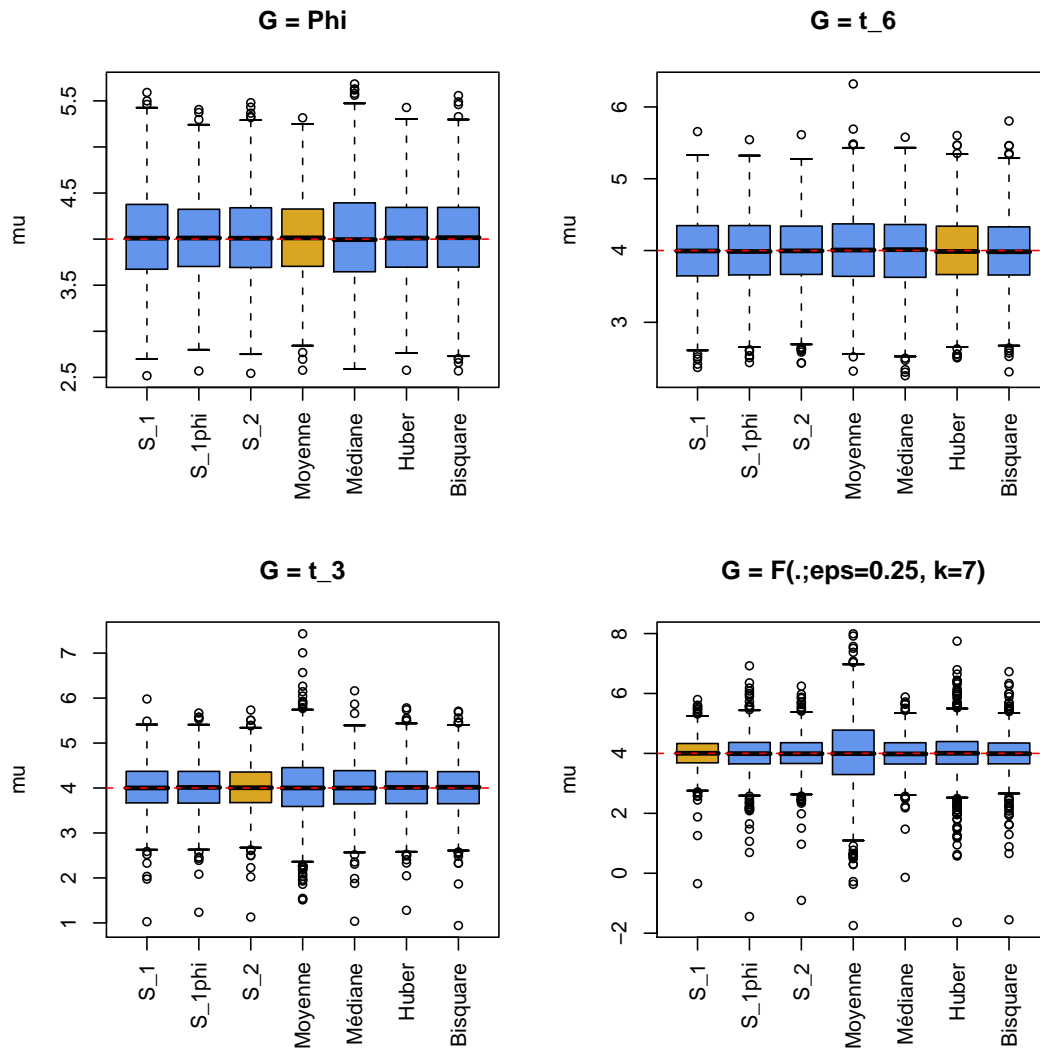


FIG. 3.10: Boxplots des résultats des simulations pour différentes loi sous-jacentes. En jaune, l'estimateur possédant l'efficacité relative maximale. La ligne rouge représente la vraie valeur du paramètre de lieu μ . Taille de l'échantillon : $n = 10$.

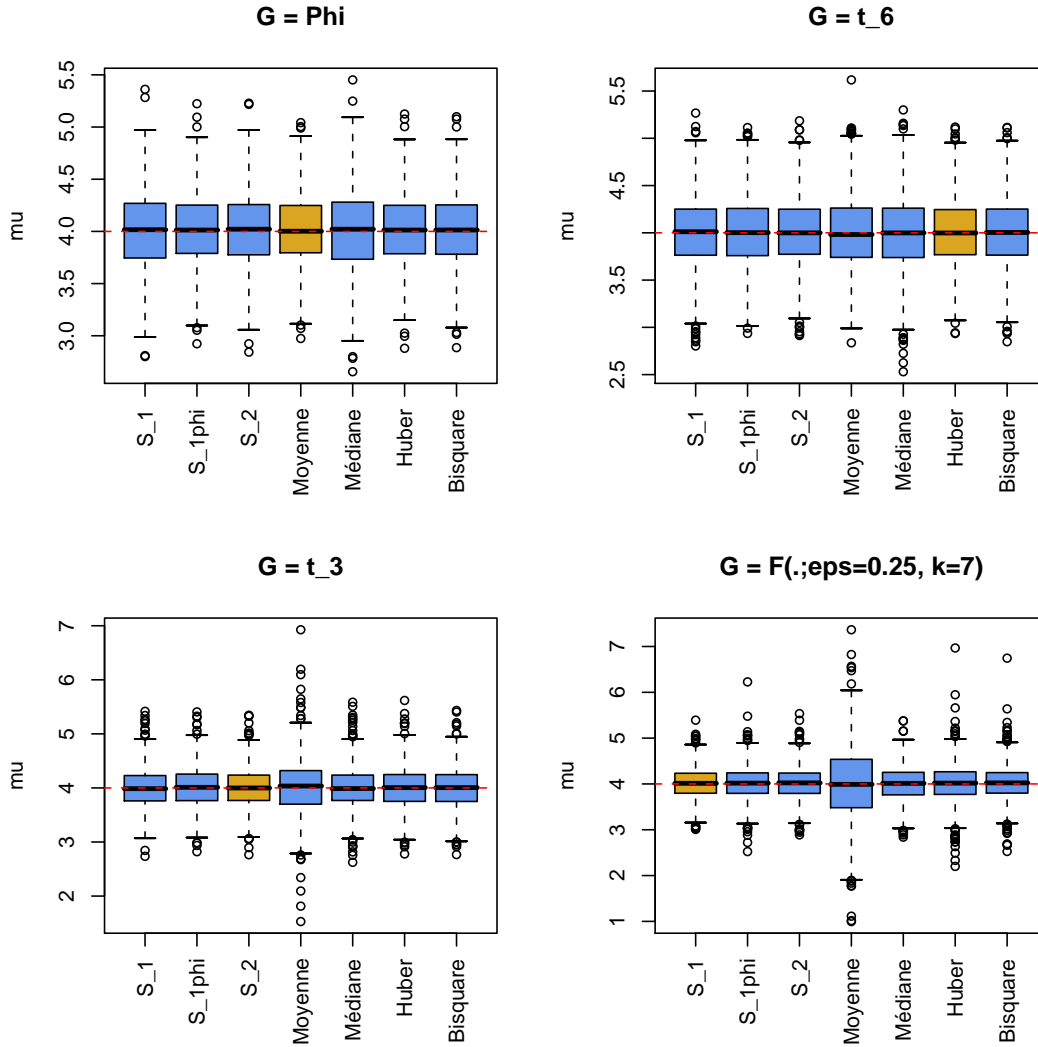


FIG. 3.11: Boxplots des résultats des simulations pour différentes loi sous-jacentes. En jaune, l'estimateur possédant l'efficacité relative maximale. La ligne rouge représente la vraie valeur du paramètre de lieu μ . Taille de l'échantillon : $n = 20$.

Les tables 3.16, 3.17 et 3.18 présentent les résultats de simulations de Monte-Carlo pour le paramètre d'échelle, similaires à celles réalisées pour le paramètre de lieu. Les estimateurs pris en considération sont les trois estimateurs de Pitman compromis S_1 , $S_{1\Phi}$ et S_2 , l'estimateur usuel de l'écart-type s , la déviation absolue médiane, et la distance interquartile empirique de l'échantillon.

3.4. Simulations dans le cas de lois t de Student

Estimateur	Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$	$\varepsilon = 0.25$	Slash
						$k = 2$	$k = 7$	
S_1	82.2	88.2	94.8	98.7	100	88.0	100	100
$S_{1\Phi}$	88.3	94.1	98.7	99.3	98.0	94.7	96.2	97.7
S_2	87.3	93.1	98.2	100	98.8	93.6	97.4	97.2
s	100	100	100	92.0	74.5	100	77.8	46.5
MAD	31.6	36.0	38.9	50.6	58.7	32.0	75.9	86.0
DIQ	36.0	36.7	41.6	51.4	58.3	35.0	69.3	81.7

TAB. 3.16: Efficacité relative (en %) du log des estimateurs d'échelle considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 5$.

Lorsque $G = \Phi$, l'estimateur habituel de l'écart-type présente à chaque fois la variance la plus faible. C'est également dans ce cas que les trois estimateurs compromis basés sur des lois t sont le moins performants. Leur efficacité relative chute lorsque la taille de l'échantillon augmente, et ce également pour $S_{1\Phi}$ qui contient pourtant la loi normale centrée et réduite.

Dans tous les autres cas, et pour toutes les tailles d'échantillon, les estimateurs compromis se comportent très bien. Parmi eux, S_2 est souvent le meilleur, notamment lorsque $n = 20$. A noter toutefois que c'est S_1 qui se comporte le mieux lorsque la distribution sous-jacente est la loi normale contaminée avec $\varepsilon = 0.25$ et $k = 7$.

Estimateur	Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$	$\varepsilon = 0.25$	Slash
						$k = 2$	$k = 7$	
S_1	78.0	88.3	95.3	99.0	100	85.1	100	86.5
$S_{1\Phi}$	86.9	93.0	97.9	98.2	98.4	92.1	89.4	86.7
S_2	85.2	94.2	100	100	99.5	91.3	93.2	86.2
s	100	100	96.6	67.1.0	44.5	100	68.9	18.0
MAD	38.7	44.5	52.1	65.9	76.6	45.8	99.4	100
DIQ	40.6	47.0	54.9	64.8	75.6	47.3	87.6	90.0

TAB. 3.17: Efficacité relative (en %) du log des estimateurs d'échelle considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 10$.

Estimateur	Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$ $k = 2$	$\varepsilon = 0.25$ $k = 7$	Slash
S_1	68.8	85.9	96.5	98.6	100	91.7	95.5	99.2
$S_{1\Phi}$	77.9	91.4	93.6	91.5	95.7	97.5	89.7	97.7
S_2	77.2	94.0	100	100	99.0	99.8	89.6	97.3
s	100	100	88.2	50.2	29.2	100	60.7	48.7
MAD	37.0	45.3	56.8	57.5	75.0	47.6	100	100
DIQ	39.0	47.2	58.9	58.7	76.2	51.1	96.1	86.6

TAB. 3.18: Efficacité relative (en %) du log des estimateurs d'échelle considérés, pour différentes distributions sous-jacentes. Taille de l'échantillon : $n = 20$.

3.5 Généralisation à la régression linéaire

Nous présentons dans ce qui suit la généralisation des estimateurs de Pitman au cas de la régression linéaire simple, ainsi que les résultats de simulations pour des estimateurs compromis basés sur des lois t de Student.

Considérons le modèle de régression linéaire habituel

$$\begin{aligned}
 \mathbf{y} &= \beta_0 + \beta_1 \mathbf{x}_1 + \dots + \beta_p \mathbf{x}_p + \sigma \cdot \mathbf{E} \\
 &= \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \sigma \begin{pmatrix} E_1 \\ \vdots \\ E_n \end{pmatrix} \\
 &= X\boldsymbol{\beta} + \sigma \cdot \mathbf{E},
 \end{aligned}$$

où

- $\mathbf{y} \in \mathbb{R}^n$ est le vecteur des réponses ;
- $\mathbf{x}_1, \dots, \mathbf{x}_p \in \mathbb{R}^n$ sont les vecteurs des covariables ;
- $X \in \mathbb{R}^{n \times (p+1)}$ est la matrice du plan d'expériences, ci-après appelée matrice de design ;
- $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p) \in \mathbb{R}^{p+1}$ sont les paramètres inconnus de la régression ;
- E_1, \dots, E_n est un échantillon indépendant provenant d'une loi F connue, supposée symétrique autour de 0.

Définition 3.5.1. Un estimateur, ou une statistique, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y})$ est dit(e) équivariant(e) pour la régression et l'échelle s'il (elle) vérifie la propriété suivante :

$$\hat{\boldsymbol{\beta}}(a\mathbf{y} + X\mathbf{b}) = \begin{pmatrix} \hat{\beta}_0(a\mathbf{y} + X\mathbf{b}) \\ \vdots \\ \hat{\beta}_p(a\mathbf{y} + X\mathbf{b}) \end{pmatrix} = a \begin{pmatrix} \hat{\beta}_0(\mathbf{y}) \\ \vdots \\ \hat{\beta}_p(\mathbf{y}) \end{pmatrix} + \mathbf{b} = a\hat{\boldsymbol{\beta}}(\mathbf{y}) + \mathbf{b},$$

pour tout $a > 0$, et pour tout $\mathbf{b} \in \mathbb{R}^{p+1}$.

Remarque 3.5.2. Soit $\hat{\boldsymbol{\beta}}$ un estimateur équivariant pour la régression, et considérons le cas où $p = 1$:

$$y_i = \beta_0 + \beta_1 x_i + \sigma \cdot E_i, \quad (i = 1, \dots, n),$$

où nous avons écrit x_i au lieu de x_{1i} .

Si les observations \mathbf{y} sont multipliées par une constante a , comme par exemple pour un changement d'unité de mesure, alors chaque estimateur $\hat{\beta}_k, k = 0, 1$, est multiplié par la même constante. Si une constante b_0 est ajoutée à toutes les observations, alors $\hat{\beta}_0$ est augmenté de cette même constante. Finalement, si l'on ajoute un multiple b_1 de x_i à chaque y_i , alors l'estimateur $\hat{\beta}_1$ est également augmenté de b_1 .

Définition 3.5.3. Un estimateur, ou une statistique $\hat{\sigma} = \hat{\sigma}(\mathbf{y})$ est dit(e) invariant(e) pour le lieu et équivariant(e) pour l'échelle s'il (elle) vérifie la propriété suivante :

$$\hat{\sigma}(a\mathbf{y} + X\mathbf{b}) = a\hat{\sigma}(\mathbf{y}),$$

pour tout $a > 0$ et pour tout $\mathbf{b} \in \mathbb{R}^{p+1}$.

Exemple 3.5.4. L'estimateur habituel des moindres carrés pour les paramètres de régression, donné par

$$\hat{\boldsymbol{\beta}}(\mathbf{y}) = (X^T X)^{-1} X^T \mathbf{y},$$

est un estimateur équivariant pour la régression et l'échelle.

De manière similaire à l'estimation du paramètre de lieu, nous introduisons alors la configuration \mathbf{c} , définie par

$$\mathbf{c} = \hat{\sigma}(\mathbf{y})^{-1} \left(\mathbf{y} - X\hat{\boldsymbol{\beta}}(\mathbf{y}) \right),$$

où $\hat{\boldsymbol{\beta}}$ est une statistique équivariante pour la régression et l'échelle, et $\hat{\sigma}$ est une statistique invariante pour le lieu et équivariante pour l'échelle.

Nous cherchons, parmi les estimateurs $\mathbf{T}(\mathbf{y}) = (T_0(\mathbf{y}), \dots, T_p(\mathbf{y}))$ de $\boldsymbol{\beta}$ équivariants pour la régression et l'échelle, celui qui minimise une quantité analogue au carré moyen de l'erreur, qui reste à déterminer. En effet, il convient de définir le carré moyen de l'erreur pour un vecteur de taille $p + 1$. Une possibilité est de considérer la trace de la matrice suivante :

$$M_F = \left[\mathbb{E}_F \left((\mathbf{T}(\mathbf{y}) - \boldsymbol{\beta}) (\mathbf{T}(\mathbf{y}) - \boldsymbol{\beta})^T \right) \right].$$

Nous posons alors le carré moyen de l'erreur de l'estimateur \mathbf{T} comme étant

$$\begin{aligned} \text{CME}_F(\mathbf{T}(\mathbf{y})) &= \text{Tr}(M_F) = \mathbb{E}_F \left((\mathbf{T}(\mathbf{y}) - \boldsymbol{\beta})^T (\mathbf{T}(\mathbf{y}) - \boldsymbol{\beta}) \right) \\ &= \sum_{k=0}^p \mathbb{E}_F \left((T_k(\mathbf{y}) - \beta_k)^2 \right). \end{aligned}$$

Remarque 3.5.5. Le choix de ce critère impose de manière implicite que les composantes du vecteur $\boldsymbol{\beta}$ sont telles que l'importance des erreurs est la même pour chacune. Autrement dit, un changement d'échelle des covariables X est nécessaire, afin que les β_k soient comparables.

En introduisant le système de coordonnées

$$\mathbf{y} = X\mathbf{b} + s\mathbf{c},$$

où $\mathbf{b} = (b_0, \dots, b_p) \in \mathbb{R}^{p+1}$, et $s > 0$, nous pouvons comme précédemment nous restreindre au carré moyen de l'erreur conditionnel

$$\text{cCME}_F(\mathbf{T}(\mathbf{y}) \mid \mathbf{c}) = \sum_{k=0}^p \mathbb{E}_F((T_k(\mathbf{y}) - \beta_k)^2 \mid \mathbf{c}).$$

En supposant sans perte de généralité que $\boldsymbol{\beta} = \mathbf{0}$, et en utilisant l'équivariance de \mathbf{T} , nous obtenons que l'estimateur de Pitman pour les paramètres de régression est donné par

$$\mathbf{T}_F(\mathbf{c}) = \begin{pmatrix} T_{F,0}(\mathbf{c}) \\ \vdots \\ T_{F,p}(\mathbf{c}) \end{pmatrix} = -\frac{1}{\mathbb{E}_F(s^2 \mid \mathbf{c})} \begin{pmatrix} \mathbb{E}_F(sb_0 \mid \mathbf{c}) \\ \vdots \\ \mathbb{E}_F(sb_p \mid \mathbf{c}) \end{pmatrix}.$$

Les espérances conditionnelles sont à calculer à l'aide de la distribution conditionnelle conjointe de s, b_0, \dots, b_p , sachant \mathbf{c} , donnée, à une constante près, par

$$f(s, b_0, \dots, b_p \mid \mathbf{c}) \propto s^{n-p-2} \prod_{i=1}^n f(sc_i + [X\mathbf{b}]_i),$$

où f est la densité de la loi F , et $[X\mathbf{b}]_i$ est la i -ème composante du vecteur $X\mathbf{b}$.

Remarque 3.5.6. Les espérances conditionnelles sont à évaluer en $p + 2$ dimensions, ce qui augmente grandement la difficulté de calcul lorsqu'une solution directe n'existe pas, comme c'est le cas pour la plupart des distributions F .

3.5.1 Estimateurs bi-optimaux pour la régression

O'Brien (1984) a généralisé la construction des estimateurs bi-optimaux présentée ci-dessus au cas de la régression linéaire. Son approche est tout à fait identique, définissant l'estimateur bi-optimal \mathbf{T} comme celui minimisant la quantité

$$\pi \text{CME}_{F_1}(\mathbf{T}) + (1 - \pi) \text{CME}_{F_2}(\mathbf{T}), \quad \pi \in [0, 1].$$

A nouveau, le résultat d'un tel problème est une combinaison linéaire des estimateurs de Pitman pour la régression correspondant aux distributions F_1 et F_2 .

Au lieu d'utiliser un compromis entre la loi normale et la loi Slash, l'estimateur de Pitman pour la régression associé à cette dernière étant impossible à calculer de manière directe, O'Brien propose un compromis basé sur une distribution semblable à la loi normale contaminée : la loi normale *k-wild*.

Définition 3.5.7. Un échantillon e_1, \dots, e_n provient d'une loi normale *k-wild* ($0 \leq k < n$) si $n - k$ des e_i proviennent d'une loi normale standard, et les k restants proviennent d'une loi normale d'écart-type 10.

Dans ce cas, les estimateurs de Pitman pour la régression sont directement calculables (voir aussi Fraser, 1979).

Nous proposons donc d'utiliser également la même approche que précédemment, en construisant un estimateur de Pitman compromis pour la régression comme suit :

$$\mathbf{T} = \frac{\sum_{j=1}^m w(F_j) \mathbf{T}_{F_j}}{\sum_{j=1}^m w(F_j)},$$

pour des distributions de compromis F_1, \dots, F_m appartenant à un ensemble de lois pré-défini \mathcal{F} , et pour $w(\cdot)$ une fonction de poids positive. Dans ce qui suit, nous présentons les résultats pour un estimateur compromis pour la régression basé sur des lois t de Student, de divers degrés de liberté.

3.5.2 Simulations dans le cas de lois t de Student

Dans cette section, nous présentons les résultats de simulations de Monte-Carlo, pour des estimateurs de Pitman compromis basés sur des lois t de Student. Comme auparavant, nous utiliserons le terme de stratégie afin de désigner un tel estimateur.

Nous allons comparer les mêmes stratégies que lors du cas du paramètre de lieu simple, à savoir les stratégies S_1 , $S_{1\Phi}$ et S_2 , construites de la manière suivante :

Stratégie	Lois
S_1	t_2
$S_{1\Phi}$	t_2 et Φ
S_2	t_2 et t_5

Les autres estimateurs pour la régression pris en compte sont :

- l'estimateur habituel des moindres carrés, noté LS ;
- l'estimateur L_1 , minimisant la somme des écarts absolus ;
- le M-estimateur associé à la fonction ψ de Huber ;
- le M-estimateur associé à la fonction du Bisquare de Tukey.

Remarque 3.5.8. Pour les M-estimateurs, les constantes et la méthode d'estimation de l'échelle sont les mêmes que dans la remarque 3.3.2.

Dans chacune des situations présentées ci-dessous, la performance de chaque estimateur $\hat{\beta}$ considéré sera donnée par les quantités suivantes : la moyenne de l'erreur L_1 , et la variance totale, définies ci-dessous.

Définition 3.5.9. Soit $\hat{\beta}$ un estimateur des paramètres de régression β , soit $i = 1, \dots, N$ l'indice d'un échantillon, et notons $\hat{\beta}_i = (\hat{\beta}_{0,i}, \dots, \hat{\beta}_{p,i})$ la valeur prise par $\hat{\beta}$ dans le cas de l'échantillon i . Nous définissons la moyenne de l'erreur L_1 totale de $\hat{\beta}$ par :

$$\text{MET}_1(\hat{\beta}) = \sum_{k=1}^p \frac{1}{N} \sum_{i=1}^N \left(|\hat{\beta}_{k,i} - \beta_k| \right).$$

Pour une collection d'estimateurs \mathbf{T}_l , $l = 1, \dots, L$ de β , nous définissons l'excès relatif de l'erreur L_1 totale de \mathbf{T}_j par :

$$\text{ER}_1(\mathbf{T}_j) = \frac{\text{MET}_1(\mathbf{T}_j)}{\min_{l=1, \dots, L} \text{MET}_1(\mathbf{T}_l)} - 1.$$

L'excès relatif de l'erreur L_1 totale représente donc une standardisation de la moyenne de l'erreur L_1 totale par rapport à l'estimateur présentant l'erreur L_1 totale la plus faible.

Définition 3.5.10. Soit $\hat{\beta}$ un estimateur des paramètres de régression β , soit $i = 1, \dots, N$ l'indice d'un échantillon, et notons $\hat{\beta}_i = (\hat{\beta}_{0,i}, \dots, \hat{\beta}_{p,i})$ la valeur prise par $\hat{\beta}$ dans le cas de l'échantillon i . La variance totale de $\hat{\beta}$ est définie par :

$$\text{TotVar}(\hat{\beta}) = \sum_{k=1}^p \frac{1}{N-1} \sum_{i=1}^N \left(\hat{\beta}_{k,i} - \bar{\hat{\beta}}_k \right)^2,$$

où $\bar{\hat{\beta}}_k = N^{-1} \sum_{i=1}^N \hat{\beta}_{k,i}$.

Pour une collection d'estimateurs \mathbf{T}_l , $l = 1, \dots, L$ de β , nous définissons l'efficacité relative totale de \mathbf{T}_j par :

$$\text{EFFT}(\mathbf{T}_j) = \frac{\min_{l=1, \dots, L} \text{TotVar}(\mathbf{T}_l)}{\text{TotVar}(\mathbf{T}_j)}.$$

L'efficacité relative totale représente donc une standardisation de la variance totale d'un estimateur, par rapport à l'estimateur présentant la variance totale la plus faible.

Designs fixes dans le cas de la régression linéaire simple

La performance des estimateurs de régression dépend de manière significative du design X . Dès lors, il convient d'étudier le comportement des différents estimateurs considérés pour des designs fixes, choisis à l'avance. Dans ce qui suit, nous nous intéresserons plus particulièrement au cas de la régression simple, pour des échantillons de taille $n = 20$. Les designs que nous utiliserons, également présentés dans Morgenthaler et Tukey (1991), sont symétriques autour de 0, et dès lors l'estimation d'un β_k n'aura qu'une influence minimale sur l'autre et pourra être négligée. Dans chaque cas, 1000 échantillons seront générés.

