Université de Montréal

Modèles Pareto hybrides pour distributions asymétriques et à queues lourdes

par

Julie Carreau

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

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

décembre 2007

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée

Modèles Pareto hybrides pour distributions asymétriques et à queues lourdes

présentée par

Julie Carreau

a été évaluée par un jury composé des personnes suivantes :

Pascal Vincent
(président-rapporteur)

Yoshua Bengio
(directeur de recherche)

Christian Léger
(membre du jury)

Hugh Chipman
(examinateur externe)

Clément Bernard
(représentant du doyen de la FES)

Thèse acceptée le:
19 décembre 2007

SOMMAIRE

Nous proposons une classe d'estimateurs de densité qui s'adaptent au cas où les observations proviennent d'une distribution asymétrique, multi-modale et qui possède des queues lourdes. Ce type de distributions survient dans des domaines tels que la finance et l'assurance. Les mélanges de gaussiennes sont des modèles non-paramétriques flexibles qui ont de bonnes propriétés d'approximation lorsque le nombre de composantes est bien choisi par rapport à la taille de l'ensemble d'entraînement. Cependant, ces modèles ont plus de difficulté à bien représenter la queue de la distribution sous-jacente lorsque celle-ci est lourde car peu d'observations se trouvent généralement dans cette région. Pour trouver une solution à ce problème, nous avons recours à la théorie des valeurs extrêmes afin d'utiliser des hypothèses paramétriques appropriées lorsqu'il est nécessaire d'extrapoler dans les régions où il n'y a pas d'observations. Plus précisément, nous nous inspirons de la méthode PoT, "Peaks-over-Threshold", qui a été développée en hydrologie pour modéliser les observations excédentaires à un seuil fixé. La distribution de ces excédents est modélisée à l'aide de la Pareto généralisée qui a la propriété d'approximer arbitrairement bien la queue de toute distribution connue. Nous proposons d'utiliser la loi Pareto hybride dans un mélange de distributions. Cette loi est une extension de la Pareto généralisée à l'axe des réels. La Pareto hybride est construite en juxtaposant une loi Normale tronquée et une loi de Pareto généralisée. Des conditions de continuité sont imposées au point de jonction. Cette loi hybride possède une queue supérieure qui a un comportement semblable à celui de la Pareto généralisée. Par ailleurs, le seuil inhérent à la méthodologie PoT est alors défini de manière implicite comme le point de jonction de la composante ayant la queue la plus lourde. Cette composante détermine l'indice de queue du

mélange. Le mélange de Pareto hybrides offre donc une façon alternative d'estimer l'indice de queue associée à des observations extrêmes. Dans de nombreuses applications, on possède de l'information ayant un pouvoir prédictif sur la variable d'intérêt. On s'intéresse alors à modéliser la densité conditionnelle de Y, la variable d'intérêt, étant donné X, le vecteur contenant l'information prédictive. Lorsque la distribution de Y étant donné X est lourde, asymétrique ou multimodale, nous proposons d'utiliser comme estimateur de densité conditionnelle, le mélange de Pareto hybrides dont les paramètres sont des fonctions de X. Ces fonctions sont modélisées à l'aide d'un réseau de neurones à une couche cachée. Les réseaux de neurones sont des modèles non-paramétriques qui permettent en principe d'approximer toute fonction continue. Des expériences sur des jeux de données artificielles et réelles démontrent que la performance en termes de log-vraisemblance du mélange de Pareto hybrides, inconditionnel et conditionnel, est supérieure à celle d'autres estimateurs de densité non-paramétriques.

Mots-clés : estimation de densité, distribution à queue lourde, loi de Pareto généralisée, valeurs extrêmes, mélange de distributions, réseau de neurones

SUMMARY

We put forward a class of density estimators that can adapt to asymmetric, multi-modal and heavy-tailed distributions. Such distributions occur in many application domains such as finance and insurance. Mixture of gaussians are flexible non-parametric density estimators that have good approximation properties when the number of components is well chosen with respect to the training set size. However, those models perform poorly on heavy-tailed data because few observations occur in the tail area. To solve this problem, we resort to extreme value theory where methods based on sound parametric assumptions have been developped to enable extrapolation beyond the range of the observations. More precisely, we build on the PoT method that was developped in hydrology where PoT stands for "Peaks-over-Threshold". The observations exceeding a given threshold are modeled by the generalized Pareto distribution. This distribution can approximate arbitrarily well the tail of most distributions. We build a new distribution, the hybrid Pareto, by stitching together a truncated Normal and a generalized Pareto distribution. We impose continuity constraints at the junction point. The hybrid Pareto is thus a smooth distribution that can be used in a mixture model. The behavior of the upper tail of the hybrid is similar to the behavior of the generalized Pareto tail. Moreover, the threshold inherent in the the PoT methodology can now be defined implicitly as the junction point of the component with the heaviest tail. This component also determines the tail index of the mixture. Hence, the hybrid Pareto mixture offers an alternate way to estimate the tail index associated with heavy-tailed data. In several applications, information that has predictive power on the variable of interest is available. In that case, we want to model the conditional density of Y given X, the vector containing predictive

information. When the distribution of Y given X is asymmetric, multi-modal and heavy-tailed, we propose to use a mixure of hybrid Paretos whose parameters are functions of X. Those functions are implemented by means of a neural network with one hidden layer. Neural neworks are non-parametric models that can, in principle, approximate any continuous function. Experiments on artificial and real data sets show that the hybrid Pareto mixture, unconditional and conditional, outperforms other density estimators in terms of log-likelihood.

Keywords : density estimation, heavy-tailed distribution, generalized Pareto distribution, extreme values, mixture of distributions, neural networks

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABBRÉVIATIONS

$\mathbb{N}$: ensemble des nombres naturels

$\mathbb{R}$: ensemble des nombres réels

$\#$ { **A**} : cardinalité de A

ξ : indice de queue

$\mathcal{MDA}(H_\xi)$: domaine d'attraction maximale de H_ξ

EZ^r: moment d'orde r

$\psi = (\xi, \mu, \sigma)$: vecteur de paramètre de la loi Pareto hybride

$\mathcal{D}_n$: ensemble d'entraînement de taille n

$\mathcal{D}_l$: ensemble de test de taille l

ummh : mélange inconditionnel avec composantes Pareto hybrides

ummg : mélange inconditionnel avec composantes gaussiennes

umml : mélange inconditionnel avec composantes Log-Normales

uparzen : estimateur de la fenêtre de Parzen

cmmh : mélange conditionnel avec composantes Pareto hybrides

cmmg : mélange conditionnel avec composantes gaussiennes

cmml : mélange conditionnel avec composantes Log-Normales

cparzen : estimateur conditionnel de la fenêtre de Parzen

Fréchet-lin-mod : jeu de données Fréchet conditionnel avec dépendance linéaire des paramètres et queue modérément lourde

Fréchet-lin-lourde : jeu de données Fréchet conditionnel avec dépendance linéaire des paramètres et queue lourde

Fréchet-sin-mod : jeu de données Fréchet conditionnel avec dépendance sinusoïdale des paramètres et queue modérément lourde

Fréchet-sin-lourde : jeu de données Fréchet conditionnel avec dépendance sinusoïdale des paramètres et queue lourde

REMERCIEMENTS

Je désire remercier tout d'abord mon directeur de recherche, Yoshua Bengio, qui m'a donné le goût de me lancer dans des études doctorales. Sa vision de la recherche, son enthousiame et sa vaste expérience scientifique ont soutenu ma motivation et m'ont permis de mener à bien ma thèse. Je remercie également Roch Roy, pour avoir éveillé mon intérêt pour les statistiques par le biais des séries chronologiques et pour m'avoir conseillé durant ma progression au doctorat. J'en profite aussi pour exprimer ma gratitude et mon affection à mes amis et ma famille qui m'ont fourni un soutien moral indispensable durant ces années d'études. Enfin, merci à mes collègues du LISA pour leur disponibilité et l'ambiance conviviale au laboratoire.

AVANT-PROPOS

La théorie des valeurs extrêmes est une branche des probabilités qui étudie le comportement des queues de distribution. Des questions typiques auxquelles on cherche à répondre sont : quelle est la probabilité qu'une variable aléatoire Z prenne des valeurs dans un ensemble C alors que peu ou pas d'observations se trouvent dans C, ou encore quel est le niveau qui est dépassé en moyenne une fois sur n où n est grand et donc on s'attend à ce que le niveau se retrouve dans la queue de la distribution. La théorie des valeurs extrêmes met en place des méthodes fondées sur un raisonnement mathématique rigoureux qui permettent d'extrapoler au-delà de la région contenant les observations. L'idée intuitive derrière ce raisonnement est que la queue de la majorité des distributions peut être classifiée selon un des trois types suivants : le type de queue légère, comme c'est le cas de la loi Normale, lourde, comme pour les lois α-stables lorsque $\alpha < 2$ ou encore finie, comme la loi Uniforme.

Par ailleurs, une branche de l'apprentissage statistique, l'estimation de densité non-paramétrique, permet de faire l'analyse de données univariées ou multivariées sans faire d'hypothèses spécifiques sur la distribution. Un modèle de densité permet de répondre à de multiples questions sur les données : y a-t-il multi-modalité, quel est le comportement de la queue de la distribution, la distribution est-elle asymétrique, etc... L'estimation de densité peut également amener à la prise de décision telle que la classification ou le regroupement de données ("clustering"). De manière générale, l'apprentissage statistique cherche à développer des algorithmes, pour différents types de tâches (pas seulement l'estimation de densité), qui apprennent à partir d'exemples. Plus précisément, ces algorithmes possèdent

un mécanisme interne flexible qui s'ajuste aux données et requièrent peu d'hypothèses a priori. Les hypothèses endossées sont typiquement très générales, elles ont trait à la continuité de la fonction à apprendre par exemple, et ne sont pas contraignantes comme les hypothèses habituelles endossées par les modèles paramétriques. De fait, de nombreux algorithmes issus de l'apprentissage statistique possèdent des propriétés d'approximation universelle. Ce qui distingue principalement un algorithme non-paramétrique d'un algorithme paramétrique est que la complexité de l'algorithme, que l'on peut généralement quantifier par le nombre de paramètres, augmente avec le nombre d'exemples disponibles.

La faiblesse des algorithmes d'estimation de densité non-paramétriques est qu'ils ne réussissent pas facilement à extrapoler au-delà de l'intervalle couvert par les données. Ces algorithmes offrent donc des résultats peu satisfaisants lorsque'il s'agit de modéliser les queues de distributions qui sont lourdes. Par ailleurs, les méthodes provenant de la théorie des valeurs extrêmes ne visent qu'à estimer la queue de la distribution et ne proposent pas de solution pour la partie centrale. L'objectif de ce travail est de jeter les ponts entre les deux domaines de la théorie des valeurs extrêmes et de l'apprentissage statistique. Pour la partie centrale de la distribution, lorsque suffisamment de données sont disponibles, nous aurons recours à un modèle d'estimation de densité non-paramétrique. Pour la queue de la distribution, lorsque des observations extrêmes sont présentes mais peu nombreuses, nous ferons appel à la théorie des valeurs extrêmes.

Cette thèse s'organise comme suit. Les chapitres 1 et 2 présentent une introduction des domaines de la théorie des valeurs extrêmes et de l'apprentissage statistique respectivement. Les chapitres 3 et 4 décrivent les contributions principales qui traitent respectivement de l'estimation de densité inconditionnelle et conditionnelle à l'aide de modèles Pareto hybrides. Finalement, le chapitre 5 conclut la thèse.

Chapitre 1

VALEURS EXTRÊMES

De manière informelle, on définit comme extrême, une observation qui se trouve très éloignée de l'ensemble des autres observations. On retrouve ce type d'observations dans de nombreux domaines tels que la finance et l'assurance. Dans le milieu financier, la distribution des rendements d'instruments financiers est un outil essentiel à la gestion de portefeuille et des risques financiers. On observe des rendements extrêmes suite à des mouvements de tendance comme la bulle technologique ou suite à des crashs boursiers majeurs. Les praticiens de la finance travaillent généralement avec le logarithme du rendement : soit p_t le prix d'un instrument financier au temps t, le log-rendement est donné par $\log(p_{t+\Delta_t}) - \log(p_t)$ où Δ_t représente un intervalle de temps. La figure 1.1 illustre les log-rendements quotidiens de l'indice boursier S&P500[1] ; on y observe des rendements extrêmes reliés au crash financier de 1987. Dans le milieu de l'assurance, la distribution des réclamations est primordiale pour que les compagnies d'assurance puissent évaluer les risques qu'elles encourent par rapport à leurs clients. La figure 1.2 illustre les réclamations d'un compagnie d'assurance danoise ; on y voit des réclamations dont le montant est de beaucoup supérieur à l'ensemble des autres réclamations.

La présence d'observations extrêmes peut être expliquée par plusieurs facteurs : des erreurs de mesure, la présence d'une autre distribution qui corrompt la distribution principale et forme avec celle-ci un mélange de distributions ou

[1]Cet indice est considéré comme le plus représentatif de l'économie américaine, il est composé de 500 compagnies majeures dans les industries principales des États-Unis.

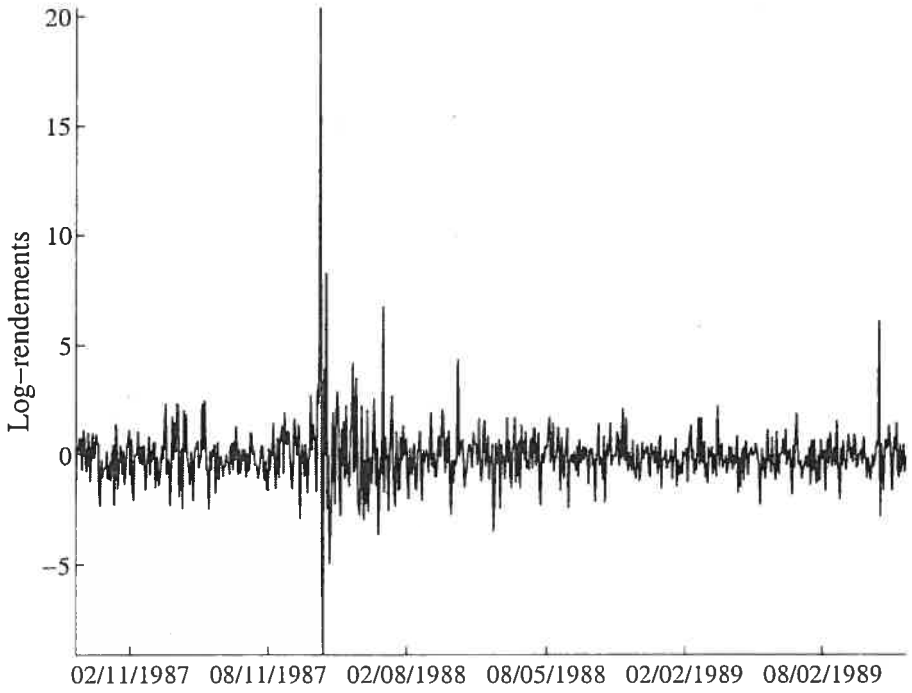

FIG. 1.1: *Log-rendements de l'indice boursier S&P 500 entre le 12/02/1986 et le 11/16/1989.*

encore le fait que la distribution sous-jacente ait une queue lourde. C'est ce dernier cas qui nous intéresse. Soit $F(z)$, la fonction de répartition de la distribution en question ; on appelle $\overline{F}(z) = 1 - F(z)$, la queue de la distribution. Par queue lourde, on entend une distribution dont plus de masse de probabilité se trouve dans la queue. Par conséquent, la probabilité d'événement se trouvant très éloigné du centre de la distribution est plus élevée. On caractérisera plus formellement le fait qu'une loi ait une queue lourde dans la sous-section suivante. Les lois α-stables avec $\alpha < 2$, la loi de Pareto, la loi de Fréchet et la loi t de Student sont des exemples de distributions à queues lourdes alors que la loi Normale et la loi exponentielle ont des queues légères.

En gestion des risques de portefeuille, la valeur-à-risque (VaR) est une mesure couramment utilisée. Il s'agit d'un quantile de la distribution des profits et des pertes du portefeuille. Soit $\Delta p_t = p_{t+\Delta_t} - p_t$, la variation de la valeur du portefeuille entre les moments t et $t+\Delta t$. La VaR de niveau 5% correspond à une perte qui survient au plus 5% du temps, autrement dit $P(\Delta p_t \leq -VaR) = 5\%$. Cependant, la VaR ne fournit qu'une borne sur la pire perte possible avec une certaine

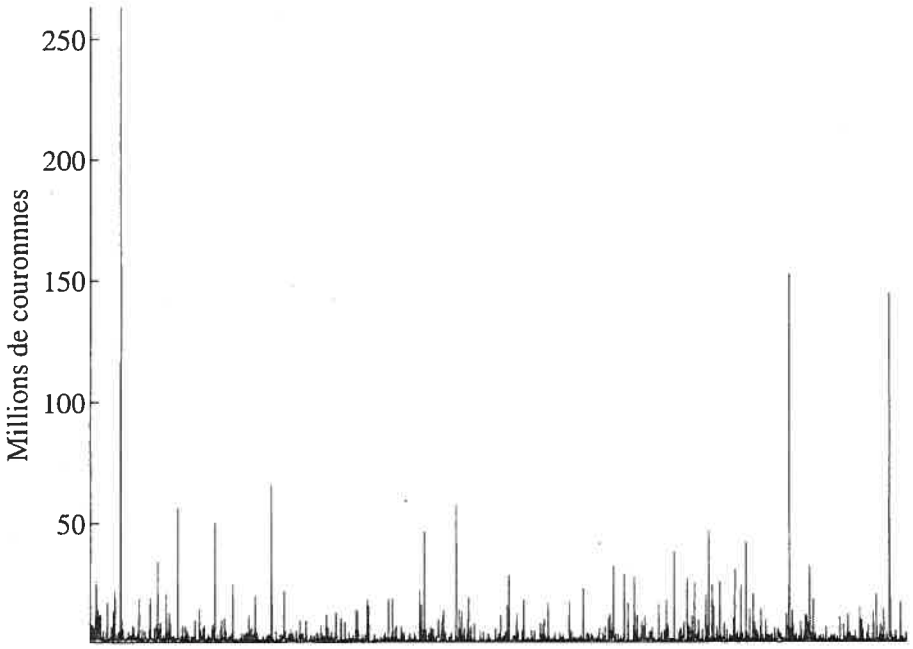

FIG. 1.2: *Réclamations reliées aux sinistres d'incendies survenus au Danemark.*

probabilité, elle ne donne aucune indication à savoir quelle peut être l'ampleur des pertes une fois la VaR dépassée. Certains auteurs ont proposé l'utilisation de l'espérance de la perte [**41**] pour remédier aux lacunes de la VaR. Il s'agit de l'espérance de la variation du portefeuille sachant que celle-ci est inférieure à la VaR : $E[\Delta p_t | \Delta p_t < -VaR]$. Par ailleurs, Mandelbrot [**32**] et Fama[**16**] ont démontré que la distribution des rendements d'instruments financiers présente des queues lourdes. Afin d'estimer la VaR et l'espérance de la perte, il est donc nécessaire de modéliser la queue inférieure de la distribution des profits et des pertes du portefeuille.

Dans le milieu de l'assurance, les assureurs se protègent contre les pertes associées aux plus grandes réclamations en ayant recours à des compagnies de réassurance. Un type de contrat de réassurance possible consiste à couvrir l'assureur contre les pertes se trouvant dans un certain intervalle, appelé l'intervalle de réassurance. Il est donc essentiel de développer de bons estimateurs de la queue de la distribution dans le but d'estimer la probabilité que des réclamations surviennent dans l'intervalle de réassurance. McNeil [**34**] utilise un modèle de distribution à queue lourde pour modéliser les réclamations dans l'intervalle de réassurance.

1.1. Théorie des valeurs extrêmes

La théorie des valeurs extrêmes et les méthodes qui en découlent [**15**] permettent la caractéristation, l'estimation et l'extrapolation des queues de distributions. Typiquement, on cherche à évaluer la probabilité qu'une variable aléatoire Z prenne une valeur dans un ensemble $C = \{z|z > u\}$ où u est grand. Dans la plupart des cas, aucune observation ne correspond à l'ensemble C. Par exemple, on aimerait estimer la probabilité que le niveau de la mer dépasse la hauteur des digues protégeant le littoral. L'estimation de quantile est une variante de ce problème. Soit F, la fonction de répartition de Z, le quantile de niveau $q \in [0, 1]$ est défini par : $z_q = \inf\{z|F(z) \geq q\}$. Soit n, le nombre d'observations dont on dispose. On cherche à estimer z_q où q est soit très petit $q < 1/n$ (quantile de la queue inférieure) soit très grand $q > 1 - 1/n$ (quantile de la queue supérieure) de sorte qu'il faille extrapoler en dehors de l'intervalle couvert par les données. Le calcul de la VaR tombe dans cette catégorie de problèmes. Nous décrirons tour à tour deux types d'approches d'estimation de la queue d'une distribution qui ont été développées dans le cadre de la théorie des valeurs extrêmes.

1.1.1. Méthode des maxima par blocs

Le premier type de méthodes a recours aux maxima d'une variable aléatoire pour faire l'inférence de la queue de la distribution. Soient $Z_1, Z_2, \ldots, Z_n$, n copies indépendantes de Z dont la fonction de répartition est F et soit $M_n = \max(Z_1, \ldots, Z_n)$, le maximum sur ces n variables. Alors, on peut modéliser la distribution de M_n par la distribution aux valeurs extrêmes généralisée. La fonction de répartition de la loi aux valeurs extrêmes généralisée est donnée par :

Définition 1.1.1 (Loi aux valeurs extrêmes généralisée).

$$H_{\xi;\mu,\beta}(z) = \begin{cases} \exp\left\{-\left(1+\xi\left(\frac{z-\mu}{\beta}\right)\right)^{-1/\xi}\right\} & si\ \xi \neq 0 \\ \exp\left\{-\exp\{-\left(\frac{z-\mu}{\beta}\right)\}\right\} & si\ \xi = 0 \end{cases}$$

où $1 + \xi\left(\frac{z-\mu}{\beta}\right) > 0$.

Le paramètre ξ est appelé paramètre de queue, il détermine l'épaisseur de la queue de la distribution. Les paramètres μ et β contrôlent respectivement l'emplacement et la dispersion de la distribution. Puisque ξ est le paramètre caractérisant la loi aux valeurs extrêmes généralisée, on note souvent $H_{\xi;\mu,\beta}$ simplement par H_ξ. Lorsque $\xi > 0$, H_ξ est une paramétrisation de la distribution de Fréchet, lorsque $\xi = 0$, H_ξ correspond à la loi de Gumbel et lorsque $\xi < 0$, H_ξ représente la loi de Weibull. L'utilisation de la loi aux valeurs extrêmes généralisée pour modéliser la distribution des maxima est justifiée par le théorème (1.1.1) dû à Fisher-Tippett [**15**]. Ce théorème stipule que si M_n, adéquatement centré et réduit, converge en distribution vers une loi H non-dégénérée[2], alors H est forcément la loi aux valeurs extrêmes généralisée. On dit alors que F (ou Z) appartient au domaine d'attraction maximale de H, ce qui se note $F \in MDA(H)$ (ou $Z \in MDA(H)$).

Théorème 1.1.1 (Fisher-Tippett). *Si $\exists c_n > 0$ et $d_n \in \mathbb{R}$ et une fonction de répartition H telle que $c_n^{-1}(M_n - d_n) \xrightarrow{d} H$[3] alors $H = H_{\xi;\mu,\beta}$ est la loi aux valeurs extrêmes généralisée.*

Pratiquement toutes les distributions connues sont dans un domaine d'attraction maximale $MDA(H_\xi)$, pour un ξ donné. Ceci permet de classifier les distributions selon leur type de queue, lourde, exponentielle ou finie, lorsque ξ est respectivement positif, nul ou négatif. Lorsque $\xi > 0$, $F \in MDA(H_\xi)$ implique qu'il existe une fonction $L(z)$ à variation lente, c'est-à-dire que

$$\lim_{z\uparrow\infty} \frac{L(tz)}{L(z)} = 1, \quad \forall t > 0, \tag{1.1.1}$$

telle que $\overline{F}(z) = z^{-1/\xi}L(z)$. Formellement, une distribution est à queue lourde si elle appartient au domaine d'attraction maximale de la loi de Fréchet. Dans ce cas, $\overline{F}(z)$ se comporte éventuellement comme $z^{-1/\xi}$. La queue de la distribution décroît donc à vitesse polynômiale. La loi de Student, les lois α-stables et la loi de Pareto appartiennent au MDA de la Fréchet. Les lois qui appartiennent au MDA de la Gumbel, auquel cas $\xi = 0$, ont des queues modérémment lourdes à légères. Pour z suffisamment grand, la décroissance de la queue de la distribution

[2]Une loi est dite dégénérée si toute la densité est concentrée en un seul point z_0, c'est-à-dire $p(z_0) = 1$.

[3]$\xrightarrow{d}$ dénote la convergence en distribution.

est exponentielle. On trouve dans ce MDA les lois Normale, Exponentielle et Log-Normale. Finalement, le MDA de la Weibull, pour lequel $\xi < 0$, contient des lois dont la queue supérieure est finie, comme les lois Uniforme et Bêta.

En pratique, on observe des réalisations de la variable aléatoire Z sur une période de temps que l'on découpe en "blocs", souvent il s'agit de blocs d'un an. On utilise ensuite les maxima des observations sur ces blocs pour estimer les paramètres de la loi aux valeurs extrêmes généralisée. Il faut choisir la taille du bloc de sorte que les maxima soient approximativement indépendants même si les observations à l'intérieur du bloc ne le sont pas. On suppose également que suffisamment d'observations surviennent dans un bloc pour qu'on puisse approximer la distribution des maxima par la distribution asymptotique. On peut donc répondre aux questions concernant les valeurs extrêmes d'un jeu de données, par exemple quelle est la probabilité que le maximum dépasse une valeur donnée, à l'aide de la distribution aux valeurs extrêmes généralisée.

1.1.2. Méthode des excès au-delà d'un seuil

Le deuxième type de méthodes utilise les observations qui se trouvent au-delà d'un seuil élevé pour modéliser la queue de la distribution sous-jacente. Cette méthodologie a d'abord été développée en hydrologie sour le nom de "Peaks-over-Threshold" (PoT). Davison et Smith [**10**] font une revue exhaustive des travaux sur cette méthode. Elle procède de la façon suivante. Soient $\{Z_1, \ldots, Z_n\}$, n copies indépendantes de la variable aléatoire Z dont la fonction de répartition est F et soient $\{V_1, \ldots, V_{n_u}\} = \{Z_i - u | Z_i > u, 1 \leq i \leq n\}$ les excédents au-delà d'un seuil u, où $n_u = \#\{Z_i > u\}$, voir la figure 1.3. Les excédents sont distribués selon la fonction de répartition excédentaire, voir la définition 1.1.2.

Définition 1.1.2 (Fonction de répartition excédentaire). *Soit Z, une variable aléatoire dont la fonction de répartition est F et soit u, un niveau de seuil fixé. Posons $V = Z - u | Z > u$, la variable aléatoire correspondant aux excédents de Z par rapport au seuil u. Alors, la fonction de répartition excédentaire est donnée par :*

$$F_u(v) = P(Z - u \leq v | Z > u) = P(V \leq v), \quad \forall v > 0.$$

La fonction de répartition excédentaire est modélisée à l'aide de la loi de Pareto généralisée dont la fonction de répartition est donnée à la définition 1.1.3.

Définition 1.1.3 (Loi de Pareto généralisée).

$$G_{\xi;\beta}(z) = \begin{cases} 1 - (1 + \frac{\xi}{\beta} z)^{-1/\xi} & si\ \xi \neq 0, \\ 1 - e^{-\frac{z}{\beta}} & si\ \xi = 0. \end{cases}$$

· *où* $z \geq 0$ *si* $\xi \geq 0$ *et* $0 \leq z \leq -\frac{\beta}{\xi}$ *si* $\xi < 0$ *et* $\beta > 0$.

L'utilisation de la loi de Pareto généralisée pour modéliser la distribution des excès est justifiée par le théorème 1.1.2 dû à Pickands [37] qui démontre que $\forall F \in MDA(H_\xi)$, la fonction de répartition des excès converge vers la distribution de Pareto généralisée lorsque le seuil u tend vers l'extrémité du support de la distribution sous-jacente. La condition $F \in MDA(H_\xi)$ est remplie en pratique par toutes les distributions connues.

Théorème 1.1.2 (Pickands). $\forall \xi \in \mathbb{R}$, $F \in MDA(H_\xi)$ *si et seulement si*

$$\lim_{u \uparrow z_F} \sup_{0 < z < z_F - u} |F_u(z) - G_{\xi,\beta(u)}(z)| = 0,$$

où $z_F = \sup\{x \in \mathbb{R} : F(z) < 1\}$.

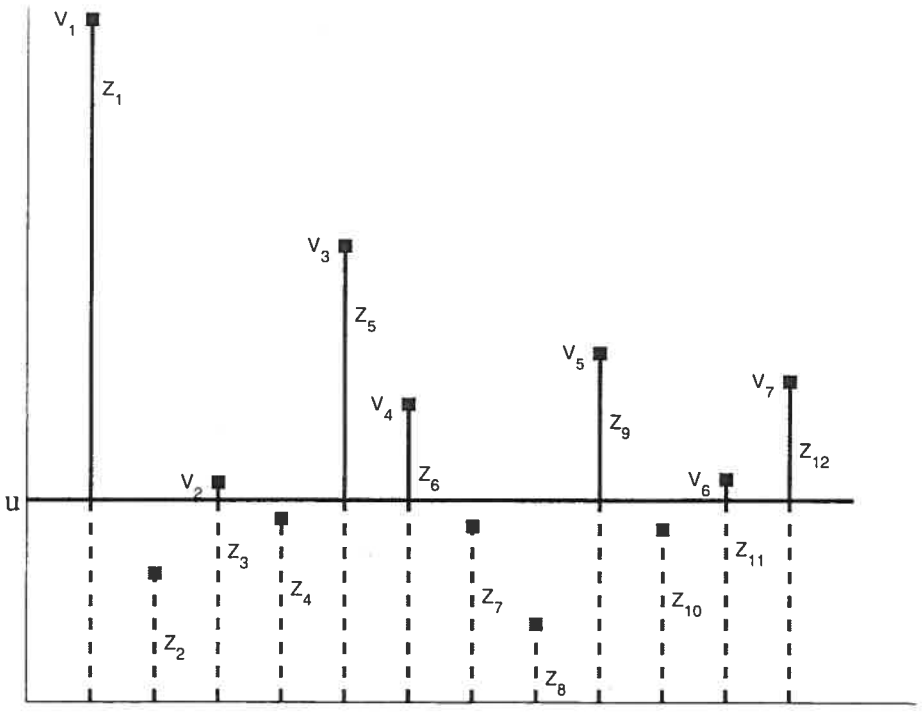

FIG. 1.3: *Observations* $Z_1, \ldots, Z_{12}$ *et excès* $V_1, \ldots, V_5$ *au-delà du seuil* u.