Design A		Design B		Design C	
Valeur x_i	h_{ii}	Valeur x_i	h_{ii}	Valeur x_i	h_{ii}
±2.375	0.1857	±2.25	0.1129	±2.00	0.100
±2.125	0.1586	±2.25	0.1129	±2.00	0.100
±1.875	0.1346	±2.00	0.0997	±2.00	0.100
±1.625	0.1135	±2.00	0.0997	±2.00	0.100
±1.375	0.0955	±2.00	0.0997	±2.00	0.100
±1.125	0.0805	±2.00	0.0997	±2.00	0.100
±0.875	0.0684	±2.00	0.0997	±2.00	0.100
±0.625	0.0594	±2.00	0.0997	±2.00	0.100
±0.375	0.0534	±1.75	0.0880	±2.00	0.100
±0.125	0.0504	±1.75	0.0880	±2.00	0.100
Design D		Design E		Design F	
Valeur x_i	h_{ii}	Valeur x_i	h_{ii}	Valeur x_i	h_{ii}
±2.74084	0.2994	±7.375	0.4411	±12.375	0.5050
±1.82455	0.1605	±2.125	0.0825	±2.125	0.0634
±1.35455	0.1109	±1.875	0.0753	±1.875	0.0604
±1.03609	0.0856	±1.625	0.0690	±1.625	0.0578
±0.79493	0.0710	±1.375	0.0636	±1.375	0.0556
±0.60077	0.0620	±1.125	0.0591	±1.125	0.0538
±0.43825	0.0564	±0.875	0.0555	±0.875	0.0523
±0.29849	0.0530	±0.625	0.0528	±0.625	0.0512
±0.17589	0.0510	±0.375	0.0510	±0.375	0.0504
±0.06669	0.0501	±0.125	0.0501	±0.125	0.0500
Design G		Design H		Design I	
Valeur x_i	h_{ii}	Valeur x_i	h_{ii}	Valeur x_i	h_{ii}
±12.375	0.4069	±12.375	0.2964	±1.84174	0.2461
±7.125	0.1683	±12.125	0.2865	±1.39220	0.1621
±1.875	0.0582	±1.875	0.0557	±1.12105	0.1227
±1.625	0.0562	±1.625	0.0542	±0.91401	0.0983
±1.375	0.0544	±1.375	0.0530	±0.74022	0.0817
±1.125	0.0529	±1.125	0.0520	±0.58638	0.0699
±0.875	0.0518	±0.875	0.0512	±0.44537	0.0615
±0.625	0.0509	±0.625	0.0506	±0.31280	0.0557
±0.375	0.0503	±0.375	0.0502	±0.18563	0.0520
±0.125	0.0500	±0.125	0.0501	±0.06152	0.0502

TAB. 3.19: Designs fixes et valeurs diagonales correspondantes de la matrice $X(X^T X)^{-1} X^T$.

Les 9 designs présentés dans la table 3.19 ont été construits afin de couvrir un large spectre, allant de situations où aucun point de levier n'existe, à des situations à points de levier multiples. Chaque point $x_i, i = 1, \dots, 20$ du design est accompagné par la valeur diagonale correspondante de la matrice $X(X^T X)^{-1} X^T$, notée h_{ii} . La taille de l'échantillon étant $n = 20$, la valeur moyenne de la diagonale de cette matrice est $p/n = 0.1$. Les points tels que la valeur diagonale est supérieure à $2p/n = 0.2$ peuvent être considérés comme des points de levier. Pour plus de détails concernant les points de levier, voir Belsley *et al.* (1980).

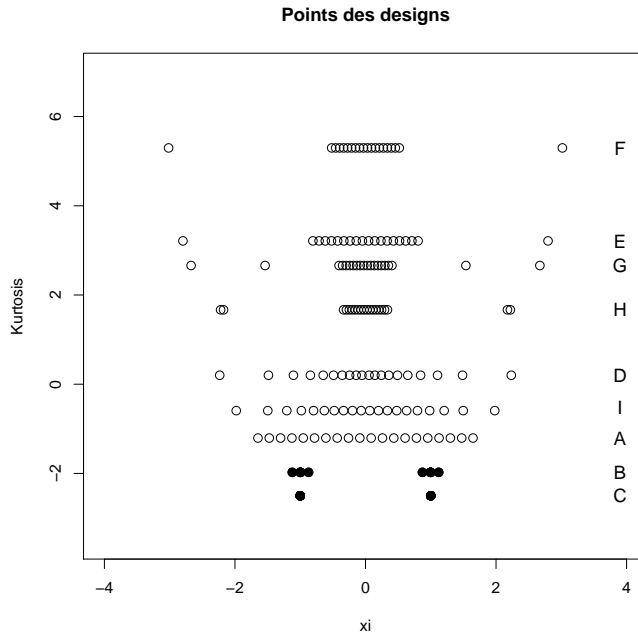


FIG. 3.12: Les différents designs et leur kurtosis respective. Les designs ont été remis à l'échelle de telle sorte que leur deuxième moment soit 1. Le design C a été dessiné à -2.5 au lieu de -2 , pour des raisons de lisibilité. Les points en noir sont des points multiples.

Le design A est composé de points répartis de manière régulière. Les designs B et C sont formés de deux groupes distincts de points. Le design D quant à lui est constitué des médianes des statistiques d'ordre pour une loi double-exponentielle, de même que le design I pour une loi normale. Les designs E, F, G et H sont des modifications du design A, qui contiennent une ou plusieurs paires de points de levier symétriques.

La table 3.20 présente les résultats pour les designs fixes, dans le cas où la loi sous-jacente des erreurs G est la loi normale standard Φ . Dans cette situation, l'estimateur des moindres carrés (LS) est le meilleur choix, quel que soit le design. Parmi les estimateurs de Pitman compromis pour la régression, $S_{1\Phi}$ est celui qui se comporte le mieux, présentant une erreur L_1 relative de l'ordre de 3% au maximum, et une perte d'efficacité totale relative comprise entre 5 et 7%. Ceci peut s'expliquer par le fait que la taille de l'échantillon, $n = 20$, est relativement importante, et ainsi le poids attribué à Φ dans le compromis $S_{1\Phi}$ est relativement important et proche de 1. S_2 se comporte quant à lui mieux que S_1 , puisqu'il comporte une distribution plus proche de Φ .

3.5. Généralisation à la régression linéaire

Estim.		A	B	C	D	E	F	G	H	I
S_1	ER ₁	11.5	13.1	11.2	11.1	9.2	5.6	9.3	7.4	11.9
	EFFT	82.4	77.2	79.3	82.2	82.9	88.6	83.8	88.5	80.8
$S_{1\Phi}$	ER ₁	2.5	3.3	2.7	2.4	3.2	1.2	2.9	1.6	3.4
	EFFT	95.1	93.0	94.0	95.6	95.1	95.1	96.9	95.8	93.8
S_2	ER ₁	6.6	7.7	6.7	6.1	5.6	2.5	5.7	4.1	7.2
	EFFT	89.3	84.8	86.9	89.4	89.8	92.9	91.2	92.0	87.5
LS	ER ₁	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	EFFT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Huber	ER ₁	1.5	2.4	2.1	1.0	1.2	0.0	2.1	1.6	2.3
	EFFT	97.2	95.1	96.1	97.4	98.0	97.5	96.9	95.8	95.9
Tukey	ER ₁	4.0	5.3	5.2	5.9	2.8	3.1	5.7	4.1	4.7
	EFFT	92.9	89.2	90.1	88.8	92.4	90.7	88.6	92.0	91.5
L1	ER ₁	23.4	27.9	30.0	31.4	19.9	30.2	31.4	28.7	29.2
	EFFT	66.4	58.4	58.9	59.3	66.4	57.4	58.5	59.0	61.6

TAB. 3.20: Excès relatif de l'erreur L_1 totale (ER₁) et efficacité relative totale (EFFT) (en %) pour les estimateurs considérés et les différents designs. Loi sous-jacente : $G = \Phi$.

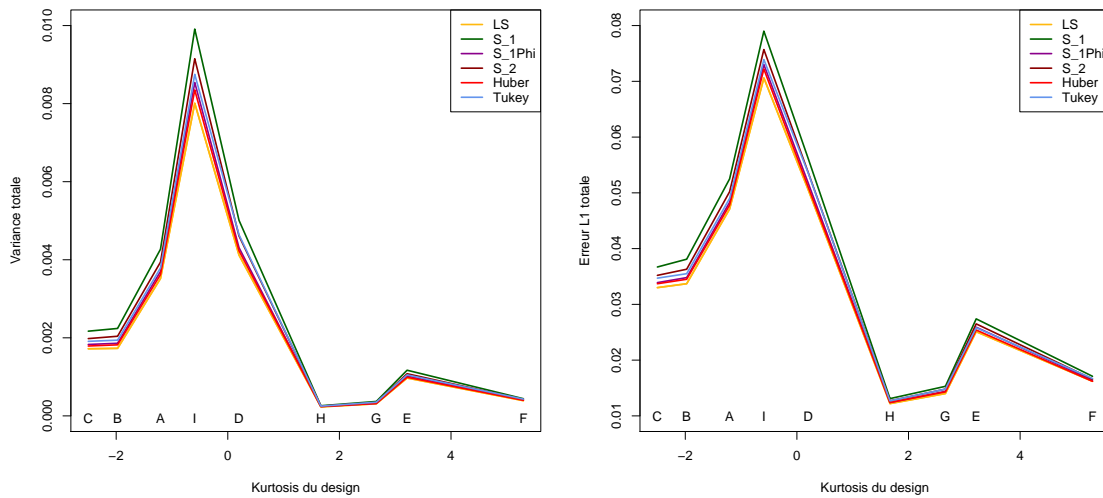


FIG. 3.13: Variance totale et erreur L_1 totale en fonction de la kurtosis des designs. Distribution sous-jacente $G = \Phi$. Le design C a été dessiné à -2.5 au lieu de -2 , pour des raisons de lisibilité.

La figure 3.13 montre le comportement de la variance totale et de l'erreur L_1 totale des estimateurs considérés en fonction de la kurtosis des designs (l'estimateur L_1 a été ignoré pour des raisons de lisibilité). Nous pouvons remarquer que le comportement est identique pour chaque estimateur. La variance totale de chacun présente une augmentation significative pour les designs A, I et D, et est globalement stable pour les

autres designs, et la présence de points de levier ne se fait sentir dans ce cas que pour les designs E et F.

La table 3.21 présente les résultats obtenus dans le cas des designs fixes lorsque la distribution sous-jacente est la loi t de Student avec 6 degrés de liberté, une distribution possédant des queues modérées. Dans ce cas, nous observons que l'estimateur des moindres carrés présente une perte d'efficacité totale de l'ordre de 10 à 12% au maximum, par rapport au M-estimateur de Huber, qui est le meilleur dans la grande majorité des cas. Les trois estimateurs compromis se comportent relativement bien. $S_{1\Phi}$ et S_2 sont très similaires, et ceci peut s'expliquer par le fait que dans la stratégie $S_{1\Phi}$ un poids relativement important semble être donné à la distribution Φ , qui est relativement proche de la distribution t_6 , tandis que dans la stratégie S_2 , c'est vraisemblablement la distribution t_5 qui a le plus d'importance. La stratégie S_1 est quant à elle légèrement en retrait, puisqu'elle ne contient que la distribution t_2 , qui possède des queues lourdes. La perte d'efficacité totale et l'erreur L_1 totale sont du même ordre que l'estimateur des moindres carrés.

Estim.		A	B	C	D	E	F	G	H	I
S_1	ER ₁	3.3	6.0	3.9	2.9	3.3	2.0	2.5	1.5	3.5
	EFFT	93.3	89.3	92.0	93.5	92.4	94.1	92.9	97.3	93.1
$S_{1\Phi}$	ER ₁	0.0	1.4	0.6	0.5	0.0	0.0	0.0	0.8	0.3
	EFFT	100.0	97.9	99.0	98.7	100.0	100.0	100.0	99.0	99.1
S_2	ER ₁	0.8	2.6	1.4	0.9	1.1	0.6	0.6	0.0	0.7
	EFFT	98.2	96.0	97.5	98.2	97.6	98.0	97.5	100.0	98.0
LS	ER ₁	3.3	5.7	5.4	4.3	1.5	2.3	1.9	3.8	5.2
	EFFT	90.3	90.1	88.3	90.5	96.0	96.0	95.1	90.7	90.2
Huber	ER ₁	1.0	0.0	0.0	0.0	0.4	0.6	0.0	0.0	0.0
	EFFT	98.5	100.0	100.0	100.0	99.2	100.0	100.0	99.6	100.0
Tukey	ER ₁	3.3	2.3	1.1	2.1	2.2	3.5	1.9	0.8	1.7
	EFFT	93.1	96.0	97.0	94.9	95.3	87.3	92.9	97.3	95.9
L1	ER ₁	14.5	23.1	22.0	16.6	17.9	22.2	12.6	23.4	18.2
	EFFT	75.1	66.1	66.2	72.8	72.0	63.2	75.0	66.4	72.2

TAB. 3.21: Excès relatif de l'erreur L_1 totale (ER₁) et efficacité relative totale (EFFT) (en %) pour les estimateurs considérés et les différents designs. Loi sous-jacente : $G = t_6$.

La figure 3.14 montre le comportement de la variance totale et de l'erreur L_1 totale en fonction de la kurtosis des designs (l'estimateur L_1 a été ignoré pour des raisons de lisibilité). A nouveau, le comportement global est le même pour chaque estimateur.

Finalement, la table 3.22 présente les résultats dans le cas où la loi sous-jacente des erreurs est la distribution de Student avec 2 degrés de liberté. Dans ce cas, l'estimateur de Pitman S_1 basé sur cette même distribution est le meilleur choix, sauf dans le cas du design A. L'estimateur compromis S_2 est quant à lui relativement proche de S_1 en

termes de performances, ne présentant qu'environ 3% d'erreur L_1 supplémentaires, et une perte d'efficacité relative de l'ordre de 8% au maximum. Comme t_2 est utilisée comme distribution de compromis dans S_2 , ces résultats sont parfaitement compréhensibles et étaient attendus. Par contre, la performance de l'estimateur compromis $S_{1\Phi}$ est plus étonnante. En effet, la distribution t_2 est également utilisée dans ce compromis, et pourtant l'estimateur présente une efficacité totale relative d'environ 50% seulement, ainsi que 30% d'erreur L_1 supplémentaires. Pourtant, dans chacun des autres cas, il nous avait semblé que le nombre d'observations était suffisamment important afin de favoriser significativement la distribution de compromis la plus proche de la distribution sous-jacente réelle.

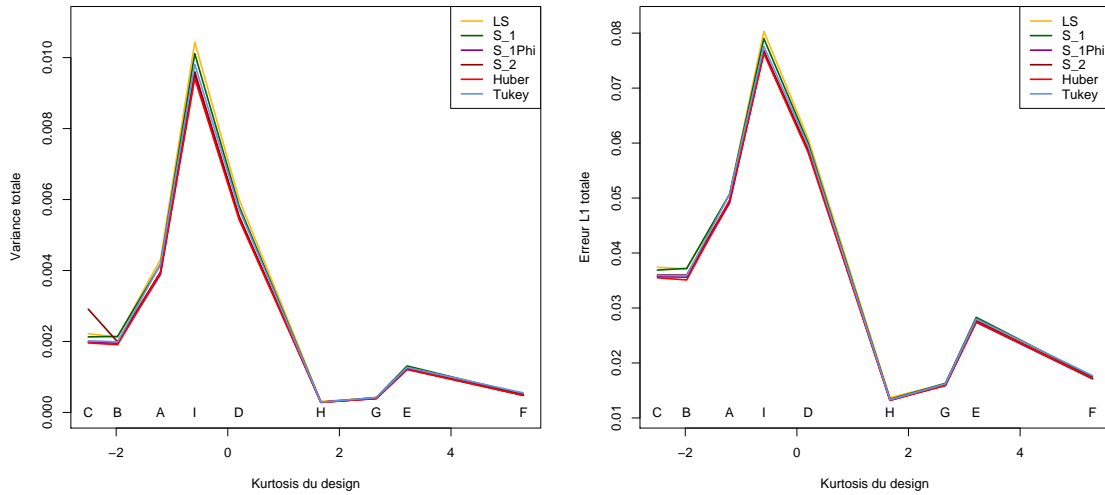


FIG. 3.14: Variance totale et erreur L_1 totale en fonction de la kurtosis des designs. Distribution sous-jacente $G = t_6$. Le design C a été dessiné à -2.5 au lieu de -2 , pour des raisons de lisibilité.

Estim.		A	B	C	D	E	F	G	H	I
S_1	ER ₁	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	EFFT	99.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$S_{1\Phi}$	ER ₁	22.3	31.3	23.3	24.7	13.5	16.1	24.1	21.5	24.9
	EFFT	59.2	42.9	59.8	55.4	70.1	60.2	42.9	61.1	49.4
S_2	ER ₁	0.0	1.3	0.6	1.3	0.6	2.4	3.4	3.5	1.6
	EFFT	100.0	96.9	98.7	98.0	97.4	97.3	92.7	94.3	97.5
LS	ER ₁	67.5	88.9	73.9	71.6	42.5	39.6	58.9	54.0	68.6
	EFFT	24.8	15.5	25.8	21.1	47.9	27.0	17.9	28.7	19.3
Huber	ER ₁	6.2	11.8	7.3	9.6	7.2	8.3	10.7	9.1	8.8
	EFFT	90.5	79.5	86.1	83.6	82.0	83.5	76.1	84.6	84.7
Tukey	ER ₁	4.3	6.7	6.4	4.9	4.7	6.1	5.7	2.5	6.1
	EFFT	93.9	87.9	90.2	90.0	86.7	79.8	83.6	97.1	87.7
L1	ER ₁	15.2	16.6	15.3	10.2	10.9	11.6	11.9	15.8	11.9
	EFFT	75.8	72.6	74.1	77.9	81.1	81.6	72.9	71.7	79.3

TAB. 3.22: Excès relatif de l'erreur L_1 totale (ER₁) et efficacité relative totale (EFFT) (en %) pour les estimateurs considérés et les différents designs. Loi sous-jacente : $G = t_2$.

La figure 3.15 montre le comportement de la variance totale et de l'erreur L_1 totale de chaque estimateur en fonction de la kurtosis des designs. Comme auparavant, le comportement global est similaire dans chaque cas. On remarque cette fois-ci que les estimateurs des moindres carrés et $S_{1\Phi}$ sont nettement moins performants.

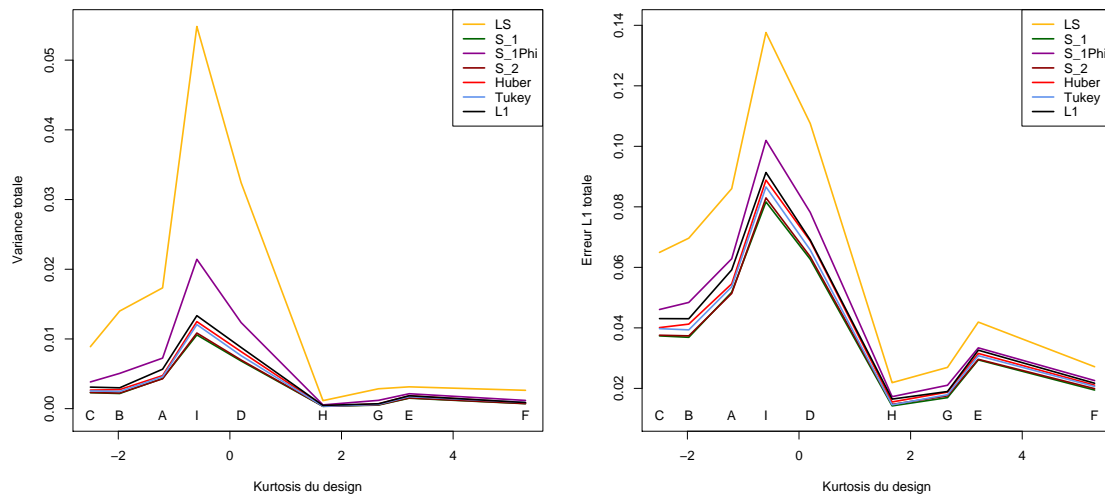


FIG. 3.15: Variance totale et erreur L_1 totale en fonction de la kurtosis des designs. Distribution sous-jacente $G = t_2$. Le design C a été dessiné à -2.5 au lieu de -2 , pour des raisons de lisibilité.

Finalement, la table 3.23 donne les estimations de l'écart-type des quantités présentées dans la table 3.22, obtenues grâce à la méthode du *jackknife*. On remarque que l'estimateur de Pitman compromis $S_{1\Phi}$ semble extrêmement sensible lorsqu'on le compare aux deux autres estimateurs compromis.

Estim.		A	B	C	D	E	F	G	H	I
S_1	ER ₁	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	EFFT	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$S_{1\Phi}$	ER ₁	3.1	4.3	3.0	3.1	2.3	3.1	4.0	2.8	3.6
	EFFT	4.9	11.7	3.5	5.2	3.9	10.2	9.3	5.3	7.4
S_2	ER ₁	0.0	0.7	0.6	0.6	0.6	0.7	0.7	0.6	0.5
	EFFT	0.0	1.2	1.0	1.3	1.4	1.8	1.6	1.2	1.0
LS	ER ₁	6.4	8.8	6.2	7.2	4.7	6.6	8.3	5.7	7.5
	EFFT	3.4	6.8	2.5	3.8	4.1	10.6	6.2	5.0	4.9
Huber	ER ₁	1.1	1.8	1.6	1.7	1.6	1.8	1.7	1.4	1.5
	EFFT	1.7	2.5	2.5	2.9	2.9	4.2	3.2	2.4	2.3
Tukey	ER ₁	1.1	1.3	1.2	1.4	1.4	2.1	1.4	1.2	1.2
	EFFT	1.8	2.1	2.1	3.3	3.3	7.6	4.7	2.1	2.1
L1	ER ₁	1.9	2.1	1.7	1.9	2.0	2.2	2.1	2.1	1.8
	EFFT	2.4	2.3	2.1	2.8	3.1	3.4	3.7	2.8	2.6

TAB. 3.23: Ecart-type estimé de l'excès relatif de l'erreur L_1 totale (ER₁) et de l'efficacité relative totale (EFFT) (en %) pour les estimateurs considérés et les différents designs. Loi sous-jacente : $G = t_2$.

Designs aléatoires

Dans cette section, nous présentons les résultats de simulations de Monte-Carlo pour des designs générés aléatoirement, avec des échantillons de tailles différentes et diverses distributions sous-jacentes. Nous présentons à nouveau le cas de la régression linéaire simple ($p = 1$), mais également le cas de la régression linéaire multiple ($p = 3$). Dans chaque cas, 1000 échantillons ont été générés. Les points du design ont quant à eux été générés de manière indépendante suivant une loi normale standard :

$$x_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1), \quad \forall i = 1, \dots, p, \quad \forall j = 1, \dots, n.$$

La table 3.24 présente les résultats dans le cas de la régression linéaire simple ($p = 1$) pour des échantillons de taille $n = 10$. Dans le cas où $G = \Phi$, l'estimateur des moindres carrés LS présente les meilleures performances, tandis que les estimateurs de Pitman compromis $S_{1\Phi}$ et S_2 se comportent relativement bien. La stratégie S_1 est nettement moins performante, puisqu'elle ne contient que la distribution t_2 , une loi à queues relativement lourdes. Pour des lois sous-jacentes avec des queues modérées, comme les lois t_{10} , t_6 ou encore la loi normale contaminée avec $\varepsilon = 0.05$ et $k = 2$, les estimateurs de Pitman compromis se comportent de manière très satisfaisante, seul S_1 étant quelque peu en difficulté dans le dernier cas. Dans le cas d'une loi à queues lourdes, comme

les lois t_3, t_2 ou la loi normale contaminée avec $\varepsilon = 0.25$ et $k = 7$, c'est alors $S_{1\Phi}$ qui est clairement la moins performante des stratégies de compromis. S_1 est la plus performante, tandis que S_2 présente tout de même un excès d'erreur L_1 de l'ordre de 7% dans le cas de la loi normale contaminée.

Estim.							$\varepsilon = 0.05$	$\varepsilon = 0.25$
		Φ	t_{10}	t_6	t_3	t_2	$k = 2$	$k = 7$
S_1	ER ₁	7.3	4.6	2.6	0.0	0.0	6.2	0.0
	EFFT	87.4	90.4	94.0	100.0	100.0	88.7	100.0
$S_{1\Phi}$	ER ₁	2.2	0.4	0.0	2.5	13.4	1.4	25.8
	EFFT	96.1	98.1	99.4	90.4	66.0	96.7	69.3
S_2	ER ₁	4.4	2.2	0.8	0.4	2.1	3.4	7.4
	EFFT	92.0	94.6	97.9	96.1	93.0	93.5	88.4
LS	ER ₁	0.0	0.9	2.0	12.5	36.9	0.6	63.0
	EFFT	100.0	98.7	95.1	70.6	37.3	97.0	42.6
Huber	ER ₁	0.9	0.0	0.1	4.2	10.4	0.0	24.2
	EFFT	98.5	100.0	100.0	81.8	74.9	100.0	65.0
Tukey	ER ₁	4.2	3.7	2.9	3.0	5.5	3.6	9.0
	EFFT	92.4	91.2	92.1	84.9	85.3	93.1	78.6
L1	ER ₁	24.5	23.3	18.0	5.0	9.0	25.4	5.9
	EFFT	62.7	65.1	69.1	79.3	84.8	63.8	90.0

TAB. 3.24: Excès relatif de l'erreur L_1 totale (ER₁) et efficacité relative totale (EFFT) (en %) pour les estimateurs considérés pour différentes lois sous-jacentes des erreurs. Cas de la régression linéaire simple ($p = 1$). Taille de l'échantillon $n = 10$.

La table 3.25 présente les résultats dans le cas de la régression linéaire simple ($p = 1$) pour des échantillons de taille $n = 20$. Globalement, les résultats sont comparables à ceux obtenus pour des échantillons plus petits. Les estimateurs de Pitman compromis se comportent bien de manière générale, sauf la stratégie $S_{1\Phi}$ dans les cas où $G = t_2$ ou $G = F(\cdot; 0.25, 7)$. Il semble que le fait de doubler la taille de l'échantillon ne résulte pas en une augmentation significative des performances des stratégies comprenant plusieurs distributions. Dans les cas extrêmes, comme $G = \Phi$ ou $G = t_2$, il semble que les stratégies $S_{1\Phi}$ et S_2 ne bénéficient pas de l'augmentation de la taille de l'échantillon. En effet, nous pouvons nous attendre à ce que la distribution la plus proche de G dans chacun des cas prenne le dessus sur la seconde distribution de compromis, ce qui ne semble pas être le cas. Cette situation se retrouve également lorsque $G = F(\cdot; 0.25, 7)$. Finalement, on trouvera dans la table 3.26 les écarts-types des quantités présentées, estimés à l'aide de la méthode du *jackknife*. Les écarts-types sont comparables lorsque la taille de l'échantillon vaut $n = 10$.

Estim.		Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$	$\varepsilon = 0.25$
							$k = 2$	$k = 7$
S_1	ER ₁	7.4	5.2	2.5	0.9	0.0	8.9	0.0
	EFFT	85.5	90.8	95.8	97.9	100.0	85.6	100.0
$S_{1\Phi}$	ER ₁	1.4	0.6	0.3	6.3	26.8	1.8	50.4
	EFFT	96.9	99.0	99.7	88.8	32.7	97.8	44.6
S_2	ER ₁	3.9	2.4	0.0	0.0	1.2	5.0	11.2
	EFFT	91.7	95.9	99.8	100.0	97.8	92.1	77.2
LS	ER ₁	0.0	1.3	5.1	25.8	71.6	0.0	124.8
	EFFT	100.0	97.1	90.2	62.1	10.5	99.8	19.3
Huber	ER ₁	0.3	0.0	0.0	4.6	7.9	0.4	31.8
	EFFT	99.0	100.0	100.0	92.1	85.1	100.0	49.5
Tukey	ER ₁	1.7	1.8	1.5	2.6	4.5	3.7	9.0
	EFFT	92.7	95.9	98.3	95.4	91.8	92.8	71.7
L1	ER ₁	23.7	15.7	15.0	15.6	10.5	21.4	16.3
	EFFT	63.5	74.8	76.6	76.1	80.2	67.2	75.7

TAB. 3.25: Excès relatif de l'erreur L_1 totale (ER₁) et efficacité relative totale (EFFT) (en %) pour les estimateurs considérés pour différentes lois sous-jacentes des erreurs. Cas de la régression linéaire simple ($p = 1$). Taille de l'échantillon $n = 20$.

Estim.		Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$	$\varepsilon = 0.25$
							$k = 2$	$k = 7$
S_1	ER ₁	1.6	1.2	1.0	0.8	0.0	1.8	0.0
	EFFT	2.5	1.8	2.1	1.1	0.0	2.0	0.0
$S_{1\Phi}$	ER ₁	0.8	0.5	0.1	1.2	5.3	0.9	3.3
	EFFT	1.5	0.9	0.9	1.9	12.7	1.1	2.1
S_2	ER ₁	1.3	0.8	0.6	0.0	0.6	1.5	1.1
	EFFT	2.1	1.2	1.5	0.0	1.0	1.5	2.2
LS	ER ₁	0.0	0.8	1.0	2.8	11.2	0.0	6.7
	EFFT	0.0	1.6	1.8	2.9	5.9	1.4	1.2
Huber	ER ₁	0.7	0.0	0.5	1.0	1.6	0.9	2.7
	EFFT	1.1	0.0	0.1	1.9	2.5	0.3	3.2
Tukey	ER ₁	1.0	0.7	1.0	1.1	1.2	1.4	1.9
	EFFT	1.7	1.6	1.5	2.4	2.0	1.8	5.7
L1	ER ₁	2.7	2.3	2.1	2.0	1.7	2.8	2.3
	EFFT	2.6	2.8	2.7	2.9	2.5	2.6	3.9

TAB. 3.26: Ecart-type estimé de l'excès relatif de l'erreur L_1 totale (ER₁) et de l'efficacité relative totale (EFFT) (en %) pour les estimateurs considérés pour différentes lois sous-jacentes des erreurs. Cas de la régression linéaire simple ($p = 1$). Taille de l'échantillon $n = 20$.

La table 3.27 présente les résultats des simulations dans le cas de la régression linéaire multiple ($p = 3$), pour des échantillons de taille $n = 10$. Lorsque $G = \Phi$, l'estima-

teur des moindres carrés présente les meilleures performances. Parmi les estimateurs de Pitman compromis, $S_{1\Phi}$ est celui qui se comporte le mieux dans ce cas, tandis que S_1 et S_2 présentent un excès relatif d'erreur L_1 de l'ordre de 7 et 5% respectivement, ainsi qu'une perte d'efficacité relative totale de l'ordre de 15 et 10% respectivement. Pour une loi sous-jacente possédant des queues modérées, les trois estimateurs de Pitman compromis se comportent de manière similaire, bien que S_1 reste le moins bon des trois. Lorsque la loi sous-jacente possède des queues lourdes, on remarque cette fois-ci que $S_{1\Phi}$ et S_2 se comportent nettement mieux que dans le cas de la régression linéaire simple, et ce même avec $n = 10$ seulement. En effet, lorsque $G = t_2$, les estimateurs $S_{1\Phi}$ et S_2 ne présentent qu'un excès relatif d'erreur L_1 de 1% au maximum, ainsi qu'une perte d'efficacité relative de 2%.

La table 3.28 présente les résultats dans le cas de la régression linéaire multiple ($p = 3$), pour des échantillons de taille $n = 20$. A nouveau, les performances des divers estimateurs sont comparables aux autres situations. Néanmoins, nous remarquons à nouveau que l'augmentation de la taille de l'échantillon n'a pas l'effet escompté sur la performance des estimateurs compromis comprenant plusieurs distributions, bien que ces derniers se comportent bien dans la plupart des situations. La table 3.29 quant à elle présente les estimations des écarts-types de ces quantités, obtenues grâce à la méthode du *jackknife*.

Estim.		Φ	t_{10}	t_6	t_3	t_2	$\varepsilon = 0.05$	$\varepsilon = 0.25$
							$k = 2$	$k = 7$
S_1	ER ₁	7.3	4.9	2.3	0.1	0.0	3.8	0.0
	EFFT	86.4	91.4	93.9	99.3	99.4	91.1	100.0
$S_{1\Phi}$	ER ₁	3.9	2.3	0.1	0.4	0.8	1.5	4.2
	EFFT	92.7	95.8	98.6	99.1	98.8	96.0	94.7
S_2	ER ₁	4.8	3.0	0.5	0.0	0.2	2.0	2.7
	EFFT	91.2	94.8	97.8	100.0	100.0	94.9	97.1
LS	ER ₁	0.0	0.2	1.5	10.0	35.6	0.2	39.0
	EFFT	100.0	99.7	96.7	78.2	25.8	99.6	61.4
Huber	ER ₁	0.3	0.0	0.0	5.1	17.3	0.0	25.6
	EFFT	99.5	100.0	100.0	87.7	33.1	100.0	71.1
Tukey	ER ₁	8.8	7.8	3.7	3.4	11.3	5.1	9.2
	EFFT	85.2	85.8	92.8	91.6	35.5	88.8	80.7
L1	ER ₁	20.2	19.1	13.3	8.8	8.7	15.7	5.5
	EFFT	68.4	70.2	76.2	84.6	82.1	74.0	83.5

TAB. 3.27: Excès relatif de l'erreur L_1 totale (ER₁) et efficacité relative totale (EFFT) (en %) pour les estimateurs considérés pour différentes lois sous-jacentes des erreurs. Cas de la régression linéaire multiple ($p = 3$). Taille de l'échantillon $n = 10$.

Estim.							$\varepsilon = 0.05$	$\varepsilon = 0.25$
		Φ	t_{10}	t_6	t_3	t_2	$k = 2$	$k = 7$
S_1	ER ₁	9.1	5.3	1.9	0.5	0.0	7.1	0.0
	EFFT	82.8	89.3	96.2	99.0	100.0	85.4	100.0
$S_{1\Phi}$	ER ₁	2.4	1.2	0.2	1.2	2.0	1.3	5.1
	EFFT	94.5	97.6	99.4	97.2	95.8	96.9	89.1
S_2	ER ₁	4.8	2.4	0.0	0.0	0.7	3.2	4.0
	EFFT	90.0	95.0	100.0	100.0	98.7	92.7	91.5
LS	ER ₁	0.0	0.7	6.1	17.6	60.8	0.1	97.0
	EFFT	100.0	98.2	87.0	61.4	26.7	99.8	27.9
Huber	ER ₁	1.2	0.0	0.0	2.9	10.2	0.0	36.1
	EFFT	96.8	100.0	99.4	94.0	79.3	100.0	52.3
Tukey	ER ₁	6.8	4.7	2.5	4.4	6.3	5.1	6.0
	EFFT	85.3	89.8	93.1	90.5	87.1	87.8	81.2
L1	ER ₁	22.8	16.1	12.0	14.8	8.2	21.3	7.4
	EFFT	66.3	72.7	80.1	75.3	83.7	67.0	86.8

TAB. 3.28: Excès relatif de l'erreur L_1 totale (ER₁) et efficacité relative totale (EFFT) (en %) pour les estimateurs considérés pour différentes lois sous-jacentes des erreurs. Cas de la régression linéaire multiple ($p = 3$). Taille de l'échantillon $n = 20$.

Estim.							$\varepsilon = 0.05$	$\varepsilon = 0.25$
		Φ	t_{10}	t_6	t_3	t_2	$k = 2$	$k = 7$
S_1	ER ₁	0.9	0.8	0.4	0.3	0.0	0.7	0.0
	EFFT	1.4	1.4	0.5	0.5	0.0	1.3	0.0
$S_{1\Phi}$	ER ₁	0.5	0.3	0.3	0.3	0.5	0.3	0.5
	EFFT	0.8	0.8	0.5	0.6	0.9	0.7	1.2
S_2	ER ₁	0.7	0.5	0.2	0.0	0.2	0.5	0.3
	EFFT	1.1	1.1	0.0	0.0	0.5	1.0	0.8
LS	ER ₁	0.0	0.5	0.9	1.9	4.5	0.5	3.7
	EFFT	0.0	1.2	2.2	7.3	3.5	1.0	1.2
Huber	ER ₁	0.3	0.0	0.4	0.7	1.1	0.0	1.9
	EFFT	0.6	0.0	1.2	1.3	2.3	0.2	1.8
Tukey	ER ₁	0.8	0.7	0.6	0.8	0.9	0.7	1.3
	EFFT	1.5	1.4	1.2	1.7	1.9	1.7	2.7
L1	ER ₁	1.5	1.4	1.2	1.2	1.1	1.4	1.3
	EFFT	1.5	1.7	1.6	1.5	1.7	1.5	2.8

TAB. 3.29: Ecart-type estimé de l'excès relatif de l'erreur L_1 totale (ER₁) et de l'efficacité relative totale (EFFT) (en %) pour les estimateurs considérés pour différentes lois sous-jacentes des erreurs. Cas de la régression linéaire multiple ($p = 3$). Taille de l'échantillon $n = 20$.

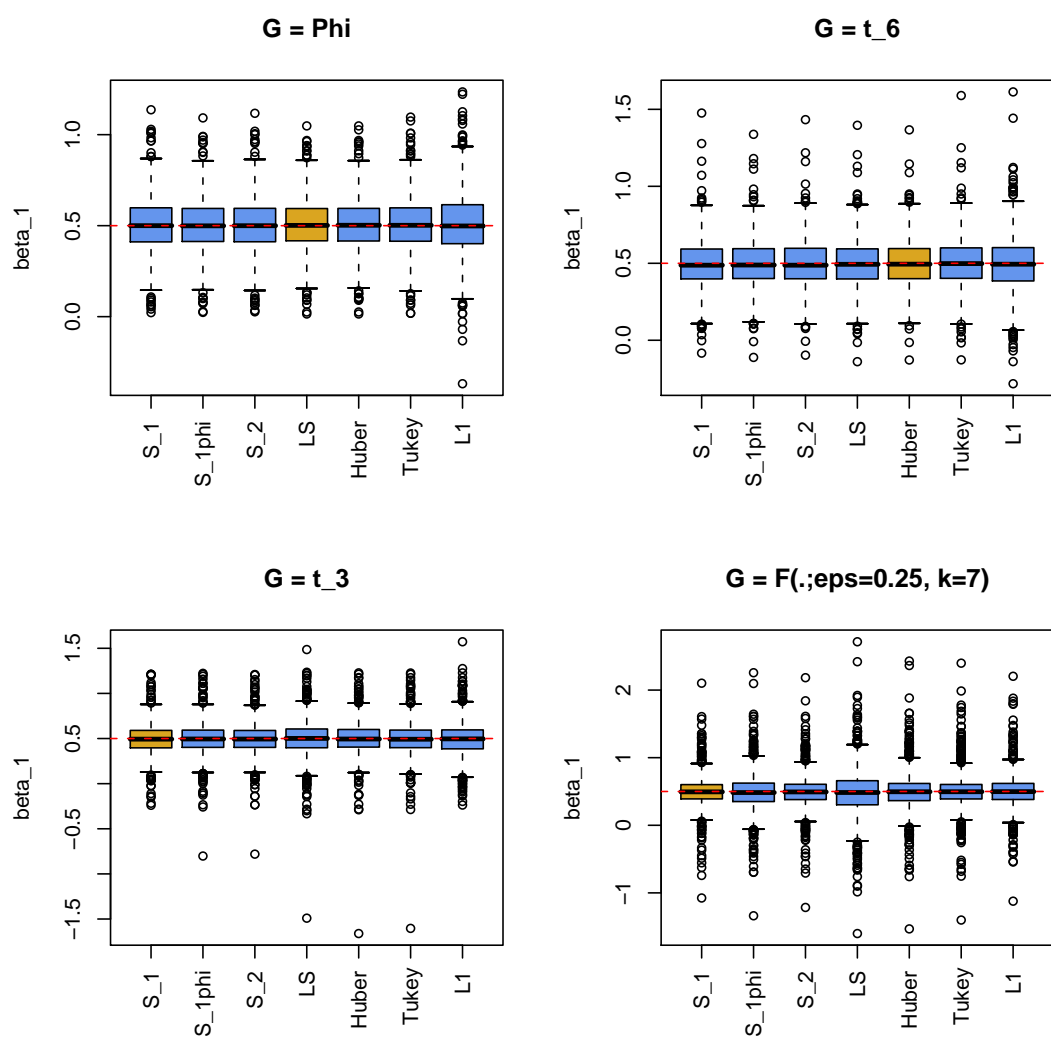


FIG. 3.16: Boxplots des résultats des simulations pour différentes loi sous-jacentes, dans le cas de la régression linéaire simple ($p = 1$). En jaune, l'estimateur possédant l'efficacité relative totale maximale. La ligne rouge représente la vraie valeur du paramètre de régression β_1 . Taille de l'échantillon : $n = 10$.

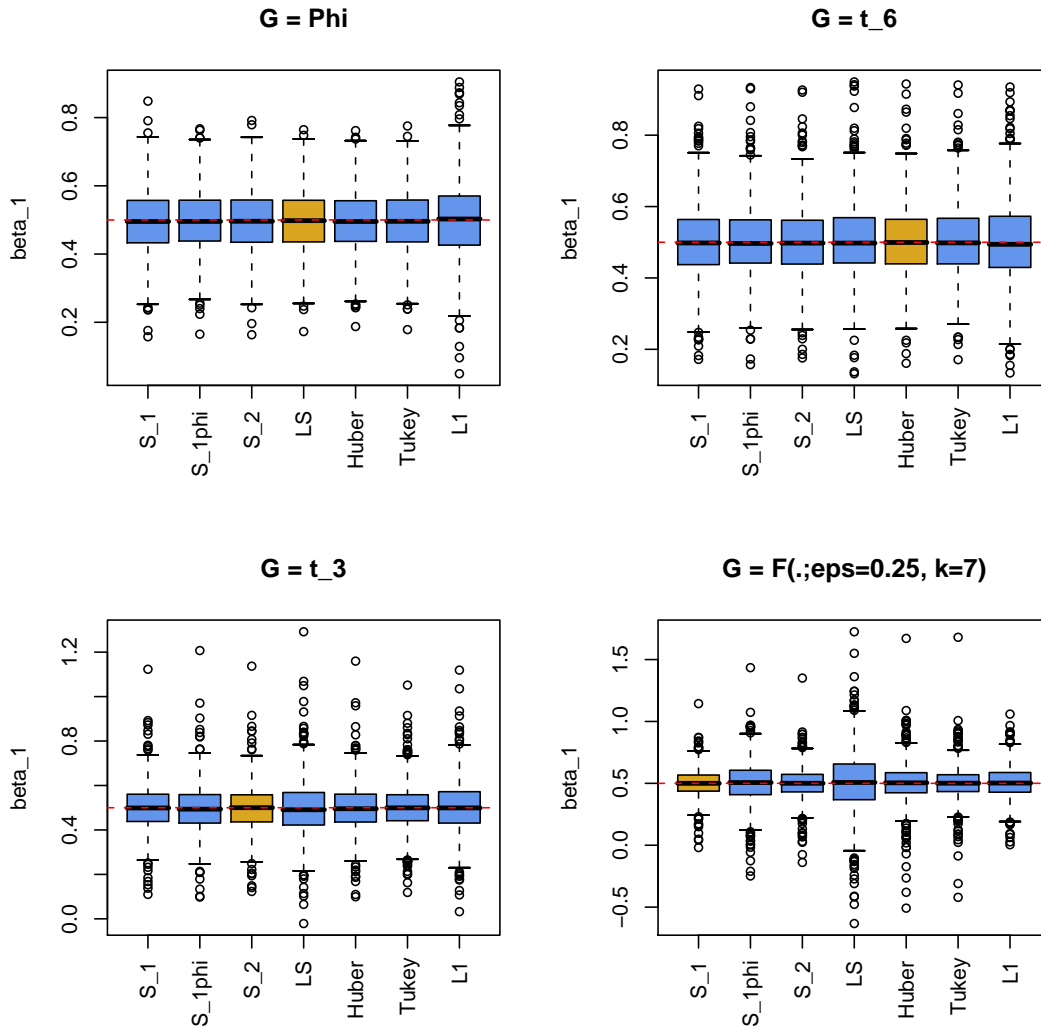


FIG. 3.17: Boxplots des résultats des simulations pour différentes loi sous-jacentes, dans le cas de la régression linéaire simple ($p = 1$). En jaune, l'estimateur possédant l'efficacité relative totale maximale. La ligne rouge représente la vraie valeur du paramètre de régression β_1 . Taille de l'échantillon : $n = 20$.

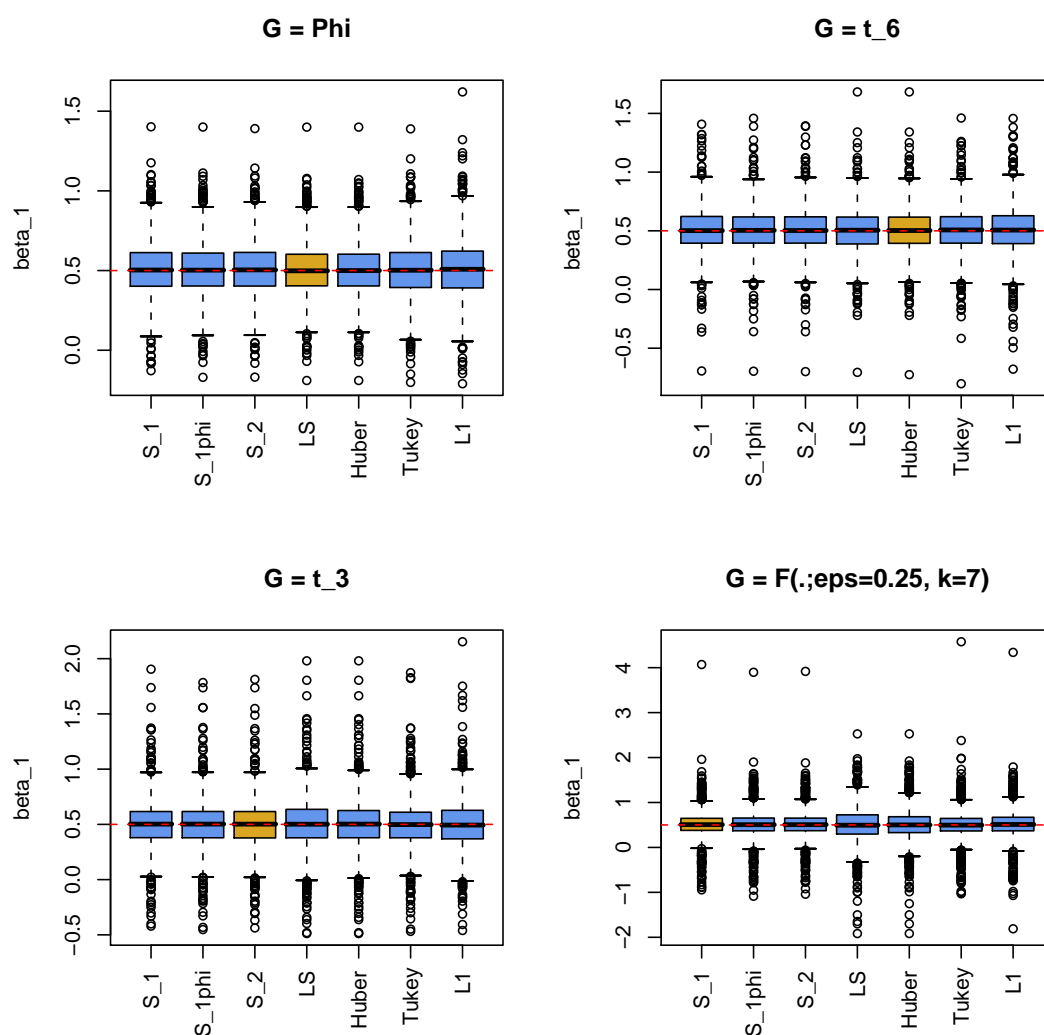


FIG. 3.18: Boxplots des résultats des simulations pour différentes loi sous-jacentes, dans le cas de la régression linéaire multiple ($p = 3$). En jaune, l'estimateur possédant l'efficacité relative totale maximale. La ligne rouge représente la vraie valeur du paramètre de régression β_1 . Taille de l'échantillon : $n = 10$.

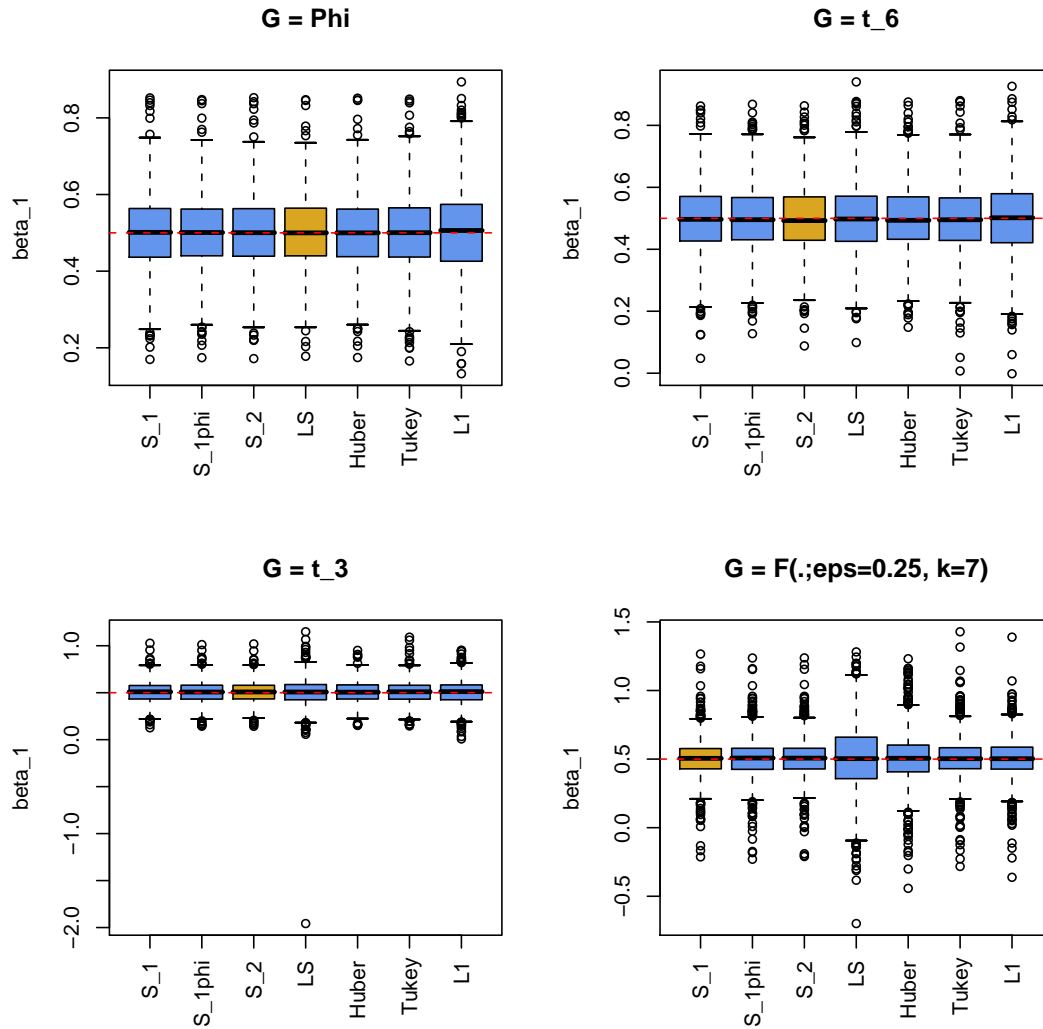


FIG. 3.19: Boxplots des résultats des simulations pour différentes loi sous-jacentes, dans le cas de la régression linéaire multiple ($p = 3$). En jaune, l'estimateur possédant l'efficacité relative totale maximale. La ligne rouge représente la vraie valeur du paramètre de régression β_1 . Taille de l'échantillon : $n = 20$.

Chapitre 4

Stratégies de compromis pour la classification

Les méthodes de classification statistiques sont utilisées afin d'identifier des groupes (*clusters*) dans un jeu de données, et ainsi de regrouper les observations qui sont similaires dans un certain sens. Il est également possible de déterminer une règle statistique permettant l'assignation de nouvelles données au groupe qui leur correspond le mieux. Ces méthodes statistiques sont utilisées dans quasiment tous les domaines, et plus particulièrement la médecine, la génétique ou encore la sociologie, et se basent sur un certain nombre de covariables présentant les différentes caractéristiques des éléments à classifier.