Le paramètre ξ de la Pareto généralisée correspond à celui de la loi aux valeurs extrêmes. Il contrôle l'épaisseur de la queue de la distribution et le nombre de moments de la loi qui sont bornés : si $Z \sim G_{\xi;\beta}$, alors $E[Z^r] < \infty \Leftrightarrow \xi < 1/r, \quad \forall r \in \mathbb{N}$. Ceci signifie entre autres que l'espérance de la Pareto généralisée existe si et seulement si $\xi < 1$ et que sa variance est finie si et seulement si $\xi < 1/2$. La densité de la loi de Pareto généralisée est illustrée à la figure 1.4 pour les queues légères à lourdes ($\xi \geq 0$) et à la figure 1.5 pour les queues finies ($\xi < 0$). Une autre motivation pour l'utilisation de la loi de Pareto généralisée comme modèle de la distribution des excès est la propriété de stabilité du seuil : si $V \sim G_{\xi;\beta}$ et $u > 0$, alors $V - u|V > u \sim G_{\xi;\beta+\xi u}$. Cette propriété caractérise la Pareto généralisée au sens où aucune autre loi ne la possède [**10**].

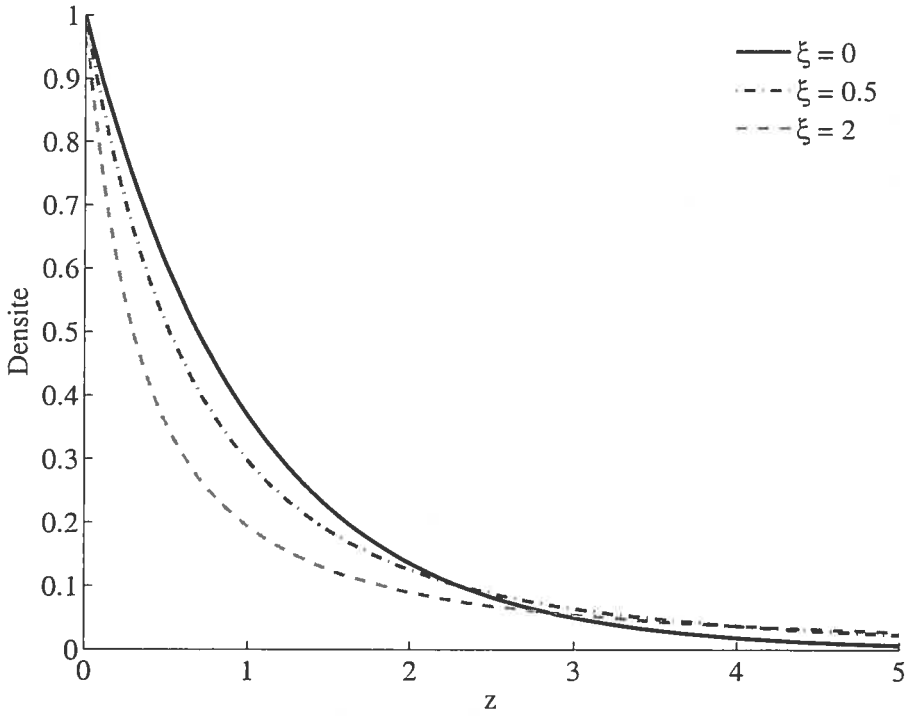

FIG. 1.4: *Densité de la Pareto généralisée, queue légère ($\xi = 0$) à lourde ($\xi = 2$).*

1.1.2.1. *Estimation des paramètres de la Pareto généralisée*

Les paramètres de la Pareto généralisée peuvent être estimés par la maximisation de la vraisemblance. Smith [**44**] a démontré que les estimateurs des paramètres de la Pareto généralisée existent pour les grands jeux de données pourvu

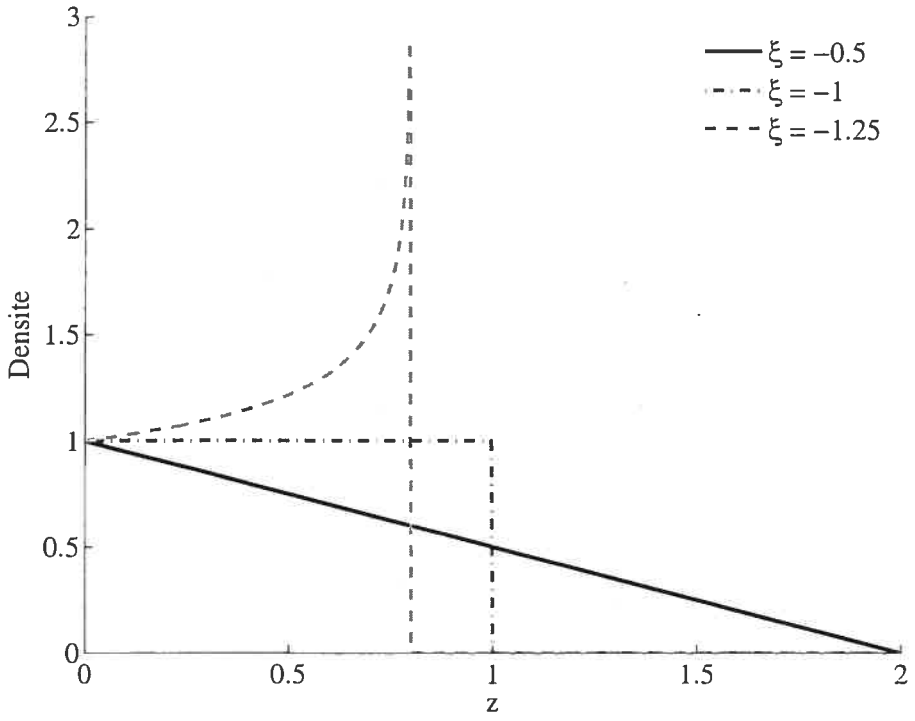

FIG. 1.5: *Densité de la Pareto généralisée, queue finies ($\xi < 0$) .*

que $\xi > -1$ et ces estimateurs sont asymptotiquement normaux et efficaces si $\xi > -1/2$. La méthode des moments probabilistes [**23**] offre une alternative viable à l'estimation par maximum de vraisemblance lorsque $\xi \geq 0$. Cependant, l'estimation par maximum de vraisemblance permet d'utiliser des modèles plus flexibles, qui permettent, par exemple, de modéliser la dépendance temporelle des données ou encore l'influence de variables explicatives.

Dans le cas de la méthodologie PoT ("Peaks-over-Threshold"), les observations $Z_1, \ldots, Z_n$ sont tirées de la distribution excédentaire, $F_u(\cdot)$ et non pas d'une loi de Pareto généralisée. Smith [**44**] a montré que l'approximation de la distribution excédentaire par la loi de Pareto généralisée introduit, dans de nombreux cas, un biais dans les estimateurs par maximum de vraisemblance de ξ et β.

1.1.2.2. *Estimateur de la queue de la distribution*

On obtient un estimateur de la queue de la distribution de F en utilisant la relation suivante, $\forall v > 0$:

$$\overline{F}(u+v) = \overline{F}(u)\overline{F}_u(v) \Leftrightarrow P(Z > u+v) = P(Z > u)P(Z-u > v|Z > u). \quad (1.1.2)$$

Tel que mentionné plus haut, la loi de Pareto généralisée sert d'approximation à la fonction de répartition excédentaire : $\overline{F}_u(v) \approx \overline{G}_{\hat{\xi}_u;\hat{\beta}_u}(v)$, où $\hat{\xi}_u$ et $\hat{\beta}_u$ sont des estimateurs des paramètres de la loi de Pareto généralisée calculés avec les excès correspondant au seuil u. La probabilité d'excéder le seuil u est généralement estimée à l'aide de la densité empirique : $\overline{F}(u) \approx n_u/n$, où n_u est le nombre d'excès et n est le nombre d'observations. L'estimateur de la queue de la distribution est alors donné par, $\forall v > 0$:

$$\widehat{F}(u+v) = 1 - \frac{n_u}{n}\left(1 + \hat{\xi}_u \frac{v}{\hat{\beta}_u}\right)^{-1/\hat{\xi}_u}.$$

1.1.2.3. *Sélection du seuil optimal*

La question qu'il reste maintenant à résoudre est celle de la sélection du seuil u. Si u est choisi trop élevé, il y aura peu d'observations qui excèdent le seuil et par conséquent, la variance des estimateurs des paramètres de la Pareto généralisée sera élevée. Par contre, si u est choisi trop bas, ces estimateurs seront biaisés puisque l'approximation de la fonction de répartition excédentaire par la Pareto généralisée n'est valable que lorsque u tend vers l'extrémité du support de la distribution z_F. Il s'agit d'un exemple du compromis biais-variance qui apparaît sous plusieurs formes dans les questions d'inférence statistique. Il est théoriquement possible de choisir u de façon optimale en quantifiant l'équilibre entre le biais et la variance. Cependant, pour y parvenir, il faut estimer l'erreur quadratique moyenne asymptotique de l'estimateur de l'indice de queue. Il faut alors faire des hypothèses supplémentaires sur le comportement de deuxième ordre de la distribution sous-jacente. Beirlant et al.[**2**] étudient le cas où $\xi > 0$ qui correspond aux queues lourdes. On a vu plus haut que si $F \in MDA(H_\xi)$, alors il existe une fonction à variation lente $L(z)$ (voir l'équation 1.1.1) telle que $\overline{F}(z) = z^{-1/\xi}L(z)$. Les conditions de deuxième ordre portent alors sur la vitesse à laquelle la fonction $L(z)$ disparait.

Embrechts et al.[**15**] proposent plutôt d'estimer les paramètres de la loi de Pareto généralisée pour plusieurs niveaux de seuil et de créer un graphe du paramètre de queue estimé $\hat{\xi}$ par rapport au niveau de seuil. Le niveau de seuil retenu sera dans la région du graphe qui est approximativement stable. Cette technique

requiert un ajustement manuel et une certaine expérience pour déterminer quel seuil est adéquat. Danielsson et de Vries [**9**] présentent une méthode d'estimation de l'erreur quadratique espérée de l'estimateur $\hat{\xi}$ qui est basée sur le bootstrap. Le seuil choisi est celui qui minimise cette estimation de l'erreur quadratique espérée. Le bootstrap de l'erreur quadratique espérée requiert un développement particulier puisque la vraie valeur de ξ n'est pas connue.

1.2. Méthodes basées sur la Pareto généralisée

L'estimation de la queue de la distribution par la méthode des excès au-delà d'un seuil est souvent préférée à la méthode des maxima par blocs car plus de données participent à l'estimation. Nous passerons en revue quelques méthodes qui ont été développées en prenant comme point de départ l'approximation de la queue de la distribution au delà d'un seuil par la loi de Pareto généralisée.

Frigessi et al.[**18**] proposent un mélange de distributions dynamique (c'est-à-dire dont la proportion des composantes du mélange varie avec la valeur de l'entrée) qui contourne le problème de la sélection du seuil dans la méthodologie liée à la loi de Pareto généralisée. Ces auteurs suggèrent d'utiliser toutes les observations pour estimer la densité complète et non pas seulement la queue de la distribution. L'estimateur proposé est un mélange ayant une composante Pareto généralisée placée en zéro et une composante à queue légère. La proportion de la Pareto généralisée dans le mélange croît en fonction de la valeur de l'entrée Z. De cette façon, la proportion agit comme un seuil graduel et la queue du mélange est gouvernée par la loi de Pareto généralisée. Frigessi et al. [**18**] démontrent la validité de leur modèle en l'applicant à l'estimation de quantiles sur des jeux de données simulés et réels.

Choulakian et Stephens [**6**] ont developé un test d'adéquation pour la loi de Pareto généralisée. Ce test peut être utilisé pour la sélection du seuil de la façon suivante. La Pareto généralisée est d'abord ajustée aux excédents d'un seuil relativement bas. On applique le test d'adéquation avec les paramètres estimés. Le niveau du seuil sera augmenté tant que le test d'adéquation ne sera pas satisfait.

Dupuis [**14**] propose une méthode de sélection robuste du seuil au-delà duquel la loi de Pareto généralisée est utilisée pour approximer la queue de la distribution. Pour un niveau de seuil donné, des estimateurs robustes des paramètres de la loi de Pareto généralisée sont obtenus à l'aide de l'estimateur robuste au biais optimal (*OBRE* : *optimal bias robust estimator*). Cette procédure robuste attribue à chaque excès un poids dans l'intervalle $[0, 1]$ qui mesure le degré de justesse du modèle en ce point. La stratégie suggérée par Dupuis [**14**] est d'utiliser *OBRE* pour un niveau de seuil relativement bas et d'estimer les poids associés aux excès. On augmente graduellement le niveau du seuil jusqu'à ce que les poids déterminés par *OBRE* soient près de 1 ce qui signifie que le modèle représente bien les données. Cette stratégie est appliquée à la modélisation de deux jeux de données réels.

McNeil et Frey [**35**] combine l'approche des excès au-delà d'un seuil avec la modélisation *GARCH* (*Generalized Autoregressive Conditionally Heteroskedastic*) pour estimer la VaR et l'espérance de perte de la distribution des rendements logarithmiques au temps $t + h$ conditionnelle à l'information disponible au temps t. L'espérance conditionnelle des rendements logarithmiques est modélisée par un processus autorégressif d'ordre 1 (AR(1)) et leur variance conditionnelle par un processus GARCH(1,1). On fait l'hypothèse que les résidus du modèle $Z_1, \ldots, Z_n$ sont des réalisations indépendantes d'une variable aléatoire Z d'espérance nulle et de variance unitaire (un "bruit blanc"). Pour estimer la queue de la distribution de Z, on utilise la méthodologie des excès au-delà d'un seuil. Soient $Z_{(1)} \geq Z_{(2)} \geq \cdots \geq Z_{(n)}$, les résidus ordonnés. Alors on fixe $u = Z_{(k+1)}$, ce qui laisse k observations pour l'estimation des paramètres de la Pareto généralisée. Une étude simulatoire sert à déterminer le choix de k. Cette étude utilise des échantillons provenant de la loi de Student. L'estimation de l'espérance et de la variance conditionnelle des rendements logarithmiques ainsi que l'approximation de la queue de la distribution des résidus par la loi de Pareto généralisée permettent de construire un estimateur de la queue de la distribution conditionnelle. À partir de cet estimateur, les auteurs calculent des estimateurs de quantiles extrêmes et de l'espérance de perte.

Beirlant et al.[**3**] proposent une extension de la Pareto généralisée qui permet de réduire le biais de l'approximation et d'utiliser une plus grande fraction des données dans l'estimation des paramètres. Cette méthode est basée sur un développement de deuxième ordre du modèle de type Pareto. La distribution *EGPD (Extended Generalized Pareto Distribution)* a trois paramètres et inclut la loi de Pareto généralisée comme cas spécial. L'approximation de la queue de la distribution par l'*EGPD* nécessite encore le choix d'une seuil approprié mais celui-ci peut être choisi plus bas tout en maintenant la qualité de l'approximation.

Chapitre 2

APPRENTISSAGE STATISTIQUE

L'apprentissage statistique élabore des algorithmes qui sont au confluent des statistiques et de l'intelligence artificielle [5]. Ces algorithmes sont en mesure de modifier leur structure interne afin de mieux apprendre des règles qui reflètent les relations entre les données. L'approche de l'apprentissage statistique diffère de celle de la statistique car l'objectif n'est pas d'identifier quel est le modèle ayant généré les données mais plutôt d'apprendre une fonction capable de mimer le processus générateur. L'idée centrale est donc que la fonction apprise par l'algorithme soit en mesure de bien prédire sur de nouveaux exemples générés par le même processus sous-jacent plutôt que d'estimer précisément ce processus. De plus, une propriété fondamentale des algorithmes issus de l'apprentissage statistique est la capacité, à la vue d'un nombre de plus en plus grand d'exemples, d'approximer de mieux en mieux le processus générateur.

2.1. APPRENTISSAGE SUPERVISÉ ET NON-SUPERVISÉ

Nous utiliserons des modèles relevant de deux types d'apprentissage statistique : l'apprentissage supervisé et non-supervisé. Dans tous les cas, on dispose d'un jeu de données $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ qui contient des observations indépendantes et identiquement distribuées (*i.i.d.*) provenant d'une variable aléatoire Z. Dans le cadre de l'apprentissage supervisé, on suppose que $Z = (X, Y)$ et qu'on dispose donc de paires d'observations $\{(X_i, Y_i)\}_{i=1}^n$. La variable d'entrée X peut être aléatoire ou déterministe ; ceci importe peu dans la modélisation. En général, $X \in \mathbb{E}^d$ où $\mathbb{E}^d$ est un espace métrique ; souvent $\mathbb{E}^d = \mathbb{R}^d$ ou encore un

sous-ensemble de $\mathbb{R}^d$. Si la sortie Y prend des valeurs discrètes, $Y \in \{1, \ldots, K\}$, il est question de classification. Étant donnée l'entrée X, Y donne l'étiquette de la classe associée à X. Par exemple, en finance, X peut être un vecteur représentant l'état de l'économie et Y, l'action associée à X, soit acheter, vendre ou maintenir la part d'un actif dans un portefeuille, auquel cas trois classes sont possibles et $K = 3$. On veut déterminer une fonction $f : \mathbb{E}^p \to \{1, \ldots, K\}$ qui classifie les exemples de $\mathcal{D}_n$. Une erreur de classification survient lorsque $f(X_i) \neq Y_i$. En général, on cherche f de sorte qu'elle minimise le nombre d'erreurs de classification.

Si $Y \in \mathbb{R}^p$, il s'agit alors de régression. Toujours dans l'exemple de finance, si X est un vecteur décrivant l'état de l'économie, Y peut représenter les rendements de p actifs sur une période de temps donnée. Dans ce cas-ci, on cherche à faire l'inférence d'une fonction f à partir du jeu de données $\mathcal{D}_n$ qui permette de prédire Y étant donné X. Peu d'hypothèses a priori sont posées sur f si ce n'est que f est une fonction lisse ce qui se traduit par le concept de proximité suivant : si x_1 et x_2 sont près au sens de la métrique de $\mathbb{E}^d$ alors $f(x_1)$ et $f(x_2)$ seront près également par rapport à la métrique de $\mathbb{E}^p$.

Dans le contexte de l'apprentissage non-supervisé, on cherche à modéliser la variable aléatoire Z elle-même. On peut vouloir chercher à regrouper les données en K sous-groupes ; c'est-à-dire que l'on veut attribuer à chaque observation Z_i une étiquette $j \in \{1, \ldots, K\}$ telle que toutes les observations du sous-groupe j sont sembables selon une métrique donnée. On appelle ce type d'apprentissage *clustering*. Ce problème est semblable à un problème de classification où la variable déterminant les étiquettes est manquante. D'une manière générale, on peut chercher à faire l'inférence de $p(\cdot)$, la fonction de densité associée à Z d'après les données $\mathcal{D}_n$. La connaissance d'un modèle de la densité génératrice permet de faire une analyse explorative des caractéristiques de la densité telles que la multimodalité, le comportement de la queue de la distribution et l'asymétrie. Un estimateur de la densité permet aussi de faire de l'analyse discriminante et de la classification.

2.2. APPRENTISSAGE ET GÉNÉRALISATION

Dans le cas paramétrique, on considère une famille de fonctions $\mathcal{F}_\theta = \{\phi_\theta(\cdot); \theta \in \mathbb{R}^k\}$ où θ est un vecteur de paramètres de longueur k. L'apprentissage ou l'entraînement du modèle consiste en un processus permettant de déterminer le vecteur $\hat{\theta}_n \in \mathbb{R}^k$ d'après les données $\mathcal{D}_n$ de sorte que la fonction $\phi_{\hat{\theta}_n}(\cdot)$ soit la mieux adaptée aux données. On appelle alors $\mathcal{D}_n$ l'ensemble d'entraînement ou d'apprentissage. L'entraînement d'un modèle se fait habituellement en minimisant un critère de coût (ou, de façon équivalente, en maximisant une mesure de performance) sur les données $\mathcal{D}_n$. Soit $l_\theta(z)$, un critère de coût. Le choix de $l_\theta(\cdot)$ dépend de la tâche à effectuer. Par exemple, il pourrait s'agir de l'erreur quadratique $l_\theta(x, y) = (y - \phi_\theta(x))^2$ en régression ou de la log-vraisemblance négative $l(z) = -\log(\phi_\theta(z))$ en estimation de densité. L'erreur de généralisation est l'espérance de la fonction de coût :

$$\mathcal{E}(\theta) = E[l_\theta(z)] = \int l_\theta(z)p(z)dz. \tag{2.2.1}$$

En pratique, on ne dispose que du jeu de données $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ généré par $p(\cdot)$ et non pas de la fonction $p(\cdot)$ elle-même. Considérons la densité empirique $\hat{p}_n(\cdot)$ qui place une masse de probabilité de $1/n$ sur chaque observation. On obtient une expression pour la densité empirique à l'aide du delta de Dirac[1] :

$$\hat{p}_n(z) = \frac{1}{n}\sum_{i=1}^{n} \delta(z - Z_i). \tag{2.2.2}$$

En utilisant la densité empirique $\hat{p}_n(\cdot)$ dans l'équation 2.2.1, on obtient l'erreur empirique moyenne :

$$\mathcal{E}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_\theta(Z_i). \tag{2.2.3}$$

L'entraînement du modèle consiste donc à déterminer $\hat{\theta}_n$ tel que l'erreur empirique moyenne soit minimisée, c'est-à-dire que $\hat{\theta}_n = \arg\min_\theta \mathcal{E}_n(\theta)$. La minimisation de $\mathcal{E}_n(\theta)$ est en général un problème d'optimisation non-linéaire et requiert des méthodes telles que la descente de gradient.

[1]Deux propriétés qui définissent le delta de Dirac sont : 1) $\delta(z - a) = 0$ si $z \neq a$, 2) $\int_{-\infty}^{+\infty} \delta(z - a)dz = 1$. En dimension d, $z = a \Leftrightarrow z^{(i)} = a^{(i)}$, $i = 1, \ldots, d$.

Lorsque $\hat{\theta}_n$ est choisi en minimisant l'erreur empirique $\mathcal{E}_n(\theta)$, $\mathcal{E}_n(\hat{\theta}_n)$ est un estimateur biaisé de l'erreur de généralisation ; $\mathcal{E}(\hat{\theta}_n)$ est en moyenne sous-estimée par $\mathcal{E}_n(\hat{\theta}_n)$. Pour obtenir un estimateur sans biais de l'erreur de généralisation, on calcule l'erreur empirique moyenne sur un ensemble de données $\mathcal{D}_l = \{Z_1, \ldots, Z_l\}$ qui est distinct de l'ensemble d'entraînement $\mathcal{D}_n$. Le but de l'apprentissage est d'obtenir une fonction $\phi_{\hat{\theta}_n}(\cdot)$ qui généralise bien, c'est-à-dire qui possède une petite erreur de généralisation. Intuitivement, la généralisation est la capacité d'une fonction $\phi_{\hat{\theta}_n}(\cdot)$ à faire une bonne prédiction sur de nouvelles données qui n'ont pas été vues à l'entraînement.

Afin de mieux comprendre le rôle de l'erreur de généralisation d'un estimateur $\phi_{\hat{\theta}_n}(\cdot)$, on la décompose généralement en la somme de trois termes : le biais carré, la variance et le bruit. Le biais est défini comme la différence systématique entre une variable aléatoire et une valeur particulière ciblée. En ce qui nous concerne, la variable aléatoire est la fonction $\phi_{\hat{\theta}_n}(\cdot)$ qui dépend de l'ensemble d'entraînement $\mathcal{D}_n$ observé. La variance mesure le caractère stochastique de l'estimateur, c'est-à-dire combien il varie en fonction de l'ensemble d'entraînement $\mathcal{D}_n$. Le bruit est la partie de l'erreur qui est propre au problème et qui ne peut être éliminée par l'algorithme d'apprentissage. La décomposition de l'erreur de généralisation la plus connue et la plus simple à dériver est celle qui a trait à la perte quadratique $l_\theta(x, y) = (y - \phi_\theta(x))^2$ dans le cadre de la régression [**19**]. On suppose donc que l'ensemble d'apprentissage est composé de paires d'observations $Z_i = (X_i, Y_i)$ et que l'on cherche une fonction $\phi_\theta(\cdot)$ telle que $\phi_\theta(X)$ permette de prédire Y. L'erreur quadratique espérée (ou erreur de généralisation pour la perte quadratique) de l'estimateur $\phi_\theta(\cdot)$ au point (x, y) s'écrit alors :

$$\begin{aligned}
E[(y - \phi_\theta(x))^2] &= \int (y - \phi_\theta(x))^2 p(x, y) dx dy \\
&= E[(y - f(x) + f(x) - \phi_\theta(x))^2] \\
&= E[(y - f(x))^2 + (f(x) - \phi_\theta(x))^2] \\
&\quad + \underbrace{E[2(y - f(x))(f(x) - \phi_\theta(x))]}_{\text{zéro}} \\
&= \underbrace{E[(y - f(x))^2]}_{\text{bruit}} + E[(f(x) - \phi_\theta(x))^2]. \qquad (2.2.4)
\end{aligned}$$

Le deuxième terme du membre de droite de l'équation 2.2.4 se développe comme suit :

$$\begin{aligned} E[(f(x)-\phi_\theta(x))^2] &= E[(f(x)-E[\phi_\theta(x)]+E[\phi_\theta(x)]-\phi_\theta(x))^2] \\ &= E[(f(x)-E[\phi_\theta(x)])^2+(E[\phi_\theta(x)]-\phi_\theta(x))^2] \\ &\quad +\underbrace{2E[(f(x)-E[\phi_\theta(x)])(E[\phi_\theta(x)]-\phi_\theta(x))]}_{\text{zéro}} \\ &= \underbrace{(f(x)-E[\phi_\theta(x)])^2}_{\text{biais carré}}+\underbrace{E[(\phi_\theta(x)-E[\phi_\theta(x)])^2]}_{\text{variance}} \end{aligned} \quad (2.2.5)$$

En combinant les équations 2.2.4 et 2.2.5, on obtient la décomposition voulue :

$$\mathcal{E}(\theta) = \text{bruit} + \text{biais}^2 + \text{variance}. \quad (2.2.6)$$

Heskes [**21**] propose une décomposition biais-variance-bruit de l'erreur de généralisation pour la log-vraisemblance négative $l(z) = -\log(\phi_\theta(z))$ et James [**25**] suggère un cadre plus large dans lequel il développe des décompositions biais-variance-bruit pour toutes les fonctions de coût symétrique.

2.3. Principe du maximum de vraisemblance

Dans le cadre de l'estimation de densité, la famille de fonction $\mathcal{F}_\theta$ doit satisfaire les propriétés suivantes : $\phi_\theta(z) \geq 0$, $\forall z$ et $\int \phi_\theta(z)dz = 1$. On cherche alors le vecteur de paramètres θ_0 tel que la fonction $\phi_{\theta_0}(\cdot)$ est une bonne approximation de la fonction de densité sous-jacente $p(\cdot)$. Pour juger de la qualité de l'approximation, on peut utiliser le critère d'information de Kullback-Leibler [**30**] :

$$\mathcal{KL}(p||\phi_\theta) = -\int p(z)\log\left(\frac{\phi_\theta(z)}{p(z)}\right)dz. \quad (2.3.1)$$

On note que $\mathcal{KL}(p||\phi_\theta) \geq 0$ avec égalité si et seulement si $\phi_\theta(z) = p(z)$, $\forall z$. Par ailleurs, à cause de la pondération par la vraie densité $p(\cdot)$, la distance de Kullback-Leibler n'est pas symétrique par rapport aux deux fonctions de densités. Selon ce critère, il est plus important que le modèle $\phi_\theta(\cdot)$ soit proche de la densité sous-jacente dans les régions de plus grande densité. On aimerait choisir θ de sorte que $\mathcal{KL}(p||\phi_\theta)$ soit minimisée ce qui revient à minimiser le critère suivant (qui est la partie du critère de Kullback-Leibler où $\phi_\theta(\cdot)$ intervient) : $\mathcal{L}(\theta) =$

$-\int p(z)\log(\phi_\theta(z))dz$. En remplaçant $p(\cdot)$ par la densité empirique $\hat{p}_n(\cdot)$ donnée à l'équation 2.2.2, le critère $\mathcal{L}(\theta)$ devient :

$$\mathcal{L}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log(\phi_\theta(Z_i)).$$

Une autre façon d'introduire le critère $\mathcal{L}_n(\theta)$ est par le biais du concept de vraisemblance. Fisher [**17**] a proposé le principe de maximum de vraisemblance comme base pour l'estimation de densité. La vraisemblance se définit comme la probabilité d'observer le jeu de données $\mathcal{D}_n$ selon le modèle ϕ_θ. Puisque les observations sont supposées i.i.d., la vraisemblance est donc donnée par :

$$P(\mathcal{D}_n;\theta) = \prod_{i=1}^{n}\phi_\theta(Z_i).$$

L'expression pour $\mathcal{L}_n(\theta)$ est donc le négatif du logarithme de la vraisemblance divisé par n, $\mathcal{L}_n(\theta) = -\log P(\mathcal{D}_n;\theta)/n$, et le vecteur de paramètres $\hat{\theta}_n$ qui maximise la vraisemblance minimise aussi le critère de Kullback-Leibler par rapport à la densité empirique.

Lorsque la densité sous-jacente appartient à la famille de modèles considérée, c'est-à-dire qu'il existe θ^* tel que $\phi_{\theta^*}(\cdot) = p(\cdot)$, sous certaines conditions, l'estimateur de maximum de vraisemblance a la propriété de convergence suivante (Wald [**45**]) : lorsque $n \to \infty$, alors $\hat{\theta}_n \overset{p.s.}{\to} \theta^*$ où $p.s.$ dénote la convergence presque sûre. Dans la plupart des cas, la famille de modèles considérée pour l'estimation de densité ne contient pas la densité génératrice. Dans ce cas, sous certaines conditions, l'estimateur par maximum de vraisemblance est convergent dans le sens suivant (White [**46**]) : lorsque $n \to \infty$, alors $\hat{\theta}_n \overset{p.s.}{\to} \theta_0$ où θ_0 dénote le vecteur de paramètres tel que $\mathcal{KL}(p||\phi_{\theta_0})$ est minimal. L'estimateur par maximum de vraisemblance $\hat{\theta}_n$ est donc un estimateur naturel pour les paramètres θ_0 qui minimise le critère de Kullback-Leibler.