Les méthodes de classification peuvent être séparées en deux groupes : les méthodes supervisées, et celles non supervisées. Si les classes des objets sont connues à l'avance, on parle d'apprentissage supervisé, ou encore d'analyse discriminante. On se base alors sur les données à disposition afin de déterminer une règle de classification. En médecine par exemple, grâce aux résultats d'analyses de patients et à la connaissance de leur état physique (malade, en bonne santé, ou autre), on peut évaluer le risque de maladie d'un nouveau patient en fonction des résultats de ses analyses personnelles. Ces méthodes, notamment l'analyse discriminante linéaire ou quadratique, se basent sur des travaux de Fisher (1936), et l'on pourra consulter McLachlan (1992) pour une synthèse de ces méthodes.

Au contraire, lorsque l'on ne connaît pas *a priori* la classe de chacune des observations, on parle d'apprentissage non supervisé. Les méthodes développées dans ce cas doivent donc « découvrir » par elles-mêmes la structure cachée des données. Une notion de distance est souvent requise, afin de quantifier la différence entre deux observations. Parmi ces méthodes, on peut citer la classification hiérarchique, consistant à fusionner des classes à chaque étape, selon une mesure de distance entre deux *clusters*. On obtient alors une suite de partitions des données, souvent représentée sous la forme d'un arbre

hiérarchique, aussi appelé dendrogramme (voir par exemple Hastie *et al.*, 2001).

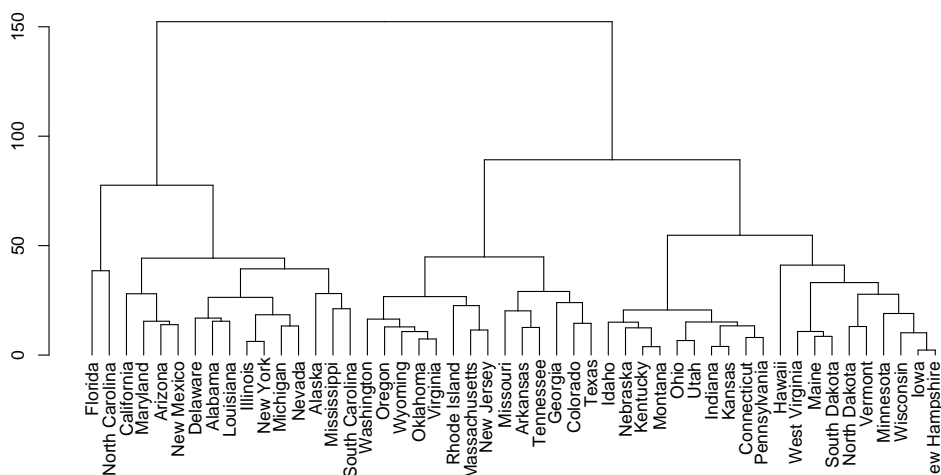


FIG. 4.1: Un dendrogramme, résultat d'une classification hiérarchique.

Parmi les méthodes non hiérarchiques, citons la méthode des k -moyennes (k -means), encore appelée méthode des centres mobiles. Cet algorithme est parmi les plus connus (voir MacQueen, 1967), et est basé sur des centroïdes, ou centres de gravité. En commençant avec k points de départ arbitraires comme centres, on regroupe les observations suivant leur distance aux centres, puis ces derniers sont mis à jour. Le processus est réitéré jusqu'à convergence. Cette méthode nécessite peu de temps de calcul, mais dépend fortement du choix initial des centres.

Finalement, nous allons étendre notre idée de stratégie de compromis à la méthode de la classification floue, également appelée *soft clustering*. Dans cette méthode, on suppose que les données ont été générées selon un modèle de mélange de distributions de probabilité dont les paramètres sont inconnus. Il s'agit alors d'estimer ces paramètres, ainsi que les paramètres du mélange, puis d'affecter chaque observation à la classe pour laquelle elle présente la plus grande probabilité d'appartenance.

4.1 Classification floue

Wolfe (1970) reformule le problème de la classification non supervisée en un problème d'estimation de paramètres d'un mélange de distributions de probabilité multivariées. Il s'agit d'identifier et de décrire les composantes du mélange, à partir d'un échantillon provenant de ce dernier, sans connaissance de la provenance exacte de chaque observation. Les distributions sont supposées unimodales et sont généralement des lois statistiques standards. Wolfe utilise la technique du maximum de vraisemblance afin

d'estimer les paramètres du mélange, et donne une solution exacte dans le cas d'un mélange de lois normales multivariées.

Notation 4.1.1. Soit $g_1(\mathbf{x}; \boldsymbol{\theta}_1), g_2(\mathbf{x}; \boldsymbol{\theta}_2), \dots, g_r(\mathbf{x}; \boldsymbol{\theta}_r)$ r lois de probabilité définies pour $\mathbf{x} \in \mathbb{R}^p$, où $\boldsymbol{\theta}_j = (\theta_{j,1}, \dots, \theta_{j,q_j})$ sont les q_j paramètres associés à la distribution g_j , $j = 1, \dots, r$. Soit également $\pi_1, \dots, \pi_r \in [0, 1]$ et tels que $\sum_{j=1}^r \pi_j = 1$, les proportions théoriques des observations dans le mélange.

On note alors le mélange des distributions g_1, \dots, g_r par

$$g(\mathbf{x}) = \sum_{j=1}^r \pi_j g_j(\mathbf{x}; \boldsymbol{\theta}_j).$$

On note également la probabilité d'appartenance à la classe j d'un vecteur \mathbf{x} par

$$P(j | \mathbf{x}) = \frac{P(j) \Pr(\mathbf{x} | j)}{P(\mathbf{x})} = \frac{\pi_j g_j(\mathbf{x}; \boldsymbol{\theta}_j)}{g(\mathbf{x})}.$$

Sur la base d'un échantillon de n vecteurs aléatoires provenant du mélange, notés

$$\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p}), \quad i = 1, \dots, n,$$

il s'agit d'estimer les paramètres du mélange, à savoir les valeurs de π_j et de $\boldsymbol{\theta}_j$, $j = 1, \dots, r$. Wolfe présente l'approche basée sur le maximum de la (log-)vraisemblance. Il s'agit de maximiser

$$\log L = \log L(\pi_1, \dots, \pi_r, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_r) = \sum_{i=1}^n \log g(\mathbf{x}_i),$$

sous la contrainte $\sum_{j=1}^r \pi_j = 1$. En utilisant le multiplicateur de Lagrange λ , on forme la fonction

$$\log \tilde{L} = \sum_{i=1}^n \log g(\mathbf{x}_i) - \lambda \left(\sum_{j=1}^r \pi_j - 1 \right),$$

et en égalant les dérivées de $\log \tilde{L}$ à 0, on obtient les équations du maximum de vraisemblance suivantes :

$$\begin{aligned} \frac{\partial \log \tilde{L}}{\partial \pi_j} &= \sum_{i=1}^n \frac{g_j(\mathbf{x}_i; \boldsymbol{\theta}_j)}{g(\mathbf{x}_i)} - \lambda = 0, \quad j = 1, \dots, r, \\ \frac{\partial \log \tilde{L}}{\partial \theta_{j,l}} &= \sum_{i=1}^n \frac{\pi_j}{g(\mathbf{x}_i)} \frac{\partial g_j(\mathbf{x}_i; \boldsymbol{\theta}_j)}{\partial \theta_{j,l}} = 0, \quad j = 1, \dots, r, \quad l = 1, \dots, q_j. \end{aligned}$$

En multipliant la première équation par π_j , et en sommant sur j , on trouve finalement que $\lambda = n$, et donc

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^n \hat{P}(j | \mathbf{x}_i),$$

où

$$\hat{P}(j | \mathbf{x}_i) = \frac{\hat{\pi}_j g_j(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_j)}{\hat{g}(\mathbf{x}_i)} \quad \text{et} \quad \hat{g}(\mathbf{x}_i) = \sum_{j=1}^r \hat{\pi}_j g_j(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_j).$$

Ainsi, l'estimateur du maximum de vraisemblance pour les proportions du mélange sont les moyennes sur l'échantillon des probabilités d'appartenance à chacune des classes. En utilisant également cette probabilité dans la seconde équation du maximum de vraisemblance, on obtient :

$$\frac{\partial \log \tilde{L}}{\partial \theta_{j,l}} = \sum_{i=1}^n P(j | \mathbf{x}_i) \frac{\partial g_j(\mathbf{x}_i; \boldsymbol{\theta}_j)}{\partial \theta_{j,l}} = 0, \quad j = 1, \dots, r, \quad l = 1, \dots, q_j.$$

Remarque 4.1.2. Si toute la population provenait d'une seule classe j , alors les équations du maximum de vraisemblance pour $\boldsymbol{\theta}_j$ seraient données par

$$\sum_{i=1}^n \frac{\partial \log g_j(\mathbf{x}_i; \boldsymbol{\theta}_j)}{\partial \theta_{j,l}} = 0, \quad l = 1, \dots, q_j.$$

Ainsi, dans le cas d'un mélange, les équations du maximum de vraisemblance sont la moyenne pondérée des expressions utilisées traditionnellement, où les poids sont les probabilités d'appartenance.

Dans le cas général, les équations du maximum de vraisemblance doivent être résolues de manière numérique. Néanmoins, lorsque les distributions du mélange sont des lois normales multivariées, une solution explicite existe.

Exemple 4.1.3. Soit $g_j(\mathbf{x}; \boldsymbol{\theta}_j) = g_j(\mathbf{x}; \boldsymbol{\mu}_j, \Sigma_j)$ la densité de la loi normale multivariée, de moyenne $\boldsymbol{\mu}_j \in \mathbb{R}^p$ et de matrice de variance-covariance $\Sigma_j \in \mathbb{R}^{p \times p}$. Cette densité est donnée par

$$g_j(\mathbf{x}; \boldsymbol{\mu}_j, \Sigma_j) = (2\pi)^{-p/2} \det(\Sigma_j)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \Sigma_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\right).$$

En prenant les dérivées de la log-vraisemblance par rapport aux éléments de $\boldsymbol{\mu}_j$ et aux éléments de l'inverse de Σ_j , les solutions suivantes apparaissent :

$$\begin{aligned}\hat{\pi}_j &= \frac{1}{n} \sum_{i=1}^n \hat{P}(j \mid \mathbf{x}_i), \quad j = 1, \dots, r; \\ \hat{\mu}_{j,l} &= \frac{1}{n\hat{\pi}_j} \sum_{i=1}^n \hat{P}(j \mid \mathbf{x}_i) x_{i,l}, \quad j = 1, \dots, r, \quad l = 1, \dots, p; \\ (\hat{\sigma}_j)_{st} &= \frac{1}{n\hat{\pi}_j} \sum_{i=1}^n \hat{P}(j \mid \mathbf{x}_i) (x_{i,s} - \hat{\mu}_{j,s})(x_{i,t} - \hat{\mu}_{j,t}), \quad j = 1, \dots, r, \quad s, t = 1, \dots, p,\end{aligned}$$

où $(\hat{\sigma}_j)_{st}$ représente l'élément de la s -ème ligne et de la t -ème colonne de l'estimateur $\hat{\Sigma}_j$ de la matrice Σ_j .

A nouveau, nous remarquons que ces équations sont des versions pondérées des équations habituelles du maximum de vraisemblance dans le cas de la loi normale multivariée, où chaque point est pondéré par sa probabilité d'appartenance à la classe en question.

Les équations ci-dessus permettent de plus l'utilisation d'une méthode itérative pour atteindre une solution, dans le sens où si l'on possède les estimateurs des paramètres $\hat{\mu}_j$ et Σ_j , $j = 1, \dots, r$, on peut alors déterminer les probabilités d'appartenance de chaque observation, puis ensuite mettre à jour les estimateurs, et ainsi de suite. Cette procédure est formalisée dans l'algorithme EM, très souvent utilisé dans la classification non supervisée, et détaillé dans ce qui suit.

4.1.1 L'algorithme EM

Proposé par Dempster *et al.* (1977), l'algorithme EM (pour Espérance-Maximisation) permet la simplification des problèmes de maximisation de la vraisemblance en introduisant une série de variables latentes que l'on ne peut pas observer. L'algorithme comporte deux phases :

- une étape Espérance, dans laquelle on détermine l'espérance conditionnelle de la log-vraisemblance en utilisant les valeurs estimées des variables latentes additionnelles ;
- une étape Maximisation, dans laquelle les estimations des paramètres sont mises à jour en maximisant la log-vraisemblance espérée trouvée précédemment.

Ces deux étapes sont effectuées en alternance, jusqu'à convergence de l'algorithme.

De manière générale et schématique, si \mathbf{X} est un ensemble de données observées, \mathbf{Z} est un ensemble de variables latentes non observables associées à \mathbf{X} , $\boldsymbol{\theta}$ un vecteur de paramètres, et $L(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z})$ une fonction de vraisemblance, alors l'algorithme EM peut être décrit comme suit.

1. Choisir une valeur initiale quelconque pour les paramètres $\hat{\boldsymbol{\theta}}^{(0)}$.
2. A la k -ème étape, calculer l'espérance de la log-vraisemblance, par rapport à la distribution conditionnelle de \mathbf{Z} sachant \mathbf{X} , avec les valeurs courantes des estimateurs $\hat{\boldsymbol{\theta}}^{(k)}$:

$$Q\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}\right) = \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \hat{\boldsymbol{\theta}}^{(k)}} (\log L(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z})).$$

3. Déterminer les nouveaux estimateurs $\hat{\boldsymbol{\theta}}^{(k+1)}$ en maximisant $Q\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}\right)$ par rapport à $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \arg \max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}\right).$$

4. Itérer jusqu'à convergence de la log-vraisemblance.

Les propriétés de l'algorithme assurent que

$$L\left(\hat{\boldsymbol{\theta}}^{(k+1)} \mid \mathbf{X}, \mathbf{Z}\right) \geq L\left(\hat{\boldsymbol{\theta}}^{(k)} \mid \mathbf{X}, \mathbf{Z}\right),$$

et donc que l'algorithme converge vers un maximum local. Afin de s'assurer une convergence globale, il convient de débiter la procédure avec plusieurs valeurs différentes pour les paramètres.

Exemple 4.1.4. Reprenons le cas d'un mélange de lois normales multivariées, et introduisons les variables latentes définies de la manière suivante :

$$\mathbf{z}_i = (z_{i,1}, \dots, z_{i,r}),$$

avec

$$z_{i,j} = \begin{cases} 1, & \text{si } \mathbf{x}_i \text{ appartient à la classe } j; \\ 0, & \text{sinon.} \end{cases}$$

Ainsi, $\mathbf{z}_1, \dots, \mathbf{z}_n$ sont indépendantes et identiquement distribuées selon une loi multinomiale avec un lancer et vecteur de probabilité $\boldsymbol{\pi} = (\pi_1, \dots, \pi_r)$, les proportions du mélange. La densité conditionnelle de \mathbf{x}_i sachant \mathbf{z}_i est alors donnée par

$$g(\mathbf{x}_i \mid \mathbf{z}_i) = \prod_{j=1}^r g_j(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j)^{z_{i,j}}.$$

En effet, connaissant la classe à laquelle appartient l'observation \mathbf{x}_i , la densité de mélange devient simplement la densité de la loi normale multivariée de cette classe. En notant $\Theta = (\boldsymbol{\pi}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_r, \Sigma_1, \dots, \Sigma_r)$, la vraisemblance totale devient

$$\begin{aligned}
L(\Theta) &= \prod_{i=1}^n g(\mathbf{x}_i, \mathbf{z}_i \mid \Theta) \\
&= \prod_{i=1}^n g(\mathbf{x}_i \mid \mathbf{z}_i, \Theta) g(\mathbf{z}_i \mid \Theta) \\
&= \prod_{i=1}^n \prod_{j=1}^r \pi_j^{z_{i,j}} g_j(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j)^{z_{i,j}}.
\end{aligned}$$

La log-vraisemblance vaut alors

$$\log L(\Theta) = \sum_{i=1}^n \sum_{j=1}^r z_{i,j} [\log \pi_j + \log g_j(\mathbf{x}_i; \boldsymbol{\theta}_j)].$$

En supposant que nous ayons à disposition une estimation des paramètres $\hat{\Theta}^{(k)}$, nous effectuons l'étape Espérance de l'algorithme :

$$\begin{aligned}
Q\left(\Theta \mid \hat{\Theta}^{(k)}\right) &= \mathbb{E}_{\mathbf{z} \mid \mathbf{x}, \hat{\Theta}^{(k)}} (\log L(\Theta)) \\
&= \sum_{i=1}^n \sum_{j=1}^r \mathbb{E}\left(z_{i,j} \mid \mathbf{x}_i, \hat{\Theta}^{(k)}\right) \left[\log \hat{\pi}_j^{(k)} + \log g_j\left(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_j^{(k)}\right)\right] \\
&= \sum_{i=1}^n \sum_{j=1}^r \hat{\tau}_{i,j}^{(k)} \left[\log \hat{\pi}_j^{(k)} + \log g_j\left(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_j^{(k)}\right)\right],
\end{aligned}$$

où

$$\begin{aligned}
\hat{\tau}_{i,j}^{(k)} &= \text{P}\left(j \mid \mathbf{x}_i, \hat{\Theta}^{(k)}\right) \\
&= \frac{\hat{\pi}_j^{(k)} g_j\left(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_j^{(k)}\right)}{\sum_{s=1}^r \hat{\pi}_s^{(k)} g_s\left(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_s^{(k)}\right)},
\end{aligned}$$

par la formule de Bayes. En effet, connaissant la valeur de \mathbf{x}_i et des paramètres $\hat{\Theta}^{(k)}$, la variable aléatoire \mathbf{z}_i suit une loi multinomiale $\text{Mult}_r\left(1, \hat{\boldsymbol{\tau}}_i^{(k)}\right)$, où $\hat{\boldsymbol{\tau}}_i^{(k)} = \left(\hat{\tau}_{i,1}^{(k)}, \dots, \hat{\tau}_{i,r}^{(k)}\right)$.

On peut à présent effectuer l'étape de Maximisation de l'algorithme, afin de mettre à jour les estimations des paramètres. Comme l'a montré Wolfe dans son article, les solutions des équations du maximum de vraisemblance sont alors données par :

$$\begin{aligned}\hat{\pi}_j^{(k+1)} &= \frac{1}{n} \sum_{i=1}^n \hat{\tau}_{i,j}^{(k)}; \\ \hat{\boldsymbol{\mu}}_j^{(k+1)} &= \frac{\sum_{i=1}^n \hat{\tau}_{i,j}^{(k)} \mathbf{x}_i}{\sum_{i=1}^n \hat{\tau}_{i,j}^{(k)}}; \\ \hat{\Sigma}_j^{(k+1)} &= \frac{\sum_{i=1}^n \hat{\tau}_{i,j}^{(k)} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{(k)} \right) \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{(k)} \right)^T}{\sum_{i=1}^n \hat{\tau}_{i,j}^{(k)}},\end{aligned}$$

et sont les versions pondérées des estimateurs habituels du maximum de vraisemblance.

4.1.2 Classification floue robustifiée

Les mélanges de lois normales ont été très souvent utilisés pour la classification de données multivariées, surtout grâce au fait de la facilité des calculs. Néanmoins, des données aberrantes dans les observations peuvent très facilement affecter les estimations des paramètres, et donc influencer la classification finale des données. En effet, les queues de la loi normale sont trop légères pour de nombreux problèmes pratiques. McLachlan et Peel (1998) proposent alors d'utiliser un mélange de lois t multivariées, afin d'obtenir une classification plus robuste.

La loi t multivariée en dimension p , de moyenne $\boldsymbol{\mu} \in \mathbb{R}^p$, de matrice de variance-covariance $\Sigma \in \mathbb{R}^{p \times p}$, et avec ν degrés de liberté, peut être définie de plusieurs manières. Sa densité de probabilité est donnée par

$$g(\mathbf{x}; \boldsymbol{\mu}, \Sigma, \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right) \det(\Sigma)^{-1/2}}{(\pi\nu)^{p/2} \Gamma\left(\frac{\nu}{2}\right) [1 + \delta(\mathbf{x}, \boldsymbol{\mu}, \Sigma)/\nu]^{\frac{1}{2}(\nu+p)}},$$

où $\delta(\mathbf{x}, \boldsymbol{\mu}, \Sigma) = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ représente la distance de Mahalanobis entre \mathbf{x} et $\boldsymbol{\mu}$, avec Σ comme matrice de variance-covariance. Néanmoins, cette manière de définir la loi multivariée t n'est pas très utile lorsqu'il s'agit de l'utiliser dans des problèmes de maximum de vraisemblance. En effet, il n'est pas possible de déterminer explicitement les solutions des équations du maximum de vraisemblance en utilisant de manière brute cette densité de probabilité.

Afin de contourner ce problème, et d'utiliser l'algorithme EM pour l'estimation des paramètres, il convient de définir la loi t multivariée comme suit. Soit \mathbf{Y} un vecteur aléatoire suivant une loi normale multivariée, de moyenne $\mathbf{0}$ et de matrice variance-covariance identité. Soit également U , une variable aléatoire indépendante de \mathbf{Y} , et distribuée selon une loi $\Gamma\left(\frac{1}{2}\nu, \frac{1}{2}\nu\right)$, où la densité de la loi $\Gamma(\alpha, \beta)$ est donnée par

$$f(u; \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha u^{\alpha-1} \exp(-\beta u), \quad \text{pour } u > 0.$$

Alors la variable $(\Sigma^{\frac{1}{2}}/\sqrt{U})\mathbf{Y} + \boldsymbol{\mu}$ suit une loi multivariée t de moyenne $\boldsymbol{\mu}$, de matrice variance-covariance Σ , avec ν degrés de liberté. Cette manière de définir cette loi va nous permettre d'utiliser l'algorithme EM, en considérant la variable U comme une variable inobservable.

Remarque 4.1.5. Si $U \sim \Gamma(\frac{1}{2}\nu, \frac{1}{2}\nu)$, alors $\nu U \sim \chi_\nu^2$, et l'on a alors l'analogie de la distribution t de Student univariée dans le cas multidimensionnel.

Considérons à présent un mélange de r distributions t multivariées :

$$g(\mathbf{x}) = \sum_{j=1}^r \pi_j g_j(\mathbf{x}; \boldsymbol{\theta}_j),$$

où π_j sont les proportions du mélange, où $g_j(\mathbf{x}; \boldsymbol{\theta}_j)$ est la densité de la loi multivariée t dans la classe j , avec $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j, \nu_j)$. Pour chaque classe, nous avons donc une moyenne, une matrice de variance-covariance, et le nombre de degrés de liberté. Introduisons encore les notations suivantes.

Notation 4.1.6. Nous notons, pour chaque observation \mathbf{x}_i , $i = 1, \dots, n$, comme auparavant,

$$\mathbf{z}_i = (z_{i,1}, \dots, z_{i,r}),$$

avec

$$z_{i,j} = \begin{cases} 1, & \text{si } \mathbf{x}_i \text{ appartient à la classe } j; \\ 0, & \text{sinon.} \end{cases}$$

Ajoutons également les variables indépendantes inobservables u_i , $i = 1, \dots, n$, associées à \mathbf{x}_i , telles que

$$u_i \mid z_{i,j} = 1 \sim \Gamma\left(\frac{1}{2}\nu_j, \frac{1}{2}\nu_j\right).$$

Ainsi, si l'on sait que \mathbf{x}_i appartient à la classe j , alors \mathbf{x}_i est distribué selon une loi normale multivariée de moyenne $\boldsymbol{\mu}_j$ et de matrice de variance-covariance Σ_j/u_i . Notons également comme auparavant

$$\Theta = (\pi_1, \dots, \pi_r, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_r, \Sigma_1, \dots, \Sigma_r)$$

l'ensemble des paramètres du mélange. Soit enfin

$$\gamma_j(u_i) = \gamma_j(u_i; \nu_j) = u_i^{\frac{1}{2}\nu_j - 1} \left(\frac{1}{2}\nu_j\right)^{\frac{1}{2}\nu_j} \frac{1}{\Gamma(\frac{1}{2}\nu_j)} \exp\left(-\frac{1}{2}\nu_j u_i\right), \quad \text{pour } u > 0$$

la densité de la loi $\Gamma(\frac{1}{2}\nu_j, \frac{1}{2}\nu_j)$, associée à la classe j , et

$$\varphi_j(\mathbf{x}_i) = \varphi_j(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j / u_i) = (2\pi)^{-\frac{p}{2}} \det(\Sigma_j)^{-\frac{1}{2}} u_i^{\frac{p}{2}} \exp\left(-\frac{u_i}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \Sigma_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\right)$$

la densité de la loi normale multivariée de dimension p , associée à la classe j .

Considérons à présent la fonction de vraisemblance pour une seule observation complétée par les variables inobservables $(\mathbf{x}_i, \mathbf{z}_i, u_i)$, pour un i fixé :

$$\begin{aligned} L_i(\Theta) &= f(\mathbf{x}_i, \mathbf{z}_i, u_i; \Theta) \\ &= f(\mathbf{z}_i \mid \Theta) f(u_i \mid \mathbf{z}_i, \Theta) f(\mathbf{x}_i \mid \mathbf{z}_i, u_i, \Theta) \\ &= \prod_{j=1}^r \pi_j^{z_{i,j}} \gamma_j(u_i)^{z_{i,j}} \varphi_j(\mathbf{x}_i)^{z_{i,j}}, \end{aligned}$$

et la log-vraisemblance, toujours pour une seule observation, vaut

$$\log L_i(\Theta) = \sum_{j=1}^r z_{i,j} [\log \pi_j + \log \gamma_j(u_i) + \log \varphi_j(\mathbf{x}_i)].$$

Avant d'effectuer l'étape Espérance de l'algorithme EM, il convient de déterminer la loi de $u_i \mid \mathbf{x}_i, \mathbf{z}_i, \Theta$. Dans la log-vraisemblance ci-dessus, considérons uniquement les termes contenant u_i :

$$\begin{aligned} \log \gamma_j(u_i) + \log \varphi_j(\mathbf{x}_i) &= \left(\frac{1}{2}\nu_j - 1\right) \log u_i + \frac{1}{2}\nu_j \log\left(\frac{1}{2}\nu_j\right) - \log \Gamma\left(\frac{1}{2}\nu_j\right) - \frac{1}{2}\nu_j u_i \\ &\quad - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log \det(\Sigma_j) + \frac{p}{2} \log u_i - \frac{u_i}{2} \delta(\mathbf{x}_i, \boldsymbol{\mu}_j, \Sigma_j) \\ &= \log u_i \left[\frac{p}{2} + \frac{1}{2}\nu_j - 1\right] + u_i \left[-\frac{1}{2}\delta(\mathbf{x}_i, \boldsymbol{\mu}_j, \Sigma_j) - \frac{1}{2}\nu_j\right] + C, \end{aligned}$$

où l'on a regroupé les termes en $\log u_i$ et les termes en u_i , et où C est une constante indépendante de u_i . On reconnaît alors la log-densité d'une loi $\Gamma(m_1, m_2)$, avec

$$m_1 = \frac{1}{2}(\nu_j + p), \quad m_2 = \frac{1}{2}(\nu_j + \delta(\mathbf{x}_i, \boldsymbol{\mu}_j, \Sigma_j)),$$

et ainsi

$$u_i \mid \mathbf{x}_i, \mathbf{z}_i, \Theta \sim \Gamma(m_1, m_2).$$

Nous pouvons à présent effectuer l'étape Espérance de l'algorithme EM, en ne considérons toujours qu'une seule observation \mathbf{x}_i :

$$\begin{aligned} Q_i(\Theta \mid \hat{\Theta}^{(k)}) &= \mathbb{E}_{\mathbf{z}, \mathbf{U} \mid \mathbf{x}, \hat{\Theta}^{(k)}}(\log L_i(\Theta)) \\ &= \mathbb{E}_{\mathbf{z} \mid \mathbf{x}, \hat{\Theta}^{(k)}} \left[\mathbb{E}_{\mathbf{U} \mid \mathbf{x}, \mathbf{z}, \hat{\Theta}^{(k)}}(\log L_i(\Theta)) \right]. \end{aligned}$$

Pour le terme $\mathbb{E}_{\mathbf{U}|\mathbf{X},\mathbf{Z},\hat{\Theta}^{(k)}}(\log L_i(\Theta))$, nous avons besoin de l'espérance suivante :

$$\mathbb{E}\left(u_i \mid \mathbf{x}_i, \mathbf{z}_i, \hat{\Theta}^{(k)}\right),$$

que nous obtenons immédiatement par

$$\begin{aligned} \mathbb{E}\left(u_i \mid \mathbf{x}_i, \mathbf{z}_i, \hat{\Theta}^{(k)}\right) &= \frac{m_1^{(k)}}{m_2^{(k)}} \\ &= \frac{\nu_j + p}{\nu_j + \delta\left(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_j^{(k)}, \hat{\Sigma}_j^{(k)}\right)}. \end{aligned}$$

Remarquons que l'espérance

$$\mathbb{E}\left(\log u_i \mid \mathbf{x}_i, \mathbf{z}_i, \hat{\Theta}^{(k)}\right)$$

n'est pas nécessaire, puisque dans la log-vraisemblance, seul le terme u_i est facteur des paramètres inconnus Θ , au travers de $\delta(\mathbf{x}_i, \boldsymbol{\mu}_j, \Sigma_j)$. Le terme $\log u_i$ n'est lui facteur que de termes constants, plus précisément p et ν_j .

Ainsi, nous obtenons finalement :

$$\begin{aligned} Q_i\left(\Theta \mid \hat{\Theta}^{(k)}\right) &= \mathbb{E}_{\mathbf{Z}|\mathbf{X},\hat{\Theta}^{(k)}}\left[\mathbb{E}_{\mathbf{U}|\mathbf{X},\mathbf{Z},\hat{\Theta}^{(k)}}(\log L_i(\Theta))\right] \\ &= \mathbb{E}_{\mathbf{Z}|\mathbf{X},\hat{\Theta}^{(k)}}\left[\sum_{j=1}^r z_{i,j} \left(\log \pi_j^{(k)} - u_i^{(k)} \frac{1}{2} \delta(\mathbf{x}_i, \boldsymbol{\mu}_j, \Sigma_j) - \frac{1}{2} \log \det(\Sigma_j) + C_j\right)\right] \\ &= \sum_{j=1}^r \mathbb{P}\left(j \mid \mathbf{x}_i, \hat{\Theta}^{(k)}\right) \left(\log \pi_j^{(k)} - u_i^{(k)} \frac{1}{2} \delta(\mathbf{x}_i, \boldsymbol{\mu}_j, \Sigma_j) - \frac{1}{2} \log \det(\Sigma_j) + C_j\right) \\ &= \sum_{j=1}^r \hat{\tau}_{i,j}^{(k)} \left(\log \pi_j^{(k)} - u_i^{(k)} \frac{1}{2} \delta(\mathbf{x}_i, \boldsymbol{\mu}_j, \Sigma_j) - \frac{1}{2} \log \det(\Sigma_j) + C_j\right), \end{aligned}$$

où

$$\hat{\tau}_{i,j}^{(k)} = \frac{\pi_j^{(k)} g_j\left(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_j^{(k)}, \hat{\Sigma}_j^{(k)}, \nu_j\right)}{\sum_{s=1}^r \pi_s^{(k)} g_j\left(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_s^{(k)}, \hat{\Sigma}_s^{(k)}, \nu_s\right)},$$

où

$$u_i^{(k)} = \frac{\nu_j + p}{\nu_j + \delta\left(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_j^{(k)}, \hat{\Sigma}_j^{(k)}\right)},$$

et où C_j est une constante indépendante de π_j , $\boldsymbol{\mu}_j$ et Σ_j . On peut à présent considérer l'échantillon \mathbf{x}_i , $i = 1, \dots, n$ dans son ensemble, pour obtenir

$$Q\left(\Theta \mid \hat{\Theta}^{(k)}\right) = \sum_{i=1}^n \sum_{j=1}^r \hat{\tau}_{i,j}^{(k)} \left(\log \pi_j^{(k)} - u_i^{(k)} \frac{1}{2} \delta(\mathbf{x}_i, \boldsymbol{\mu}_j, \Sigma_j) - \frac{1}{2} \log \det(\Sigma_j) + C_j \right).$$

Nous pouvons maintenant effectuer l'étape Maximisation de l'algorithme EM. En utilisant l'approche du multiplicateur de Lagrange, on obtient directement la mise à jour des proportions du mélange :

$$\pi_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \hat{\tau}_{i,j}^{(k)}.$$

En dérivant directement $Q\left(\Theta \mid \hat{\Theta}^{(k)}\right)$ par rapport à $\boldsymbol{\mu}_j$, on obtient facilement la mise à jour des moyennes des classes :

$$\hat{\boldsymbol{\mu}}_j^{(k+1)} = \frac{\sum_{i=1}^n \hat{\tau}_{i,j}^{(k)} u_i^{(k)} \mathbf{x}_i}{\sum_{i=1}^n \hat{\tau}_{i,j}^{(k)} u_i^{(k)}}.$$

Enfin, en utilisant les relations suivantes, démontrées dans Henderson et Searle (1981), ou encore Searle (1982), chap. 12,

$$\frac{\partial \text{Tr}(AB)}{\partial A} = B + B^T - \text{diag}(B)$$

et

$$\frac{\partial \log \det(A)}{\partial A} = 2A^{-1} - \text{diag}(A^{-1}),$$

pour A une matrice symétrique, on obtient la mise à jour des matrices de variance-covariance des classes :

$$\hat{\Sigma}_j^{(k+1)} = \frac{\sum_{i=1}^n \hat{\tau}_{i,j}^{(k)} u_j^{(k)} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{(k)}) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{(k)})^T}{\sum_{i=1}^n \hat{\tau}_{i,j}^{(k)}}.$$

Remarque 4.1.7. Notons que dans la démarche présentée ci-dessus, les degrés de liberté des lois t multivariées associées à chaque classe, ν_j , $j = 1, \dots, r$, sont considérés comme connus et fixés. Les seuls paramètres inconnus sont les moyennes et les matrices de variance-covariance des distributions, ainsi que les proportions du mélange.

Cette manière de faire nous permet d'obtenir des solutions explicites pour les équations du maximum de vraisemblance, mais il est également possible de ne faire aucune supposition sur ces degrés de liberté. Les degrés de liberté ν_j , $j = 1, \dots, r$, sont alors d'autant plus de paramètres du modèle, à estimer dans l'étape maximisation de l'algorithme EM. Dans ce cas, il n'est plus possible d'obtenir des solutions explicites, et une optimisation numérique est nécessaire à chaque étape.

4.2 Stratégie de compromis pour la classification floue

Nous désirons à présent développer une stratégie basée sur un compromis de modèles pour la classification floue. Cette dernière reposant intégralement sur l'approche du maximum de vraisemblance, notre démarche sera très directe, puisque nous allons supposer plusieurs modèles sous-jacents possibles, déterminer dans chaque cas les probabilités d'appartenance de chaque observation aux différentes classes, puis combiner ces probabilités en donnant un poids à chaque modèle.

L'idée de combiner des modèles de classification floue a été évoquée par Baudry *et al.* (2010), avec cependant une approche différente. Dans cet article, Baudry et Raftery proposent de combiner les *clusters* obtenus dans le cas d'un mélange de lois normales multivariées, en utilisant un critère d'entropie. Dans la plupart des cas, le critère BIC est utilisé pour déterminer le nombre total de classes voir (voir Fraley et Raftery, 1998). Néanmoins, en pratique, il se peut qu'une classe présentant une distribution non gaussienne soit mal interprétée. En effet, une telle classe peut alors être considérée comme étant elle-même un mélange de lois normales. Le critère BIC vise quant à lui à déterminer le nombre de lois présentes dans le mélange plutôt que le nombre de classes proprement dit, et cela conduit souvent à une surestimation du nombre de *clusters*.

Baudry et Raftery proposent une méthode consistant à agréger successivement des classes, en minimisant une mesure d'entropie qui pénalise le critère BIC. Une suite de classifications est alors formée, en partant du modèle préconisé par le critère BIC, et à chaque étape deux classes sont fusionnées. Chaque solution combinée approche les données de manière équivalente à la solution du BIC, car la vraisemblance ne change pas. Seuls le nombre et la définition des classes sont différents d'étape en étape. La procédure s'arrête lorsque l'on peut constater que la baisse d'entropie lors de la fusion de deux classes devient relativement faible.

Cette méthode préserve les avantages de la classification utilisant des mélanges de lois normales, notamment une bonne adéquation aux données. De plus, elle permet d'éviter la surestimation du nombre de classes, et ne fait appel qu'aux probabilités d'appartenance aux classes, et peut donc être facilement appliquée sans modification à des mélanges d'autres distributions de probabilité.

Nous allons à présent appliquer notre idée de compromis de modèles au cas de la classification floue. Soit $g = g(\mathbf{x}; \Theta)$ la loi sous-jacente des données. g est donc un mélange des densités $g_1(\mathbf{x}; \theta_1), \dots, g_r(\mathbf{x}; \theta_r)$, avec π_1, \dots, π_r les proportions du mélange.

Au lieu de supposer que g est un certain mélange de lois, par exemple un mélange de lois normales, ou un mélange de lois t multivariées, nous allons nous donner plusieurs possibilités pour le mélange, et faire un compromis entre différents modèles. Soit donc \mathcal{F} un ensemble de mélanges. Pour $f_k \in \mathcal{F}$, nous avons que

$$f_k = f_k(\mathbf{x}; \Theta_k)$$

est un mélange entre les r densités de probabilité

$$f_{k,1}(\mathbf{x}; \boldsymbol{\theta}_{k,1}), \dots, f_{k,r}(\mathbf{x}; \boldsymbol{\theta}_{k,r}),$$

avec comme proportions

$$\pi_{k,1}, \dots, \pi_{k,r},$$

et où $\Theta_k = (\pi_{k,1}, \dots, \pi_{k,r}, \boldsymbol{\theta}_{k,1}, \dots, \boldsymbol{\theta}_{k,r})$. Notons immédiatement que le nombre de classes doit être le même dans chaque mélange $f_k \in \mathcal{F}$. En effet, il serait impossible de faire un compromis dans le cas où une classe existerait dans un modèle mais pas dans un autre. Nous ne pourrions utiliser aucune des deux informations, puisque dans un cas un modèle nous donnerait une probabilité d'appartenance à une classe n'existant pas dans l'autre, et dans l'autre cas, nous aurions des probabilités d'appartenance ajustées sans tenir compte d'une classe supplémentaire possible.

Au contraire, il n'est pas nécessaire que les paramètres des distributions de chaque mélange soit comparables, ou qu'ils possèdent la même signification. En effet, contrairement au cas des estimateurs de Pitman compromis, présentés ci-dessus, nous ne cherchons pas au final à obtenir une estimation des paramètres, mais bien une classification des données. Dès lors, nous construisons notre compromis de modèle non pas comme un compromis des paramètres, mais plutôt comme un compromis de la règle de classification de chaque modèle, et plus particulièrement sur les probabilités d'appartenance aux classes.

Notons $\hat{P}_k(j \mid \mathbf{x}_i)$ la probabilité (*a posteriori*) que l'observation \mathbf{x}_i appartienne à la classe j , sous le modèle f_k . Plus particulièrement, nous avons que

$$\hat{P}_k(j \mid \mathbf{x}_i) = \frac{\hat{\pi}_{k,j} f_{k,j}(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k)}{f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k)}$$

grâce à la règle de Bayes. Cette probabilité est donc obtenue après l'estimation de tous les paramètres Θ_k du mélange f_k . La règle de classification associée à ce mélange est alors de placer l'observation \mathbf{x}_i dans la classe avec la plus grande probabilité d'appartenance. En se basant seulement sur ces probabilités, nous définissons à présent une règle de classification compromis.

Définition 4.2.1. Soit $f_1(\mathbf{x}; \Theta_1), \dots, f_m(\mathbf{x}; \Theta_m)$ des mélanges de distributions à r composantes. La probabilité d'appartenance compromis de l'observation \mathbf{x}_i à la classe j , basée sur f_1, \dots, f_m , est donnée par

$$P(j \mid \mathbf{x}_i, f_1, \dots, f_m) = \frac{\sum_{k=1}^m w(f_k) \hat{P}_k(j \mid \mathbf{x}_i)}{\sum_{k=1}^m w(f_k)},$$

où $w(\cdot)$ est une fonction de poids non-négative, et est donc une moyenne pondérée des probabilités d'appartenance sous chaque modèle du compromis.

La règle de classification compromis basée sur f_1, \dots, f_m est alors de classer l'observation \mathbf{x}_i dans la classe possédant la plus grande probabilité d'appartenance compromis.

Comme mentionné précédemment, nous voyons que nous ne combinons pas les paramètres des différents modèles, mais bien les probabilités d'appartenance *a posteriori*. Les paramètres des différents modèles du compromis peuvent donc avoir des significations différentes (par exemple la matrice de variance-covariance dans le cas normal multivarié ou dans le cas t multivarié), voire pas la même dimension, si par exemple on force les matrices de variance-covariance à être égales à l'identité dans un des modèles du compromis, ou si l'on désire estimer en plus le nombre de degrés de liberté des lois t multivariées.

Néanmoins, une autre difficulté apparaît avec ce type de compromis, et apparaît également lorsque l'on utilise une approche bayésienne pour la classification floue. En effet, la vraisemblance se trouve être invariante à un changement des identifiants des groupes. Autrement dit, pour le cas de deux *clusters*, en nommant le groupe 1 en « groupe 2 » et inversement, la valeur de la vraisemblance ne change pas. Le problème d'identifiabilité des groupes (*label switching* en anglais) se pose alors.

Une stratégie habituelle consiste à imposer des contraintes d'identifiabilité, comme par exemple

$$\pi_1 < \pi_2 < \dots < \pi_r,$$

ou encore

$$\mu_1 < \mu_2 < \dots < \mu_r,$$

pour des cas univariés. Néanmoins, Stephens (2000) montre que ce genre d'approche ne résoud pas systématiquement le problème, et introduit une approche basée sur la minimisation d'une fonction de perte. Une possibilité pourrait être de procéder comme suit :

1. Classifier les observations selon le modèle F_1 ;
2. Pour les autres modèles, classer également les observations dans les groupes. Pour chaque classe, comparer la somme des probabilités d'appartenance de ses observations aux groupes définis dans le modèle F_1 . On choisira alors les identifiants des classes de telle sorte que cette somme soit maximale.

Dans ce qui suit, nous nous intéresserons qu'à des cas comprenant deux groupes relativement bien séparés dans \mathbb{R}^2 , et le problème d'identification des classes pourra être traité relativement simplement. Néanmoins, pour des cas plus complexes, cette question est bien présente et nécessite la plus grande attention.

La procédure de classification compromis peut être résumée comme suit :

1. Pour chaque mélange de compromis f_k , $k = 1, \dots, m$, estimer les paramètres du modèle, grâce à l'algorithme EM par exemple :

$$\hat{\Theta}_k = \arg \max_{\Theta_k} \prod_{i=1}^n f_k(\mathbf{x}_i; \Theta_k);$$

2. Pour chaque observation \mathbf{x}_i , $i = 1, \dots, n$, et pour chaque mélange de compromis, calculer les probabilités d'appartenance aux classes :

$$\hat{P}_k(j | \mathbf{x}_i) = \frac{\hat{\pi}_{k,j} f_{k,j}(\mathbf{x}_i; \hat{\Theta}_k)}{f_k(\mathbf{x}_i; \hat{\Theta}_k)}, \quad j = 1, \dots, r;$$

3. Déterminer le poids à donner à chaque modèle de compromis :

$$w(f_1), \dots, w(f_m);$$

4. Calculer les probabilités d'appartenance compromis :

$$P(j | \mathbf{x}_i, f_1, \dots, f_m) = \frac{\sum_{k=1}^m w(f_k) \hat{P}_k(j | \mathbf{x}_i)}{\sum_{k=1}^m w(f_k)}, \quad i = 1, \dots, n, \quad j = 1, \dots, r,$$

et classer les observations en conséquence.

La fonction de poids $w(\cdot)$ se doit de représenter la crédibilité de chaque modèle par rapport aux autres. Comme dans le cas des estimateurs de Pitman compromis, on peut par exemple choisir la vraisemblance profil comme mesure d'adéquation des données à notre modèle :

$$w(f_k) = L(f_k) = \prod_{i=1}^n f_k(\mathbf{x}_i; \hat{\Theta}_k).$$

On notera également dans certains cas $w(F_k) = L(F_k)$, où F_k est la fonction de répartition du mélange f_k .

Exemple 4.2.2. La figure 4.2 montre un exemple concret de classification compromis.

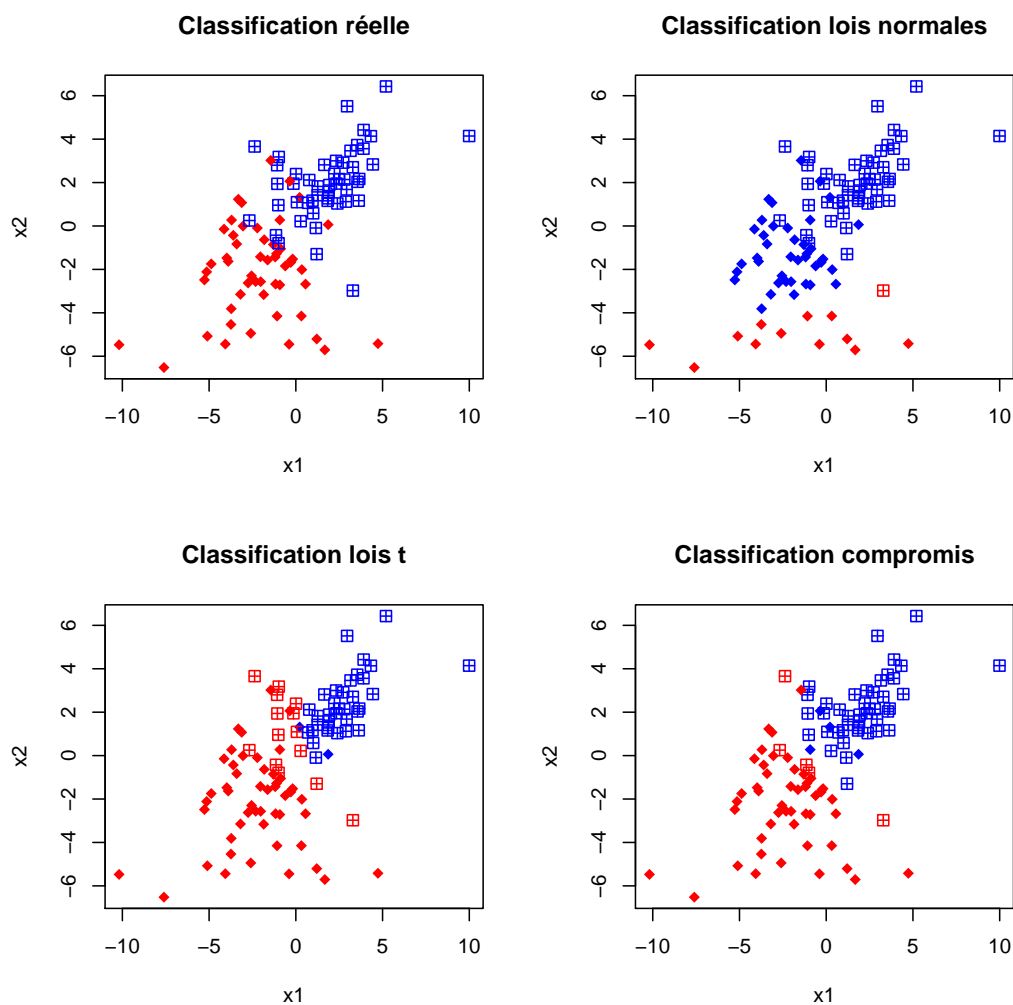


FIG. 4.2: Exemple de stratégie de classification compromis.

En haut à gauche, les deux groupes de données réelles, générées selon un mélange de lois multivariées t . En haut à droite, le résultat de la classification floue en supposant que la loi sous-jacente est un mélange de lois normales. On remarque que l'algorithme n'a pas su différencier la majorité des deux groupes, en considérant les données dispersées du bas de l'image comme étant un groupe à part entière, avec une structure de covariance nettement différente du reste des observations. Lorsque l'on utilise des lois multivariées t , on obtient la classification présentée en bas à gauche, qui parvient à gérer les points mentionnés précédemment. Enfin, en bas à droite, la classification compromis entre les deux cas précédents donne les meilleurs résultats, avec une bonne séparation des deux groupes.

Comme dans le cas des estimateurs de Pitman compromis, si l'on suppose que la fonction de poids est la vraisemblance profil du modèle, nous pouvons facilement déterminer le comportement asymptotique de la stratégie de compromis pour la régression. Ce résultat est présenté dans la proposition suivante.

Proposition 4.2.3. *Soit*

$$g = g(\mathbf{x}; \Theta) = \sum_{j=1}^r \pi_j g_j(\mathbf{x}; \boldsymbol{\theta}_j)$$

la distribution sous-jacente des observations, où $\Theta = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_r)$ représente les paramètres, et notons $G = G(\mathbf{x}; \Theta)$ sa fonction de répartition. Soit pour $k \in \{1, \dots, m\}$

$$f_k = f_k(\mathbf{x}; \Theta_k) = \sum_{j=1}^r \pi_{k,j} f_{k,j}(\mathbf{x} \mid \boldsymbol{\theta}_{k,j})$$

une distribution de compromis, où $\Theta_k = (\boldsymbol{\pi}_k, \boldsymbol{\theta}_{k,1}, \dots, \boldsymbol{\theta}_{k,2})$, et notons $F_k = F_k(\mathbf{x}; \Theta_k)$ sa fonction de répartition. Soit $\hat{\Theta}_k^n$ l'estimateur des paramètres Θ_k basé sur un échantillon de taille n , et notons

$$\hat{\Theta}_k = \lim_{n \rightarrow \infty} \hat{\Theta}_k^n.$$

Alors, lorsque $n \rightarrow \infty$, la règle de classification floue compromis basée sur les distributions f_1, \dots, f_m est la règle de classification utilisant le modèle f_k tel que

$$k = \arg \min_{l \in \{1, \dots, m\}} D \left(G(\mathbf{x}; \Theta) \parallel F_l(\mathbf{x}; \hat{\Theta}_l) \right),$$

où $D(\cdot \parallel \cdot)$ est la divergence de Kullback-Leibler.

Preuve Il convient donc d'étudier le comportement asymptotique de la fonction de poids, et donc plus particulièrement de la vraisemblance profil $L(F_k)$. Soit

$$l(F_k) = \log L(F_k) = \sum_{i=1}^n \log f_k(\mathbf{x}_i; \hat{\Theta}_k^n).$$

Ainsi :

$$\begin{aligned} \frac{1}{n} l(F_k) &\rightarrow \int \log f_k(\mathbf{x} \mid \hat{\Theta}_k) g(\mathbf{x}; \Theta) d\mathbf{x} \\ &= \int \log g(\mathbf{x}; \Theta) g(\mathbf{x} \mid \Theta) d\mathbf{x} - D \left(G(\mathbf{x}; \Theta) \parallel F_k(\mathbf{x}; \hat{\Theta}_k) \right), \end{aligned}$$

lorsque $n \rightarrow \infty$. Contrairement au cas des estimateurs de Pitman compromis, présenté précédemment, il n'est ici pas possible de simplifier l'expression, car nous ne pouvons

pas affirmer que les paramètres Θ et Θ_k ont la même signification. Ainsi, en posant par souci de simplification $\hat{F}_k = F(\mathbf{x}; \hat{\Theta}_k)$, nous avons que

$$\frac{1}{n}l(F_k) \rightarrow c(G) - D(G||\hat{F}_k),$$

lorsque $n \rightarrow \infty$, où $c(G)$ est une constante ne dépendant que de la distribution sous-jacente G . Pour n grand, on a donc

$$L(F_k) \sim \exp(nc(G) - nD(G||\hat{F}_k)).$$

Comme $c(G)$ est la même constante pour chaque distribution de compromis $f_k, k = 1, \dots, m$, et comme $\exp(-nD(G||\hat{F}_k)) \rightarrow 0 \forall k = 1, \dots, m$, (pour autant que F_k ne soit pas une transformation de G) lorsque $n \rightarrow \infty$, il en découle que

$$\frac{w(f_k)}{\sum_{l=1}^m w(f_l)} \rightarrow \begin{cases} 1, & \text{si } D(G||\hat{F}_k) \leq D(G||\hat{F}_l), \forall l = 1, \dots, m; \\ 0, & \text{sinon,} \end{cases}$$

lorsque $n \rightarrow \infty$. Ainsi, la règle de classification compromis converge vers celle associée à la distribution de compromis f_k telle que \hat{F}_k est la plus proche de G au sens de la distance de Kullback-Leibler. □

Dans ce qui suit, nous présentons le résultats de diverses simulations et comparons les résultats des méthodes de classification suivantes : classification floue utilisant un mélange de lois normales, classification floue utilisant un mélange de lois multivariée t , stratégie de classification compromis, et méthode des k - moyennes. Plusieurs situations sous-jacentes sont étudiées, et dans chaque cas nous comparerons les différentes méthodes sur le taux d'erreur de classification moyen de chacune.