2.4. Estimation de densité non-paramétrique

Un estimateur est dit non-paramétrique lorsque son niveau de complexité, c'est-à-dire le nombre de paramètres libres, croît avec la taille de l'ensemble d'entraînement. L'approche non-paramétrique pose peu d'hypothèses a priori sur la

densité génératrice $p(\cdot)$ et ne suppose pas une forme fonctionnelle spécifique pour celle-ci. Le type d'hypothèses qui est habituellement endossé par cette approche concerne la régularité de $p(\cdot)$ au sens où si deux observations Z_1 et Z_2 sont proches dans $\mathbb{E}^d$, alors $p(Z_1)$ et $p(Z_2)$ sont proches dans $\mathbb{R}^+$. Inversement, l'approche paramétrique fait l'hypothèse que la densité ayant généré les données appartient à une famille de distributions spécifiques, comme celle des lois Normales. Un estimateur paramétrique possède un petit nombre de paramètres mais ce nombre est fixe, quelle que soit la taille de l'ensemble d'entraînement. Si la densité génératrice n'appartient pas à la famille paramétrique considérée, l'estimation paramétrique peut introduire un biais substantiel.

2.4.1. Estimateurs à noyaux

La densité empirique $\hat{p}_n(z) = 1/n \sum_{i=1}^{n} \delta(Z_i - z)$ est un estimateur non-paramétrique particulier. La fonction de répartition empirique est donnée par :

$$\hat{F}_n(z) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Z_i \leq z\}}, \tag{2.4.1}$$

où $Z_i \leq z \Leftrightarrow Z_i^{(j)} \leq z^{(j)} \quad j = 1, \ldots, d$. Le théorème de Glivenko-Cantelli [**4**] démontre la convergence de la fonction de répartition empirique vers la fonction de répartition du modèle générateur F :

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{p.s.} 0, \quad \text{lorsque } n \to \infty.$$

On appelle "noyau" une fonction $K : \mathbb{E}^d \to \mathbb{R}$ telle que $K(u) \geq 0$, $\forall u$ et $\int K(u)du = 1$. Dans le cas du noyau gaussien : $K(u) = 1/(2\pi)^{d/2} \exp\{-||u||^2/2\}$. En remplaçant la fonction delta $\delta(\cdot)$ par un noyau $K(\cdot)$ dans l'expression pour la densité empirique, on obtient une famille d'estimateurs non-paramétriques appelés les estimateurs à noyaux. Soit h la largeur de fenêtre du noyau. L'estimateur à noyaux place un noyau de largeur h sur chaque point de l'ensemble $\mathcal{D}_n$ et estime la densité à un nouveau point $z \in \mathbb{E}^d$ en faisant la moyenne des contributions de chaque noyau en ce point, cette moyenne étant pondérée par le facteur $1/h^d$ pour que la densité intègre à 1. L'estimateur à noyaux se formule alors de la façon

suivante :

$$\tilde{p}_h(z) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{h}\right).$$

La masse de probabilité associée à chaque point est partagée par un voisinage dont la taille est déterminée par h, la largeur de fenêtre du noyau. Lorsque $h \to 0$, chaque noyau tend vers un delta de Dirac et l'estimateur à noyaux s'approche de la fonction de densité empirique. La densité empirique d'un échantillon contenant cinq observations est illustrée à la figure 2.1 par des lignes verticales (si on se base sur le delta de Dirac, les lignes devraient aller vers l'infini). Selon $\hat{p}_n(\cdot)$, chaque observation a donc une masse de probabilité de 1/5 et tout autre point a une masse de probabilité de 0. Centré sur chaque observation, un noyau gaussien, avec $h = 0.5$ et pondéré par 1/5, est dessiné en pointillé. L'estimateur de densité à noyaux résultant de l'addition de ces noyaux pondérés est montré par la ligne pleine au dessus des observations.

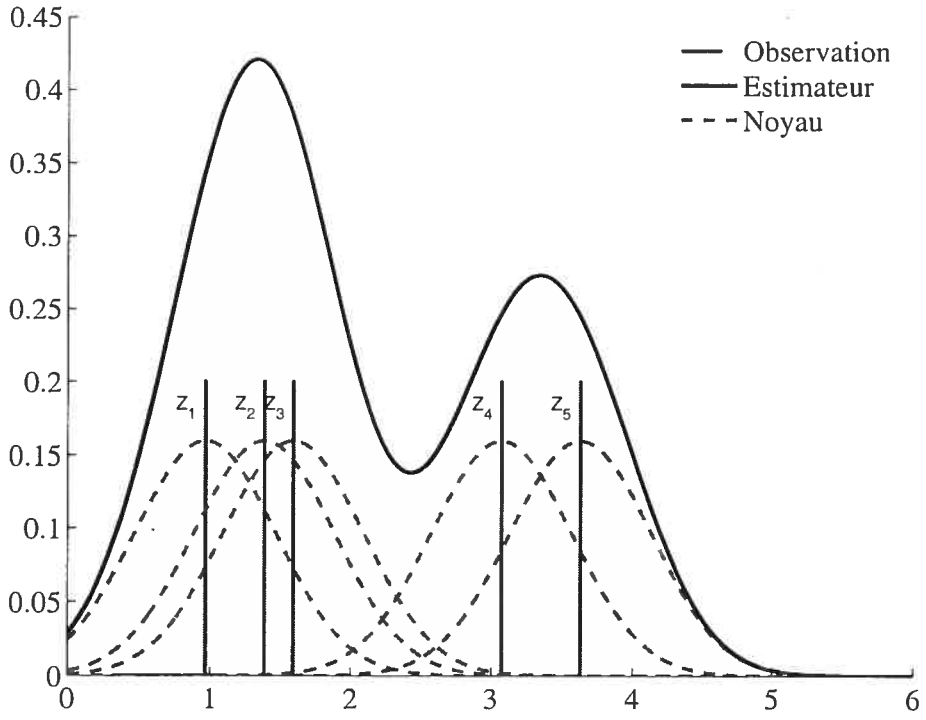

FIG. 2.1: *Densité empirique associée à* $\mathcal{D}_n = \{Z_1, \ldots, Z_5\}$ *et estimateur à noyaux où* $K(u) = 1/\sqrt{2\pi}\exp\{-u^2/2\}$ *et* $h = 0.5$.

On peut voir les emplacements des noyaux, c'est-à-dire les observations, comme les paramètres des estimateurs à noyaux. Le nombre de paramètres des estimateurs à noyaux croît donc avec la taille de l'ensemble d'entraînement, ce qui est typique des modèles non-paramétriques. La largeur de fenêtre h est un hyperparamètre qui contrôle la complexité du modèle et affecte le biais et la variance qui composent l'erreur de généralisation de l'estimateur (voir l'équation 2.2.6). Lorsque h est grand, les noyaux sont très étendus et la densité résultante est très lisse (on distingue peu ou pas les bosses des noyaux individuels). Ceci est illustré dans le panneau du bas de la figure 2.2 ; l'estimateur à noyaux dénote un biais marqué par rapport à la densité sous-jacente mais possède peu de variance. Lorsque h est petit, les noyaux sont très étroits et l'estimateur de densité contient beaucoup de structures. C'est le cas dans le panneau du haut de la figure 2.2 où l'estimateur à noyaux affiche une grande variance mais peu de biais dans le voisinage des points de l'ensemble d'entraînement. Idéalement, la largeur de fenêtre doit minimiser simultanément le biais et la variance de l'estimateur à noyaux afin d'obtenir un estimateur semblable à celui du panneau du milieu de la figure 2.2. Devroye et Gyorfi **[12]** ont montré que les estimateurs à noyaux sont fortement convergents dans la norme L_1.

2.4.2. Mélanges de distributions

Bien que très flexibles et possédant des propriétés de convergence intéressantes, les estimateurs à noyaux peuvent s'avérer lourds en temps de calcul car les n observations de l'ensemble d'entraînement sont requises pour pouvoir évaluer la densité en un nouveau point z. Les mélanges de distributions sont un type de modèles non-paramétriques qui offrent un compromis entre les modèles paramétriques et les modèles à noyaux. Un mélange de distributions, dont la formule générale est donnée à l'équation 2.4.2, estime la densité en un point z en faisant la moyenne pondérée de m composantes. Chaque composante $p_j(z;\theta_j)$ est une fonction de densité de paramètres θ_j. Les poids π_j du mélange doivent être tels que $\pi_j \geq 0$ et $\sum_{j=1}^{m} \pi_j = 1$. On regroupe les paramètres du mélange dans le vecteur

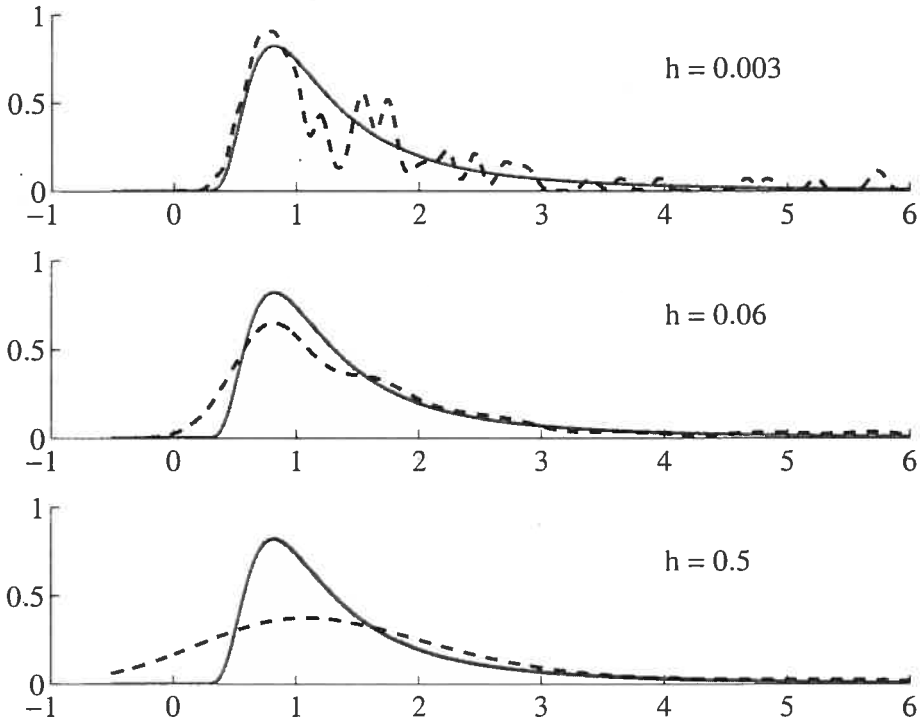

FIG. 2.2: *Estimateur à noyaux en pointillé avec différentes largeurs de fenêtres, de haut en bas : $h = 0.003$, $h = 0.06$, $h = 0.5$. L'ensemble d'apprentissage contient* 100 *observations provenant de la densité dessiné en trait plein.*

$\theta = \{\pi_1, \ldots, \pi_m, \theta_1, \ldots, \theta_m\}$.

$$\phi_\theta(z) = \sum_{j=1}^{m} \pi_j p_j(z; \theta_j) \tag{2.4.2}$$

On peut voir l'estimateur à noyaux comme un mélange de distributions particulier tel que $m = n$, $\pi_j = 1/m$ et chaque composante est un noyau de largeur h centré sur une observation Z_j. Cependant, habituellement dans un mélange le nombre de composantes est de beaucoup inférieur au nombre d'observations, $m \lll n$, et le nombre de paramètres d'un mélange est beaucoup plus petit que celui d'un estimateur à noyaux. Aussi, le vecteur de paramètres d'un mélange de distributions est déterminé en maximisant la log-vraisemblance sur les données $\mathcal{D}_n$. La méthode *EM* (Expectation-Maximisation) [**11**] a été proposée comme alternative à la maximisation directe de la log-vraisemblance. Elle permet de modifier itérativement le vecteur de paramètres tout en s'assurant que la log-vraisemblance augmente à chaque itération. Cependant, cette méthode est avantageuse s'il existe

une solution analytique à l'estimation par maximum de vraisemblance des paramètres de chaque composante $p_j(\cdot;\theta_j)$ comme c'est le cas par exemple pour la loi Normale. Des extensions de EM ont été développées [**31**] pour le cas où, comme pour la loi de Student, certains paramètres ont une solution analytique et d'autres non.

Le nombre de composantes m est un hyper-paramètre qui contrôle la complexité du mélange de distributions. Comme la largeur de fenêtre pour l'estimateur à noyaux, le nombre de composantes affecte de façon inverse le biais et la variance du mélange de distributions. Plus m est grand et plus l'estimateur produit par le mélange de distributions sera bosselé, il aura plus de variance et le biais diminuera autour des points de l'ensemble d'apprentissage. Inversement, lorsque m diminue, l'estimateur du mélange de distributions est plus lisse, le biais augmente et la variance diminue. Tout comme dans le cas de l'estimateur à noyaux, m doit être choisi de façon à minimiser simultanément le biais et la variance de l'estimateur.

Le mélange de lois Normales est le type de mélange le plus souvent utilisé. Lorsqu'on permet au nombre de composantes de croître lentement par rapport au nombre d'observations, le mélange de Normales est un estimateur de densité convergent dans la norme L_1 [**40**]. Les mélanges de distributions offrent donc des estimateurs de densité ayant des propriétés de convergence intéressantes et qui sont d'une complexité modérée. La contrepartie de ces avantages est que ces estimateurs requièrent un apprentissage substantiellement plus long que celui des estimateurs paramétriques ou des estimateurs à noyaux.

2.5. Réseau de neurones artificiels

Les réseaux de neurones artificiels sont une famille de fonctions qui permettent de représenter des fonctions non-linéaires qui transforment une entrée de dimension d en une sortie de dimension p. Pour y parvenir, un réseau de neurones combine des fonctions non-linéaires d'une variable, appelées fonctions d'activations. La figure 2.3 représente un réseau de neurones ayant une couche cachée. Il est possible d'insérer d'autres couches cachées entre l'entrée et la sortie. Plusieurs

autres configurations sont possibles à condition que le graphe ne contienne pas de boucle et donc que la fonction calculée par le réseau de neurones puisse être écrite de manière explicite.

2.5.1. Réseau de neurones à une couche cachée

La fonction calculée par le réseau de neurones de la figure 2.3 peut être décrite de la façon suivante. Soit n_h, le nombre de neurones dans la couche cachée. Chacun de ces neurones reçoit en entrée une combinaison linéaire, notée a_j, de l'entrée $x \in \mathbb{R}^d$ plus un terme constant appelé *biais* (à ne pas confondre avec le biais statistique introduit plus tôt) :

$$a_j = \sum_{i=1}^{d} v_{j,i} x_i + v_{j,0}, \tag{2.5.1}$$

où $v_{j,i}$, $i = 1 \ldots d$ et $j = 1 \ldots n_h$, est un poids de la couche cachée connectant le $j^{\text{ième}}$ neurone caché avec la $i^{\text{ième}}$ composante de l'entrée et $v_{j,0}$ est le biais associé au $j^{\text{ième}}$ neurone caché. Chacun des neurones j de la couche cachée transforme non-linéairement a_j : $z_j = g(a_j)$. La fonction $g(\cdot)$ est appelée fonction d'activation. Son rôle est d'introduire une non-linéarité dans la fonction calculée par le réseau de neurones. Plusieurs choix sont possibles, on adoptera ici la tangente hyperbolique $g(\cdot) = \tanh(\cdot)$, illustrée à la figure 2.4. La tangente hyperbolique prend ses valeurs dans l'intervalle $[-1, 1]$, elle est linéaire dans le voisinage de zéro et aux asymptotes alors qu'elle est non-linéaire aux voisinages des points d'inflexion. Lorsque x est grand en valeur absolue, $\tanh(x)$ se comporte comme la fonction de "heavy-side" ou fonction échelon. La tangente hyperbolique a l'avantage d'être différentiable. Soit p, le nombre de neurones dans la couche de sortie. De la même façon que les neurones de la couche cachée, chaque neurone de la couche de sortie reçoit en entrée une combinaison linéaire, notée b_k, des sorties de la couche précédente, c'est-à-dire des sorties z_j de la couche cachée, plus un terme constant de biais :

$$b_k = \sum_{j=1}^{n_h} w_{k,j} z_j + w_{k,0}, \tag{2.5.2}$$

où $w_{k,j}$, $k = 1, \ldots, p$ et $j = 1, \ldots, n_h$, est le poids de la couche de sortie connectant le $k^{\text{ième}}$ neurone de sortie avec la $j^{\text{ième}}$ unité cachée et $w_{k,0}$ est le biais associé au $k^{\text{ième}}$ neurone de sortie. Chaque neurone de sortie transforme son entrée à l'aide d'une fonction d'activation : $f_k(x) = \tilde{g}(b_k)$, où $\tilde{g}(\cdot)$ peut être différente de $g(\cdot)$. Le choix de la fonction d'activation de la couche de sortie $\tilde{g}(\cdot)$ dépend de la tâche accomplie par le réseau de neurones. S'il s'agit de régression, la fonction que le réseau de neurones tente d'approximer prend ses valeurs dans $\mathbb{R}^p$. La couche de sortie sera alors linéaire et $\tilde{g}(\cdot)$ sera simplement l'identité. S'il s'agit de classification, on peut interpréter chaque sortie $f_k(x)$ du réseau de neurones comme la probabilité que l'entrée soit de la classe C_k. Ces probabilités sont ensuite utilisées pour classifier une entrée x : on lui attribuera la classe C_k telle que $f_k(x) \geq f_j(x)$, $1 \leq j \leq p$. Pour que les sorties $f_k(x)$ soient interprétées comme des probabilités, il faut que $f_k(x) \geq 0$ et $\sum_{j=1}^{l} f_j(x) = 1$. Ceci peut être accompli par la fonction *softmax* : $f_k(x) = \tilde{g}_k(b) = e^{b_k} / \sum_{j=1}^{p} e^{b_j}$. S'il n'y a que deux classes, alors le réseau de neurones n'a qu'une sortie qui est interprétée comme la probabilité que l'entrée soit de classe C_1. L'entrée x sera de la classe C_1 si $f(x) \geq 0.5$ et de la classe C_2 sinon. Dans ce cas, la fonction d'activation utilisée en sortie est $f(x) = \tilde{g}(b) = 1/\left(1 + e^{-b}\right)$. Cette fonction, appelée "sigmoïde" à cause de sa forme en "S", est en fait une transformation linéaire de la tangente hyperbolique. La sigmoïde prend ses valeurs dans l'intervalle $[0, 1]$.

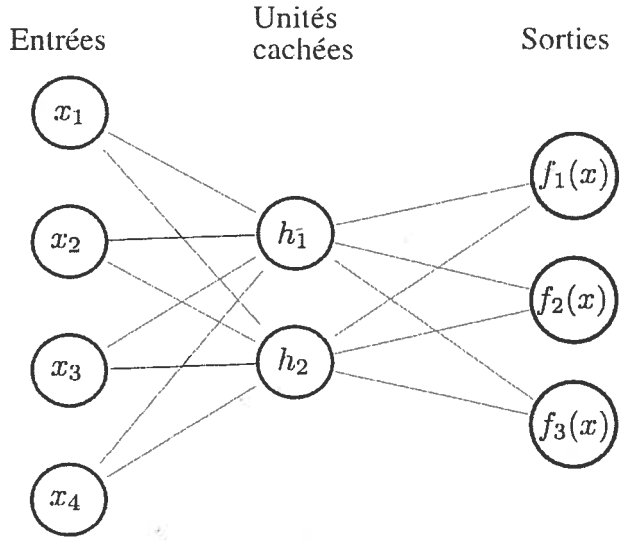

FIG. 2.3: *Réseau de neurones avec une couche cachée*

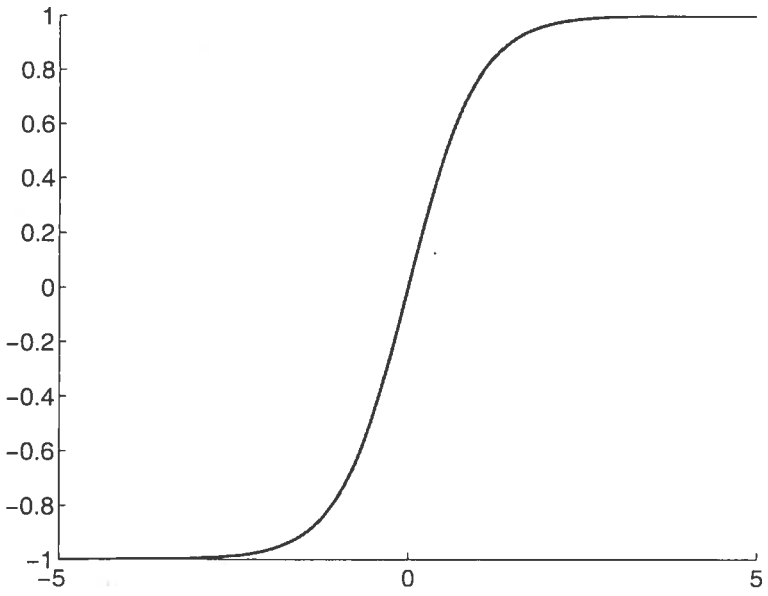

FIG. 2.4: *Tangente hyperbolique*

2.5.2. Propriété d'approximation

Kolmogorov [29] a démontré un résultat intéressant en lien avec l'approximation de fonction par un réseau de neurones. Par construction, il a prouvé que toute fonction continue de plusieurs variables à support compact peut être représentée comme une superposition d'un petit nombre de fonctions d'une variable. Bien que ce théorème n'ait pas d'utilité pratique, il confirme l'idée intuitive de chercher à approximer une fonction de plusieurs variables par plusieurs fonctions d'une variable. Dans le cas précis des réseaux de neurones à une couche cachée, Hornik [22] a montré que si la fonction d'activation de la couche cachée est continue, bornée et non constante, alors ces réseaux de neurones sont des "approximateurs" universels pourvu qu'un nombre suffisant d'unités cachées soit disponible. Donc en principe, un réseau de neurones ayant une couche cachée et la tangente hyperbolique comme fonction d'activation de la couche cachée peut approximer arbitrairement bien toute fonction continue à support compact. Cependant, cela ne nous dit pas comment choisir les paramètres et, en général, trouver ceux qui minimisent l'erreur empirique est NP-difficile.

2.5.3. Apprentissage

Les poids ou paramètres du réseau de neurones sont déterminés par la minimisation de l'erreur empirique moyenne $\mathcal{E}_n$ (voir l'équation 2.2.3) sur un ensemble d'apprentissage $\mathcal{D}_n$. Les méthodes d'optimisation requièrent généralement le calcul du gradient et parfois le calcul de la matrice hessienne. Rumelhart et al.[**43**] ont proposé une méthode de calcul efficace du gradient de l'erreur empirique moyenne qui s'appelle la rétropropagation d'erreurs. Cette méthode permet de calculer le gradient avec un nombre d'opérations similaire à celui requis pour le calcul de l'erreur empirique moyenne. Ce nombre d'opérations est proportionnel au nombre de poids dans le réseau de neurones. On développe ici les calculs de rétropropagation de l'erreur propres aux réseaux de neurones ayant une couche cachée dont la fonction d'activation est la tangente hyperbolique puisque c'est ce type de réseaux qu'on utilisera plus loin. Soit (x, y) un exemple de l'ensemble d'entraînement et soit $l_\theta(x, y)$ la fonction de coût utilisée pour l'entraînement. Dans le cas du réseau de neurones à une couche cachée, le vecteur de paramètres θ contient les poids de la couche cachée et de la couche de sortie, c'est-à-dire $\theta = \{v_{j,i} | j = 1, \ldots, n_h, \quad i = 0, \ldots, d\} \vee \{w_{k,j} | k = 1, \ldots, l \quad j = 0, \ldots, n_h\}$. Puisque $\mathcal{E}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_\theta(X_i, Y_i)$ et que la dérivée d'une somme est la somme des dérivées, il suffit de développer les formules pour le calcul de gradient de $l_\theta(x, y)$ en un point (x, y). La rétropropagation d'erreurs est basée sur le principe de la dérivation en chaîne : les dérivées sont calculées étape par étape et ensuite multipliées entre elles pour obtenir le gradient. Développons d'abord les dérivées nécessaires au calcul du gradient par rapport aux poids de la couche de sortie :

$$\frac{\partial l_\theta}{\partial w_{k,j}} = \sum_{i=1}^{p} \frac{\partial l_\theta}{\partial f_i(x)} \frac{\partial f_i(x)}{\partial b_k} \frac{\partial b_k}{\partial w_{k,j}}.$$

Les dérivées partielles $\partial l_\theta / \partial f_i(x)$ et $\partial f_i(x) / \partial b_k$ dépendent du choix de la fonction de coût $l_\theta(x, y)$ et de la fonction d'activation de la couche de sortie $\tilde{g}(\cdot)$. Puisque celles-ci varient selon la tâche du réseau de neurones, on développera ici explicitement uniquement les dérivées qui ont trait aux niveaux inférieurs du réseau de neurones. Dans ce cas-ci, on obtient : $\partial b_k / \partial w_{k,j} = z_j$.

On développe ensuite les dérivées partielles requises pour le calcul du gradient par rapport aux poids de la couche cachée :

$$\frac{\partial l_\theta}{\partial v_{j,i}} = \sum_{k=1}^{p} \frac{\partial l_\theta}{\partial f_k(x)} \sum_{m=1}^{p} \frac{\partial f_k(x)}{\partial b_m} \frac{\partial b_m}{\partial z_j} \frac{\partial z_j}{\partial a_j} \frac{\partial a_j}{\partial v_{j,i}}.$$

En développant les dérivées partielles, on obtient :

$$\begin{aligned} \frac{\partial b_m}{\partial z_j} &= w_{m,j} \\ \frac{\partial z_j}{\partial a_j} &= \frac{\partial}{\partial a_j} \tanh(a_j) = 1 - \tanh^2(a_j) = 1 - z_j^2 \\ \frac{\partial a_j}{\partial v_{j,i}} &= x_i. \end{aligned}$$

Dans un réseau de neurones ayant une couche cachée, le nombre d'unités cachées n_h est l'hyper-paramètre qui contrôle la complexité du modèle et le nombre de paramètres libres. Il y a $(d+1)n_h$ poids dans la couche cachée et $(n_h+1)p$ poids dans la couche de sortie pour un total de $n_h(d+1+p)+l$ paramètres libres. Plus il y a d'unités cachées, plus le nombre de paramètres libres est grand et plus le réseau de neurones est capable d'apprendre les données d'entraînement parfaitement. L'estimateur produit par le réseau de neurones a donc plus de variance. Inversement, lorsque le nombre d'unités cachées diminue, le réseau de neurones capte de moins en moins bien les fluctuations de la fonction sous-jacente et le biais de l'estimateur est plus grand.

2.6. Complexité et choix d'hyper-paramètre

On a vu que dans les familles de modèles étudiées, il y a toujours un hyper-paramètre qui contrôle la complexité du modèle et qui doit être choisi de façon à minimiser simultanément les propriétés conflictuelles du biais et de la variance. Dans le cas des estimateurs à noyaux, l'hyper-paramètre en question est la largeur de fenêtre h. Pour les mélanges de distributions, il s'agit du nombre de composantes m. Comme mentionné dans la section précédente, pour les réseaux de neurones à une couche cachée, c'est le nombre d'unités cachées n_h qui affecte la complexité.

Les paramètres d'un modèle sont déterminés par minimisation de l'erreur empirique moyenne sur les données d'entraînement $\mathcal{D}_n$. Typiquement, plus la complexité d'un modèle augmente, plus l'erreur d'entraînement diminue. Une courbe d'erreur d'entraînement typique en fonction de la complexité est dessinée à la figure 2.5. Éventuellement, cette courbe atteint un minimum qui est déterminé par le nombre d'exemples dans l'ensemble d'entraînement et par le bruit intrinsèque du problème (voir la sous-section 2.2). Si on tente de choisir le niveau de complexité d'un modèle en minimisant l'erreur d'entraînement, ceci conduira à la sélection d'un modèle qui est suffisamment complexe pour "apprendre par coeur" l'ensemble d'apprentissage. Par exemple, considérons, dans un contexte de régression, le choix du nombre d'unités cachées n_h d'un réseau de neurones à une couche cachée. Si n_h est suffisamment grand, le réseau de neurones aura assez de paramètres libres pour qu'il existe un choix de valeurs de paramètres produisant une fonction qui passe exactement par les points de l'ensemble d'entraînement $\mathcal{D}_n$. Cette fonction aura donc une erreur quadratique moyenne de zéro sur les données de $\mathcal{D}_n$. En estimation de densité, le choix de la largeur de fenêtre h de l'estimateur à noyaux qui minimise la log-vraisemblance négative moyenne sur l'ensemble d'entraînement mène à un estimateur qui est une distribution discrète avec des sauts aux points de $\mathcal{D}_n$. Ceci est atteint lorsque $h \to 0$ et que chaque noyau devient un delta de Dirac centré sur un des points de $\mathcal{D}_n$. Dans ce cas, la vraisemblance sur l'ensemble d'entraînement tend vers l'infini. Pour le mélange de distributions, l'optimisation des paramètres du mélange sur l'ensemble d'entraînement peut mener à une vraisemblance infinie si une des composantes s'effondre sur un des points de $\mathcal{D}_n$. Pour un mélange de Normales, ceci se produit lorsqu'une des composantes j est centrée sur une observation et que son écart-type tend vers zéro, c'est-à-dire $\mu_j = Z_i$ pour un i donné et $\sigma_j \to 0$. Lorsque le nombre de composantes m augmente, l'effondrement de composantes sur une observation est de plus en plus possible.