4.2.1 Simulations dans le cas de mélanges de lois elliptiques

Nous présentons ici les résultats de simulations dans le cas de mélanges de lois elliptiques. Nous considérons 6 situations, dans lesquelles les centres et les structures de covariance des groupes changent. Dans chaque situation, nous générons les données soit avec un mélange de lois normales, soit avec un mélange de lois t multivariées avec 3 degrés de liberté. Les figures 4.3 et 4.4 présentent graphiquement ces situations, sous la forme d'ellipses de confiance. Pour chaque distribution, on a tracé les régions de confiance aux niveaux 50%, 90% et 95%, afin de représenter l'allure générale de chacune. Dans chacune des situations, nous avons généré 1000 échantillons de taille 30, 50 et 100 observations. Sauf mention explicite du contraire, les proportions du mélange ont été fixées à $\pi_1 = \pi_2 = 0.5$. Nous avons ensuite appliqué les méthodes de classification floue suivantes :

- classification floue utilisant des mélanges de lois normales ;
- classification floue utilisant des mélanges de lois t avec 3 degrés de liberté ;
- stratégie de compromis entre les deux méthodes ci-dessus.

Nous avons également pris en considération la méthode non-hiérarchique 2-moyennes. Nous comparons enfin les taux moyens de mauvaise classification de chacune des méthodes. Si l'on note par ρ_i , $i = 1, \dots, 1000$, le taux de mauvaise classification d'une certaine méthode pour l'échantillon i , alors le taux moyen de mauvaise classification de cette dernière est donné simplement par

$$\rho = \frac{1}{1000} \sum_{i=1}^{1000} \rho_i.$$

Les résultats sont présentés dans les tables 4.1 à 4.6 ci-dessous.

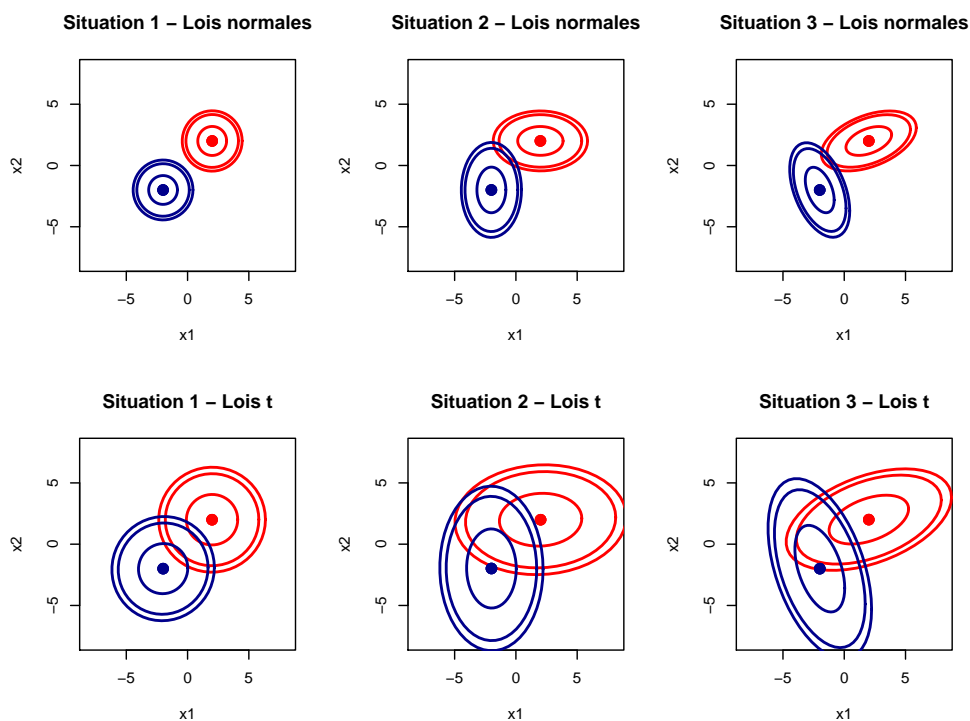


FIG. 4.3: Description des situations 1, 2 et 3.

4.2. Stratégie de compromis pour la classification floue

G	Méthode	n=30		n=50		n=100	
		Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	Normales	0.47	1.41	0.38	0.98	0.36	0.61
	Multiv. t	0.86	2.00	0.46	1.00	0.37	0.62
	Compromis	0.56	1.73	0.39	1.00	0.36	0.61
	2-moyennes	0.24	0.92	0.25	0.73	0.27	0.52
t	Normales	5.28	6.71	5.33	6.71	5.15	6.16
	Multiv. t	5.31	5.47	4.25	3.31	3.62	6.08
	Compromis	5.45	6.00	4.39	3.87	3.63	6.16
	2-moyennes	3.35	3.46	3.42	2.44	3.38	1.73

TAB. 4.1: Moyenne et écart-type (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 1.

G	Méthode	n=30		n=50		n=100	
		Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	Normales	2.81	3.61	2.72	3.16	2.11	1.73
	Multiv. t	3.91	5.56	2.92	3.31	2.12	1.70
	Compromis	3.18	4.69	2.73	3.16	2.11	1.73
	2-moyennes	1.81	2.44	1.96	2.00	1.73	1.34
t	Normales	10.05	8.94	9.92	9.05	11.66	11.35
	Multiv. t	9.42	7.00	7.91	4.89	6.79	3.00
	Compromis	9.89	8.00	8.25	6.08	6.85	3.31
	2-moyennes	6.98	5.19	6.77	4.24	6.42	2.64

TAB. 4.2: Moyenne et écart-type (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 2.

G	Méthode	n=30		n=50		n=100	
		Moy.	Var.	Moy.	Var.	Moy.	Var.
Normales	Normales	2.74	3.74	1.97	2.44	1.62	1.41
	Multiv. t	3.44	4.69	2.23	2.82	1.69	1.41
	Compromis	2.91	4.00	2.00	2.44	1.62	1.41
	2-moyennes	2.65	3.31	2.28	2.23	2.28	1.41
t	Normales	8.69	8.12	9.06	8.12	9.43	9.84
	Multiv. t	7.48	6.00	6.68	4.89	5.55	2.44
	Compromis	8.00	7.14	6.78	5.09	5.56	2.44
	2-moyennes	6.76	5.00	7.06	4.79	6.51	2.64

TAB. 4.3: Moyenne et variance (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 3.

Dans la situation 1 (table 4.1), les deux groupes possèdent une matrice de variance-covariance égale à la matrice identité. Lorsque la loi sous-jacente est un mélange de

lois normales, les groupes sont très bien séparés, et la classification floue utilisant des distributions normales est la mieux adaptée. En utilisant des lois t multivariées, la classification est légèrement moins bonne dans le cas de petits échantillons, mais lorsque n augmente, cette méthode devient équivalente à la précédente. La méthode de compromis se comporte de la même manière, avec néanmoins de meilleurs résultats que la méthode basée uniquement sur des lois t lorsque $n = 30$.

Lorsque la loi sous-jacente est un mélange de lois t , les deux groupes sont moins bien séparés, et la méthode basée sur les lois normales est mise en difficulté lorsqu'on la compare aux méthodes utilisant des lois t multivariées. Ceci est surtout visible avec de grands échantillons, même si les résultats présentent une grande variabilité dans tous les cas. La méthode de compromis se comporte comme précédemment et présente des résultats équivalents à une classification utilisant uniquement des lois t lorsque $n = 30$ déjà.

Dans la situation 2 (table 4.2), les deux groupes sont formés chacun à l'aide d'une matrice de variance-covariance diagonale. Lorsque la loi sous-jacente est un mélange de lois normales, la stratégie de compromis se comporte mieux que la classification utilisant uniquement des lois t , et se rapproche rapidement des performances de la classification utilisant des lois normales lorsque la taille de l'échantillon augmente.

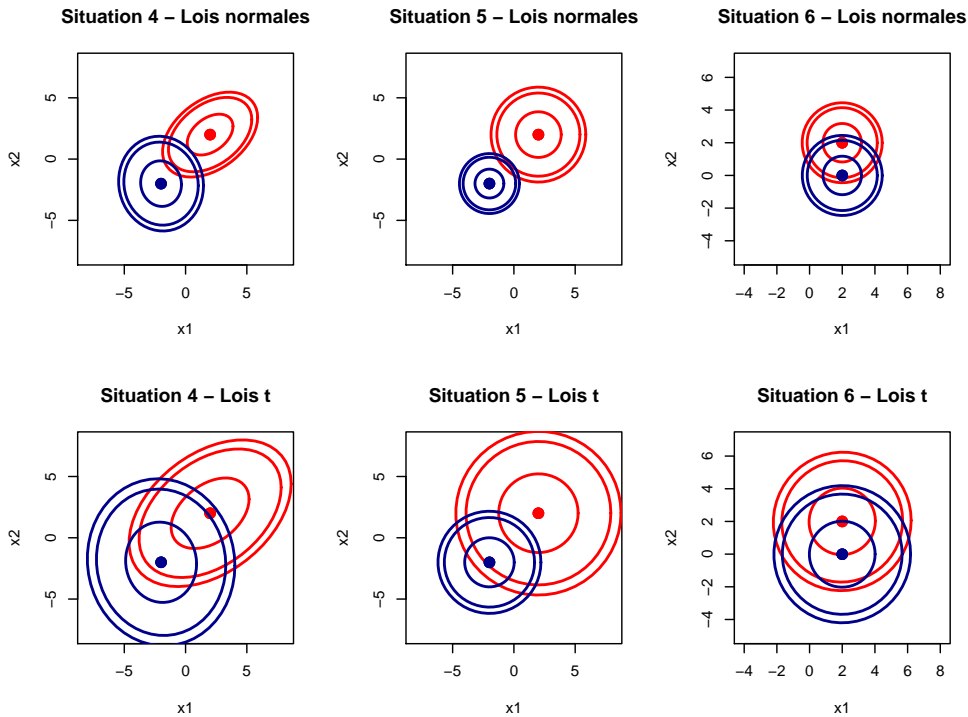


FIG. 4.4: Description des situations 4, 5 et 6.

Lorsque la loi sous-jacente présente une plus grande dispersion, la classification utilisant des lois normales est à nouveau en difficulté, et l'utilisation de lois t multivariées ou d'une stratégie de compromis est très efficace. Lorsque $n = 100$, ces deux méthodes sont comparables à la méthode des 2-moyennes, alors que la classification utilisant des lois normales présente un taux d'erreur moyen deux fois supérieur, ainsi qu'une variabilité nettement plus élevée.

G	Méthode	n=30		n=50		n=100	
		Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	Normales	7.88	8.06	6.93	5.91	5.39	3.61
	Multiv. t	9.77	9.32	7.93	7.14	5.48	3.46
	Compromis	8.89	8.94	7.10	6.24	5.38	3.61
	2-moyennes	4.69	4.12	4.60	3.16	4.45	2.00
t	Normales	16.82	12.72	20.84	14.31	26.69	16.06
	Multiv. t	14.89	10.14	13.31	8.18	10.67	4.79
	Compromis	16.65	12.04	16.19	11.70	11.86	7.54
	2-moyennes	9.99	6.40	9.71	5.00	9.33	3.31

TAB. 4.4: Moyenne et écart-type (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 4.

Dans les situations 3 et 4 (tables 4.3 et 4.4), les deux groupes présentent des structures de covariance différentes. A nouveau, la stratégie de compromis présente des performances comprises entre celles d'une classification basée uniquement sur des lois normales et celles d'une classification basée uniquement sur des lois t . Lorsque la loi sous-jacente est un mélange de lois normales, la stratégie de compromis se comporte rapidement comme la première méthode, tandis que lorsque la loi sous-jacente est plus dispersée, la stratégie de compromis devient équivalente à la deuxième méthode de classification. On remarque dans la situation 3 que la méthode des 2-moyennes donne des résultats moins bons que les méthodes de classification floue, du fait de la dépendance des variables x_1 et x_2 . Ceci est moins visible dans la situation 4, probablement car ces deux variables semblent non-corrélées dans le groupe dessiné en bleu.

G	Méthode	n=30		n=50		n=100	
		Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	Normales	3.73	6.16	2.60	3.61	1.71	1.41
	Multiv. t	4.61	7.41	3.01	4.00	1.76	1.41
	Compromis	4.16	7.07	2.63	3.61	1.71	1.41
	2–moyennes	4.03	4.79	3.78	3.31	3.19	2.00
t	Normales	9.49	9.94	10.78	10.53	12.66	12.76
	Multiv. t	10.31	10.00	7.36	5.09	6.18	2.82
	Compromis	10.57	10.44	8.23	7.34	6.35	3.31
	2–moyennes	7.08	6.71	6.62	4.12	6.48	3.16

TAB. 4.5: Moyenne et écart-type (en %) du taux d’erreur de classification pour les différentes méthodes prises en considération. Situation 5.

G	Méthode	n=30		n=50		n=100	
		Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	Normales	24.34	11.09	26.53	10.53	26.91	9.94
	Multiv. t	26.58	11.26	26.18	10.24	24.94	9.16
	Compromis	26.11	11.57	26.91	10.04	26.8	9.84
	2–moyennes	18.33	8.88	17.23	6.08	16.48	3.87
t	Normales	32.77	11.48	38.58	9.43	43.78	5.83
	Multiv. t	31.21	11.09	30.76	10.39	29.81	9.84
	Compromis	32.73	11.26	35.42	10.34	35.93	10.39
	2–moyennes	24.83	10.72	24.02	9.43	22.39	7.14

TAB. 4.6: Moyenne et écart-type (en %) du taux d’erreur de classification pour les différentes méthodes prises en considération. Situation 6.

La situation 5 (table 4.5) est une modification de la situation 1 dans laquelle le groupe dessiné en rouge est plus dispersé que le groupe bleu, et que la proportion des données dans le mélange est respectivement de $2/3$ et $1/3$, contrairement aux autres situations. Lorsque la loi sous-jacente est un mélange de lois normales, les trois méthodes de classification floue se comportent de manière équivalente. Lorsque G est un mélange de lois t multivariées, la classification floue utilisant des lois normales est en difficulté, tandis que les autres méthodes donnent des performances semblables, particulièrement lorsque l’échantillon est grand.

Dans la situation 6, les centres des deux groupes ont été rapprochés, et ces derniers sont quasiment confondus lorsque la loi sous-jacente est un mélange de lois t . Les résultats des méthodes de classification s’en ressentent, même si les méthodes de classification floues sont comparables dans le cas d’une loi sous-jacente mélangeant des lois normales.

Lorsque ce n'est plus le cas, une classification floue basée sur des lois t multivariées est alors préférable, et la stratégie de compromis présente un taux d'erreur plus élevé d'environ 5%. Il s'agit de la seule situation dans laquelle la stratégie de compromis ne se rapproche vraisemblablement pas de la meilleure classification.

4.2.2 Simulations dans le cas de mélanges de lois non-elliptiques

Nous présentons ici les résultats de simulations dans le cas de mélanges de lois non-elliptiques. Plus précisément, nous considérons 3 situations dans lesquelles les observations sont générées de la manière suivante. Pour chaque cas, deux groupes sont formés d'un certain nombre de points fixes. Nous sélectionnons aléatoirement des points, la moitié dans chaque groupe, et y ajoutons un bruit aléatoire. Nous les classifions ensuite à l'aide des méthodes suivantes :

- classification floue utilisant des mélanges de lois normales ;
- classification floue utilisant des mélanges de lois t avec 3 degrés de liberté ;
- stratégie de compromis entre les deux méthodes ci-dessus ;
- méthode des 2-moyennes.

La figure 4.5 montre la forme générale des groupes dans chacune des situations, avant l'ajout du bruit.

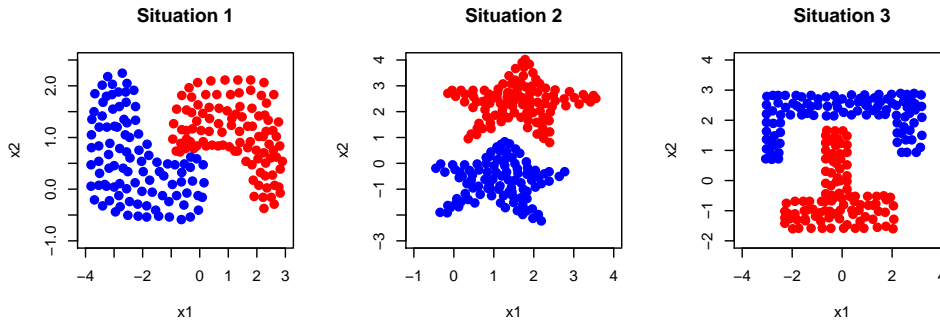


FIG. 4.5: Forme générale des groupes avant l'ajout d'un bruit aléatoire, pour les situations 1, 2 et 3.

Nous comparons enfin les taux de mauvaise classification moyens de chacune des méthodes, et les résultats sont présentés dans les tables 4.7 à 4.9 ci-dessous.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	11.04	8.24	11.46	7.79	12.08	7.05
Multiv. t	12.85	9.34	12.52	8.59	12.01	7.40
Compromis	11.60	8.67	11.56	7.91	12.02	7.05
2–moyennes	9.43	4.92	9.65	3.72	9.63	2.36

TAB. 4.7: Moyenne et écart-type (en %) du taux d’erreur de classification pour les différentes méthodes prises en considération. Situation 1.

La situation 1 (table 4.7) correspond à deux groupes en forme de « L » imbriqués. Après ajout de bruit et sélection aléatoire des points, les méthodes de classification floues, ainsi que la méthode compromis donnent des résultats équivalents, quelle que soit la taille de l’échantillon. La méthode des 2–moyennes présente des performances légèrement meilleures, notamment une variabilité plus faible dans le taux de mauvaise classification.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	4.10	5.64	3.39	4.33	2.30	2.19
Multiv. t	6.30	7.78	4.33	5.45	2.78	2.72
Compromis	4.87	6.94	3.46	4.59	2.30	2.19
2–moyennes	1.55	2.25	1.39	1.51	1.31	1.04

TAB. 4.8: Moyenne et écart-type (en %) du taux d’erreur de classification pour les différentes méthodes prises en considération. Situation 2.

Dans la situation 2 (table 4.8), les deux groupes sont relativement bien séparés, et en forme d’étoiles. Dans ce cas, la stratégie de compromis semble devenir équivalente à la méthode utilisant des lois normales, tandis que la méthode utilisant des lois t multivariées présente des performances légèrement inférieures. A nouveau, la méthode des 2–moyennes est la plus performante, et de loin la moins variable.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	19.57	11.90	20.40	10.87	20.13	8.92
Multiv. t	18.17	11.44	17.40	10.52	14.88	8.42
Compromis	18.54	11.56	19.04	10.41	19.04	8.64
2–moyennes	21.60	12.25	23.70	11.68	25.43	10.31

TAB. 4.9: Moyenne et écart-type (en %) du taux d’erreur de classification pour les différentes méthodes prises en considération. Situation 3.

Finalement, dans la situation 3 (table 4.9), les deux groupes ont des formes géométriques très particulières qui mettent en difficulté toutes les méthodes de classification

floue, ainsi que la méthode des 2–moyennes. Néanmoins, pour de grands échantillons ($n = 100$), la méthode utilisant des lois t multivariées est de loin la meilleure, mais la stratégie de compromis ne semble pas la privilégier. Nous pouvons donc remarquer que le fait de baser les poids sur la vraisemblance totale ne conduit pas forcément à une meilleure classification des données.

4.2.3 Simulations dans le cas de compromis sur la structure de covariance

Nous présentons ici les résultats de simulations dans le cas de compromis entre diverses structures de covariance. Nous considérons 6 situations dans lesquelles les centres et les structures de covariance des groupes changent. Dans chaque situation, nous générons les données avec un mélange de lois normales. Les figures 4.6 et 4.7 présentent graphiquement ces situations, sous la forme d’ellipses de confiance. Pour chaque distribution, on a tracé les régions de confiance aux niveaux 50%, 90% et 95%, afin de représenter l’allure générale de chacune.

Dans chacune des situations, nous avons généré 1000 échantillons de taille 30, 50 et 100 observations. Sauf mention explicite du contraire, les proportions du mélange ont été fixées à $\pi_1 = \pi_2 = 0.5$. Nous avons ensuite appliqué les méthodes de classification suivantes :

- classification floue utilisant des mélanges de lois normales, sans aucune restriction concernant la covariance des groupes (notée *Normales*) ;
- classification floue utilisant des mélanges de lois normales, en ajustant une matrice de variance diagonale pour chacun des deux groupes (notée *Normales diag*) ;
- classification floue utilisant des mélanges de lois normales, en exigeant une matrice de covariance égale à l’identité pour chacun des deux groupes (notée *Normales id*) ;
- stratégie de compromis entre les trois méthodes ci-dessus ;
- méthode des 2–moyennes.

Nous comparons enfin les taux de mauvaise classification moyens de chacune des méthodes, et les résultats sont présentés dans les tables 4.10 à 4.15 ci-dessous.

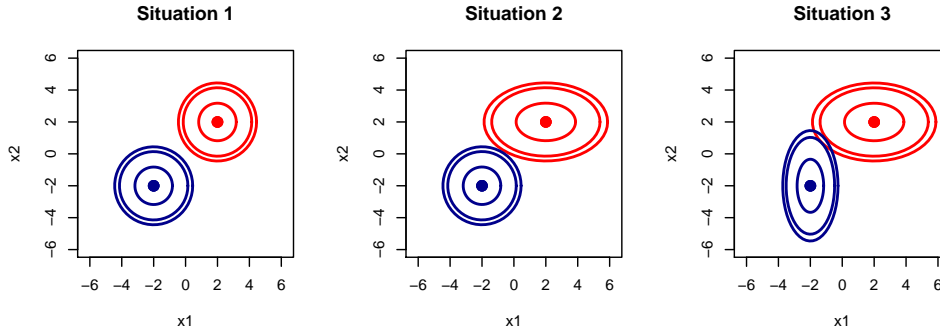


FIG. 4.6: Description des situations 1, 2 et 3.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	0.41	1.00	0.37	0.92	0.29	0.55
Normales diag	0.29	0.98	0.31	0.83	0.26	0.51
Normales id	0.23	0.85	0.24	0.71	0.23	0.47
Compromis	0.43	1.00	0.38	0.92	0.29	0.55
2-moyennes	0.23	0.85	0.24	0.71	0.23	0.48

TAB. 4.10: Moyenne et écart-type (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 1.

Dans la situation 1 (table 4.10), les deux groupes sont bien séparés, et les matrices de variance-covariance sont toutes les deux égales à la matrice identité. Dans ce cas, la troisième méthode fournit les meilleurs résultats, quelle que soit la taille de l'échantillon. En forçant les matrices de variance-covariance à être diagonales, les résultats sont semblables mais légèrement moins bons, de même lorsque l'on ne pose aucune restriction, comme dans la première méthode. La stratégie de compromis quant à elle se comporte comme cette dernière, quelle que soit la taille de l'échantillon, et ne converge donc apparemment pas vers la classification utilisant des lois normales de matrice de covariance identité.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	1.29	2.50	1.01	1.61	0.75	0.91
Normales diag	1.02	2.00	0.82	1.30	0.68	0.81
Normales id	1.27	2.21	1.11	1.64	1.01	1.04
Compromis	1.25	2.46	0.95	1.51	0.73	0.90
2-moyennes	1.28	2.21	1.07	1.58	0.99	1.03

TAB. 4.11: Moyenne et écart-type (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 2.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	1.92	2.86	1.65	2.14	1.27	1.26
Normales diag	1.59	2.46	1.43	1.78	1.20	1.14
Normales id	1.47	2.34	1.35	1.70	1.22	1.14
Compromis	1.9	2.84	1.62	2.07	1.24	1.18
2-moyennes	1.49	2.34	1.35	1.67	1.22	1.14

TAB. 4.12: Moyenne et écart-type (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 3.

Dans la situation 2 (table 4.11), un des groupes présente une matrice de covariance diagonale, et le second une matrice de covariance identité. Dans la situation 3 (table 4.12), les deux groupes possèdent une matrice de covariance diagonale. Dans le premier cas, la méthode Normales diag donne les meilleurs résultats, suivie par la stratégie de compromis. La méthode Normales id semble en difficulté, alors que sans aucune restriction, les résultats sont semblables lorsque l'échantillon est de grande taille. Contrairement à la situation 1, la stratégie de compromis semble cette fois adopter le même comportement que la méthode correspondant le plus à la situation sous-jacente.

Nous pouvions nous attendre à des résultats similaires dans la situation 3, mais lorsque l'échantillon est de taille $n = 30$ ou $n = 50$, la méthode Normales id donne les meilleures performances. Lorsque $n = 100$, la méthode Normales diag possède un léger avantage, mais toutes les méthodes sont relativement proches.

De manière générale, on aurait pu penser que dans les situations 1 à 3 la méthode utilisant des lois normales sans aucune restriction présenterait des meilleurs résultats, équivalents à ceux de la méthode utilisant des lois normales avec covariances diagonales. Comme cela ne semble pas être le cas, nous réalisons que l'estimation des paramètres supplémentaire semble coûteuse en termes de mauvaises classifications des données *a posteriori*. Cependant, les différences s'atténuent lorsque l'échantillon est relativement grand.