Intuitivement, il est peu probable que les modèles décrits ci-haut performent bien sur de nouvelles données. En effet, de tels modèles manifestent beaucoup

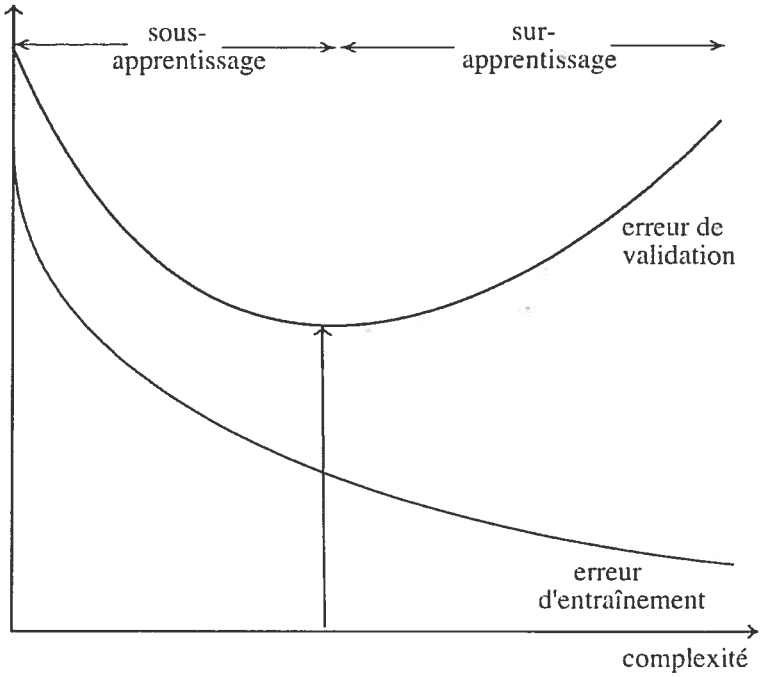

FIG. 2.5: *Sélection de modèles par minimisation de l'erreur empirique moyenne en validation.*

de variance puisqu'ils ont capté les variations particulières de l'ensemble d'entraînement $\mathcal{D}_n$ plutôt que la relation sous-jacente aux données. Étant donné que l'objectif de l'apprentissage statistique est de développer des algorithmes qui généralisent bien, le niveau de complexité d'un modèle devrait être choisi afin de minimiser l'erreur de généralisation. Mais, l'erreur empirique moyenne mesurée sur $\mathcal{D}_n$ de la fonction dont les paramètres ont été estimés sur $\mathcal{D}_n$ est un estimateur optimiste de l'erreur de généralisation (voir la sous-section 2.2). Pour obtenir un estimateur non-biaisé de l'erreur de généralisation, il faut mesurer l'erreur empirique moyenne sur un ensemble d'observations $\mathcal{D}_l = \{Z_1, \ldots, Z_l\}$, appelé ensemble de validation, qui est distinct de l'ensemble d'entraînement $\mathcal{D}_n$. Une courbe de l'erreur empirique sur l'ensemble de validation en fonction de la complexité d'un algorithme est tracée à la figure 2.5. Typiquement, cette courbe est plus élevée que la courbe de l'erreur d'entraînement. Dans la phase de sous-apprentissage, l'erreur de validation diminue d'abord avec la complexité jusqu'à ce qu'elle ait atteint le niveau de complexité optimale indiquée par la flèche. En sous-apprentissage, le potentiel de l'algorithme n'est pas exploité au maximum. La courbe de l'erreur de validation remonte ensuite et passe dans la phase de

sur-apprentissage. Dans cette phase, la complexité est trop élevée et l'algorithme tente d'apprendre parfaitement les données. Puisque la complexité est choisie de façon à minimiser l'erreur empirique moyenne sur l'ensemble de validation, celle-ci devient à son tour un estimateur optimiste de l'erreur de généralisation. Pour estimer l'erreur de généralisation de l'algorithme, on utilise un troisième jeu de données, appelé ensemble de test, distinct de l'ensemble d'entraînement et de l'ensemble de validation. On estime l'erreur de généralisation de l'algorithme, dont les paramètres ont été adaptés à l'ensemble d'entraînement et le niveau de complexité a été choisi sur l'ensemble de validation, en calculant l'erreur empirique moyenne sur l'ensemble de test.

Chapitre 3

MODÈLES PARETO HYBRIDES : DENSITÉ INCONDITIONNELLE

Les méthodes d'estimation de la queue de la distribution dérivées par la théorie aux valeurs extrêmes décrites à la section 1.1 s'appliquent au cas où l'on dispose d'observations univariées. L'objectif est donc d'estimer la queue de la distribution de $F_Z(\cdot)$ où Z est une variable aléatoire univariée. Des extensions au cas où Z est une variable aléatoire multivariée ont été développées dans le cadre de la théorie aux valeurs extrêmes multivariées [7]. Dans le cas univarié, on s'intéresse donc à l'estimation de $\overline{F}_Z(z) = 1 - F_Z(z) = P(Z > z)$, où z est grand. Dans la plupart des applications, z est si élevé qu'aucune observation ne se trouve dans cette région de la queue de la distribution. La théorie aux valeurs extrêmes propose des hypothèses paramétriques fondées sur un raisonnement mathématique rigoureux qui permettent d'extrapoler au-delà de l'intervalle contenant les données.

Des méthodes d'estimation de densité complètement non-paramétriques ne peuvent pas fournir de bons résultats dans des régions où il n'y a pas d'observations. La figure 3.1 montre la fonction de répartition empirique $\hat{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i \leq z\}}$ pour le cas où les observations Z_i sont échantillonnées à partir d'une loi de Fréchet dont la fonction de répartition est $\Phi_\xi(z) = \exp\left(-z^{-1/\xi}\right)$ avec $\xi = 1/2$. Le fait que $\hat{F}_n(\cdot)$ augmente d'abord rapidement puis ralentisse avant d'atteindre la limite $\hat{F}_n(z) = 1$ est typique des lois à queue lourde. Soient $Z_{(1)} \geq Z_{(2)} \geq \cdots \geq Z_{(n-1)} \geq Z_{(n)}$, les observations ordonnées. La fonction de

quantile empirique $F_n^{\leftarrow}(\cdot)$ est définie comme la fonction inverse de $\hat{F}_n(\cdot)$:

$$F_n^{\leftarrow}(q) = Z_{(k)} \quad \text{pour } 1 - \frac{k}{n} < q \leq 1 - \frac{k-1}{n}.$$

Pour les niveaux de quantiles extrêmes, $1 - 1/n < q \leq 1$, le quantile empirique vaut $Z_{(1)}$ puisqu'aucune autre information n'est disponible. Cependant, la plupart du temps, le quantile recherché est fortement sous-estimé. Dans l'exemple de la figure 3.1, les quantiles de niveau $q > 0.9$ sont estimés par $F_{10}^{\leftarrow}(q) = Z_{(1)} = 4.0130$ alors que les quantiles du modèle générateur sont donnés par $x_{0.95} = 4.4154$, $x_{0.99} = 9.9749$ et $x_{0.999} = 31.6149$.

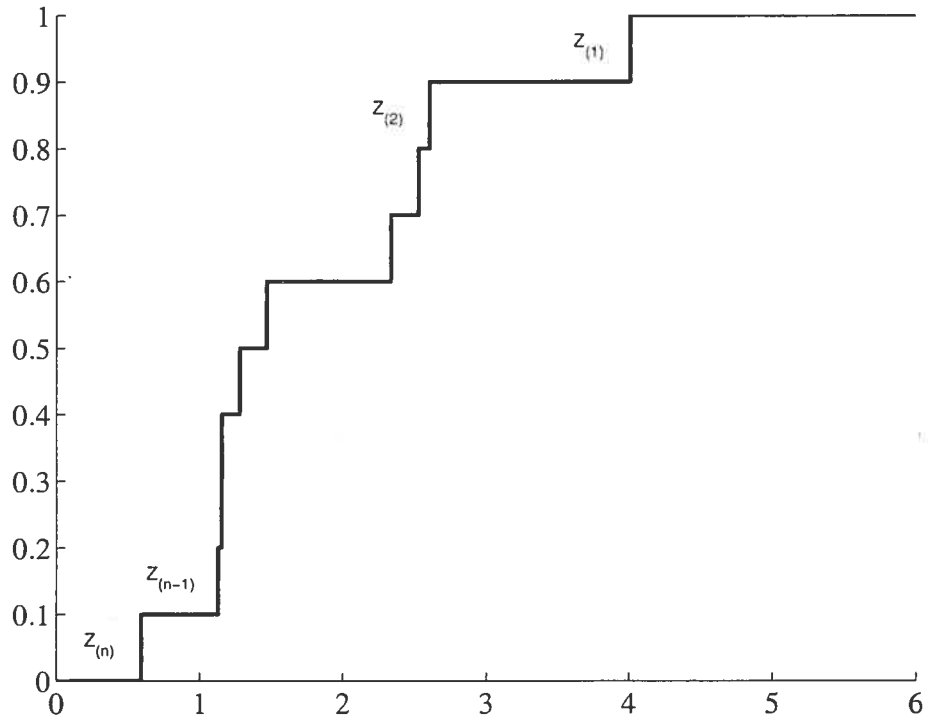

FIG. 3.1: *Fonction de répartition empirique pour $n = 10$; les $Z_{(i)}$ sont les observations ordonnées provenant d'une loi de Fréchet dont l'indice de queue est $\xi = 1/2$. Chaque $Z_{(i)}$ correspond à une marche de hauteur $1/n$.*

L'utilisation d'un mélange de distributions permet de lisser les prédictions de la fonction de répartition empirique. Lorsque les observations proviennent d'une loi dont la queue de la distribution est lourde, l'apprentissage de la densité par un mélange de Gaussiennes peut fournir de bons résultats empiriques. Une des composantes du mélange sera placée autour des observations les plus extrêmes et aura un grand écart-type pour tenter de bien représenter la densité dans la

queue de la distribution. Cependant, l'estimateur du mélange de Gaussiennes ne tiendra compte que des observations extrêmes présentes dans les données et sous-estimera la queue de la distribution au-delà de la région couverte par les données (ceci est particulièrement vrai pour les petits jeux de données). Ce phénomène est illustré à la figure 3.2 ; un mélange de Gaussiennes ayant cinq composantes est entraîné sur 10 000 données provenant d'une loi de Fréchet dont le paramètre de queue est $\xi = 1/2$. La ligne pleine représente la densité logarithmique du modèle générateur tandis que la ligne pointillée représente la densité logarithmique apprise par le mélange de Gaussiennes. Le panneau du haut de la figure 3.2 représente les densités sur un intervalle couvert par les données d'entraînement (de la moyenne échantillonale à l'observation maximale) alors que le panneau du bas représente les densités au-delà de l'observation maximale, donc dans une région sans données d'entraînement. Alors que l'estimateur de densité du mélange de Gaussiennes approxime raisonnablement bien la densité du modèle générateur sur l'intervalle couvert par les données, celui-ci sous-estime sévèrement la queue de la distribution lorsqu'il doit extrapoler.

Une autre difficulté survient lors de l'estimation de la densité d'une loi asymétrique comme la Fréchet. L'utilisation de composantes symétriques comme la Gaussienne entraîne dans ce cas la sur-estimation de la queue inférieure.

3.1. Loi Pareto hybride

Notre objectif est de développer un estimateur de densité *global*, qui permette d'estimer les moments, les quantiles et autres caractéristiques de la distribution, que ceux-ci relèvent de la partie centrale de la distribution ou de ses extrémités. Pour que cet estimateur soit en mesure de bien refléter le cas où la distribution sous-jacente a une queue lourde, inférieure ou supérieure, nous aurons recours à la théorie des valeurs extrêmes dont un survol est donné à la section 1.1. Plus précisément, nous proposons d'utiliser le mélange de distributions comme estimateur de densité et de subvenir aux problèmes liés à la présence de valeurs extrêmes en utilisant un type de composante apparenté à la loi de Pareto généralisée. L'utilisation de la loi de Pareto généralisée telle quelle dans un mélange pose problème.

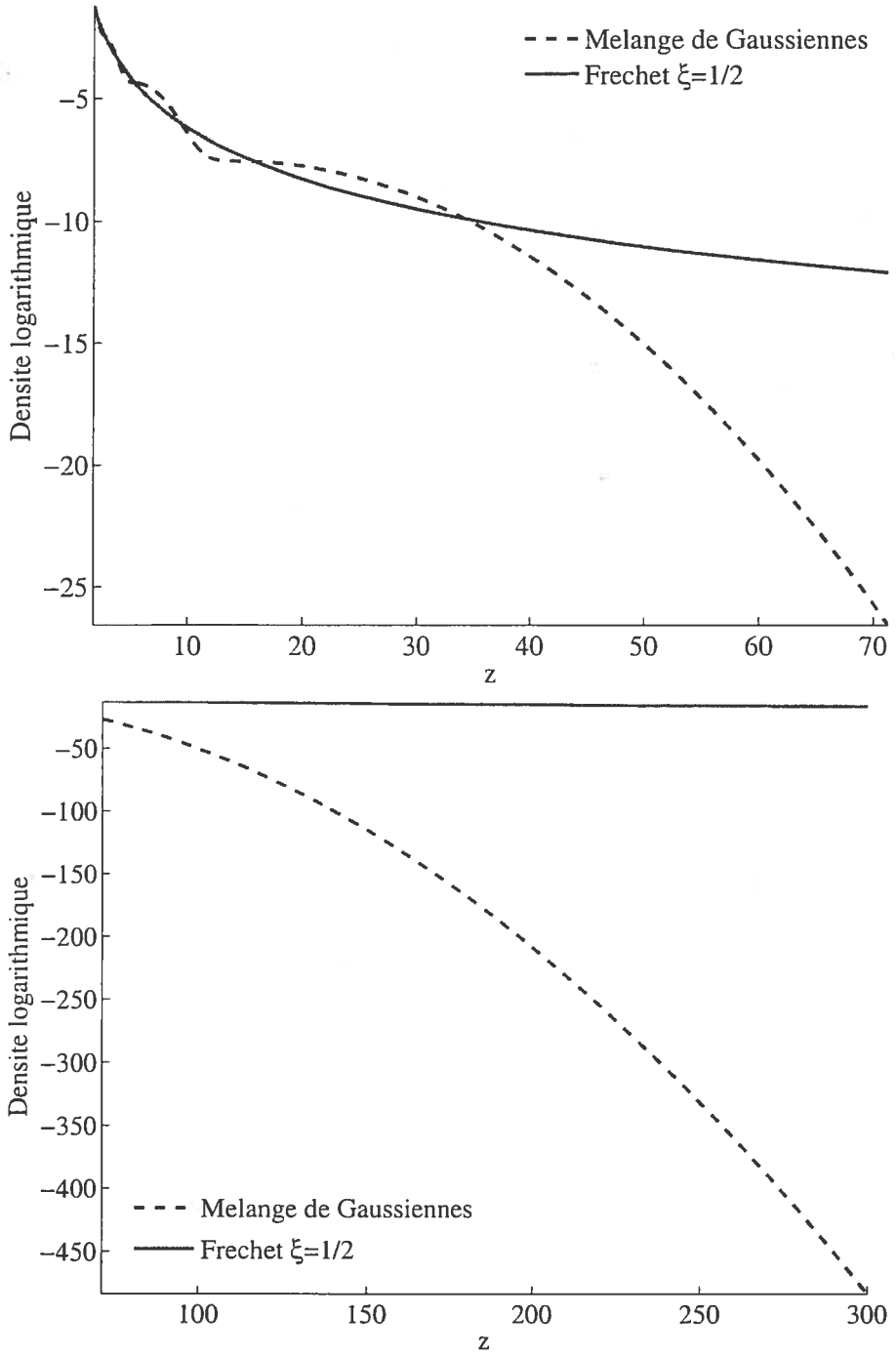

FIG. 3.2: *Mélange de cinq Gaussiennes entraîné sur 10 000 données provenant d'une loi de Fréchet avec $\xi = 1/2$. Dans le panneau du haut, l'intervalle représenté comprend les données plus grandes que la moyenne empirique alors que dans celui du bas, l'intervalle représenté ne contient aucune observation.*

D'abord, cette loi ne représente que la queue d'une distribution (voir les figures 1.4 et 1.5), elle vaut zéro sous le seuil ce qui complique l'apprentissage des paramètres par la maximisation de la log-vraisemblance du mélange. Par ailleurs, ceci ne résoud pas la question de la sélection d'un seuil adéquat. Nous proposons plutôt une nouvelle loi, la loi *Pareto hybride*, qui est construite en juxtaposant une loi Normale tronquée à une loi de Pareto généralisée et en repondérant pour s'assurer que la densité intègre à un, voir la figure 3.3. La Pareto hybride est donc définie sur tout l'axe des réels, la queue inférieure de la distribution est légère (c'est la loi Normale) et l'épaisseur de la queue supérieure est contrôlée par la valeur de l'indice de queue ξ, hérité de la loi de Pareto généralisée. En inversant la loi Pareto hybride, c'est-à-dire en ayant la partie Pareto généralisée pour la queue inférieure et la partie normale pour la queue supérieure, on obtient une loi qui peut prendre en compte des valeurs extrêmes négatives.

3.1.1. Dérivation

Soit α le point de jonction entre la loi Normale et la loi de Pareto généralisée. Ce point de jonction correspond donc au seuil au-delà duquel la Pareto généralisée est utilisée. Afin de construire la Pareto hybride, nous imposons deux contraintes en α : la continuité de la densité et de sa dérivée. Ces contraintes imposent des restrictions quant à l'emplacement du seuil. La densité de la loi Normale d'espérance μ et d'écart-type σ est donnée par :

$$f_{\mu;\sigma}(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right).$$

La densité de la loi de Pareto généralisée située au-delà du seuil α et dont l'indice de queue est ξ et le paramètre de dispersion est β est donnée par :

$$g_{\xi;\beta}(z-\alpha) = \begin{cases} \frac{1}{\beta}(1+\frac{\xi}{\beta}(z-\alpha))^{-1/\xi-1} & \text{si } \xi \neq 0, \\ \frac{1}{\beta}e^{-\frac{(z-\alpha)}{\beta}} & \text{si } \xi = 0. \end{cases} \tag{3.1.1}$$

où $z \geq \alpha$ lorsque $\xi \geq 0$ et $\alpha \leq z \leq \alpha - \beta/\xi$ lorsque $\xi < 0$. On dispose donc initialement de cinq paramètres, μ, σ, ξ, β et α. Comme nous imposons deux contraintes de continuité, il ne reste que trois paramètres libres. Nous choisissons de garder μ, σ et ξ comme paramètres libres alors que β et α deviennent des fonctions de

ces trois paramètres. De cette façon, l'indice de queue ξ sera déterminé par les données lors de l'apprentissage ainsi que l'emplacement μ et la dispersion σ, ce qui semble intuitivement raisonnable. À partir des deux contraintes de continuité, nous développons des formules explicites pour exprimer β, α et γ, le facteur de normalisation, en fonction de ξ, μ et σ. Nous traitons ici le cas $\xi > 0$ pour simplifier l'exposition, les autres cas ne nécessitent que des ajustements mineurs. La contrainte de continuité de la densité signifie que $f_{\mu;\sigma}(\alpha) = g_{\xi;\beta}(0)$ ce qui donne :

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = \frac{1}{\beta}.$$

En isolant l'exponentielle, nous obtenons :

$$\exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = \frac{\sqrt{2\pi}\sigma}{\beta}. \tag{3.1.2}$$

La deuxième contrainte de continuité de la dérivée signifie que $f'_{\mu;\sigma}(\alpha) = g'_{\xi;\beta}(0)$, d'où nous tirons :

$$-\frac{(\alpha-\mu)}{\sqrt{2\pi}\sigma^3} \exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = -\frac{(1+\xi)}{\beta^2}. \tag{3.1.3}$$

Nous remplaçons ensuite l'expression pour l'exponentielle trouvée en (3.1.2) dans l'équation (3.1.3) :

$$\frac{1+\xi}{\beta} = \frac{\alpha-\mu}{\sigma^2} \Leftrightarrow \alpha = \mu + \frac{\sigma^2}{\beta}(1+\xi). \tag{3.1.4}$$

Afin d'éliminer α qui est un paramètre dépendant, on le remplace dans l'équation (3.1.2) par l'expression trouvée en (3.1.4). En ré-arrangeant, on obtient :

$$\begin{aligned} \exp\left(-\frac{\sigma^2(1+\xi)^2}{2\beta^2}\right) = \frac{\sigma\sqrt{2\pi}}{\beta} &\Leftrightarrow \frac{1}{\sqrt{2\pi}} = \frac{\sigma}{\beta} \exp\left(\frac{\sigma^2(1+\xi)^2}{2\beta^2}\right) \\ &\Leftrightarrow \frac{1}{2\pi} = \frac{\sigma^2}{\beta^2} \exp\left(\frac{\sigma^2(1+\xi)^2}{\beta^2}\right). \end{aligned}$$

Pour que le terme qui multiplie l'exponentielle soit le même que l'argument de l'exponentielle, on multiplie la dernière équation de chaque côté par $(1+\xi)^2$:

$$\frac{(1+\xi)^2}{2\pi} = \frac{\sigma^2(1+\xi)^2}{\beta^2} \exp\left(\frac{\sigma^2(1+\xi)^2}{\beta^2}\right) \tag{3.1.5}$$

Pour résoudre l'équation (3.1.5), on se sert de la fonction $W(\cdot)$ de Lambert. Étant donné z, $W(z) = w$ est tel que w satisfait $z = we^w$. Il existe un algorithme efficace

permettant de trouver la racine de $z - we^w$ (voir [8]). Dans le cas présent, posons $z = \kappa(\xi)$, où

$$\kappa(\xi) = (1+\xi)^2/(2\pi). \tag{3.1.6}$$

Nous obtenons alors une expression pour β comme fonction de ξ et de σ en résolvant l'équation (3.1.5) :

$$W(\kappa(\xi)) = \frac{\sigma^2(1+\xi)^2}{\beta^2} \Leftrightarrow \beta(\xi,\sigma) = \frac{\sigma(1+\xi)}{\sqrt{W(\kappa(\xi))}}. \tag{3.1.7}$$

Afin d'éliminer β, qui est le deuxième paramètre dépendant dans l'expression pour α, on remplace β dans l'équation (3.1.4) par l'expression trouvée en (3.1.7). Nous obtenons alors une expression pour α comme fonction de ξ, μ et σ :

$$\alpha(\xi,\mu,\sigma) = \mu + \sigma\sqrt{W(\kappa(\xi))}. \tag{3.1.8}$$

Le facteur de normalisation γ pondère la densité de sorte que la densité de l'hybride intègre à 1. L'expression pour γ est donnée par :

$$\begin{aligned}
\gamma(\xi) &= \int_{-\infty}^{\alpha} f_{\mu;\sigma}(z)dz + \int_{\alpha}^{\infty} g_{\xi;\beta}(z-\alpha)dz && (3.1.9)\\
&= F_{\mu;\sigma}(\alpha) + 1 && (3.1.10)\\
&= 1 + \frac{1}{2}\left(1 + Erf\left(\sqrt{W(\kappa(\xi))/2}\right)\right) && (3.1.11)
\end{aligned}$$

où $F_{\mu;\sigma}(\cdot)$ est la fonction de répartion de la loi Normale d'espérace μ et d'écart-type σ. La fonction d'erreur $Erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$ peut être approximée de façon arbitrairement précise par des méthodes standards (voir [**39**]). Soit $\psi = (\xi,\mu,\sigma)$, le vecteur de paramètre de la loi Pareto hybride. La densité de la Pareto hybride est donnée par :

$$h_\psi(z) = \begin{cases} \frac{1}{\gamma} f_{\mu;\sigma}(z) & \text{si } z \le \alpha, \\ \frac{1}{\gamma} g_{\xi;\beta}(z-\alpha) & \text{si } z > \alpha \end{cases}$$

où β est donné par (3.1.7), α par (3.1.8) et γ par (3.1.9) pour le cas où $\xi > -1$[1].

[1]Lorsque $\xi \le -1$, un facteur de sign$(1+\xi)$ doit être inséré dans les formules pour β, α et γ. Par ailleurs, le développement de la Pareto hybride nécessite un cas spécial lorsque $\xi = 0$ puisque la Pareto généralisée est alors une exponentielle. La formule générale de la Pareto hybride est cependant la même.

3.1.2. Moments de la Pareto hybride

La loi Pareto hybride est donc une extension de la loi de Pareto généralisée à l'axe des réels. Le seuil au-delà duquel la loi de Pareto généralisée représente la queue de la distribution est le point de jonction avec la loi Normale α qui est défini comme une fonction des trois paramètres de la Pareto hybride (voir l'équation (3.1.8)). Soit H_ψ, la fonction de répartition de la Pareto hybride et soit Z une variable aléatoire qui suit cette loi. Lorsque $z > \alpha$, la queue de la Pareto hybride est donnée par :

$$\overline{H}_\psi(z) = \frac{1}{\gamma}\overline{G}_{\xi;\beta}(z-\alpha) \Leftrightarrow P(Z > z) = P(Z > \alpha)P(Z - \alpha > z - \alpha | Z > \alpha).$$

Cette équation est équivalente à l'équation (1.1.2) qui relie la fonction de répartition dans la queue de la distribution à la fonction de répartition excédentaire. Dans ce cas-ci, le seuil est donné par α et la probabilité d'excéder le seuil est $1/\gamma$. Ces deux quantités, α et γ, sont des fonctions de ψ, le vecteur de paramètres de la Pareto hybride. Puisque la queue de l'hybride se comporte, à un facteur de normalisation près, exactement comme la queue de la loi de Pareto généralisée, l'existence des moments de la loi Pareto hybride est contrôlée par la valeur de l'indice de queue : soit Z, une variable aléatoire de loi Pareto hybride de paramètre $\psi = (\xi, \mu, \sigma)$, alors $E[Z^r] < \infty \Leftrightarrow \xi < 1/r$. En particulier, l'espérance existe si et seulement si $\xi < 1$ et vaut :

$$E[Z] = \begin{cases} \frac{1}{\gamma}(F_{\mu;\sigma}(\alpha) + \frac{\beta}{1-\xi}) & \text{si } \xi \neq 0 \\ \frac{1}{\gamma}(F_{\mu;\sigma}(\alpha) + \beta) & \text{si } \xi = 0 \end{cases}$$

La densité de la loi Pareto hybride est illustrée dans le panneau du haut de la figure 3.3 avec pour le vecteur de paramètres $\psi = (0.4, 0, 1)$ ce qui implique que l'espérance et la variance existent. La densité logarithmique de la Pareto hybride est tracée dans le panneau du bas de la figure 3.3 pour différentes valeurs de l'indice de queue. La densité logarithmique permet de faire ressortir les différentes épaisseurs des queues de distribution associées aux indices de queue.

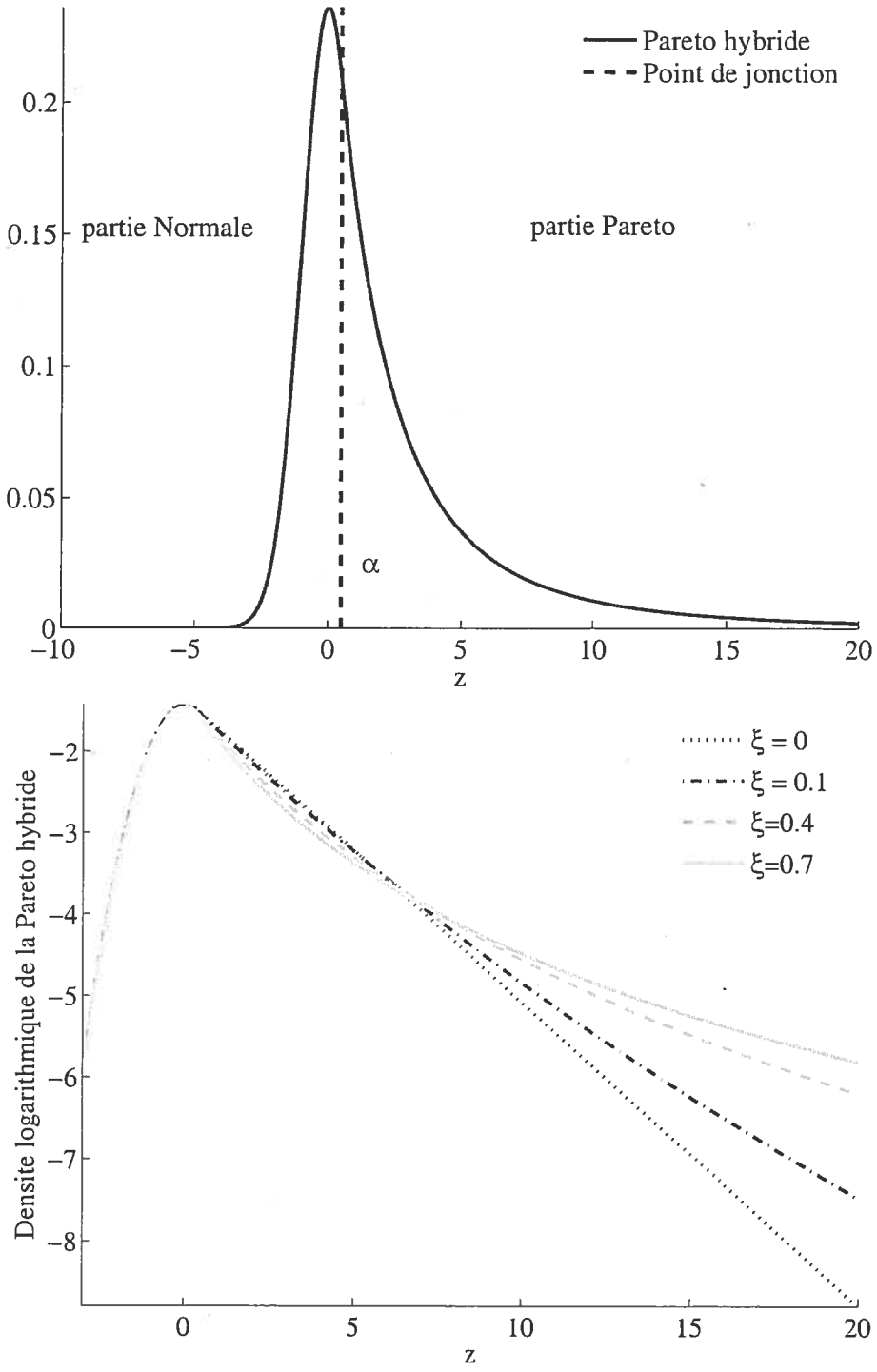

FIG. 3.3: *Panneau du haut : densité de la loi Pareto hybride de paramètres* $\psi = (0.4, 0, 1)$ *et point de jonction* $\alpha(\xi, \mu, \sigma)$. *Panneau du bas : densité logarithmique de la Pareto hybride pour différents indices de queue positifs. Dans tous les cas,* $\mu = 0$ *et* $\sigma = 1$.