De plus, lorsque $n = 30$, on remarque que les performances de la stratégie de compromis sont systématiquement très proches de celles de la méthode utilisant des lois normales sans restriction. Il semble donc qu'en utilisant cette dernière, nous parvenons à une vraisemblance totale plus élevée, puisque nous permettons plus de liberté dans les paramètres, et cette méthode reçoit un poids important dans la stratégie de compromis. Cependant, cela ne se traduit pas forcément par une meilleure classification.

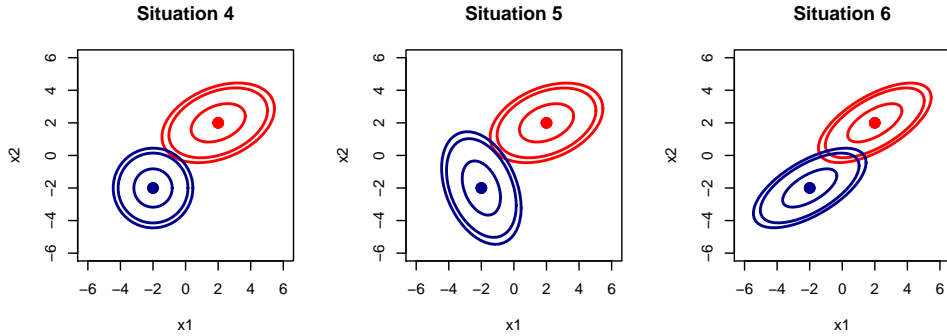


FIG. 4.7: Description des situations 4, 5 et 6.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	1.84	3.00	1.46	2.04	1.14	1.09
Normales diag	1.46	2.44	1.31	1.78	1.19	1.18
Normales id	1.63	2.44	1.46	1.76	1.53	1.34
Compromis	1.82	3.00	1.47	2.07	1.14	1.09
2-moyennes	1.63	2.64	1.43	1.78	1.51	1.34

TAB. 4.13: Moyenne et écart-type (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 4.

Dans la situation 4 (table 4.13), le groupe bleu possède une matrice de variance-covariance égale à l'identité, tandis que le groupe rouge présente une structure de corrélation entre les deux variables. La méthode Normal id donne les moins bons résultats, comme on pouvait s'y attendre. Les deux autres méthodes donnent des résultats comparables, et on remarque que la stratégie de compromis est très proche de la méthode utilisant des lois normales sans restriction. Il semble que dès que l'un des groupes possède une structure de corrélation, le poids associé à cette méthode dans la stratégie de compromis est très important, et ce même pour des petits échantillons.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	1.75	3.14	1.45	2.16	1.15	1.09
Normales diag	1.97	3.00	1.95	2.32	1.74	1.44
Normales id	1.55	2.42	1.54	1.89	1.46	1.22
Compromis	1.77	3.24	1.45	2.16	1.16	1.09
2-moyennes	1.55	2.42	1.52	1.87	1.45	1.22

TAB. 4.14: Moyenne et écart-type (en %) du taux d'erreur de classification pour les différentes méthodes prises en considération. Situation 5.

Méthode	n=30		n=50		n=100	
	Moy.	ET.	Moy.	ET.	Moy.	ET.
Normales	4.79	5.91	3.99	4.28	2.99	2.21
Normales diag	3.82	4.28	3.57	2.98	3.17	2.07
Normales id	3.83	3.78	3.73	2.79	3.61	1.97
Compromis	4.79	5.91	3.99	4.28	2.99	2.21
2–moyennes	3.81	3.81	3.74	2.79	3.63	2.00

TAB. 4.15: Moyenne et écart-type (en %) du taux d’erreur de classification pour les différentes méthodes prises en considération. Situation 6.

Dans les situations 5 et 6 (tables 4.14 et 4.15), les deux groupes présentent une structure de covariance entre les deux variables. Dans ce cas, la dernière remarque faite pour la situation 4 ci-dessus est exacerbée, et la stratégie de compromis donne les mêmes résultats que la méthode utilisant des lois normales sans restriction. On remarque néanmoins que cela ne se traduit pas forcément par un meilleur taux de classification lorsque l’échantillon est relativement de petite taille.

4.2.4 Remarques concernant les résultats obtenus

Les résultats des simulations présentés ci-dessus peuvent paraître surprenants, dans le sens où la méthode 2–moyennes présente parfois de meilleures performances que les méthodes probabilistes. En effet, on pouvait s’attendre à ce que ces dernières puissent s’adapter plus facilement à la forme elliptique des jeux de données considérés dans la section 4.2.1. Les figures 4.8 et 4.9 ci-dessous illustrent les problèmes que peuvent rencontrer les méthodes de classification floues.

Dans les deux cas présentés, la présence d’observations très éloignées des groupes principaux influence énormément l’estimation des matrices de variance-covariance dans le cas de la classification utilisant des lois normales. Dès lors, il peut arriver que cette méthode choisisse de former deux groupes centrés relativement au même point : un groupe pour la plupart des observations et un second contenant les données éloignées, en faisant « exploser » le déterminant de la matrice de variance-covariance de ce dernier.

Une telle solution présente malheureusement une vraisemblance plus élevée que la vraie solution, et même en indiquant les vraies valeurs des paramètres comme point de départ de l’algorithme EM, ce dernier va converger vers cette solution dégénérée. Remarquons également que la méthode de classification utilisant des lois t multivariées n’est pas à l’abri d’un tel problème. Ainsi, les méthodes de classification floue ne semblent pas compétitives pour des échantillons de taille relativement petite, surtout lorsqu’il est possible d’avoir des observations très éloignées, comme c’est le cas avec des lois t multivariées.

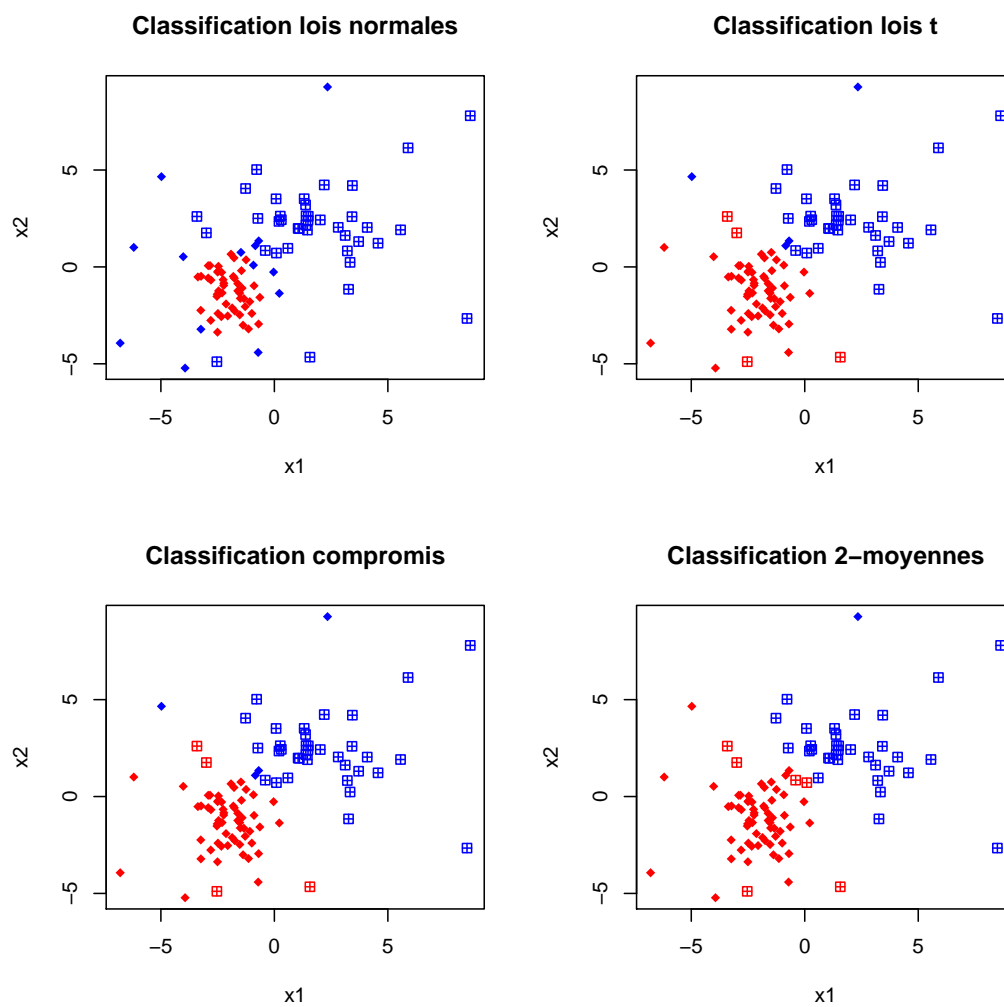


FIG. 4.8: Exemple de problème rencontré par les méthodes de classification floue.

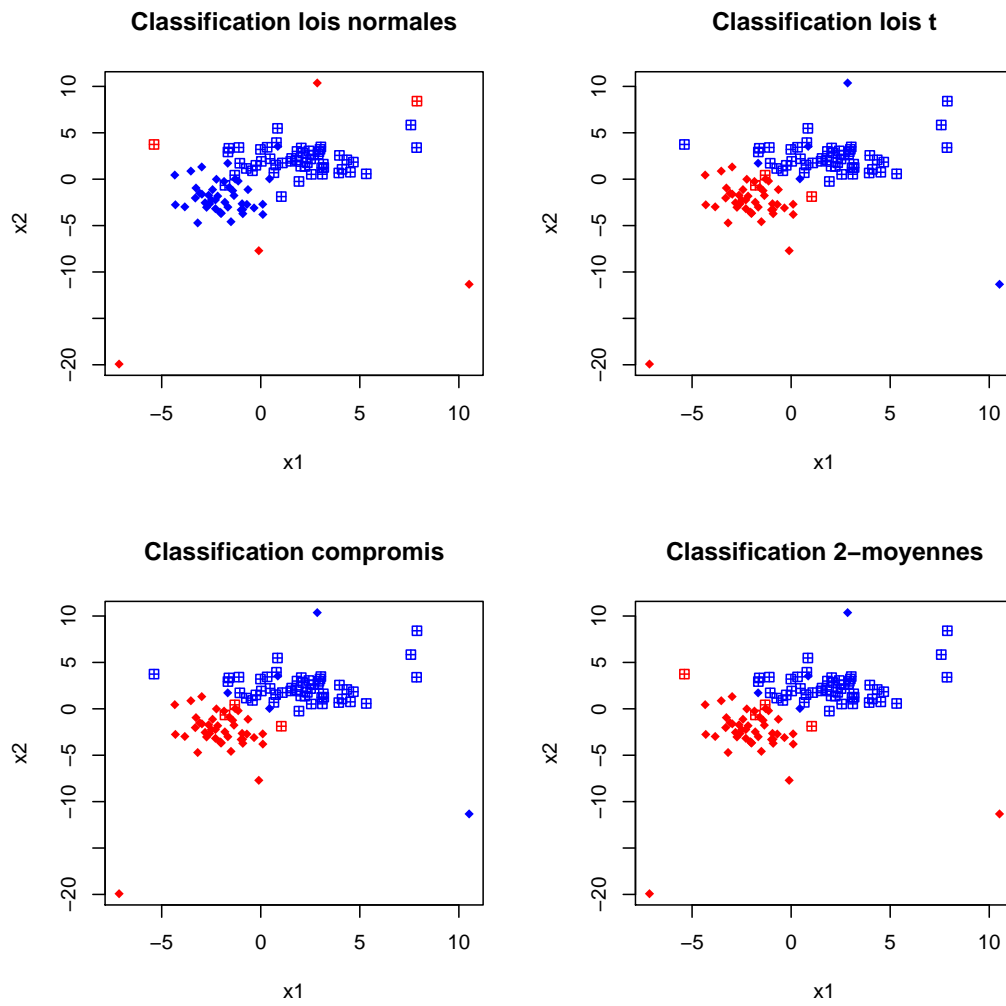


FIG. 4.9: Exemple de problème rencontré par les méthodes de classification floue.

Chapitre 5

Conclusion

Le but de ce travail était d'étudier une alternative à la sélection de modèles classique, en développant des stratégies de compromis entre différents modèles, dans une optique de robustesse.

Dans le chapitre 2, nous avons donné un aperçu des méthodes de sélection de modèles basées sur des scores, comme le critère AIC ou le critère BIC, ainsi qu'une présentation d'une approche alternative, les compromis de modèles. Nous avons présenté de manière générale deux types de compromis, les compromis bayésiens et les compromis fréquentistes, et c'est cette dernière approche qui a été développée dans la suite du travail et appliquée à l'estimation des paramètres de lieu et d'échelle, à l'estimation des paramètres de régression, et à la classification floue.

Dans le chapitre 3, nous avons tout d'abord présenté en détail les estimateurs de Pitman des paramètres de lieu et d'échelle, en donnant plusieurs exemples, notamment dans le cas d'une loi normale contaminée. Nous avons également démontré que les estimateurs de Pitman étaient asymptotiquement équivalents aux estimateurs du maximum de vraisemblance classiques, propriété qui aura été utilisée dans la suite du travail.

Nous avons ensuite construit dans un premier temps l'estimateur de Pitman compromis pour le paramètre de lieu, en utilisant une approche fréquentiste : cet estimateur est défini comme étant la moyenne pondérée des estimateurs de Pitman associés à plusieurs distributions sous-jacentes possibles. Le poids donné à chaque distribution est égal à la vraisemblance profil observée, qui donne une mesure de la qualité de l'ajustement. Par la suite, nous avons appliqué la même construction pour former un estimateur compromis du paramètre d'échelle. Pour ce faire, nous avons dû imposer une condition supplémentaire, à savoir normaliser la dispersion des distributions de compromis, afin d'assurer que le paramètre d'échelle ait la même signification dans chacun des modèles.

En étudiant le comportement asymptotique de la vraisemblance profil, et donc des poids attribués à chaque modèle, nous avons démontré que la distance de Kullback-

Leibler jouait un rôle important dans le comportement asymptotique des estimateurs compromis que nous avons construits. Cette notion de distance entre distributions nous a également permis de déterminer une stratégie afin de choisir les distributions de compromis, suivant une approche minimax.

Par la suite, nous avons comparé les performances de ces estimateurs avec d'autres estimateurs de lieu et d'échelle classiques et robustes, dans plusieurs situations différentes. Nous avons construit des estimateurs compromis basés sur des distributions normales contaminées et sur des lois t de Student. Les estimateurs compromis ont montré de très bons résultats en général, présentant une perte d'efficacité relative limitée dans la plupart des cas. Les estimateurs construits sur un compromis de lois normales contaminées se révèlent par contre inefficaces pour des lois à queues extrêmement lourdes, comme la loi Slash par exemple, du fait de leur structure sous-jacente gaussienne.

Nous avons ensuite utilisé la généralisation des estimateurs de Pitman au cas de la régression linéaire, et nous avons construit des estimateurs compromis similaires pour les paramètres de la régression. Nous avons également entrepris des simulations de Monte-Carlo, avec dans un premier temps des designs fixes, mettant en avant la problématique des points de leviers, et dans un deuxième temps des designs aléatoires. Les performances des estimateurs compromis étaient similaires au cas des paramètres de lieu et d'échelle.

Une question intéressante que l'on pourrait étudier à propos des estimateurs compromis tels que nous les avons construits est leur notion d'optimalité. Comme nous l'avons vu, les estimateurs compromis convergent vers les estimateurs du maximum de vraisemblance associés à la distribution la plus proche de la distribution sous-jacente réelle, au sens de la distance de Kullback-Leibler. Ainsi, nous pouvons dire que ces estimateurs compromis sont asymptotiquement optimaux si la distribution sous-jacente réelle fait partie des distributions de compromis utilisées. Il serait intéressant d'étudier d'autres types d'optimalité, comme par exemple chercher à construire des estimateurs compromis optimaux pour des petits échantillons plutôt.

Il serait également intéressant de déterminer la relation entre la distance maximale (au pire cas) entre une distribution de compromis et la vraie distribution des erreurs, et la variance asymptotique de l'estimateur compromis. En effet, nous avons utilisé cette distance afin de choisir le nombre de distributions de compromis. Il pourrait être utile de déterminer une éventuelle relation avec la variance de l'estimateur de Pitman compromis.

Dans le chapitre 4, nous avons étendu notre notion de stratégie de compromis au problème de la classification, et plus particulièrement à la méthode de classification floue. Nous avons présenté l'algorithme EM, très souvent utilisé pour l'estimation des paramètres dans la classification floue. Cette méthode se basant sur des modèles de mélanges de distributions, nous avons pu supposer divers mélanges sous-jacents, et construire une probabilité d'appartenance aux classes en réalisant un compromis fréquentiste. Il nous

était par contre impossible de construire des estimateurs compromis des paramètres du mélange, car ceux-ci ne possèdent pas la même signification suivant les queues des distributions sous-jacentes. Il convient également de se préoccuper de l'identifiabilité des classes sous les différents modèles.

Plusieurs simulations de Monte-Carlo ont été menées, afin de juger les performances de divers types de stratégies de compromis pour la classification floue. Tout d'abord, nous avons étudié une stratégie faisant un compromis entre une distribution sous-jacente étant un mélange de lois normales et une distribution sous-jacente étant un mélange de lois t multivariées, avec degrés de liberté fixes. Nous avons étudié diverses situations pour deux groupes dans \mathbb{R}^2 , en générant les observations suivant ces deux types de mélanges. Dans la plupart des cas, la stratégie de compromis présente des résultats équivalents à la méthode de classification floue correspondant au type de mélange sous-jacent : si ce dernier est formé avec des lois normales, alors la stratégie de compromis se comporte comme la méthode utilisant des lois normales, et si le mélange est formé de lois t multivariées, alors la stratégie de compromis se comporte comme la classification utilisant des lois t .

Cette stratégie de compromis a également été étudiée dans le cas de lois sous-jacentes non-elliptiques. Dans ce cas, ses performances sont comparables aux deux autres méthodes, qui sont rapidement mises en difficulté lorsque la forme des *clusters* devient particulière.

Nous avons également étudié une stratégie de compromis totalement différente, basée sur des lois normales dont la structure de covariance changeait. Nous avons construit une stratégie de compromis entre un mélange de lois normales sans aucune restriction sur la matrice de variance-covariance, des lois normales avec matrices de variance-covariance diagonales, et des lois normales avec matrices de variance-covariance égales à l'identité. Nous avons alors pu remarquer que cette stratégie ne se comportait pas forcément comme celle présentant le taux de mauvaise classification le plus faible. Il semble donc que le fait de baser les poids sur la vraisemblance totale ne garantit pas une meilleure classification.

L'application de l'idée de compromis à la classification donne donc des résultats mitigés. Il serait intéressant de poursuivre dans cette voie, car ce domaine des statistiques est extrêmement vaste et d'une importance grandissante. On pourrait par exemple construire le compromis en utilisant une autre manière de combiner les probabilités d'appartenance, ou encore construire des compromis basés sur des méthodes non-probabilistes, en déterminant une nouvelle manière d'attribuer un poids à chacune des classifications considérées.

Intégration numérique : méthode de Gauss-Legendre

Dans ce travail, afin de déterminer les différents estimateurs de Pitman utilisés, il était impératif d'utiliser une méthode d'intégration numérique pour approximer les différentes intégrales nécessaires. La méthode que nous avons choisie est la méthode de quadrature de Gauss, qui est une méthode d'intégration numérique exacte pour des polynômes de degré $2N - 1$, avec N points pris sur le domaine d'intégration.

De manière générale, cette méthode estime les intégrales du type

$$\int_a^b f(x)\omega(x)dx$$

par la somme pondérée

$$\sum_{i=1}^N w_i f(x_i),$$

où $\omega(x)$ est une fonction de poids non-négative, qui peut être introduite pour assurer l'intégrabilité de f , et où les x_i , $i = 1, \dots, N$, sont des réels distincts appelés *nœuds*, et où les w_i , $i = 1, \dots, N$, sont appelés coefficients de quadrature, ou encore *poids*.

En forçant la règle de quadrature à être exacte pour

$$f(x) = 1, x, x^2, \dots, x^{2N-1},$$

et en regardant les x_i et les w_i , $i = 1, \dots, N$, comme étant autant de paramètres, on obtient alors un ensemble de $2N$ équations non-linéaires, données par

$$\begin{aligned}
 w_1 + w_2 + \dots + w_N &= c_0 \\
 w_1 x_1 + w_2 x_2 + \dots + w_N x_N &= c_1 \\
 w_1 x_1^2 + w_2 x_2^2 + \dots + w_N x_N^2 &= c_2 \\
 &\vdots \\
 w_1 x_1^{2N-1} + w_2 x_2^{2N-1} + \dots + w_N x_N^{2N-1} &= c_{2N-1}.
 \end{aligned}$$

Pour le problème le plus classique, où le domaine d'intégration est $[-1, 1]$, et où $\omega(x) = 1$, la règle de quadrature est appelée la méthode de Gauss-Legendre. Les N nœuds sont alors les racines du N -ième polynôme de Legendre $P_N(x)$, et les poids sont obtenus par l'une ou l'autre des égalités suivantes :

$$w_i = \frac{-2}{(N+1)P'_N(x_i)P_{N+1}(x_i)} = \frac{4}{NP_{N+2}(x_i)P'_N(x_{i+2})}.$$

Les nœuds et les poids pour un grand nombre d'ordres peuvent être obtenus dans Abramowitz et Stegun (1964), page 875 et suivantes, mais sont également disponibles dans plusieurs librairies mathématiques, dans de nombreux langages de programmation.

Dans notre cas, nous devons évaluer des intégrales du type

$$I(g) = \int_0^\infty \int_{-\infty}^\infty g(s, t) f(s, t \mid \mathbf{c}) dt ds,$$

où $f(s, t \mid \mathbf{c})$ est la densité conjointe conditionnelle. Afin d'appliquer la quadrature de Gauss-Legendre, nous devons donc ramener le domaine d'intégration à $[-1, 1] \times [-1, 1]$, en utilisant le changement de variables suivant :

$$\begin{aligned}
 S &: [-1, 1] \rightarrow [0, \infty[\\
 u &\mapsto a \left(\frac{1+u}{1-u} \right)^b \\
 T &: [-1, 1] \rightarrow]-\infty, \infty[\\
 v &\mapsto c + d \log \left(\frac{1+v}{1-v} \right),
 \end{aligned}$$

où $a, c \in \mathbb{R}$, $b, d > 0$ sont des constantes telles que les nœuds de la méthode se situent pour la plupart là où la densité conjointe conditionnelle est « intéressante ». Nous avons donc

$$I(g) = \int_{-1}^1 \int_{-1}^1 g(S(u), T(v)) f(S(u), T(v) \mid \mathbf{c}) S'(u) T'(v) du dv,$$

où

$$S'(u) = ab \left(\frac{1+u}{1-u} \right)^{b-1} \frac{2}{(1-u)^2},$$

$$T'(v) = d \frac{2}{1-v^2},$$

sont les dérivées des fonctions S et T . $I(g)$ peut alors être approximée par

$$I(g) \approx \sum_{j=1}^N \sum_{k=1}^N w_j w_k g(S(u_j), T(v_k)) f(S(u_j), T(v_k) \mid \mathbf{c}) S'(u_j) T'(v_k).$$

Annexe **B**

Fonctions C₊₊ et R

Les diverses intégrales intervenant dans le calcul des estimateurs de Pitman ont été approximées à l'aide de la quadrature gaussienne présentée précédemment. Cette méthode a été programmée à l'aide des langages C₊₊ et R simultanément. Il est en effet possible de faire appel à des routines programmées en C₊₊ à partir du logiciel libre R, afin notamment d'accélérer l'exécution. Pour ce faire, il convient de stocker les fonctions C₊₊ dans une librairie dynamique et de charger cette dernière dans R.

Dans ce qui suit, nous ne donnerons que les fonctions utilisées pour le calcul des estimateurs de Pitman pour les paramètres de lieu et d'échelle dans le modèle simple. Pour les estimateurs de Pitman dans le cas de la régression, les fonctions sont semblables.