3.1.3. Estimateurs de quantiles

Outre les estimateurs du maximum de vraisemblance, il est possible de dériver des estimateurs de ψ à partir d'un estimateur initial de l'indice de queue ξ et de quantiles de la distribution. Lorsque $\xi > 0$, l'estimateur de Hill (voir **[15]**) permet d'estimer l'indice de queue des lois qui sont dans le domaine d'attraction maximale de la Fréchet. Les estimateurs des paramètres de la Pareto généralisée basés sur la méthode des moments ou la méthode des moments probabilistes **[23]** fournissent une autre façon d'estimer l'indice de queue sans avoir recours à la maximisation de la log-vraisemblance. Ces estimateurs tout comme l'estimateur de Hill nécessitent la sélection d'un seuil. Cependant, puisque nous ne cherchons qu'un estimateur initial, nous proposons d'utiliser comme règle que le seuil choisi fasse en sorte que 10% des données participent à l'estimation. Deux quantiles de la distribution Pareto hybride ont une formulation particulièrement simple :

$$P(Z \leq \alpha) = \frac{F_{0,1}\left(\sqrt{W(\kappa(\xi))}\right)}{1 + F_{0,1}\left(\sqrt{W(\kappa(\xi))}\right)} \tag{3.1.12}$$

$$P(Z \leq \mu) = \frac{1}{2}\left(\frac{1}{1 + F_{0,1}\left(\sqrt{W(\kappa(\xi))}\right)}\right), \tag{3.1.13}$$

où $F_{0,1}$ est la fonction de répartition de la loi Normale Standard. En supposant que l'on dispose d'un estimateur $\hat{\xi}$ pour l'indice de queue, on peut calculer une estimation des niveaux de quantiles représentés par α et μ qui sont donnés par les membres de droite des équations (3.1.12) et (3.1.13) respectivement. Soit $\hat{\alpha}$ et $\hat{\mu}$, les quantiles empiriques correspondant aux niveaux estimés pour α et μ. Un estimateur de σ peut être calculé en isolant σ dans l'équation (3.1.8) :

$$\hat{\sigma} = \frac{(\hat{\alpha} - \hat{\mu})}{\sqrt{W(\kappa(\xi))}}.$$

Les estimateurs découlant de la méthode décrite ci-haut peuvent être utilisés comme valeurs initiales pour l'estimation par maximum de vraisemblance. Cependant, certaines restrictions s'appliquent à l'utilisation de ces estimateurs. Si $\xi \leq 0$, l'estimateur de Hill ne peut être utilisé. Les estimateurs basés sur les méthodes de moments existent seulement si la variance est finie, c'est-à-dire si

$\xi < 1/2$. Par ailleurs, ces méthodes peuvent parfois donner lieu à des estimateurs qui ne sont pas compatibles avec les données. C'est le cas lorsque $\hat{\xi} < 0$ et que la plus grande observation ne satisfait pas la contrainte du domaine (voir la définition 1.1.3), c'est-à-dire que $Z_{(1)} > \hat{\alpha} - \hat{\beta}/\hat{\xi}$. Finalement, si le jeu de données est très petit, les quantiles empiriques servant d'estimateurs à α et μ auront souvent la même valeur ce qui donnera lieu à un estimateur $\hat{\sigma} = 0$ qui n'est pas acceptable.

3.1.4. Estimation par maximum de vraisemblance

Nous avons testé à l'aide d'une simulation Monte Carlo le comportement asymptotique des estimateurs de maximum de vraisemblance des paramètres de la Pareto hybride. Un ensemble d'entraînement $\mathcal{D}_n$ et de test $\mathcal{D}_l$ sont générés selon la loi Pareto hybride de vecteur de paramètres ψ où n est choisi de plus en plus grand et l est fixé à 10 000. Soit $\hat{\psi}_n$, l'estimateur de maximum de vraisemblance de ψ calculé à partir de l'ensemble d'entraînement contenant n observations. Les estimateurs obtenus par la méthode de quantiles décrite ci-haut sont utilisés comme point de départ dans l'optimisation de la log-vraisemblance. La justesse avec laquelle la densité estimée par $h_{\hat{\psi}_n}(\cdot)$ reflète les données est mesurée sur l'ensemble de test en calculant la log-vraisemblance relative, c'est-à-dire la différence moyenne en terme de log-vraisemblance entre la densité du modèle générateur et la densité estimée :

$$\mathcal{R}_l(\widehat{\psi}_n; \psi) = \frac{1}{l} \sum_{i=1}^{l} (\log h_{\psi}(Z_i) - \log h_{\widehat{\psi}_n}(Z_i)).$$

La log-vraisemblance relative correspond à une estimation empirique de $\mathcal{KL}(h_{\psi} || h_{\widehat{\psi}_n})$ à un facteur près. Plus $\mathcal{R}_l(\widehat{\psi}_n; \psi)$ est petit, plus la densité estimée est près du modèle générateur. Pour chaque valeur de n, 100 ensembles d'entraînement sont générés et l'estimateur de maximum de vraisemblance $\widehat{\psi}_n^i$ est calculé pour chacun de ces ensembles. Ceci nous permet de rapporter la log-vraisemblance relative moyenne en test $(1/100 \sum_{i=1}^{100} \mathcal{R}_l(\widehat{\psi}_n^i; \psi))$, de calculer des intervalles de confiance autour de cette moyenne et d'estimer le biais carré et la variance des paramètres estimés. Pour le cas où $\psi = (0.7, 0, 1)$, la log-vraisemblance relative moyenne en test est donnée à la figure 3.4. Un intervalle de confiance de niveau 5 % est

tracé au-dessus de chaque barre. On observe la décroissance vers zéro de la log-vraisemblance relative moyenne ce qui indique que la densité estimée reflète de plus en plus fidèlement la densité génératrice lorsque la taille de l'ensemble d'entraînement augmente. De plus, le rétrécissement de l'intervalle de confiance indique que la variance de la log-vraisemblance relative diminue avec n. Le biais carré et la variance des paramètres estimés de la Pareto hybride sont illustrés à la figure 3.5. Le biais fluctue pour n petit puis décroît alors que la variance décroît régulièrement. Des résultats semblables pour d'autres valeurs de ψ sont fournis dans l'annexe B. Ces simulations Monte Carlo suggèrent que la méthode de maximum de vraisemblance pour l'estimation des paramètres de l'hybride Pareto a bien le comportement asymptotique attendu.

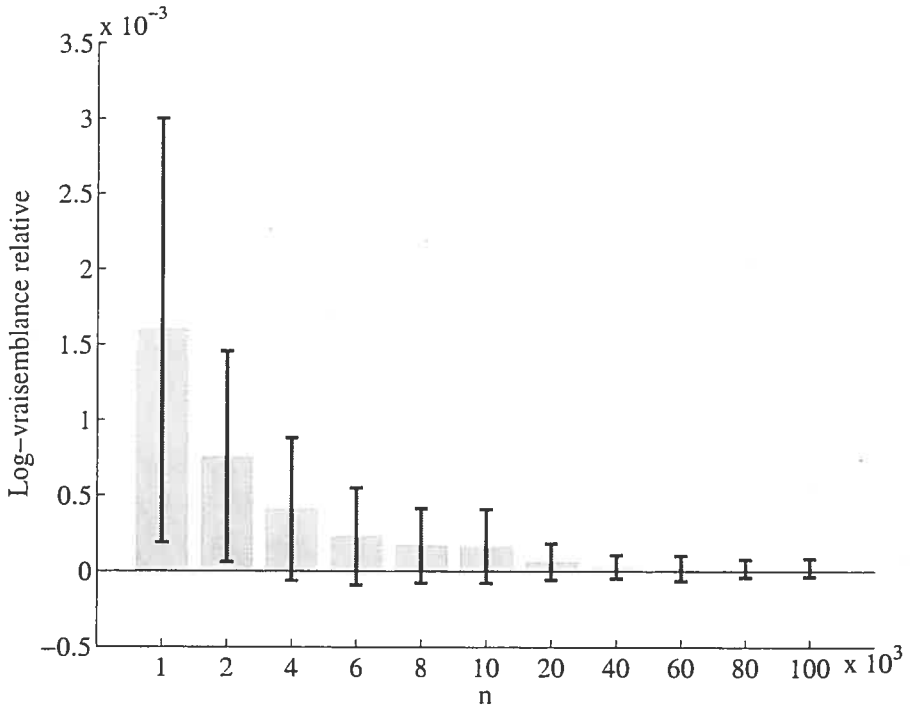

FIG. 3.4: *Estimation du vecteur de paramètres $\psi = (0.7, 0, 1)$ de la Pareto hybride : convergence de l'estimateur de maximum vraisemblance en termes de log-vraisemblance relative à la densité génératrice sur l'ensemble de test lorsque la taille n de l'ensemble d'entraînement augmente.*

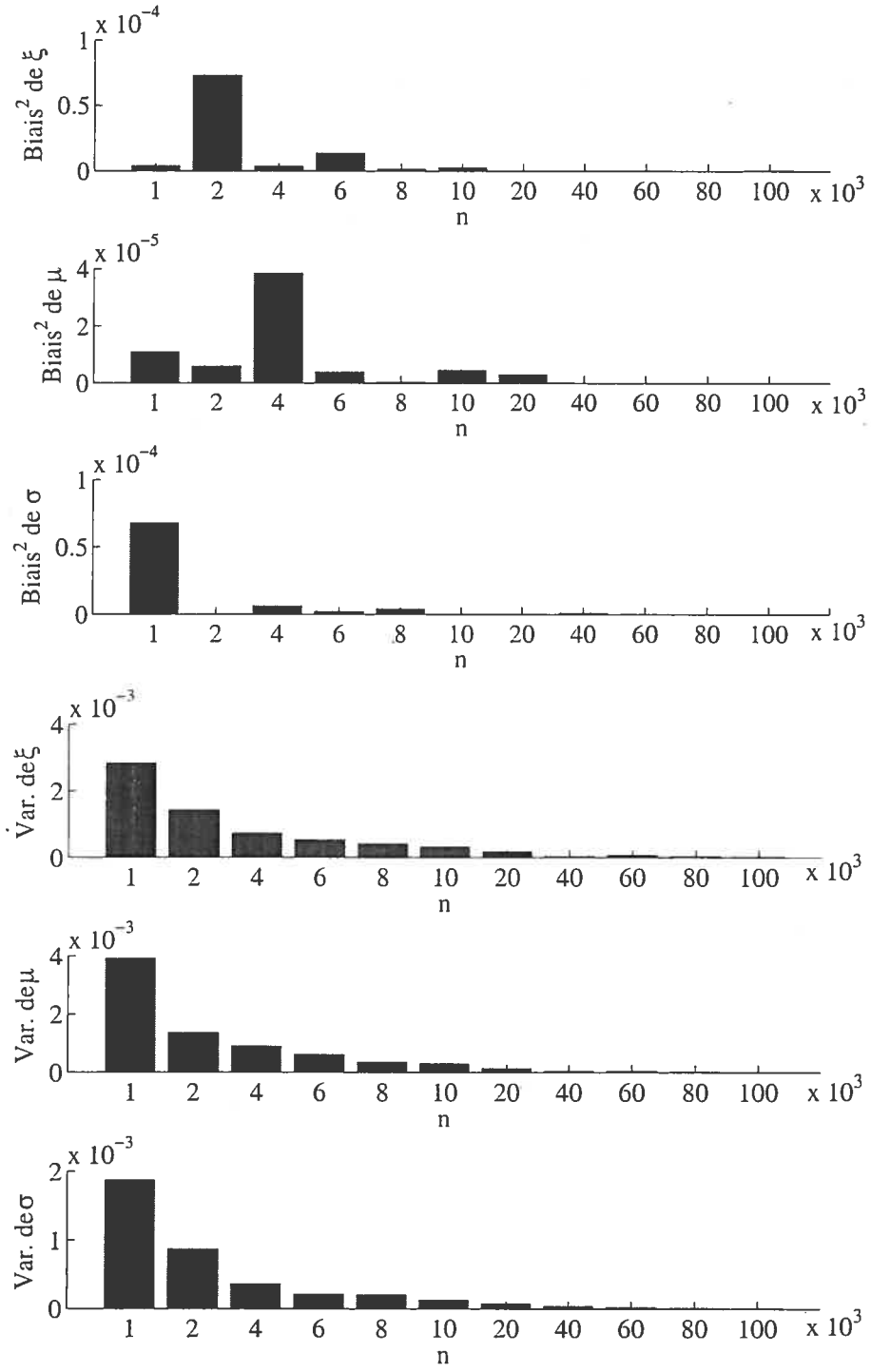

FIG. 3.5: *Biais carré (panneau du haut) et variance (panneau du bas) des estimateurs de maximum de vraisemblance des paramètres de la Pareto hybride avec* $\psi = (0.7, 0, 1)$ *lorsque la taille* n *de l'ensemble d'entraînement augmente.*

3.2. Mélange de Pareto hybrides

La distribution Pareto hybride peut être utilisée dans un mélange de plusieurs façons. Lorsque la distribution sous-jacente est étirée à droite ("right skewed"), seuls des extrêmes positifs sont présents. Nous proposons dans ce cas d'utiliser un mélange dont chaque composante est une Pareto hybride. Inversement, lorsque la distribution sous-jacente est étirée à gauche ("left skewed"), les extrêmes observés sont négatifs. Nous proposons alors de prendre comme composantes du mélange des Pareto hybrides inversées, c'est-à-dire que la partie Pareto généralisée se trouve dans la queue inférieure et la partie normale dans la queue supérieure. Si la distribution sous-jacente possède des queues lourdes des deux côtés, alors des extrêmes positifs et négatifs peuvent se produire. Dans ce cas, un mélange contenant des composantes Pareto hybrides et Pareto hybrides inversées sera en mesure de bien modéliser les deux queues de la distribution, en incluant le cas où celles-ci sont d'épaisseur différente. Soient $h_\psi(z)$ et $H_\psi(z)$ la densité et la fonction de répartition de la loi Pareto hybride. Soient également $\tilde{h}_\psi(z)$ et $\widetilde{H}_\psi(z)$ les fonctions de densité et de répartition de la loi Pareto hybride inversée. On a alors les relations suivantes :

$$\tilde{h}_\psi(z) = h_\psi(-z) \quad \widetilde{H}_\psi(z) = 1 - H_\psi(-z).$$

L'apprentissage de paramètres de toutes les variantes de mélange de Pareto hybrides se fait par la maximisation de la log-vraisemblance sur l'ensemble d'entraînement tel que discuté à la section 2.4. Cependant, dans le cas de composantes Pareto hybrides (inversées ou non), l'algorithme *EM* ne s'applique pas car il n'existe pas de solution analytique à la phase de maximisation. Il n'est pas non plus possible d'utiliser une extension de *EM* car les paramètres de la Pareto hybride sont inter-dépendants et il n'existe pas de façon de séparer les paramètres μ et σ de la loi Normale, pour lesquels une solution analytique existe, du paramètre ξ provenant de la Pareto généralisée. Nous utilisons donc, pour maximiser la log-vraisemblance, une routine d'optimisation numérique basée sur le gradient conjugué.

On peut également considérer modéliser la densité d'observations contenant des valeurs extrêmes avec un mélange ayant plusieurs composantes Normales et soit une hybride Pareto dans le cas d'extrêmes positifs, soit une hybride inversée dans le cas d'extrêmes négatifs, soit les deux dans le cas d'extrêmes positifs et négatifs. Ce type de mélange pourrait en principe être en mesure de bien modéliser une distribution à queue lourde puisque la composante hybride Pareto se chargerait d'extrapoler là où les données manquent. La complexité du mélange s'en trouverait réduite (moins de paramètres à apprendre) et l'apprentissage pourrait se faire de façon plus efficace (il existe une solution analytique à la phase de maximisation de *EM* pour les composantes normales). Cependant, des expériences préliminaires démontrent que ce type de mélange mixte ne donne pas d'aussi bons résultats qu'un mélange contenant uniquement des composantes Pareto hybrides. Les expériences ont été réalisées sur des jeux de données synthétiques générés à partir de la loi de Fréchet. Nous avons comparé un mélange de Pareto hybrides standards avec un mélange de composantes Normales et d'une composante Pareto hybride. La forme asymétrique de la distribution génératrice est peut être une raison pour laquelle le mélange contenant les composantes Normales, donc symétriques, réussit moins bien à apprendre la densité génératrice. Par ailleurs, il est possible que le fait d'utiliser plusieurs hybrides dans le mélange permette de réduire d'une part le biais lié à l'approximation par la Pareto généralisée et d'autre part, le biais dû aux contraintes sur l'emplacement du seuil dans la construction de l'hybride. C'est pourquoi nous avons décidé de nous concentrer sur des mélanges n'utilisant que des composantes de type Pareto hybrides.

3.2.1. Composante dominante et seuil implicite

Puisque nous considérons l'utilisation de mélange de distributions comme estimateur de densité, une question naturelle qui survient est de savoir quel est l'indice de queue du mélange en tant que distribution, ce qui est équivalent à savoir à quel domaine d'attraction maximale le mélange appartient (voir la section 1.1). Pour répondre à cette question, nous aurons recours au concept de *queue dominante* introduit par Kang et Serfozo [**28**] :

Définition 3.2.1 (Queue dominante d'un mélange). *Soient $F_1, \ldots, F_m$ les fonctions de répartition associées aux composantes d'un mélange. On dit que la queue de la distribution de F^* domine celles des F_i si pour $i = 1 \ldots m$:*

$$\lim_{x \to \infty} \frac{\overline{F}_i(x)}{\overline{F}^*(x)} = r_i,$$

où $0 \leq r_i < \infty$.

Kang et Serfozo [**28**] ont démontré que la composante d'un mélange dont la queue est dominante détermine le domaine d'attraction maximale du mélange et par conséquent, l'indice de queue associé au mélange. Autrement dit, soit F la fonction de répartition du mélange et soit F_{i^*} la fonction de répartition de la composante ayant la queue dominante. Alors :

$$F \in MDA(H_\xi) \Leftrightarrow F_{i^*} \in MDA(H_\xi).$$

Dans le cas d'un mélange de lois Normales, la queue dominante du mélange est la composante i^* telle que $\sigma_{i^*} = \max_i \sigma_i$ et $\mu_{i^*} > \mu_i$ pour tout $i \neq i^*$ tel que $\sigma_i = \sigma_{i^*}$. Ceci signifie que le mélange de lois Normales appartient au même domaine d'attraction maximale que la loi Normale, c'est-à-dire celui de la loi de Gumbel qui comprend les distributions dont les queues sont légères à modérément lourdes. L'indice de queue associée au mélange est donc $\xi = 0$.

Dans le cas du mélange de Pareto hybrides, la composante i^* ayant la queue dominante est celle dont la queue est la plus lourde, c'est-à-dire que i^* est tel que $\xi_{i^*} = \max_i \xi_i$ et $\beta_{i^*} > \beta_i$ si $\xi_{i^*} = \xi_i$ et $i \neq i^*$. Le domaine d'attraction maximale du mélange de Pareto hybrides est donc celui de la composante dominante, $MDA(\xi_{i^*})$. Si $\xi_{i^*} > 0$, il s'agit du domaine d'attraction de la Fréchet qui comprend les distributions à queues lourdes. Il semble donc raisonnable d'utiliser un mélange de Pareto hybrides lorsque la distribution sous-jacente est à queues lourdes.

L'utilisation d'un mélange de Pareto hybrides permet de contourner la question de la sélection du seuil inhérente à la méthodologie PoT (voir la sous-section 1.1.2). En effet, le seuil peut être défini implicitement comme étant la point de jonction α_{i^*} de la composante dominante. Le seuil devient donc une fonction des paramètres de la composante dominante (voir l'équation (3.1.8)). Comme les

paramètres du mélange sont appris en maximisant la log-vraisemblance sur l'ensemble d'entraînement, le seuil est alors déterminé de manière indirecte par les données.

3.3. ÉTUDE SIMULATOIRE

À l'aide de jeux de données synthétiques, nous avons étudié le comportement du mélange de Pareto hybrides en le comparant à celui d'autres estimateurs. Nous avons généré des données à partir de la loi de Fréchet d'indice de queue $\xi = 0.2$ et $\xi = 0.5$. Ces densités sont illustrées à la figure 3.6. Les paramètres d'emplacement et de dispersion ont été fixés à $\mu = 0$ et $\sigma = 1$ dans tous les cas. Des valeurs incrémentales de n, la taille de l'ensemble d'entraînement, sont utilisées (de 100 à 20 000) alors que la taille de l'ensemble de test l est fixée à 10 000. Pour chaque taille d'ensemble d'entraînement n, un nombre b_n de paires d'ensembles d'entraînement et de test $(\mathcal{D}_n, \mathcal{D}_l)$ a été généré. Nous sommes donc en mesure de calculer la performance moyenne ainsi que la variance de chaque estimateur. Puisqu'il y a plus de variabilité dans les petits jeux de données et que l'entraînement est plus long pour les grands jeux de données, nous avons utilisé plus de paires $(\mathcal{D}_n, \mathcal{D}_l)$ lorsque n est petit que lorsque n est grand. Plus précisément, $b_n = 100$ pour $100 \leq n \leq 800$, $b_n = 60$ pour $1000 \leq n \leq 8000$ et $b_n = 30$ pour $n > 10000$.

3.3.1. Entraînement et critères d'évaluation

Le but de l'étude est double : premièrement, vérifier s'il est avantageux d'utiliser le mélange de Pareto hybrides par rapport à d'autres types d'estimateurs de densité et deuxièmement, comparer l'estimation de la queue d'une distribution par un mélange de Pareto hybrides avec la méthode PoT. Pour répondre au premier objectif, nous comparons le mélange de Pareto hybrides avec les estimateurs suivants :

(1) *Le mélange de Normales* : le type de mélange le plus commun, qui possède de bonnes propriétés de convergence si le nombre de composantes est bien choisi par rapport à la taille de l'ensemble d'entraînement [**40**].

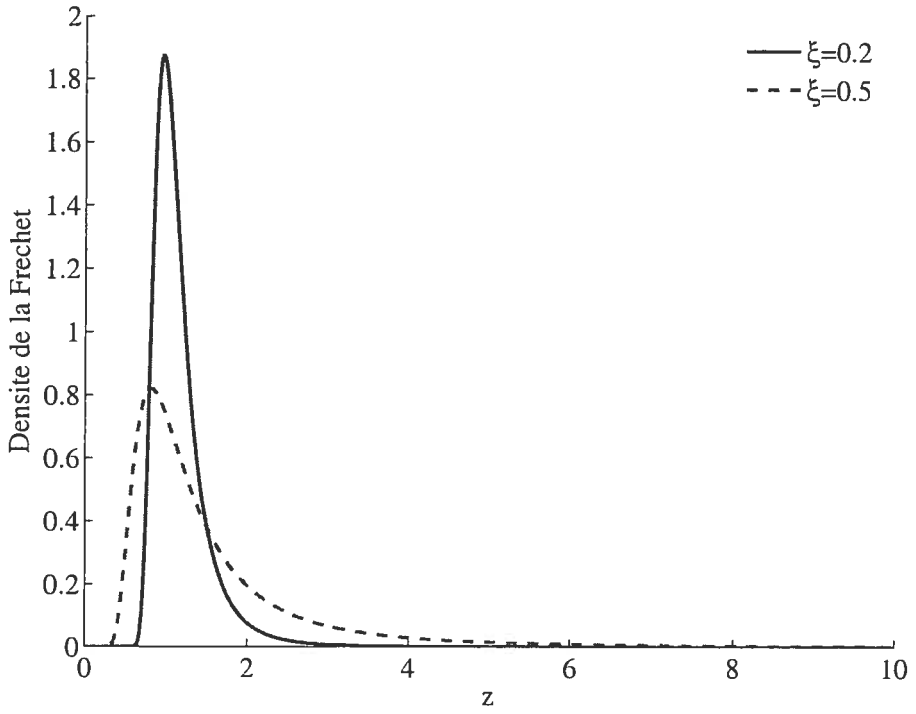

FIG. 3.6: *Densité de la Fréchet pour différentes valeurs de l'indice de queue ξ.*

(2) *Le mélange de Log-Normales* : la loi Log-Normale est une alternative robuste à la loi Normale. Elle est définie de la façon suivante : Y est de loi Log-Normale si et seulement si $Z = \log Y$ est distribuée selon la loi Normale. La loi Log-Normale est asymétrique et n'est définie que sur l'axe positif ce qui en restreint l'application aux données positives. Cette loi a une queue modérément lourde. Elle appartient toutefois au domaine d'attraction maximale de la Gumbel qui contient aussi la loi Normale. Un mélange avec des composantes Log-Normales est équivalent à un mélange de Normales sur le logarithme des données. Soient $f(\cdot)$ et $g(\cdot)$, les fonctions de densité de la loi Normale et Log-Normale respectivement. On a la relation suivante : $\log g(x) = \log f(\log(x)) - \log(x)$. Cette relation se transfère aux mélanges.

(3) *L'estimateur "fenêtre de Parzen"* qui est un estimateur à noyaux gaussiens : cet estimateur est entièrement non-paramétrique et est convergent dans la norme L_1 **[12]**.

Tous les mélanges de distributions sont entraînés par maximisation de la log-vraisemblance sur l'ensemble d'entraînement. Comme l'optimisation peut atteindre des minima locaux, l'initialisation des mélanges est relancée cinq fois et l'optimisation est reprise en conséquence. Les paramètres qui correspondent à l'erreur d'entraînement la plus petite sont retenus. Les mélanges sont initialisés à l'aide d'un algorithme de "clustering" tel que k-means ou k-medians [**38**] afin de regrouper les données en m sous-groupes. Les paramètres de chaque composante sont alors estimés à l'aide des données du sous-groupe correspondant. Les poids du mélange ("priors") sont initialisés en prenant la proportion des données dans chaque sous-groupe.

Ces estimateurs sont comparés en termes de la log-vraisemblance relative à la densité génératrice et en termes d'estimation de quantiles extrêmes. La log-vraisemblance relative est donnée par :

$$\mathcal{R}_l(\theta) = -\frac{1}{l}\sum_{i=1}^{l} \log\left(\frac{\phi_\theta(z_i)}{p(z_i)}\right), \tag{3.3.1}$$

où la somme est sur l'ensemble de test $\mathcal{D}_l$, $p(\cdot)$ est la densité du modèle générateur et $\phi_\theta(\cdot)$ est la densité de l'estimateur. Le critère $\mathcal{R}_l$ est l'équivalent empirique du critère de Kullback-Leibler à un facteur près, voir l'équation (2.3.1). Plus la log-vraisemblance relative est petite, mieux l'estimateur $\phi_\theta(\cdot)$ reflète la densité génératrice. Nous calculons aussi des estimations de quantiles extrêmes à partir des estimateurs de densités. Soit F la fonction de répartition de la distribution qui nous intéresse. Un quantile de niveau q de cette distribution est la valeur z_q telle que $F(z_q) = q$, si F est croissante. Dans le cas des estimateurs de densité que nous considérons, il n'existe pas une forme analytique pour exprimer l'inverse de la fonction de répartition. Nous approximons donc les quantiles en cherchant la racine de $F(z_q) - q$. Nous calculons des quantiles de niveau $q = 0.99$, $q = 0.999$ et $q = 0.9999$. Les quantiles estimés sont standardisés, c'est-à-dire divisés par le quantile de même niveau de la distribution génératrice. De cette façon, la valeur ciblée pour l'estimation des quantiles est 1, quel que soit le niveau q. Ces quantiles standardisés sont ensuite comparés à l'aide de la racine carrée de l'erreur

quadratique moyenne (*RMSE : Root Mean Square Error*) :

$$RMSE(\hat{z}_q) = \sqrt{\text{biais}^2 + \text{variance}} = \sqrt{(E[\hat{z}_q] - 1)^2 + E[(\hat{z}_q - E[\hat{z}_q])^2]}, \quad (3.3.2)$$

où $\hat{z}_q$ est le quantile estimé et standardisé de niveau q.

Pour atteindre le deuxième objectif, nous comparons l'indice de queue et les quantiles extrêmes tels qu'estimés par le mélange de Pareto hybrides avec les estimateurs résultants de la méthode PoT. Nous utilisons aussi le critère *RMSE* pour ces comparaisons. Le seuil au-delà duquel la Pareto généralisée est ajustée aux données est choisi comme étant un quantile empirique de niveau q_{PoT} qui est considéré comme un hyper-paramètre. Dans ce cas-ci, choisir le niveau de quantile en maximisant la log-vraisemblance sur un ensemble de validation n'est pas une procédure adéquate. En effet, la log-vraisemblance favorise les observations au centre de la distribution donc maximiser ce critère entraîne la sélection du seuil le plus bas possible. Nous déterminons plutôt le niveau de quantile optimal à l'aide du test d'adéquation de la Pareto généralisée développé par Choulakian et Stephens [**6**]. Le niveau de quantile de départ est choisi bas, puis il est augmenté petit à petit jusqu'à ce que le test d'adéquation soit satisfait par les paramètres estimés. Pour ces simulations, le niveau de quantile de départ est de 0.01. Si le test d'adéquation n'est pas satisfait lorsque le niveau de quantile 0.95 est atteint, ce niveau est retenu comme seuil.

Tous les autres hyper-paramètres, le nombre de composantes pour les mélanges et la largeur de fenêtre pour l'estimateur de la fenêtre de Parzen, ont été sélectionnés en maximisant la log-vraisemblance sur un ensemble de validation. Le nombre de composantes exploré varie entre une et vingt alors que la largeur de fenêtre prend des valeurs entre 10^{-10} et un. Nous utilisons un t-test unilatéral de niveau 5% pour déterminer si un modèle offre une performance supérieure à un autre modèle. L'ensemble d'entraînement $\mathcal{D}_n$ est séparé en deux : 80 % des données servent à l'apprentissage des paramètres et les 20% restantes servent d'ensemble de validation pour la sélection de modèles.