Fonctions C₊₊ de la librairie dynamique

Les bibliothèques GSL (*Gnu Scientific Library*) sont libres de droits et contiennent un grand nombre de routines mathématiques, comme des générateurs de nombres aléatoires, des méthodes d'interpolation, et autres fonctions spéciales. Elles peuvent être obtenues à l'adresse www.gnu.org/software/gsl/.

```
//Inclusion des bibliothèques, notamment gsl_randist qui contient diverses  
    lois de probabilité.
```

```
#include <stdio.h>  
#include <math.h>  
#include <gsl/gsl_randist.h>
```

La fonction `transformation` applique les changements de variables

$$S : [-1, 1] \rightarrow [0, \infty]$$

$$u \mapsto a \left(\frac{1+u}{1-u} \right)^b$$

$$T : [-1, 1] \rightarrow]-\infty, \infty[$$

$$v \mapsto c + d \log \left(\frac{1+v}{1-v} \right),$$

et évalue également les dérivées de ces transformations.

```
//Résultat sur R: une liste dont le 7e élément est le vecteur S, le 8e le
    vecteur T, le 9e le vecteur Sprime, et le 10e le vecteur Tprime
//Arguments: nodes -- les noeuds; Nb -- le nombre de noeuds; a,b,c,d -- les
    paramètres des transformations; S, T, Sprime, Tprime -- vecteurs de taille
    Nb qui recevront les résultats.
```

```
void transformation(double *nodes, int *Nb, double *a, double *b, double *c,
    double *d, double *S, double *T, double *Sprime, double *Tprime)

{   int i;
    double temp;

    for (i=0; i < *Nb; i++) {

        temp = nodes[i];

        S[i] = (*a)*pow((1+temp)/(1-temp),(*b));
        T[i] = (*c) + (*d)*log((1+temp)/(1-temp));
        Sprime[i] = (*a)*(*b)*((double)2/pow((double)1-temp,2))*pow(((double)
            1+temp)/((double)1-temp),((b)-1));
        Tprime[i] = (*d)*((double)2/((double)1-pow(temp,2)));
    }
}
```

Les fonctions suivantes évaluent la densité conjointe conditionnelle aux nœuds de la quadrature, qui auront donc été transformés auparavant. Pour chaque type de densité sous-jacente (normale contaminée, t de Student, Slash), il existe une fonction différente. Pour le cas de la loi normale, il suffit de fixer la fraction de contamination à 0 et d'utiliser la fonction de la loi normale contaminée.

```
//Résultat sur R: une liste dont le 8e élément doit être considéré comme une
    matrice Nb x Nb contenant les évaluations.
//Arguments: S,T -- les vecteurs des noeuds transformés, où l'évaluation sera
    faite; Nb -- le nombre de noeuds; -- eps,k -- les paramètres pour la loi
    normale contaminée; df -- degrés de liberté pour la loi t de Student; c --
```

```

    le vecteur de la configuration; n -- la taille de l'échantillon; res --
    vecteur de taille Nb^2 qui recevra les résultats.

//Pour la loi normale contaminée (et la loi normale standard)
void jointdensity_cnorm(double *S, double *T, int *Nb, double *eps, double *k,
    double *c, int *n, double *res)

{
    int i;
    int j;
    int l;

    double Stemp;
    double Ttemp;
    double ctemp;
    double p=0.39894;

    for (i=0; i < *Nb; i++) {
        Stemp = S[i];

        for (j=0; j < *Nb; j++) {
            Ttemp = T[j];

            for (l=0; l < *n; l++) {
                ctemp = c[l];

                res[i*(Nb) + j] = res[i*(Nb) + j] * ((1-eps)*
                    gsl_ran_gaussian_pdf(Stemp*(Ttemp + ctemp), 1) + *eps*
                    gsl_ran_gaussian_pdf(Stemp*(Ttemp + ctemp), *k));
            }

            res[i*(Nb) +j] = res[i*(Nb) + j]*pow(Stemp, *n-1);
        }
    }
}

```

```
//Pour la loi Slash
```

```
void jointdensity_slash(double *S, double *T, int *Nb, double *c, int *n,
    double *res)

{   int i;
    int j;
    int l;

    double Stemp;
    double Ttemp;
    double ctemp;
    double p=0.39894;

    for (i=0; i < *Nb; i++) {

        Stemp = S[i];

        for (j=0; j < *Nb; j++) {

            Ttemp = T[j];

            for (l=0; l < *n; l++) {

                ctemp = c[l];

                if (Stemp*(Ttemp + ctemp) == 0) {

                    res[i*(*Nb) + j] = res[i*(*Nb) + j] * ((double)1/((double)
                        )2*p));

                }

                else {

                    res[i*(*Nb) + j] = res[i*(*Nb) + j] * pow(Stemp*(Ttemp +
                        ctemp),-2) * (p - gsl_ran_gaussian_pdf(Stemp*(Ttemp +
                        ctemp),1));

                }

            }

            res[i*(*Nb) +j] = res[i*(*Nb) + j]*pow(Stemp, *n-1);

        }

    }

}
```

```

//Pour la loi t de Student

void jointdensity_student(double *S, double *T, int *Nb, double *df, double *c
, int *n, double *res)

{   int i;
    int j;
    int l;

    double Stemp;
    double Ttemp;
    double ctemp;
    double p=0.39894;

    for (i=0; i < *Nb; i++) {

        Stemp = S[i];

        for (j=0; j < *Nb; j++) {

            Ttemp = T[j];

            for (l=0; l < *n; l++) {

                ctemp = c[l];

                res[i*( *Nb) + j] = res[i*( *Nb) + j] * gsl_ran_tdist_pdf(Stemp
                    *(Ttemp + ctemp), *df);

            }

            res[i*( *Nb) +j] = res[i*( *Nb) + j]*pow(Stemp, *n-1);

        }

    }
}

```

Enfin, la fonction `iterate` effectue la quadrature et retourne les approximations des intégrales désirées.

```

//Résultats sur R: une liste dont le 8e élément est la constante d'intégration
de la densité conjointe conditionnelle, le 9e est proportionnel à  $E(s^2|c)$ 
), le 10e est proportionnel à  $E(ts^2|c)$ , le 11e est proportionnel à  $E(t^2
s^2|c)$  et le 12e est proportionnel à  $E(-\log s|c)$ .

```

```

//Arguments: S, T, Sprime, Tprime -- les vecteurs des noeuds transformés et la
dérivée des transformations; w -- le vecteur des poids pour la quadrature
; Nb -- le nombre de noeuds; eval -- la matrice des évaluations aux noeuds
transformés; cste, int1, int2, int3, int4 -- nombres réels qui vont
recevoir le résultat des 5 intégrales à approximer.

```

```
void iterate(double *S, double *T, double *Sprime, double *Tprime, int *Nb,
             double *w, double *eval, double *cste, double *int1, double *int2, double
             *int3, double *int4)

{
    int i;
    int j;

    *cste = 0;
    *int1 = 0;
    *int2 = 0;
    *int3 = 0;
    *int4 = 0;

    double wi;
    double wj;
    double Stemp;
    double Ttemp;
    double Sprimetemp;
    double Tprimetemp;
    double e;
    double lStemp;
    double ftemp;

    for (i=0; i < *Nb; i++) {

        wi = w[i];
        Stemp = S[i];
        Sprimetemp = Sprime[i];
        lStemp = -log(Stemp);

        for (j=0; j < *Nb; j++) {

            wj = w[j];
            Ttemp = T[j];
            Tprimetemp = Tprime[j];
            e = eval[( *Nb)*i + j];
            ftemp = wi*wj*Sprimetemp*Tprimetemp*e;

            *cste = *cste + ftemp;
            *int1 = *int1 + ftemp*pow(Stemp,2);
            *int2 = *int2 + ftemp*pow(Stemp,2)*Ttemp;
            *int3 = *int3 + ftemp*pow(Stemp*Ttemp, 2);
            *int4 = *int4 + ftemp*lStemp;

        }
    }
}
```

La fonction COF effectue les itérations pour le calcul d'un intervalle de confiance pour le paramètre de lieu.

```
//Résultats sur R: une liste dont le 8e élément est l'intégrale partielle sur
  t de -crit à l'infini

//Arguments: T, Sprime, Tprime -- les vecteurs des noeuds transformés et leur
  dérivée; Nb -- le nombre de noeuds; w -- les poids de la quadrature; eval
  -- les évaluations de la densité conjointe conditionnelle aux noeuds
  transformés; crit -- moins la borne inférieure de l'intégrale sur t; res
  -- un nombre réel qui contiendra le résultat de l'intégrale.

void COF(double *T, double *Sprime, double *Tprime, int *Nb, double *w, double
  *eval, double *crit, double *res)
{
  int i;
  int j;

  *res = 0;

  double wi;
  double wj;
  double Ttemp;
  double Sprimetemp;
  double Tprimetemp;
  double e;

  for (i=0; i < *Nb; i++) {

    wi = w[i];
    Sprimetemp = Sprime[i];

    for (j=0; j < *Nb; j++) {

      wj = w[j];
      Ttemp = T[j];
      Tprimetemp = Tprime[j];

      e = eval[( *Nb)*i + j];

      if (Ttemp > -*crit) {

        *res = *res + wi*wj*Sprimetemp*Tprimetemp*e;

      }
    }
  }
}
```

Fonction R

Il s'agit maintenant d'appeler les routines C++ à partir du logiciel R. Pour ce faire, il s'agit avant tout de charger la librairie dynamique contenant les routines, puis d'y faire appel en utilisant la fonction `.C`. Le premier argument de cette fonction est une chaîne de caractères correspondant au nom de la routine C++ à appeler, et les arguments suivants sont les arguments de la routine.

La fonction `pitman` ci-dessous calcule les estimateurs des paramètres de lieu et d'échelle pour diverses densités sous-jacentes des erreurs, ainsi qu'un intervalle de confiance bilatéral à 95% pour le paramètre de lieu. Les arguments sont :

- `data` : le vecteur des observations $y_i, i = 1, \dots, n$;
- `density` : une chaîne de caractères donnant le type de la densité sous-jacente, pouvant être `norm` (loi normale), `cnorm` (loi normale contaminée), `slash` (loi Slash) ou `student` (loi t de Student) ;
- `confidence` : un booléen indiquant si l'intervalle de confiance pour le paramètre de lieu doit être calculé ;
- `N` : le nombre de nœuds pour la quadrature gaussienne ;
- `...` : les paramètres additionnels pour les densités, par exemple ε et k pour la loi normale contaminée.

```
pitman <- function(data, density, confidence = FALSE, N=80, ...) {  
  
  #Calcul de la configuration c  
  #-----  
  
  c_i <- (data - mean(data))/(var(data)^(1/2))  
  
  #Choix de la densité sous-jacente  
  #-----  
  
  if (density == "norm") {  
  
    epsilon <- 0  
    k <- 1  
  
    f <- function(x) {return(dnorm(x,0,1))} }  
  
  else if (density == "cnorm") {  
  
    if (is.null(list(...)$epsilon) | is.null(list(...)$k)) {stop("Missing  
      arguments: epsilon and/or k.", call.=F)}  
  
    epsilon <- list(...)$epsilon  
    k <- list(...)$k
```

```

f <- function(x) {return((1-epsilon)*dnorm(x,0,1) + epsilon*dnorm(x,0,k))}
}

else if (density == "student") {

  if (is.null(list(...)$df)) {stop("Missing argument: df.", call.=F)}

  df <- list(...)$df

  f <- function(x) {return(dt(x, df))} }

else if (density == "slash") {

  f <- function(x) {if (x == 0) {return(((2*pi)^(-1/2))*(1/2))}
    else {return((1/(x^2))*(((2*pi)^(-1/2)) - dnorm(x,0,1)))}} }

else {stop("Density not supported", call.=F)}

#Moins le log de la densité conjointe conditionnelle
#-----

logf <- function(arg) {

  if (arg[1] < 0) {return(Inf)}

  else {

    temp <- (length(c_i) - 1)*log(arg[1])

    for (i in 1:length(c_i)) {

      temp <- temp + log(f(arg[1]*(arg[2] + c_i[i]))) }

    return(-temp) } }

#Optimisation (recherche du mode de la densité)
#-----

best <- c(var(data)^(1/2),0)
val <- Inf

smin = 1/(max(data) - min(data))
smax = max(data) - min(data)

up <- round(smax) + 2

for (l in 1:up) {

  stemp <- l*(smax-smin)/up + smin

```

```
if (logf(c(stemp,0)) < Inf) {  
  m <- optim(c(stemp, 0), logf)  
  if (m$value < val) {best <- m$par  
    val <- m$val } }  
}  
  
#Paramètres des changements de variables  
#-----  
  
a <- best[1]  
c <- best[2]  
b <- 1/(length(c_i)^(1/2))  
d <- 1/(length(c_i)^(1/2))  
  
#Noeuds et poids de la quadrature, disponibles dans la librairie statmod  
#-----  
  
library(statmod)  
  
G <- gauss.quad(N, "legendre")  
nodes <- G$nodes  
weights <- G$weights  
  
#Chargement de la librairie dynamique  
#-----  
  
dyn.load("pitman.dll")  
  
#Changements de variables  
#-----  
  
c.transfo <- .C("transformation", as.double(nodes), as.integer(N), as.double  
  (a), as.double(b), as.double(c), as.double(d), as.double(rep(0,N)), as.  
  double(rep(0,N)), as.double(rep(0,N)), as.double(rep(0,N)))  
  
S <- c.transfo[[7]]  
T <- c.transfo[[8]]  
Sprime <- c.transfo[[9]]  
Tprime <- c.transfo[[10]]  
  
#Evaluations de la densité conjointe conditionnelle  
#-----  
  
if (density == "norm") {  
  M <- matrix(.C("jointdensity_cnorm", as.double(S), as.double(T), as.integer(  
    N), as.double(0), as.double(1), as.double(c_i), as.integer(length(c_i)),  
    as.double(rep(1,N^2)))[[8]], N, N)  
  }  
}
```

```

else if (density == "cnorm") {

M <- matrix(.C("jointdensity_cnorm", as.double(S), as.double(T), as.integer(
  N), as.double(epsilon), as.double(k), as.double(c_i), as.integer(length(
  c_i)), as.double(rep(1,N^2))))[[8]], N, N)
  }

else if (density == "student") {

M <- matrix(.C("jointdensity_student", as.double(S), as.double(T), as.
  integer(N), as.double(df), as.double(c_i), as.integer(length(c_i)), as.
  double(rep(1,N^2))))[[7]], N, N)
  }
else if (density == "slash") {

M <- matrix(.C("jointdensity_slash", as.double(S), as.double(T), as.integer(
  N), as.double(c_i), as.integer(length(c_i)), as.double(rep(1,N^2)))
  [[6]], N, N)
  }

else {stop("Density not supported", call.=F)}

#Itérations de la quadrature
#-----

c.iter <- .C("iterate", as.double(S), as.double(T), as.double(Sprime), as.
  double(Tprime), as.integer(N), as.double(weights), as.double(M), as.
  double(0), as.double(0), as.double(0), as.double(0), as.double(0))

cste <- c.iter[[8]] #Constante d'intégration
s2 <- c.iter[[9]] #Proportionnel à l'intégrale de s^2 f(s,t|c)
ts2 <- c.iter[[10]] #Proportionnel à l'intégrale de s^2 t f(s,t|c)
t2s2 <- c.iter[[11]] #Proportionnel à l'intégrale de s^2 t^2 f(s,t|c)
los <- c.iter[[12]] #Proportionnel à l'intégrale de -log(s) f(s,t|c)

#Résultats renvoyés par la fonction
#-----

est <- list(T_c = - ts2/s2, T_y = (-ts2/s2)*(var(data)^(1/2)) + mean(data),
  logS_c = los/cste, logS_y = (los/cste) + log(var(data)^(1/2)))

#Intervalle de confiance pour le paramètre de lieu
#(basé sur une approximation normale)
#-----

if (confidence == TRUE) {

t_start <- -ts2/s2

h <- 1/100

```

```
start <- .C("COF", as.double(T), as.double(Sprime), as.double(Tprime), as.
  integer(N), as.double(weights), as.double(M), as.double(t_start), as.
  double(0))[[8]]/cste

droite <- .C("COF", as.double(T), as.double(Sprime), as.double(Tprime), as.
  integer(N), as.double(weights), as.double(M), as.double(t_start + h), as.
  .double(0))[[8]]/cste
gauche <- .C("COF", as.double(T), as.double(Sprime), as.double(Tprime), as.
  integer(N), as.double(weights), as.double(M), as.double(t_start - h), as.
  .double(0))[[8]]/cste

qtn <- qnorm(c(gauche, start, droite), mean=0, sd=1)

approx <- lm(c(t_start - h, t_start, t_start + h) ~ qtn)

coeffs <- as.vector(approx$coefficients)

bbc <- qnorm(c(0.025, 0.975), mean=0, sd=1)*coeffs[2] + coeffs[1]

bornes <- bbc*(var(data)^(1/2)) + mean(data)

#Déchargement de la librairie dynamique
#-----

dyn.unload("pitman.dll")

#Résultats renvoyés par la fonction
#-----

return(list(Estimators = est, Confidence = list(T_y = bornes))) }

else {dyn.unload("pitman.dll")

  return(list(Estimators = est))}

}
```

Bibliographie

- Abramowitz, M. et Stegun, I. A. (1964) *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.
- Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pp. 267–281. Budapest : Akadémiai Kiadó.
- Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automatic Control* **AC-19**, 716–723. System identification and time-series analysis.
- Akaike, H. (1977) On entropy maximization principle. In *Applications of statistics (Proc. Sympos., Wright State Univ., Dayton, Ohio, 1976)*, pp. 27–41. Amsterdam : North-Holland.
- Akaike, H. (1978) A new look at the Bayes procedure. *Biometrika* **65**(1), 53–59.
- Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K. et Gottardo, R. (2010) Combining mixture components for clustering. *Journal of Computational and Graphical Statistics* **19**(2), 332–353.
- Belsley, D. A., Kuh, E. et Welsch, R. E. (1980) *Regression diagnostics : identifying influential data and sources of collinearity*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics.
- Box, G. E. P. (1953) Non-normality and tests on variances. *Biometrika* **40**, 318–335.
- de Bruijn, N. G. (1981) *Asymptotic methods in analysis*. Third edition. New York : Dover Publications Inc.
- Burnham, K. P. et Anderson, D. R. (2002) *Model selection and multimodel inference*. Second edition. New York : Springer-Verlag. A practical information-theoretic approach.

BIBLIOGRAPHIE

- Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **158**(3), pp. 419–466.
- Cover, T. M. et Thomas, J. A. (2006) *Elements of information theory*. Second edition. Hoboken, NJ : Wiley-Interscience [John Wiley & Sons].
- Creasy, M. A. (1954) Symposium on interval estimation : Limits for the ratio of means. *Journal of the Royal Statistical Society. Series B. Methodological* **16**, 186–194.
- Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**(1), 1–38. With discussion.
- Draper, D. (1995) Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. Ser. B* **57**(1), 45–97. With discussion and a reply by the author.
- Easton, G. S. (1985) *Finite Sample and Asymptotic Approaches to Compromise Estimation Including Compromise Maximum Likelihood Estimators*. Ph.D. thesis, Princeton University, Dept. of Statistics. Unpublished.
- Easton, G. S. (1986) Compromise maximum likelihood estimators for location. Technical report, University of Chicago, Graduate School of Business, Statistics Research Center.
- Easton, G. S. (1991) Compromise maximum likelihood estimators for location. *Journal of the American Statistical Association* **86**(416), 1051–1064.
- Fieller, E. C. (1954) Symposium on interval estimation : Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B. Methodological* **16**, 175–185.
- Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222**, pp. 309–368.
- Fisher, R. A. (1925) Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.
- Fisher, R. A. (1930) Inverse probability. *Proceedings of the Cambridge Philosophical Society* **26**, 528–535.
- Fisher, R. A. (1934) Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **144**(852), 285–307.
- Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.
- Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*. Hafner Publishing Company, New York.

- Fraley, C. et Raftery, A. E. (1998) How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal* **41**, 578–588.
- Fraser, D. A. S. (1979) *Inference and linear models*. New York : McGraw-Hill International Book Co. Advanced Book Program.
- Freedman, D. A. (1963) On the asymptotic behavior of Bayes' estimates in the discrete case. *Annals of Mathematical Statistics* **34**, 1386–1403.
- Hampel, F. R. (1971) A general qualitative definition of robustness. *Ann. Math. Statist.* **42**, 1887–1896.
- Hampel, F. R. (1974) The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383–393.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. et Stahel, W. A. (1986) *Robust statistics*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. New York : John Wiley & Sons Inc. The approach based on influence functions.
- Hartigan, J. A. (1965) The asymptotically unbiased prior distribution. *Annals of Mathematical Statistics* **36**, 1137–1152.
- Hastie, T., Tibshirani, R. et Friedman, J. (2001) *The elements of statistical learning*. Springer Series in Statistics. New York : Springer-Verlag. Data mining, inference, and prediction.
- Henderson, H. V. et Searle, S. R. (1981) On deriving the inverse of a sum of matrices. *SIAM Rev.* **23**(1), 53–60.
- Hjort, N. L. et Claeskens, G. (2003) Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98**(464), 879–899.
- Hjort, N. L. et Claeskens, G. (2008) *Model selection and model averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge : Cambridge University Press.
- Hoeting, J. A., Madigan, D., Raftery, A. E. et Volinsky, C. T. (1999) Bayesian model averaging : a tutorial. *Statist. Sci.* **14**(4), 382–417. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors.
- Huber, P. J. (1964) Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.
- Huber, P. J. (1981) *Robust statistics*. New York : John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics.
- Hurvich, C. M. et Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika* **76**(2), 297–307.

BIBLIOGRAPHIE

- Hurvich, C. M. et Tsai, C.-L. (1995) Relative rates of convergence for efficient model selection criteria in linear regression. *Biometrika* **82**(2), 418–425.
- Kim, S.-H. et Cohen, A. S. (1998) On the behrens-fisher problem : A review. *Journal of Educational and Behavioral Statistics* **23**(4), 356–377.
- Kullback, S. et Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Statistics* **22**, 79–86.
- Leamer, E. E. (1978) *Specification searches*. New York : Wiley-Interscience [John Wiley & Sons]. Ad hoc inference with nonexperimental data, Wiley Series in Probability and Mathematical Statistics.
- Lehmann, E. L. et Casella, G. (1998) *Theory of point estimation*. Second edition. Springer Texts in Statistics. New York : Springer-Verlag.
- MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pp. Vol. I : Statistics, pp. 281–297. Berkeley, Calif. : Univ. California Press.
- Mallows, C. (1998) The zeroth problem. *Amer. Statist.* **52**(1), 1–9.
- McLachlan, G. J. (1992) *Discriminant analysis and statistical pattern recognition*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. New York : John Wiley & Sons Inc. A Wiley-Interscience Publication.
- McLachlan, G. J. et Peel, D. (1998) Robust cluster analysis via mixtures of multivariate t -distributions. In *Advances in pattern recognition (Sydney, 1998)*, volume 1451 of *Lecture Notes in Comput. Sci.*, pp. 658–666. Berlin : Springer.
- McQuarrie, A. D. R. et Tsai, C.-L. (1998) *Regression and time series model selection*. River Edge, NJ : World Scientific Publishing Co. Inc.
- Morgenthaler, S. (1986) Asymptotics for configural location estimators. *The Annals of Statistics* **14**(1), 174–187.
- Morgenthaler, S. et Tukey, J. W. (eds) (1991) *Configural polysampling*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. New York : John Wiley & Sons Inc. A route to practical robustness, A Wiley-Interscience Publication.
- O'Brien, F. L. (1984) *Polyefficient and Polyeffective Simple Linear Regression Estimators and the Biweight Regression Estimator*. Ph.D. thesis, Princeton University, Dept. of Statistics. Unpublished.
- Pitman, E. J. G. (1939) The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika* **30**(3/4), 391–421.

- Pregibon, D. et Tukey, J. W. (1981) Assessing the behavior of robust estimates in small samples : Introduction to configural polysampling. Technical report, Princeton University, Dept. of Statistics.
- Rao, J. S. et Tibshirani, R. (1997) The out-of-bootstrap method for model averaging an selection. Technical report, University of Toronto, Dept. of Statistics.
- Regal, R. R. et Hook, E. B. (1991) The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* **10**(5), 717–721.
- Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.* **6**(2), 461–464.
- Searle, S. R. (1982) *Matrix algebra useful for statistics*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. Chichester : John Wiley & Sons Ltd.
- Shao, J. et Tu, D. S. (1995) *The jackknife and bootstrap*. Springer Series in Statistics. New York : Springer-Verlag.
- Sin, C.-Y. et White, H. (1996) Information criteria for selecting possibly misspecified parametric models. *J. Econometrics* **71**(1-2), 207–225.
- Stephens, M. (2000) Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62**(4), 795–809.
- Stigler, S. M. (2010) The changing history of robustness. *The American Statistician* **64**(4), 277–281.
- Takeuchi, K. (1976) Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematic Sciences)* **153**, 12–18.
- Tierney, L. et Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**(393), 82–86.
- Tukey, J. W. (1960) A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, pp. 448–485. Stanford, Calif. : Stanford Univ. Press.
- Tukey, J. W. (1962) The future of data analysis. *Ann. Math. Statist.* **33**, 1–67.
- Tukey, J. W. (1981) Some advanced thoughts on the data analysis involved in configural polysampling directed towards high performance estimates. Technical report, Princeton University, Dept. of Statistics.
- Tukey, J. W. (1987) Configural polysampling. *SIAM Rev.* **29**(1), 1–20.
- Wel, J. (1975) Least squares fitting of an elephant. *Chemtech* **Feb.**, 128–129.
- Wolfe, J. H. (1970) Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* **5**, 329–350.
- Yang, Y. (2001) Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96**(454), 574–588.

FOURNIER Nicolas

Vieille Route 1
1124 Gollion
+41 (0)21 861 44 46
+41 (0)79 387 51 17

23.09.1983
Célibataire
Suisse
nicolas.fournier@gmail.com

FORMATION

- 2007-2011 **Thèse de doctorat en statistiques robustes à l'École Polytechnique Fédérale de Lausanne (EPFL)**
Développement de méthodes robustes pour la régression linéaire et la classification, basées sur des compromis de modèles
- Présentations orales lors de l'International Conference on Robust Statistics 2008 (Antalya, Turquie), 2009 (Parme, Italie) et 2010 (Prague, République Tchèque)
- Publication : Compromise Pitman Estimators, J. of Stat. Planning and Inference (2011)
- 2002-2007 **Bachelor et Master of Science en ingénierie mathématique à l'EPFL**
Spécialisation : statistiques médicales, statistiques génétiques

DISTINCTIONS ACADÉMIQUES

- 2007 **Prix de la Fondation Annaheim**
Récompense un projet de Master ou de semestre de haut niveau consacré au rapprochement des sciences de la vie et de l'informatique (bio-informatique, systèmes bio-inspirés)

PROJETS ACADÉMIQUES

- 2007 **Expression génétique, analyse de survie et puissance sous cross-validation**
Développement et étude d'un test génétique pour la réapparition de tumeurs cancéreuses, basé sur des données de puces ADN (projet de Master)
- 2006 **Fragilité dans l'étude de données de survie**
Etude d'un modèle de risque à composante d'hétérogénéité (projet de semestre)

EXPÉRIENCE PROFESSIONNELLE

- 2005-2011 **Assistant à l'EPFL**
Cours de Probabilités et Statistiques, sections de Mathématiques, Sciences et technologies du vivant, Génie mécanique et Génie chimique, niveau Bachelor et Master, classes de 10 à 120 étudiants, remplacements pour cours ex-cathedra
- 2010 **Consulting statistique pour la Fédération Internationale de Basket-ball Amateur (FIBA)**
Etude d'un système de ranking pour les joueurs de street basket-ball, en collaboration avec l'Academy of Sports Science and Technology (AISTS)
- 2000-2001 **Réalisation du site web www.romanel-sur-morges.ch**
En collaboration avec une commission communale créée à cet effet
Responsable technique, design, contenu et mises à jour

LANGUES

Français : langue maternelle
Anglais : bonnes connaissances orales et écrites
Allemand : connaissances niveau maturité, rapidement perfectibles

CENTRES D'INTÉRÊT

Tennis, badminton, golf, trekking
Cuisine