3.3.2. Résultats des simulations

3.3.2.1. *Log-vraisemblance et hyper-paramètres sélectionnés*

Les résultats complets pour la statistique de log-vraisemblance relative à la densité génératrice sont fournis dans les tableaux 3.1 et 3.2 pour la Fréchet d'indice de queue $\xi = 0.2$ et $\xi = 0.5$ respectivement. Lorsque la performance d'un algorithme est significativement meilleure, la log-vraisemblance relative est en caractères gras. La moyenne des hyper-paramètres sélectionnés sur les b_n répétitions de la simulation se trouvent dans le tableau 3.3. On désigne par **ummh**, **ummg** et **umml**, les mélanges à composantes Pareto hybrides, Gaussiennes et Log-Normales respectivement. On dénote l'estimateur de la fenêtre de Parzen par **uparzen**. Tous les mélanges de distributions convergent en terme de log-vraisemblance relative, c'est-à-dire que la log-vraisemblance relative diminue en moyenne et en erreur standard lorsque la taille de l'ensemble d'entraînement augmente. Cependant, la log-vraisemblance relative du mélange de Pareto hybrides est généralement plus petite que celle des autres modèles. L'écart est d'autant plus prononcé que la taille de l'ensemble d'entraînement est petite. Pour ces deux jeux de données, l'estimateur de la fenêtre de Parzen obtient de mauvais résultats. Ceci met en évidence les limites des estimateurs non-paramétriques classiques en présence d'observations extrêmes. Pour les données Fréchet avec indice de queue $\xi = 0.2$, la performance du mélange de Pareto hybrides est significativement meilleure que celle de tous les autres estimateurs. Pour les données Fréchet $\xi = 0.5$ la situation est un peu différente ; les différences entre la performance du mélange de Pareto hybrides et de Log-Normales sont moins souvent significatives. Pour ces deux jeux de données, le mélange de Pareto hybrides utilise moins de composantes. Puisque la queue de la distribution est plus lourde lorsque $\xi = 0.5$, les mélanges de distributions ont besoin en moyenne de plus de composantes pour modéliser la densité génératrice. Les résultats de ces simulations pour la méthode PoT sont présentés dans le tableau 3.4. Le niveau de quantile qui détermine le seuil, est choisi de plus en plus grand au fur et à mesure que plus de données

d'entraînement deviennent disponibles. Ceci est cohérent avec le dilemme biais-variance lié à la sélection du seuil : puisque plus de données sont disponibles, la sélection d'un seuil plus élevé permet de réduire le biais de l'approximation de la queue de la distribution par la Pareto généralisée tout en contrôlant la variance des estimateurs.

n	ummh	ummg	umml	uparzen
100	**0.047 (0.0028)**	0.22 (0.023)	0.1 (0.017)	4.5e+17 (3.8e+16)
200	**0.025 (0.00088)**	0.11 (0.0038)	0.049 (0.003)	2.5e+17 (1.8e+16)
400	**0.013 (0.00035)**	0.08 (0.0016)	0.026 (0.0012)	1.6e+17 (1.1e+16)
800	**0.0071 (0.00014)**	0.056 (0.0012)	0.017 (0.00016)	1.3e+17 (7.2e+15)
1000	**0.0059 (0.00012)**	0.042 (0.0011)	0.017 (0.00018)	9.2e+16 (3.5e+15)
2000	**0.0035 (3.5e-05)**	0.018 (0.00044)	0.013 (0.00014)	9e+16 (3.1e+15)
4000	**0.0021 (1.4e-05)**	0.011 (0.00019)	0.0079 (0.00011)	3.2e+16 (9e+14)
8000	**0.0016 (7.8e-06)**	0.0066 (0.0001)	0.0026 (3.9e-05)	1.8e+16 (4.5e+14)
10000	**0.0015 (7.9e-06)**	0.0061 (0.00011)	0.002 (3e-05)	2.6e+16 (5e+14)
20000	**0.00092 (4.9e-06)**	0.0025 (1.5e-05)	0.001 (4e-06)	1.3e+16 (2e+14)

TAB. 3.1: *Log-vraisemblance (err. std) relative à la densité génératrice sur l'ensemble de test. Plus ce critère est petit, meilleure est la performance de l'estimateur. Les performances significativement meilleures sont en caractères gras. Les données sont générées d'après une loi de Fréchet d'indice de queue $\xi = 0.2$. n représente la taille de l'ensemble d'entraînement.*

n	ummh	ummg	umml	uparzen
100	0.07 (0.005)	1.6 (0.23)	0.089 (0.012)	3.7e+20 (1e+20)
200	0.036 (0.0024)	0.84 (0.083)	0.038 (0.002)	2.6e+20 (3.4e+19)
400	**0.019 (0.00046)**	0.55 (0.063)	0.024 (0.00043)	3.2e+20 (4.4e+19)
800	0.011 (0.00019)	0.29 (0.022)	0.019 (0.00021)	4.4e+20 (5.7e+19)
1000	**0.0098 (0.00013)**	0.23 (0.018)	0.017 (0.00017)	1.5e+20 (9.6e+18)
2000	**0.0066 (5.3e-05)**	0.21 (0.011)	0.012 (0.00013)	2.4e+20 (1.5e+19)
4000	**0.0048 (3.2e-05)**	0.21 (0.01)	0.0071 (0.0001)	1e+20 (6.3e+18)
8000	0.0032 (1.4e-05)	0.13 (0.0058)	0.0027 (3.6e-05)	9.3e+19 (2.7e+18)
10000	0.0026 (1.5e-05)	0.072 (0.0031)	0.002 (3e-05)	1.2e+20 (2.5e+18)
20000	0.0021 (8.7e-06)	0.1 (0.0026)	**0.0011 (4.2e-06)**	8.4e+19 (2.2e+18)

TAB. 3.2: *Log-vraisemblance (err. std) relative à la densité génératrice sur l'ensemble de test. Plus ce critère est petit, meilleure est la performance de l'estimateur. Les performances significativement meilleures sont en caractères gras. Les données sont générées d'après une loi de Fréchet d'indice de queue $\xi = 0.5$. n représente la taille de l'ensemble d'entraînement.*

n	m_{ummh}	m_{ummg}	m_{umml}	σ_{uparzen}
100	2.5	2	2	0.0023
200	2	2.1	2.3	0.0077
400	2	2.9	2.3	0.0072
800	2	3.3	2.5	0.0044
1000	2.2	4	2.5	0.002
2000	2	4.1	2.9	0.0028
4000	2.3	4.8	3.3	0.002
8000	2.2	5.9	4.4	0.0019
10000	2.3	6.8	4.2	0.0027
20000	2.8	7.5	4.3	0.002

n	m_{ummh}	m_{ummg}	m_{umml}	σ_{uparzen}
100	2.1	2.6	2.1	0.0027
200	2	3.2	2	0.0093
400	2	4.5	2.1	0.017
800	2	5	2.3	0.0071
1000	2.2	5.5	2.3	0.0067
2000	2.1	6	2.9	0.0052
4000	2.4	7	3.6	0.012
8000	3	8.5	3.9	0.0068
10000	3.9	9.1	4.6	0.013
20000	3.9	9.3	4.5	0.02

TAB. 3.3: *Hyper-paramètres moyens sélectionnés pour les données Fréchet d'indice de queue $\xi = 0.2$ dans le panneau de gauche et d'indice de queue $\xi = 0.5$ dans le panneau de droite, n représente la taille de l'ensemble d'entraînement. Pour les mélanges, m est le nombre de composantes et pour l'estimateur de la fenêtre de Parzen, σ_{uparzen} est la largeur de fenêtre.*

n	u	q_{PoT}	$\mathcal{R}_l$ (*err.std.*)
100	0.92	0.2	0.034 (0.0028)
200	0.95	0.26	0.02 (0.0013)
400	1	0.34	0.01 (0.00051)
800	1	0.39	0.0086 (0.00033)
1000	1	0.42	0.0066 (0.00035)
2000	1.1	0.52	0.0013 (0.00024)
4000	1.2	0.59	0.002 (0.00019)
8000	1.2	0.64	0.004 (0.00015)
10000	1.2	0.66	0.003 (0.00018)
20000	1.2	0.7	0.0012 (0.00012)

n	u	q_{PoT}	$\mathcal{R}_l$ (*err.std.*)
100	0.67	0.096	0.034 (0.0031)
200	0.82	0.2	0.019 (0.0011)
400	0.85	0.24	0.012 (0.00081)
800	0.97	0.32	0.005 (0.0003)
1000	0.96	0.32	0.0053 (0.0003)
2000	1.1	0.41	0.0021 (0.00026)
4000	1.2	0.5	0.00068 (0.00019)
8000	1.4	0.57	0.00037 (0.00012)
10000	1.4	0.57	0.0043 (0.00015)
20000	1.6	0.64	0.0031 (0.00011)

TAB. 3.4: *Log-vraisemblance relative à la densité génératrice $\mathcal{R}_l$ (voir l'équation 3.3.1) sur les excès et niveau de quantile q_{PoT} moyen sélectionné pour la méthode PoT sur les données Fréchet d'indice de queue $\xi = 0.2$ dans le panneau de gauche et $\xi = 0.5$ dans le panneau de droite, n représente la taille de l'ensemble d'entraînement. Le seuil moyen correspondant au quantile q_{PoT} est u.*

3.3.2.2. *Densités apprises*

Nous avons tracé les courbes des densités apprises par les différents estimateurs pour une des expériences avec un ensemble d'entraînement de taille 100 et une des expériences avec un ensemble d'entraînement de taille 1000. Les figures 3.7 et 3.10 illustrent la densité génératrice et les estimateurs dans la partie centrale de la distribution pour les données Fréchet avec $\xi = 0.2$ et $\xi = 0.5$ respectivement. Les estimateurs approximent la densité raisonnablement bien dans cette région bien que de nombreux artéfacts des mélanges soient visibles pour le petit jeu de données. L'estimation en général devient plus lisse et plus proche de la densité génératrice lorsque le nombre de points d'entraînement augmente. Les figures 3.8 et 3.11 montrent la densité estimée par les différents modèles dans la queue supérieure de la distribution pour les cas où $\xi = 0.2$ et $\xi = 0.5$ respectivement. On peut comparer la méthode PoT avec les autres modèles uniquement dans cette région. Le mélange de Gaussiennes sous-estime grandement la queue lorsque l'ensemble d'entraînement est petit et que la queue est plus lourde ($\xi = 0.5$). Le mélange de Log-Normales produit de meilleurs résultats mais sous-estime tout de même systématiquement la queue supérieure. L'estimateur de la fenêtre de Parzen n'est pas en mesure d'extrapoler au-delà des données et prédit donc une densité pratiquement nulle dans cette région. L'estimateur de la méthode PoT et le mélange de Pareto hybrides approximent très bien la queue de la Fréchet pour les deux indices de queue. Finalement, les figures 3.9 et 3.12 illustrent les densités estimées dans la queue inférieure de la distribution. La densité de la Fréchet est zéro dans cette région. Par construction, le mélange de Log-Normales prédit une densité nulle sur l'axe négatif et fournit donc dans ce cas une approximation adéquate. Puisque l'estimateur de la fenêtre de Parzen ne réussit pas à extrapoler au-delà de l'intervalle couvert par les données, il prédit aussi très peu de densité dans cette région. La différence de performance importante se trouve plutôt entre le mélange de Pareto hybrides et de Gaussiennes. La queue inférieure du mélange de Pareto hybrides décroît rapidement alors que celle du mélange de Gaussiennes est beaucoup lente à décroître. La Gaussienne étant symétrique, la composante du mélange ayant la queue la plus lourde permet de mieux modéliser les observations

extrêmes de la queue supérieure mais du même coup, elle sur-estime la queue inférieure.

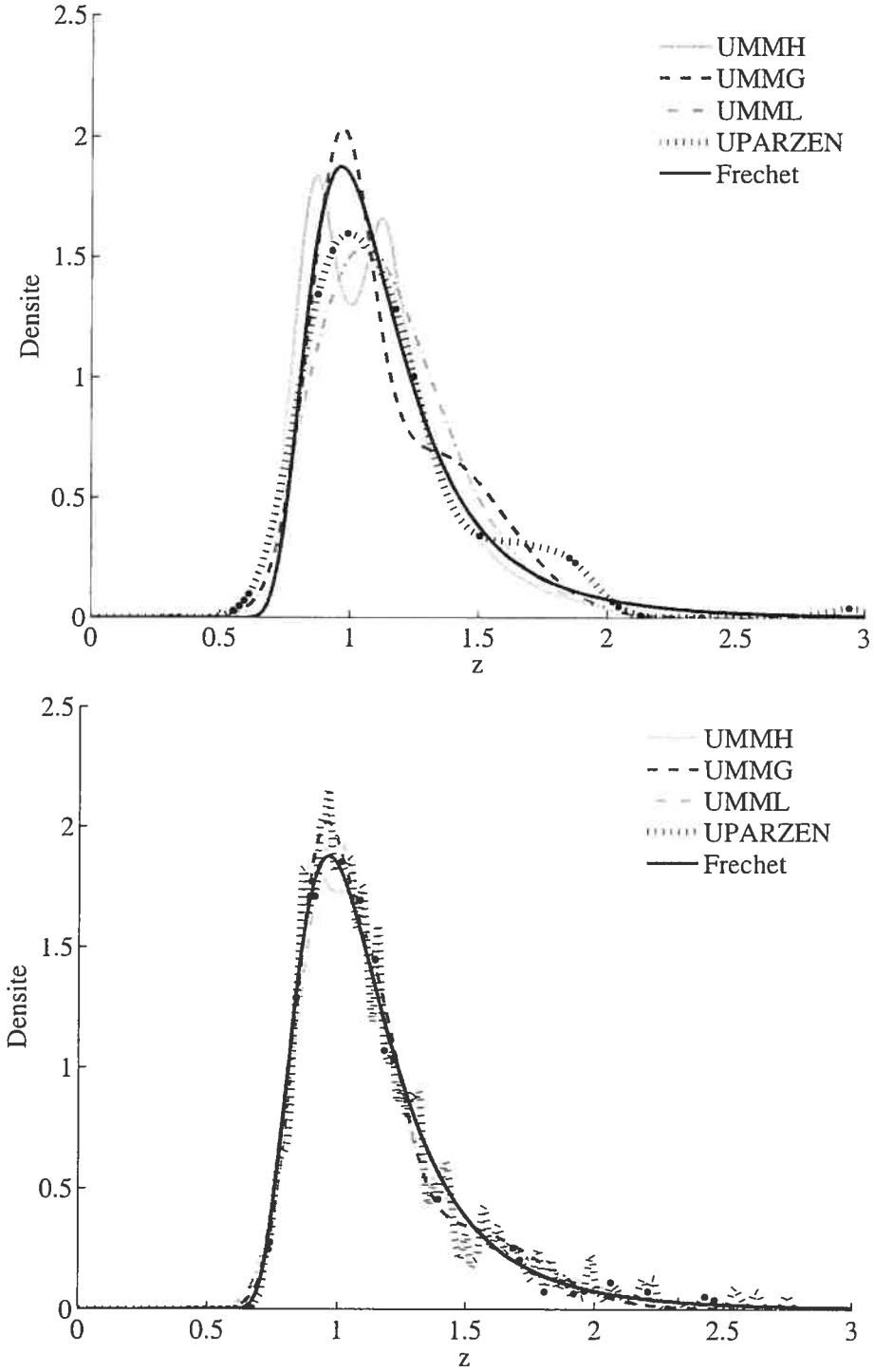

FIG. 3.7: *Densité estimée dans la partie centrale (99% de l'ensemble d'entraînement) pour les données Fréchet avec $\xi = 0.2$. Ensemble d'entraînement de 100 observations dans le panneau du haut et de 1000 observations dans le panneau du bas.*

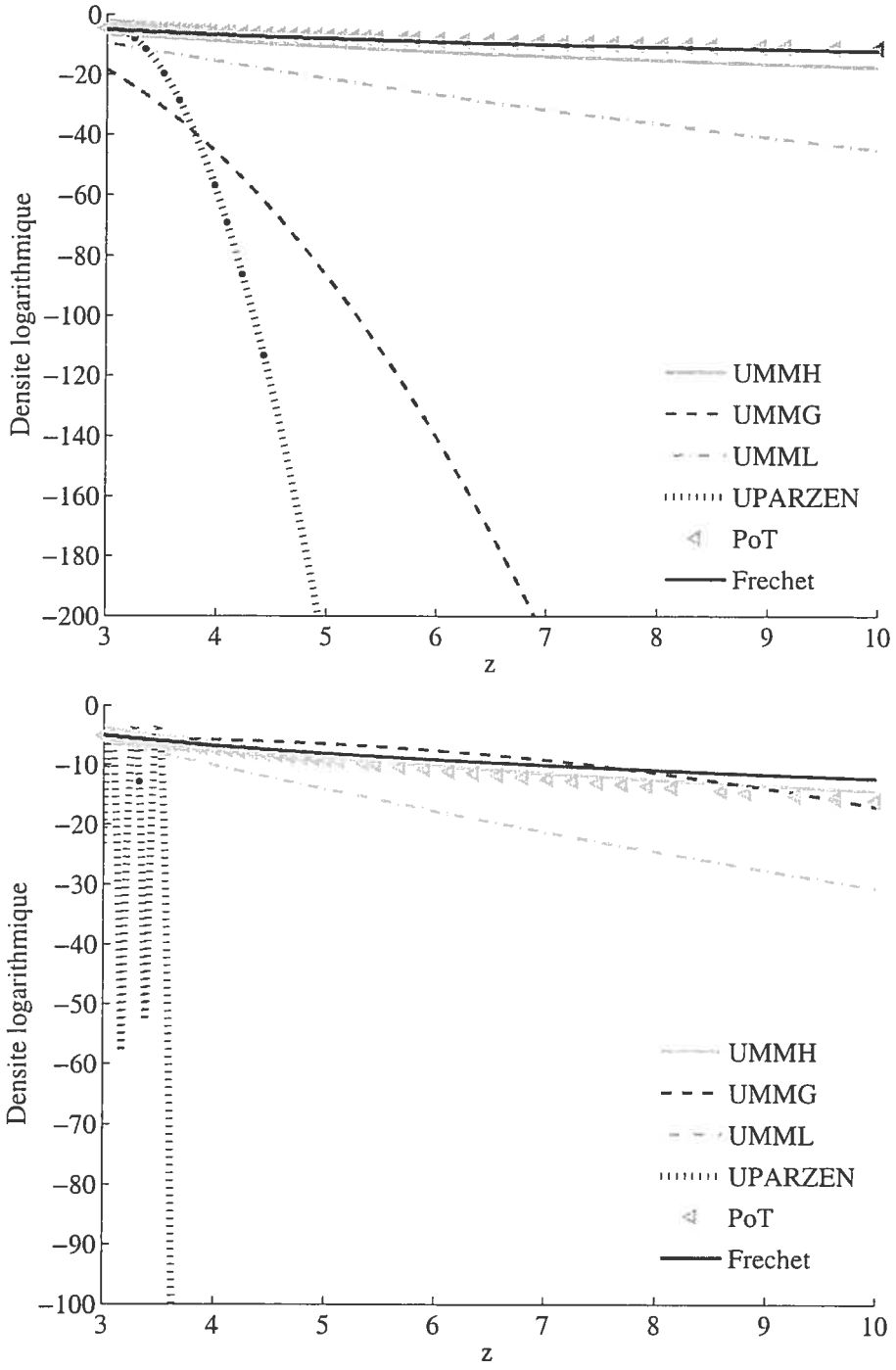

FIG. 3.8: *Densité estimée dans la queue supérieure de la distribution (moins de 1% de l'ensemble d'entraînement) pour les données Fréchet avec* $\xi = 0.2$. *Ensemble d'entraînement de 100 observations dans le panneau du haut et de 1000 observations dans le panneau du bas.*

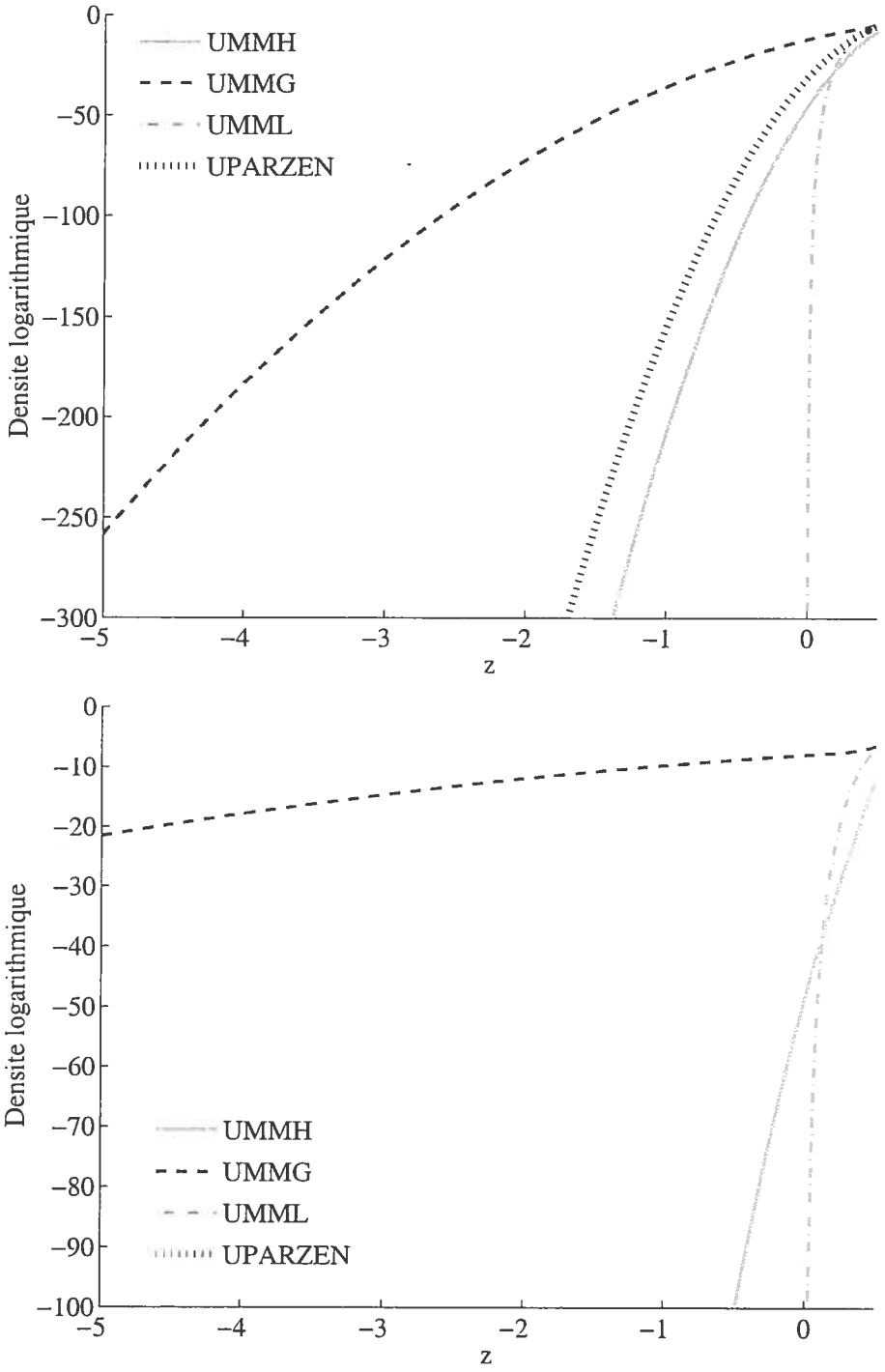

FIG. 3.9: *Densité estimée dans la queue inférieure de la distribution (pas d'observations) pour les données Fréchet avec* $\xi = 0.2$. *Ensemble d'entraînement de 100 observations dans le panneau du haut et de 1000 observations dans le panneau du bas.*

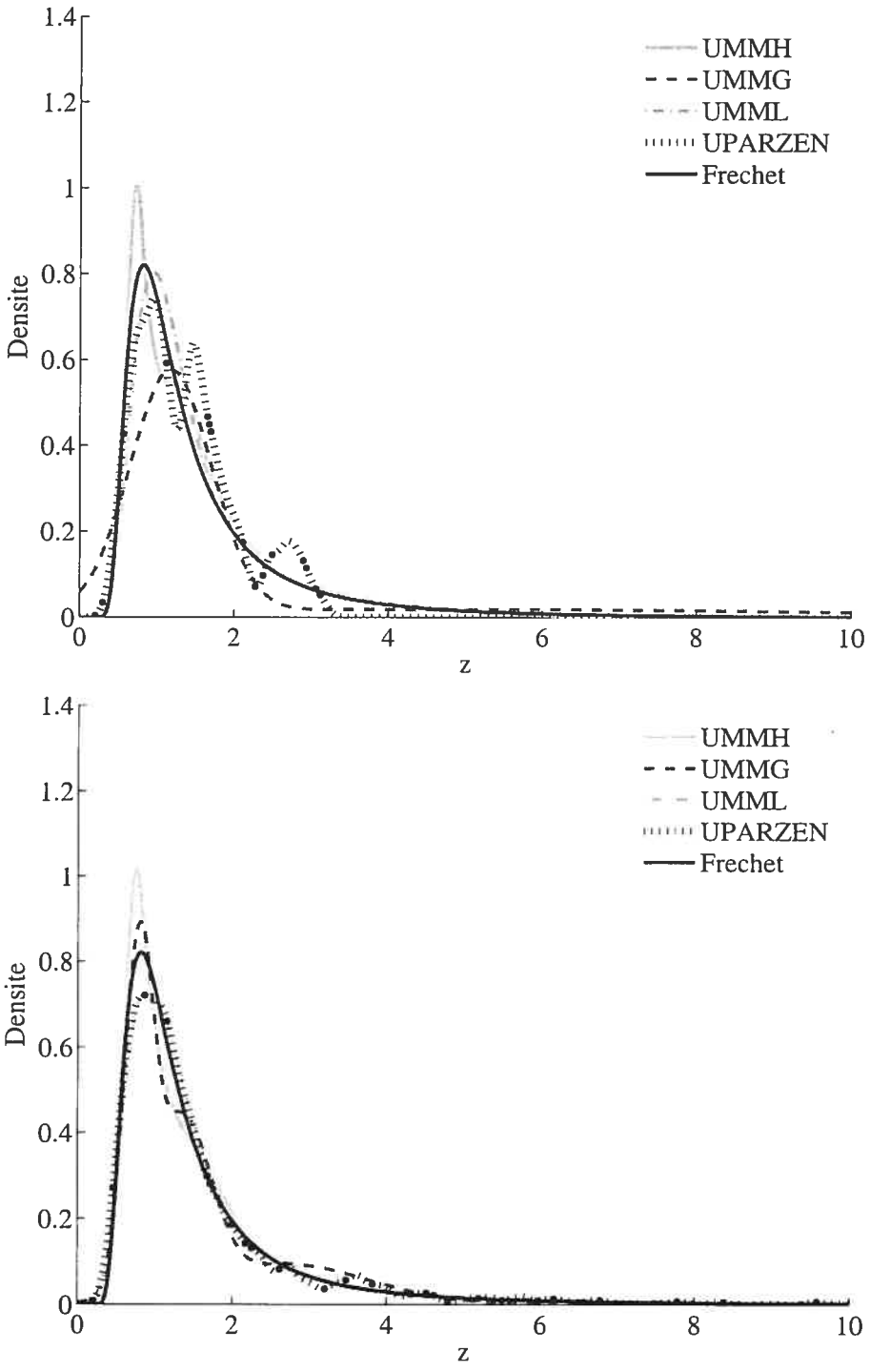

FIG. 3.10: *Densité estimée dans la partie centrale (99% de l'ensemble d'entraînement) pour les données Fréchet avec* $\xi = 0.5$. *Ensemble d'entraînement de 100 observations dans le panneau du haut et de 1000 observations dans le panneau du bas.*

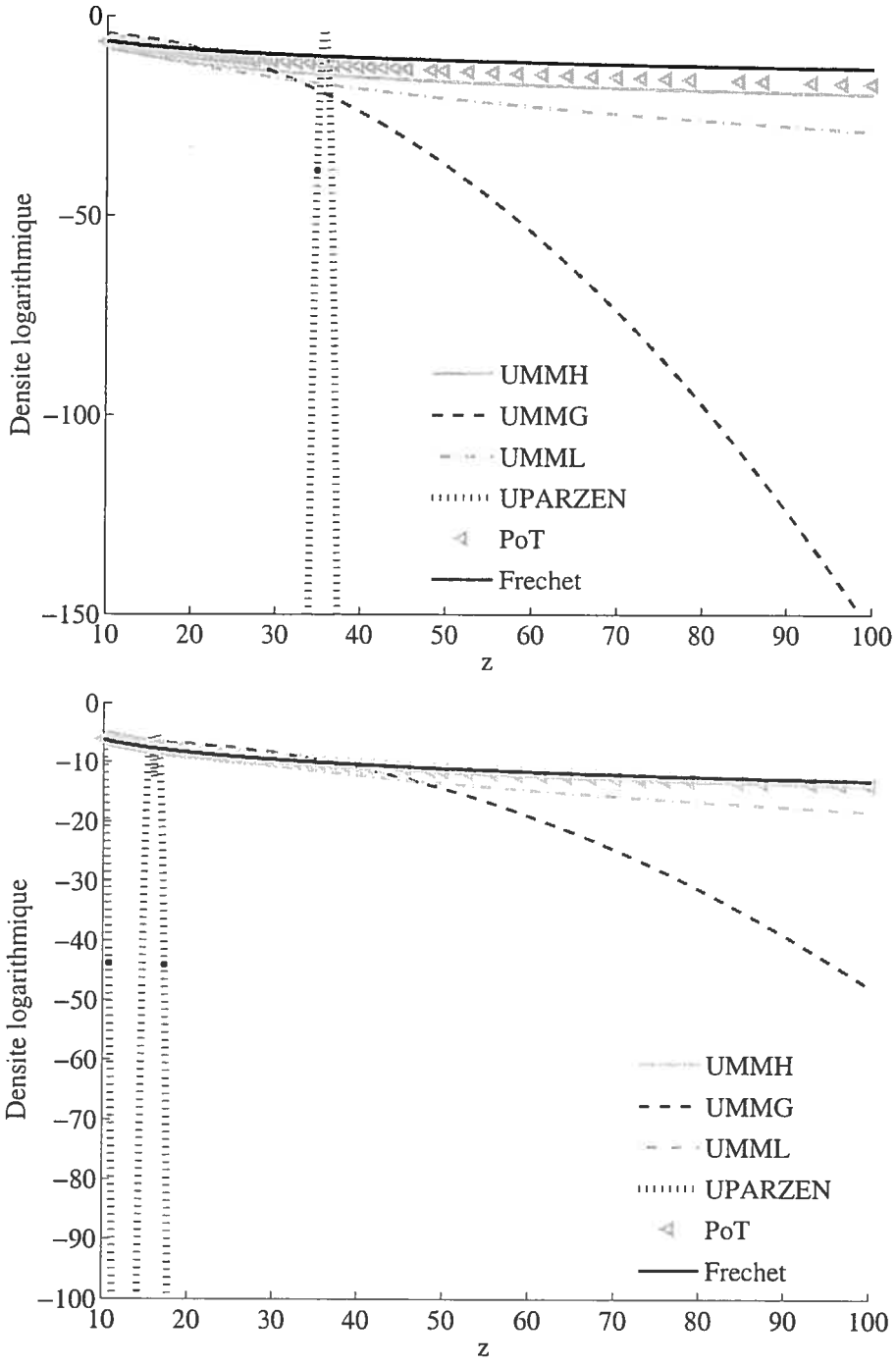

FIG. 3.11: *Densité estimée dans la queue supérieure de la distribution (moins de 1% de l'ensemble d'entraînement) pour les données Fréchet avec* $\xi = 0.5$. *Ensemble d'entraînement de 100 observations dans le panneau du haut et de 1000 observations dans le panneau du bas.*

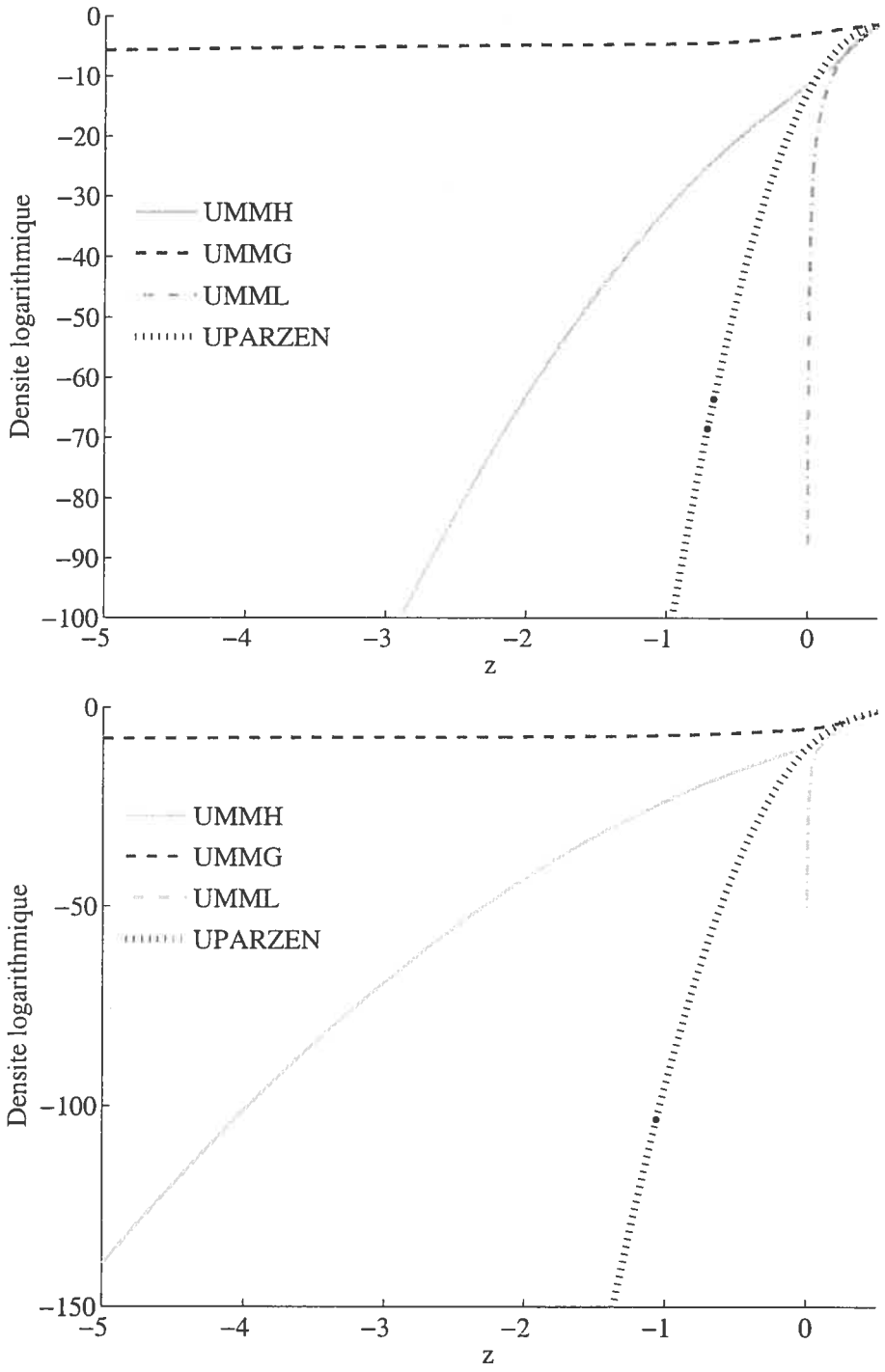

FIG. 3.12: *Densité estimée dans la queue inférieure de la distribution (pas d'observations) pour les données Fréchet avec* $\xi = 0.5$. *Ensemble d'entraînement de 100 observations dans le panneau du haut et de 1000 observations dans le panneau du bas.*

3.3.2.3. *Quantiles et indices de queue estimés*

Pour les données Fréchet d'indice de queue $\xi = 0.2$, les figures 3.13, 3.14 et 3.15 représentent les quantiles estimés standardisés pour les niveaux $q = 0.99$, $q = 0.999$ et $q = 0.9999$ respectivement. La *RMSE* correspondant à ces quantiles est donnée dans les tableaux 3.5, 3.6 et 3.7. Pour les données Fréchet d'indice de queue $\xi = 0.5$, les quantiles estimés standardisés pour les trois niveaux sont représentés dans les figures 3.16, 3.17 et 3.18 et la *RMSE* correspondante est fournie dans les tableaux 3.8, 3.9 et 3.10. Pour les petits jeux de données (moins de 1000 observations) et en particulier lorsque la queue de la distribution est plus lourde ($\xi = 0.5$), le mélange de Pareto hybrides donne des estimations de quantiles dont la variance, et par conséquent la *RMSE*, est très élevée. Ceci s'observe surtout pour les niveaux de quantiles plus élevés. Peu de composantes sont choisies pour constituer le mélange et celles-ci sont centrées là où la majeure partie des données se trouvent. Pour tenir compte des observations extrêmes présentes, l'indice de queue d'une des composantes doit être très grand. Les quantiles extrêmes se trouvent alors largement sur-estimés. Ceci pourrait expliquer, du moins en partie, le fait que la *RMSE* oscille plutôt que de décroître régulièrement. En termes de log-vraisemblance, le critère qui est maximisé à l'entraînement, le mélange de Pareto hybrides produit de meilleurs résultats que les autres modèles dans tous les cas. On constate donc que dans certains cas, le critère de maximum de vraisemblance ne fournit pas de bons estimateurs de quantiles extrêmes. Cependant, pour les jeux de données contenant au moins 1000 observations, le mélange de Pareto hybrides fournit des estimateurs de quantiles qui sont généralement plus précis que ceux fournis par les autres méthodes, incluant la méthode PoT. Lorsque la taille de l'ensemble d'entraînement augmente, les mélanges de Gaussiennes et de Log-Normales fournissent des estimateurs de quantiles de plus en plus comparables à ceux du mélange de Pareto hybrides. Pour ce qui est de la méthode PoT, la précision en terme de *RMSE* est raisonnablement bonne pour les données Fréchet d'indice de queue $\xi = 0.2$. La précision se dégrade cependant lorsque le niveau de quantile augmente, ce qui se remarque particulièrement pour les données Fréchet $\xi = 0.5$. Il serait sans doute possible d'ajuster la sélection du seuil pour obtenir une

meilleure performance de la méthode PoT. Nous nous sommes cependant limités à une méthode automatique simple à mettre en oeuvre qui consiste à déterminer le seuil à l'aide d'un test d'adéquation de la Pareto généralisée. L'estimateur de la fenêtre de Parzen fournit des quantiles estimés raisonnables lorsque la queue n'est pas trop lourde $\xi = 0.2$ et le niveau de quantile pas trop élevé $q = 0.99$. Autrement, la *RMSE* peut être très grande.

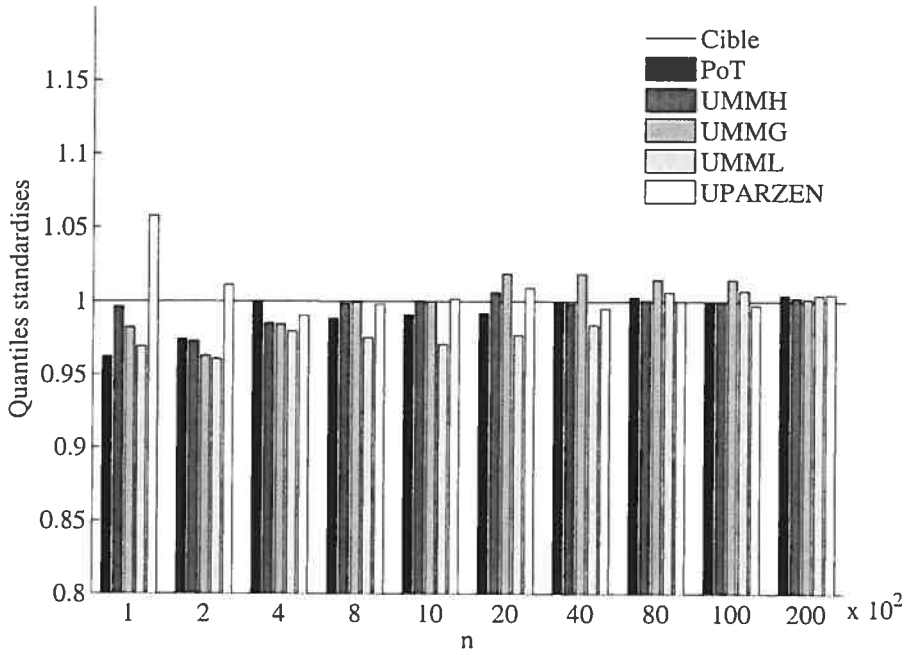

FIG. 3.13: *Pour différentes tailles d'ensemble d'entraînement n : moyenne des quantiles estimés standardisés de niveau 0.99 pour les données Fréchet avec $\xi = 0.2$.*

n	PoT	ummh	ummg	umml	uparzen
100	0.29	0.18	0.28	0.29	0.68
200	0.21	0.11	0.15	0.14	0.35
400	0.19	0.081	0.11	0.11	0.24
800	0.14	0.059	0.091	0.076	0.18
1000	0.13	0.049	0.071	0.077	0.16
2000	0.084	0.046	0.074	0.052	0.1
4000	0.057	0.031	0.055	0.041	0.077
8000	0.048	0.017	0.043	0.032	0.053
10000	0.033	0.017	0.032	0.028	0.044
20000	0.031	0.013	0.024	0.021	0.038

TAB. 3.5: *RMSE correspondant aux quantiles estimés de la figure 3.13. Le quantile de la distribution génératrice est $z_{0.99} = 2.5094$.*

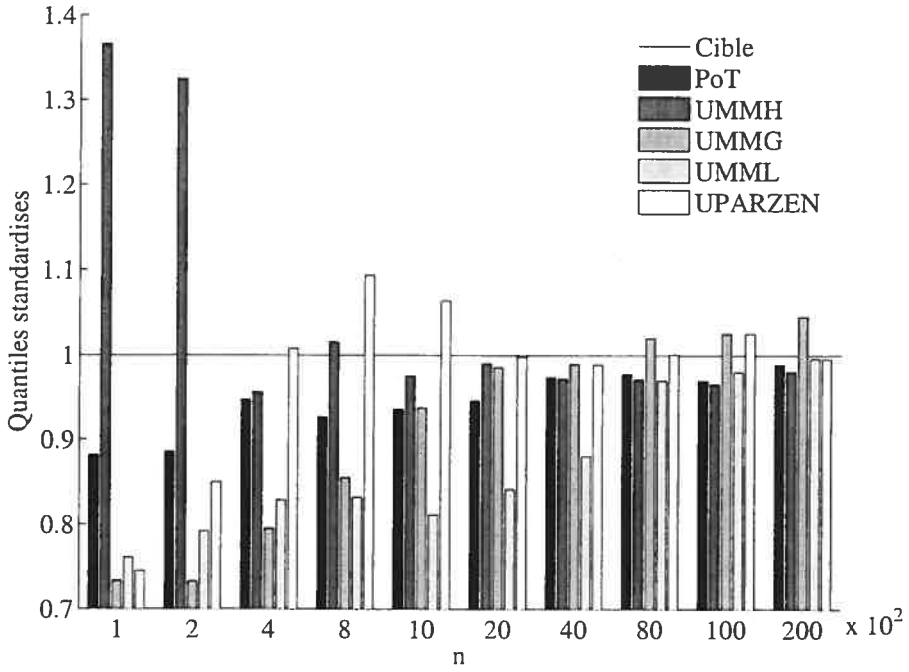

FIG. 3.14: *Pour différentes tailles d'ensemble d'entraînement n : moyenne des quantiles estimés standardisés de niveau 0.999 pour les données Fréchet avec $\xi = 0.2$.*

n	PoT	ummh	ummg	umml	uparzen
100	0.9	2	0.36	0.35	1.1
200	0.7	2.4	0.31	0.29	1.3
400	0.75	0.18	0.3	0.24	1.6
800	0.54	0.28	0.25	0.24	1.1
1000	0.52	0.13	0.27	0.22	1.2
2000	0.36	0.13	0.19	0.19	0.48
4000	0.25	0.084	0.14	0.16	0.32
8000	0.18	0.058	0.097	0.087	0.3
10000	0.15	0.063	0.11	0.089	0.36
20000	0.14	0.043	0.095	0.063	0.18

TAB. 3.6: *RMSE correspondant aux quantiles estimés de la figure 3.14. Le quantile de la distribution génératrice est* $z_{0.999} = 3.9807$.

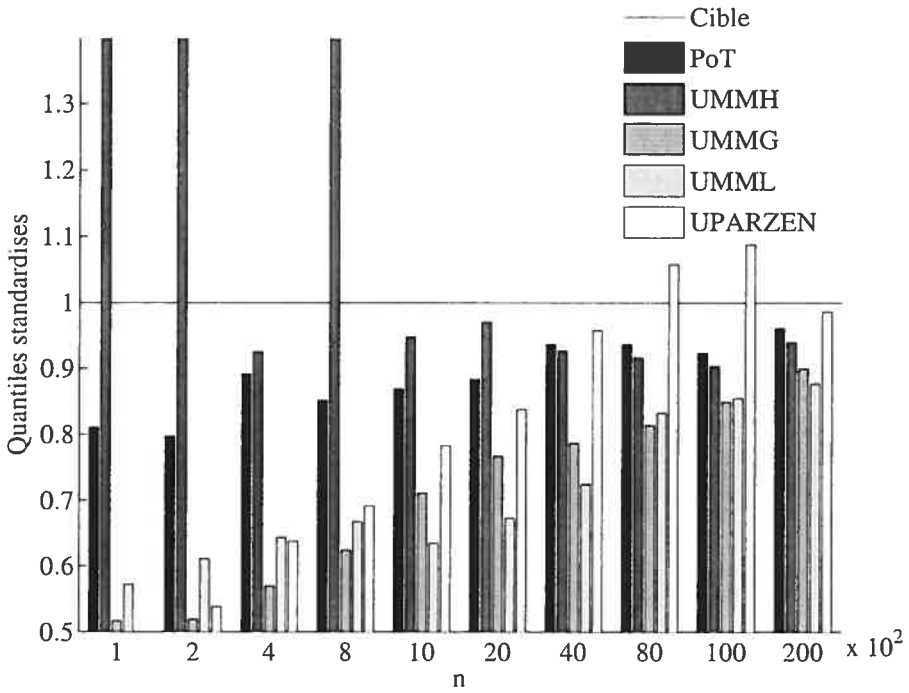

FIG. 3.15: *Pour différentes tailles d'ensemble d'entraînement n : moyenne des quantiles estimés standardisés de niveau 0.9999 pour les données Fréchet avec $\xi = 0.2$.*

n	PoT	ummh	ummg	umml	uparzen
100	2.9	1.1e+03	0.52	0.48	1.2
200	2.3	5.5e+02	0.5	0.44	1.3
400	2.3	0.36	0.47	0.38	1.6
800	1.5	7.6	0.42	0.41	1.1
1000	1.5	0.25	0.4	0.38	1.9
2000	0.99	0.27	0.32	0.35	1.5
4000	0.83	0.17	0.27	0.31	1.7
8000	0.5	0.13	0.22	0.2	1.5
10000	0.47	0.13	0.22	0.19	1.5
20000	0.44	0.11	0.17	0.16	1

TAB. 3.7: *RMSE correspondant aux quantiles estimés de la figure 3.15. Le quantile de la distribution génératrice est* $z_{0.9999} = 6.3095$.

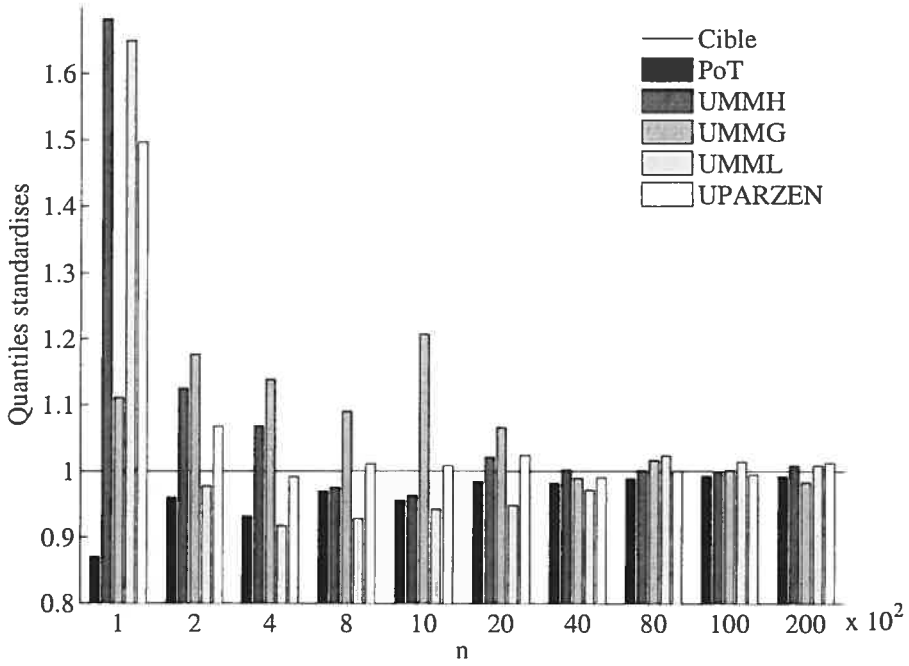

FIG. 3.16: *Pour différentes tailles d'ensemble d'entraînement n : moyenne des quantiles estimés standardisés de niveau 0.99 pour les données Fréchet avec $\xi = 0.5$.*

n	PoT	ummh	ummg	umml	uparzen
100	2.5	3.2	0.92	6	15
200	2.7	1.4	0.73	0.39	4
400	1.8	0.35	0.48	0.23	2.6
800	1.3	0.16	0.31	0.18	1.9
1000	1.1	0.13	0.75	0.15	1.6
2000	1.2	0.11	0.39	0.14	1.1
4000	0.57	0.07	0.14	0.11	0.75
8000	0.41	0.051	0.34	0.069	0.53
10000	0.6	0.042	0.09	0.067	0.44
20000	0.3	0.038	0.068	0.055	0.38

TAB. 3.8: *RMSE correspondant aux quantiles estimés de la figure 3.16. Le quantile de la distribution génératrice est $z_{0.99} = 9.9749$.*

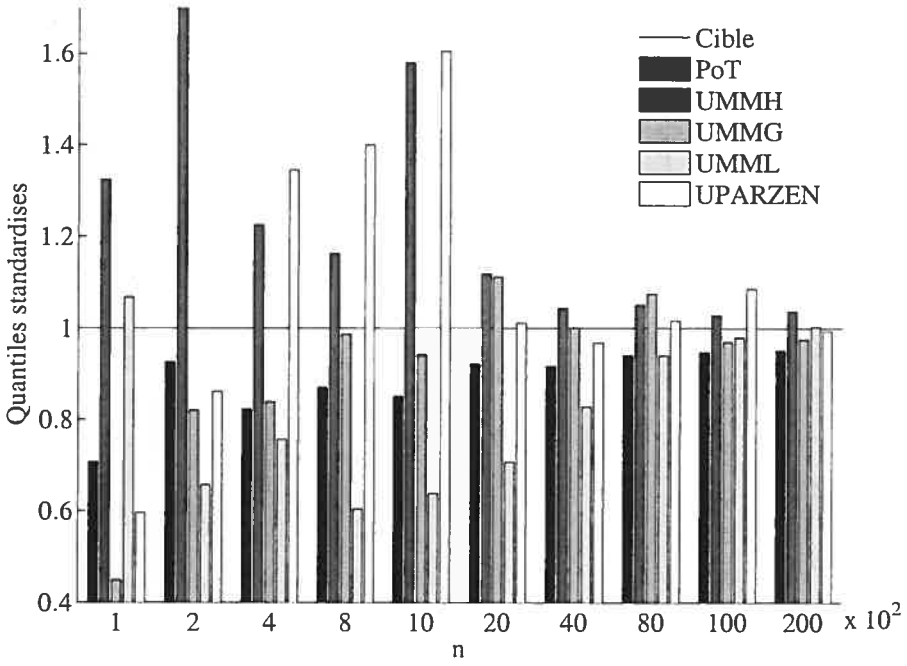

FIG. 3.17: *Pour différentes tailles d'ensemble d'entraînement n : moyenne des quantiles estimés standardisés de niveau 0.999 pour les données Fréchet avec $\xi = 0.5$.*

n	PoT	ummh	ummg	umml	uparzen
100	14	1.9	0.68	2.4	21
200	22	5.6	2.2	0.73	46
400	12	0.97	0.81	1.7	64
800	7.7	1	0.96	0.46	32
1000	7.3	3.2	0.79	0.45	75
2000	8.6	0.35	0.98	0.43	12
4000	4.1	0.24	0.42	0.33	6.4
8000	3.2	0.19	0.32	0.21	6.1
10000	4.1	0.14	0.26	0.21	7.6
20000	2.4	0.11	0.29	0.16	3.7

TAB. 3.9: *RMSE correspondant aux quantiles estimés de la figure 3.17. Le quantile de la distribution génératrice est $z_{0.999} = 31.6149$.*

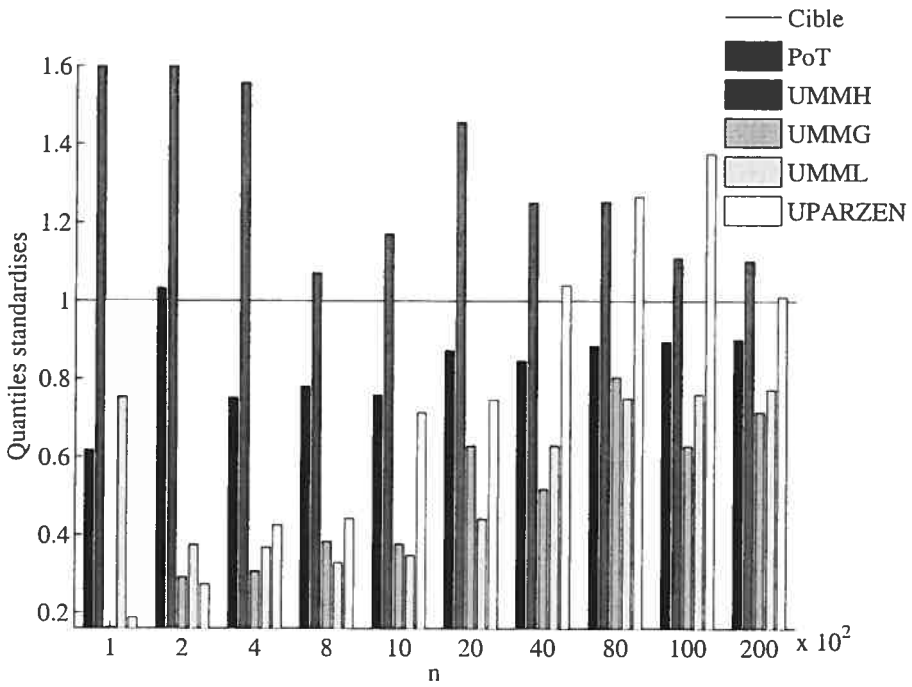

FIG. 3.18: *Pour différentes tailles d'ensemble d'entraînement n : moyenne des quantiles estimés standardisés de niveau 0.9999 pour les données Fréchet avec $\xi = 0.5$.*

n	PoT	ummh	ummg	umml	uparzen
100	76	21	0.85	2.1	21
200	1.9e+02	7.1e+02	0.99	0.81	46
400	61	3.1	0.74	0.83	64
800	36	0.92	0.69	0.69	32
1000	35	1.2	0.68	0.69	1.2e+02
2000	50	1.1	0.81	0.63	65
4000	21	0.77	0.61	0.62	1.1e+02
8000	17	0.65	0.56	0.64	78
10000	21	0.43	0.45	0.39	87
20000	12	0.36	0.61	0.36	39

TAB. 3.10: *RMSE correspondant aux quantiles estimés de la figure 3.18. Le quantile de la distribution génératrice est $z_{0.9999} = 99.9975$.*

Les estimateurs d'indices de queues provenant du mélange de Pareto hybrides et de la méthode PoT sont donnés aux figures 3.19 et 3.20 pour les données Fréchet avec $\xi = 0.2$ et $\xi = 0.5$ respectivement. On trouve aussi dans les panneaux du bas des figures la *RMSE* correspondant aux estimateurs. En général, l'estimateur de l'indice de queue provenant du mélange d'hybrides est environ équivalent à celui de la méthode PoT. En particulier lorsque $\xi = 0.2$, l'estimateur du mélange d'hybrides a tendance a être plus grand, pour les raisons énumérées lors de la discussion sur l'estimation de quantiles ci-dessus. Aussi, la *RMSE* de l'estimateur du mélange d'hybrides est généralement plus grande que celle de l'estimateur de la méthode PoT. Ceci est en grande partie dû au facteur de variance de la *RMSE*. En particulier pour les données Fréchet avec $\xi = 0.2$, on n'observe pas de convergence claire de l'estimation de l'indice de queue. Celui-ci semble être sous-estimé ; l'estimation, tant du mélange d'hybrides que de la méthode PoT, le situe plus près de 0.15 que de 0.2, la vraie valeur. Par ailleurs, la *RMSE* ne semble pas décroître vers zéro mais plutôt tendre vers une valeur asymptotique positive. Étant donné que la Pareto généralisée, et donc la Pareto hybride, est une approximation de la fonction de répartition excédentaire, il existe de nombreux cas où les estimateurs de maximum de vraisemblance sont biaisés, voir Smith [**44**]. L'estimation de l'indice de queue sert à déterminer l'épaisseur de la queue de la distribution sous-jacente. Elle est aussi centrale, dans la méthode PoT, pour obtenir des estimateurs de la queue de la distribution et des quantiles extrêmes. Pour le mélange de Pareto hybrides, une estimation précise de l'indice de queue de la distribution sous-jacente n'est pas si importante. En effet, l'estimation de la queue de la distribution et donc des quantiles extrêmes, dépend de toutes les composantes du mélanges, pas seulement de la composante dominante (composante qui détermine le seuil implicite et l'indice de queue du mélange, voir la sous-section 3.2.1). En ce qui a trait à l'estimation de quantiles extrêmes, nous avons vu que le mélange de Pareto hybrides fournit dans presque tous les cas des estimateurs plus précis que ceux de la méthode PoT. Par ailleurs, lorsque l'on compare la densité estimée dans la queue supérieure, les deux méthodes fournissent toutes deux de

bons résultats. Le mélange de Pareto hybrides est donc une alternative valide à la méthode PoT en ce qui concerne l'estimation de la queue de la distribution.

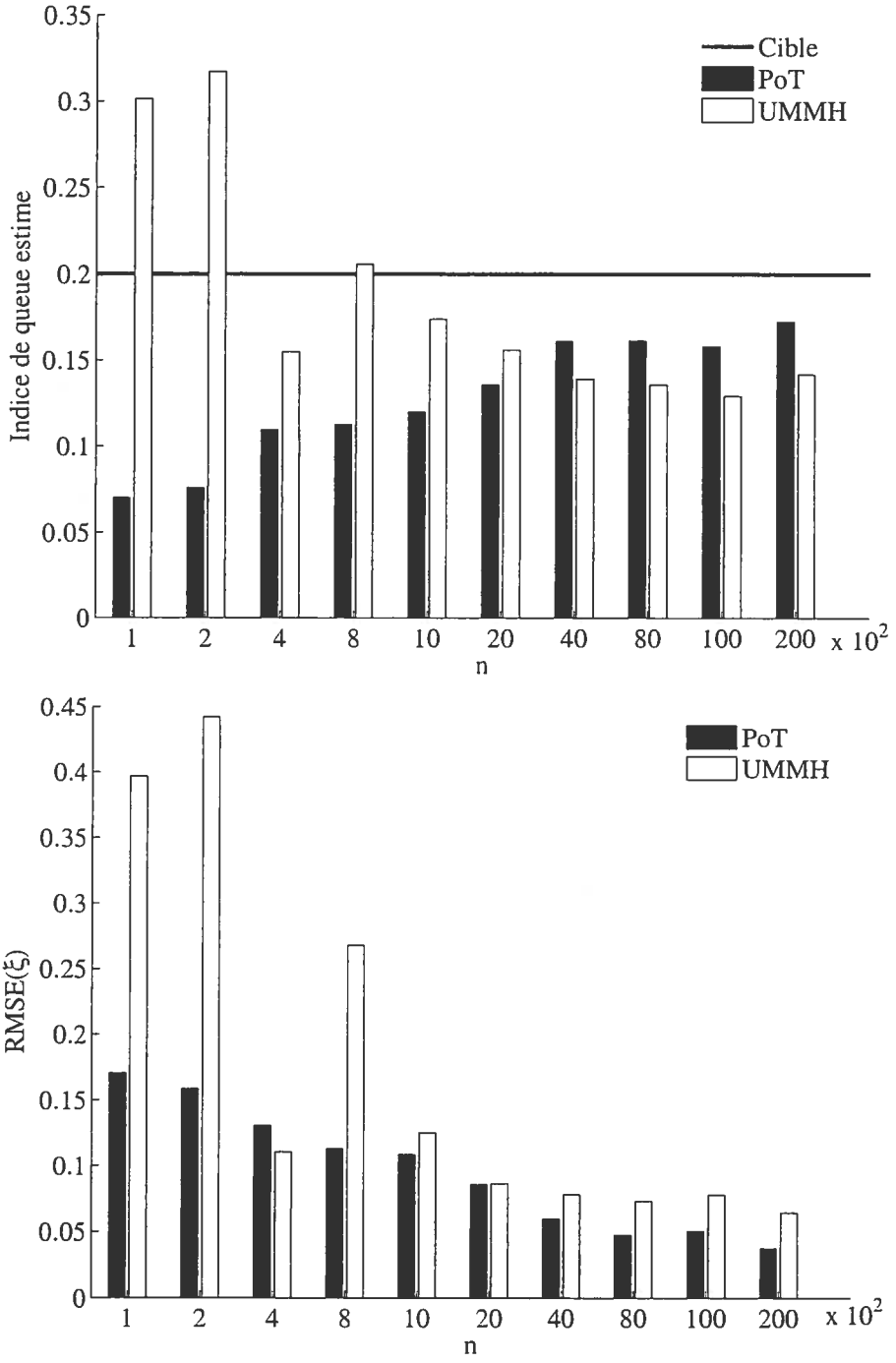

FIG. 3.19: *Indice de queue estimé dans le panneau du haut et RMSE correspondante dans le panneau du bas. Les données sont générées par la loi de Fréchet d'indice de queue* $\xi = 0.2$.

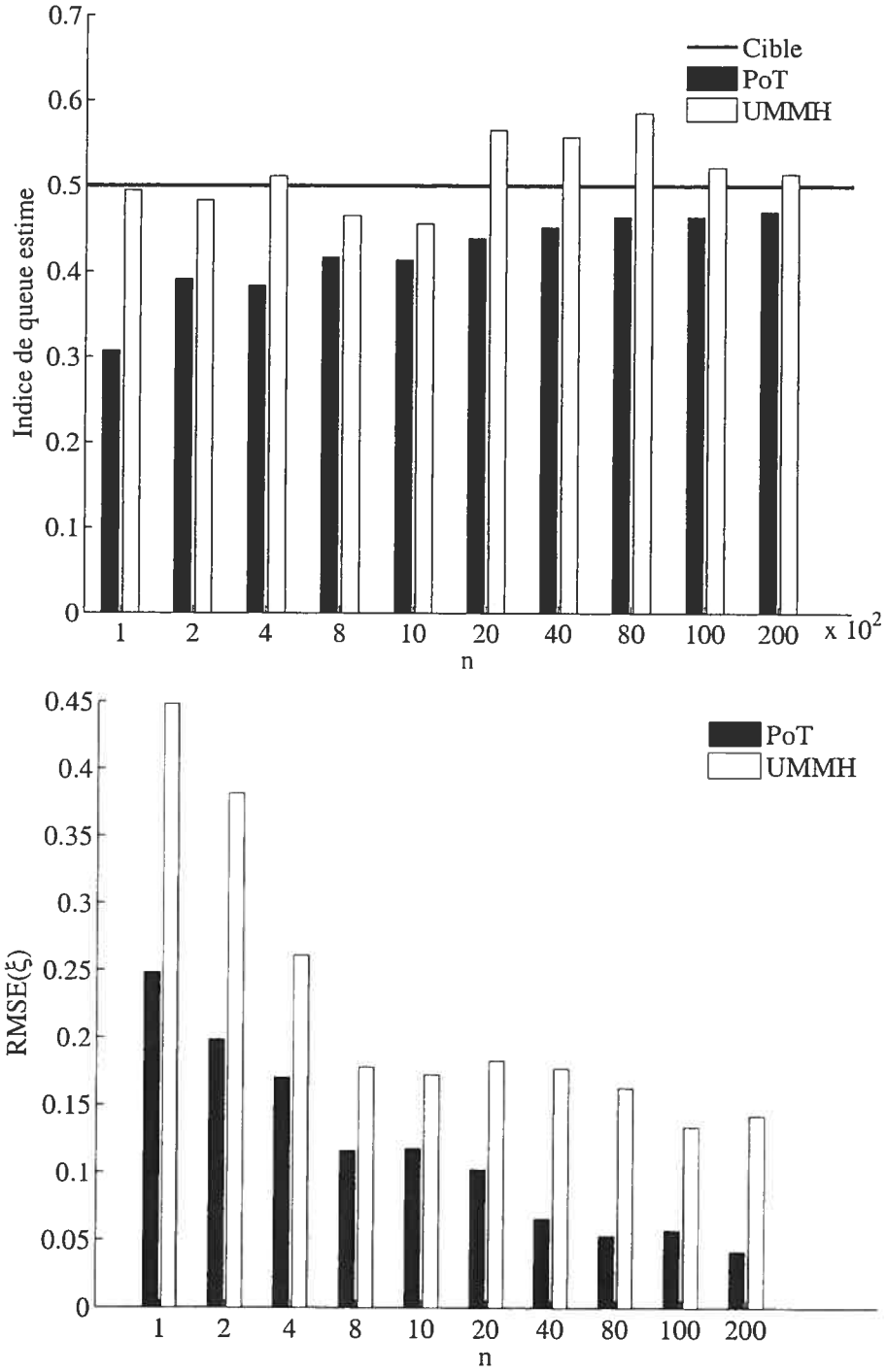

FIG. 3.20: *Indice de queue estimé dans le panneau du haut et RMSE correspondante dans le panneau du bas. Les données sont générées par la loi de Fréchet d'indice de queue* $\xi = 0.5$.

3.4. RÉCLAMATIONS DANOISES

Nous avons ensuite comparé le mélange de Pareto hybrides sur un jeu de données réelles. Il s'agit des réclamations danoises d'assurance contre le feu qui sont disponibles avec le logiciel *R*. Ces données consistent en une série chronologique irrégulière de 2167 observations que nous considérerons comme indépendantes. McNeil [**34**] a utilisé ce jeu de données pour illustrer la méthodologie PoT. Un histogramme du logarithme des réclamations danoises est tracé à la figure 3.21.

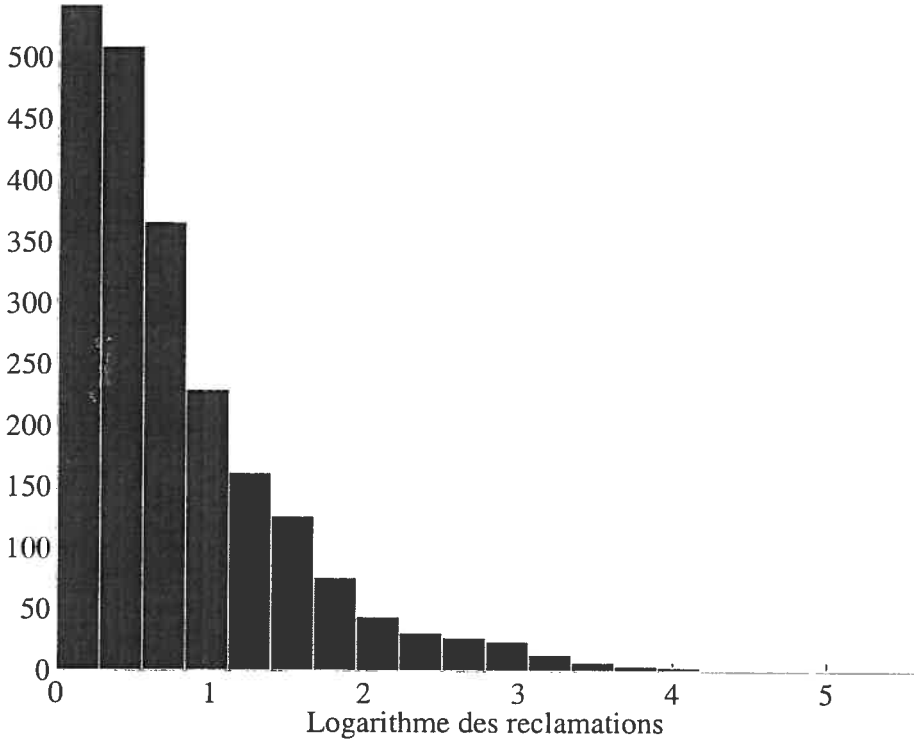

FIG. 3.21: *Histogramme du logarithme des réclamations pour les données danoises d'assurance contre le feu.*

3.4.1. Entraînement et critères d'évaluation

Tout comme dans l'étude simulatoire de la section précédente, nous allons comparer le mélange de Pareto hybrides en tant qu'estimateur de densité globale et en tant qu'estimateur de la queue de la distribution. Comme estimateurs de densité alternatifs, nous utiliserons à nouveau les mélanges de Gaussiennes et de Log-Normales ainsi que l'estimateur de la fenêtre de Parzen. Puisque la distribution génératrice des données est inconnue, nous mesurons la performance en

termes de log-vraisemblance relative au mélange de Pareto hybrides sur un ensemble de test. Soit z, un point de l'ensemble de test, ce critère de performance s'écrit :

$$\mathcal{R}(z) = \log(\phi_\theta^{\text{ummh}}(z)) - \log(\phi_\theta^{\text{alt}}(z)),$$

où $\phi_\theta^{\text{ummh}}$ est l'estimateur du mélange de Pareto hybrides et ϕ_θ^{alt} est un estimateur alternatif. Nous comparerons également ces estimateurs de densité globaux en termes d'estimation de quantiles extrêmes. Dans ce cas-ci, nous ne pouvons pas calculer l'erreur quadratique moyenne puisque les vrais quantiles ne sont pas connus. Nous utilisons plutôt un test binomial développé dans McNeil et Frey [**35**]. Soit $1_{\{Z>z_q\}}$, la fonction indicatrice de l'événement $Z > z_q$, une violation du quantile z_q. Alors $1_{\{Z>z_q\}}$ est une variable aléatoire de loi Bernouilli $Be(1-q)$. Soit $\mathcal{D}_l = \{Z_1, \ldots, Z_l\}$ un ensemble de test dont les observations sont indépendantes et identiquement distribuées. Par conséquent, la variable aléatoire $\sum_{i=1}^{l} 1_{\{Z_i>z_q\}}$ suit une loi Binômiale $B(l, 1-q)$. On formule l'hypothèse nulle suivante : si la méthode proposée estime correctement les quantiles, alors la statistique empirique $\sum_{i=1}^{l} 1_{\{Z_i>\hat{z}_q\}}$ suit aussi la loi Binômiale $B(l, 1-q)$. Nous avons donc testé cette hypothèse nulle pour tous les modèles et tous les niveaux de quantiles estimés contre l'hypothèse alternative que le modèle ait un biais systématique, c'est-à-dire qu'elle produise trop ou trop peu de violations. Nous utilisons le test binômial du logiciel gratuit R qui en propose une version exacte.

Finalement, nous comparerons les performances du mélange de Pareto hybrides en tant qu'estimateur de la queue de la distribution par rapport à la méthode PoT. Nous calculons, pour les deux méthodes, les quantiles extrêmes auxquels nous appliquons le test binômial et les estimateurs d'indice de queue que nous comparons à ceux obtenus par McNeil [**34**] .

Plusieurs tailles d'ensembles d'entraînement sont utilisées afin d'étudier le comportement des estimateurs ; elles correspondent à 50%, 75% et 90% du jeu de données original. Nous utilisons la même procédure pour la sélection des hyperparamètres que dans l'étude simulatoire. Le niveau de quantile q_{PoT} qui détermine le seuil de la méthode PoT est choisi à l'aide du test d'adéquation de la Pareto généralisée de Choulakian et Stephens [**6**]. Le nombre de composantes des mélanges

et la largeur de fenêtre de l'estimateur de la fenêtre de Parzen sont choisis sur un ensemble de validation constitué de 20% de l'ensemble d'entraînement. Nous utilisons un t-test unilatéral de niveau 5% pour mesurer une différence significative de performance entre deux modèles.

3.4.2. Résultats des expériences

Le tableau 3.11 donne la log-vraisemblance relative au mélange de Pareto hybrides sur l'ensemble de test ainsi que l'erreur standard pour toutes les tailles d'ensemble d'entraînement. Les hyper-paramètres sélectionnés se trouvent dans le tableau 3.12. Toutes les entrées du tableau 3.11 des colonnes correspondant aux mélanges de Gaussiennes et de Log-Normales sont significativement positives. Ceci signifie que le mélange de Pareto hybrides donne de meilleurs résultats en termes de log-vraisemblance que les autres types de mélanges. La largeur de fenêtre de l'estimateur de la fenêtre de Parzen est choisie très petite, ce qui rend la densité de ce modèle très irrégulière car un noyau très étroit est centré sur chaque observation. La vraisemblance est très grande autour des points de l'ensemble d'entraînement et négligeable ailleurs. Sur l'ensemble de test, la performance de l'estimateur de la fenêtre de Parzen est donc mauvaise et extrêmement variable.

n	ummg	umml	uparzen
1084	0.075 (0.0098)	0.038 (0.01)	6.4e+20 (4.5e+20)
1625	0.13 (0.011)	0.12 (0.012)	1.3e+21 (5.2e+20)
1950	0.13 (0.01)	0.13 (0.012)	1.4e+19 (4.6e+18)

TAB. 3.11: *Log-vraisemblance moyenne relative au mélange de Pareto hybrides (ummh) sur l'ensemble de test pour les données danoises d'assurance contre le feu avec taille d'ensemble d'entraînement n. Lorsque la log-vraisemblance relative est positive, cela signifie que le mélange d'hybrides donne de meilleurs résultats que le modèle alternatif envisagé.*

McNeil [**34**] a ajusté la Pareto généralisée à différents niveaux de seuils allant de 3.14 à 22 pour un total de 30 modèles différents. McNeil mentionne également que d'après les analyses exploratives, la Pareto généralisée pourrait être ajustée à l'ensemble des données. Plus le seuil est bas, plus les quantiles estimés sont

n	m_{ummh}	m_{ummg}	m_{umml}	σ	u	q
1084	2	4	2	1e-10	1.4694	0.3
1625	2	4	2	1e-10	1.4417	0.3
1950	2	4	2	1e-10	1.4201	0.3

TAB. 3.12: *Hyper-paramètres choisis en validation pour les données danoises d'assurance contre le feu avec taille d'ensemble d'entraînement n.*

grands. Le méthode de sélection du seuil que nous utilisons, basée sur le test d'adéquation de la Pareto généralisée de Choulakian et Stephens [6] fournit des niveaux de seuil plutôt bas comme on peut le voir dans le tableau 3.12. Cette méthode sélectionne le plus bas niveau de seuil tel que le test d'adéquation soit satisfait.

Les figures 3.22, 3.23 et 3.24 illustrent les quantiles estimés par les cinq méthodes pour les niveaux de quantiles $q = 0.99$, $q = 0.999$ et $q = 0.9999$ respectivement. Les estimateurs de l'indice de queue pour le mélange de Pareto hybrides et la méthode PoT sont illustrés à la figure 3.25. McNeil fournit des estimations de l'indice de queue et des quantiles de niveau $q = 0.999$ et $q = 0.9999$ calculées sur le jeu de données original pour différentes valeurs du seuil de le méthode PoT. L'estimateur de l'indice de queue va de 0.5 à 0.72, celui du quantile de niveau $q = 0.999$ varie entre 95 et 147 et celui du quantile de niveau $q = 0.9999$ se trouve entre 306 et 770. Nos estimations basées sur la méthode PoT se trouvent sur la limite supérieure de l'intervalle de valeurs trouvées par McNeil car le seuil sélectionné est relativement bas. Le mélange de Pareto hybrides fournit des estimations de l'indice de queue un peu plus élevées que celles de la méthode PoT, ce que l'on observait aussi sur les jeux de données synthétiques. Les quantiles estimés sont donc également plus grands. Les estimations de quantiles des mélanges de Gaussiennes et de Log-Normales et de l'estimateur de la fenêtre de Parzen sont plus petits et donc plus conservateurs.

Les résultats du test binômial, c'est-à-dire la valeur P et l'intervalle de confiance de niveau 95%, pour les quantiles estimés se trouvent dans le tableau 3.13. Pour accepter l'hypothèse nulle que le modèle estime correctement les quantiles, il faut que la valeur P dépasse un certain seuil, que l'on choisit typiquement à 5%. On

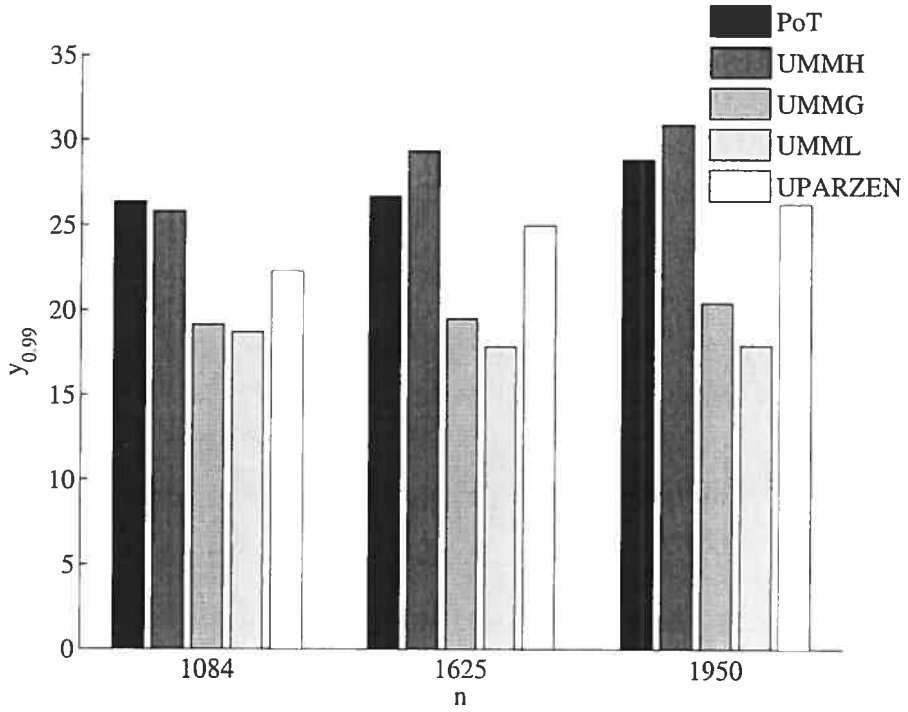

FIG. 3.22: *Données danoises d'assurance contre le feu : estimation des quantiles de niveau* 0.99 *avec taille d'ensemble d'entraînement* n.

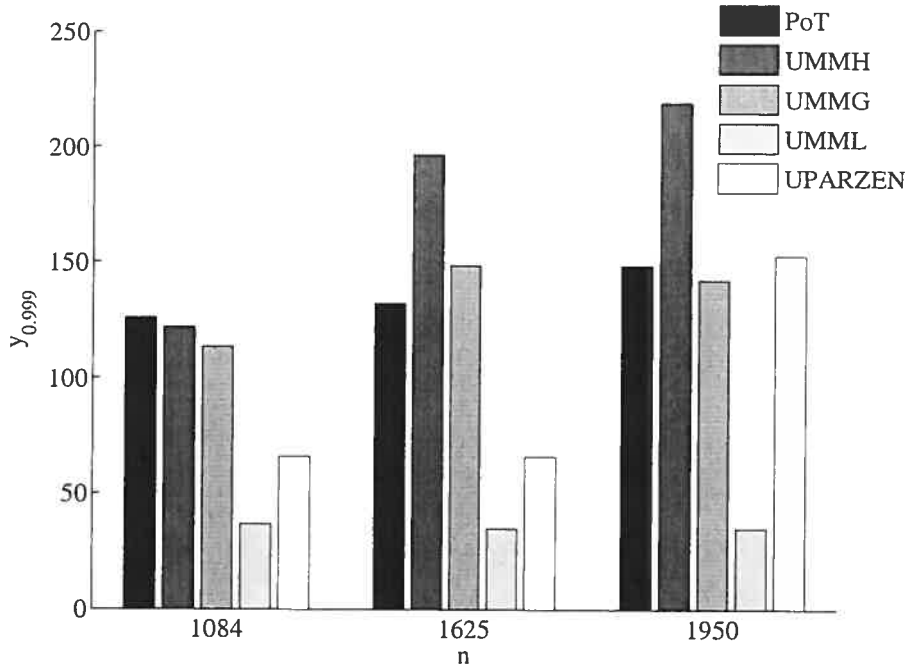

FIG. 3.23: *Données danoises d'assurance contre le feu : estimation des quantiles de niveau* 0.999 *avec taille d'ensemble d'entraînement* n.

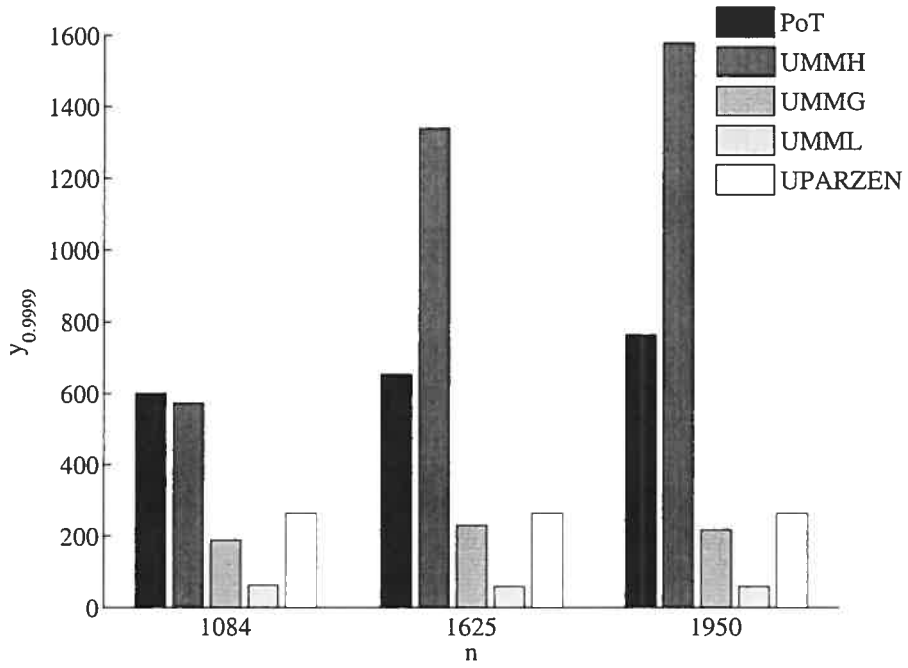

FIG. 3.24: *Données danoises d'assurance contre le feu : estimation des quantiles de niveau* 0.9999 *avec taille d'ensemble d'entraînement* n.

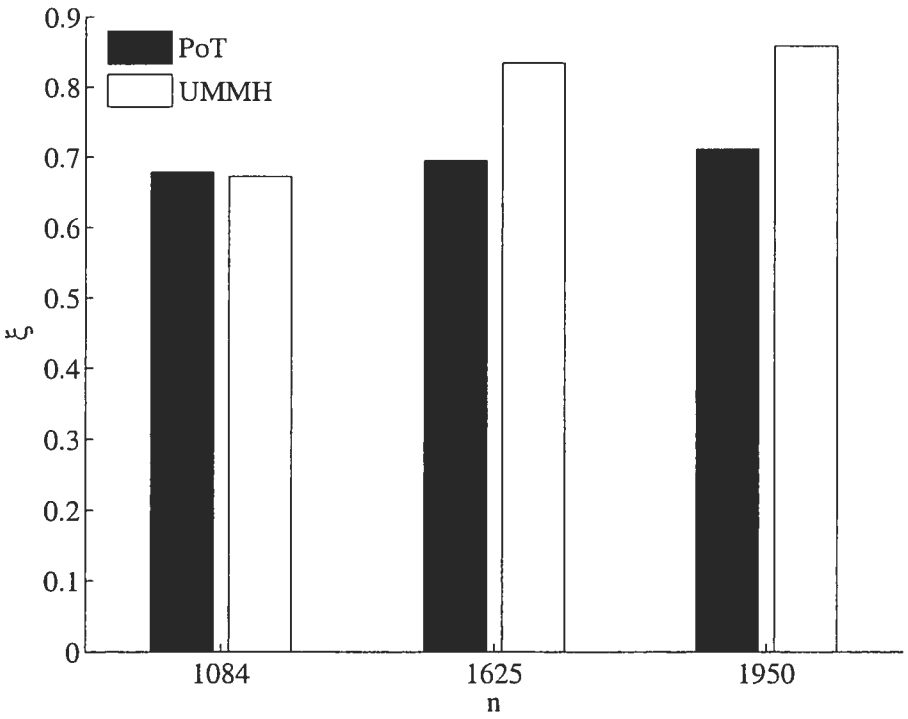

FIG. 3.25: *Données danoises d'assurance contre le feu : estimation de l'indice de queue avec taille d'ensemble d'entraînement* n.

s'attend également à ce que la vraie proportion de violations $1 - q$ soit comprise dans l'intervalle de confiance. Lorsqu'aucune violation ne se produit, ce qui est le cas lorsque le niveau de quantile est très grand ou lorsque l'ensemble de test est très petit, on ne peut pas vraiment tirer de conclusions de ce test. En effet, la valeur P est 1 et l'intervalle de confiance est très large. L'hypothèse nulle ne peut donc pas être rejetée mais le test ne semble pas très significatif. Cependant, lorsque des violations surviennent, le test binômial ne rejette jamais ni le mélange Pareto hybride ni la méthode PoT. Nous avons marqué en caractères gras dans le tableau 3.13 les cas où le modèle sur-estime le nombre de violations, ce qui équivaut à la sous-estimation du quantile associé. Les estimations du mélange de Gaussiennes sont rejetées par le test binômial en deux endroits. Celles du mélange de Log-Normales sont rejetées en plusieurs endroits alors que l'estimateur de la fenêtre de Parzen donne de bons résultats selon ce test.

n_{test}	$1-q=0.01$	$1-q=0.001$	$1-q=0.0001$
ummh			
1083	0.3554 (0.007 0.0216)	0.2948 (0.0002 0.0067)	1 (0.00 0.0034)
542	0.6672 (0.0041 0.0239)	1 (0.00 0.0068)	1 (0.00 0.0068)
217	0.73 (0.0001 0.0254)	1 (0.00 0.0169)	1 (0.00 0.0169)
ummg			
1083	**0.0020 (0.0128 0.0306)**	0.2948 (0.0002 0.0067)	1 (0.00 0.0034)
542	**0.0038 (0.0128 0.041)**	0.4186 (0.00 0.0010)	1 (0.00 0.0068)
217	0.4824 (0.0029 0.0399)	0.1952 (0.0001 0.0254)	1 (0.00 0.0169)
umml			
1083	**0.0020 (0.0128 0.0306)**	**0.0051 (0.0015 0.0107)**	**0.005453 (0.0002 0.0067)**
542	**0.0005 (0.0156 0.0452)**	**0.0023 (0.0020 0.0188)**	**0.001414 (0.0004 0.0133)**
217	0.1740 (0.0050 0.0465)	0.1952 (0.0001 0.0254)	0.02147 (0.0001 0.0254)

suite à la page suivante

suite de la page précédente			
n_{test}	$1-q=0.01$	$1-q=0.001$	$1-q=0.0001$
uparzen			
1083	0.04422 (0.0099 0.0261)	0.2948 (0.0002 0.0067)	1 (0.00 0.0034)
542	0.2721 (0.0064 0.0289)	0.1031 (0.0004 0.0133)	1 (0.00 0.0068)
217	1 (0.0011 0.0329)	1 (0.00 0.0169)	1 (0.00 0.0169)
PoT			
1083	0.4454 (0.0064 0.0204)	0.2948 (0.0002 0.0067)	1 (0.00 0.0034)
542	0.5105 (0.0052 0.0264)	0.1031 (0.0004 0.0133)	1 (0.00 0.0068)
217	0.73 (0.0001 0.0254)	1 (0.00 0.0169)	1 (0.00 0.0169)

TAB. 3.13: *Valeurs P et intervalle de confiance de niveau 95% pour le test binômial sur le nombre de violations des estimateurs de quantiles. Sous l'hypothèse nulle, le nombre de violations suit une loi Binômiale $B(n_{test}, 1-q)$, où n_{test} est la taille de l'ensemble de test. Si l'hypothèse nulle est appropriée, la vraie proportion de violations $1-q$ devrait être contenue dans l'intervalle de confiance. Les modèles rejetés par ce test sont marqués en caractères gras.*

3.5. Conclusion

Les événements extrêmes se manifestent dans de nombreux contextes. Des conditions extrêmes de la mer peuvent provoquer des inondations. Les compagnies d'assurance doivent évaluer la probabilité de grandes réclamations afin de se protéger adéquatement avec une compagnie de ré-assurance. Les gestionnaires de portefeuille doivent estimer les risques reliés à de larges variations dans le prix des instruments financiers composant leur portefeuille. La théorie des valeurs extrêmes établit un cadre mathématique formel permettant de modéliser les queues de distributions, précisément la région où les événements extrêmes se produisent. La méthode PoT, "Peaks-over-Threshold", a été développée dans ce cadre. Les excédents au-delà d'un seuil sont modélisés par la Pareto généralisée.

Le mélange de Gaussiennes est un estimateur non-paramétrique fréquemment utilisé. Lorsque le nombre de composantes est choisi de façon adaptative selon la taille de l'ensemble d'entraînement, le mélange de Gaussiennes possède de bonnes propriétés de convergence. Cependant, il n'est pas en mesure de bien modéliser des distributions à queues épaisses. Ceci est particulièrement frappant pour les petits jeux de données. Afin de combiner les avantages de la méthode PoT et des mélanges de distributions, nous avons construit une nouvelle distribution, la Pareto hybride, qui juxtapose une Gaussienne avec une Pareto généralisée tout en imposant des contraintes de continuité. La Pareto hybride est une extension continue de la Pareto généralisée à l'axe des réels et peut donc être utilisée comme composante dans un mélange. Le mélange de Pareto hybrides permet de contourner le problème de sélection du seuil inhérent à la méthodologie PoT. En effet, le seuil est défini comme le point de jonction de la composante dominante. Il est donc une fonction des paramètres du mélange qui sont appris en maximisant la vraisemblance des données.

Une différence centrale entre le mélange d'hybrides et la méthode PoT est que pour le mélange, toutes les données, pas uniquement les excédents, participent à l'estimation de la queue de la distribution. Celle-ci est alors approximée par une combinaison de lois de Pareto généralisée plutôt que par une seule. Ceci permet probablement d'atténuer le biais de l'approximation de la queue de la

distribution par la Pareto généralisée. En effet, l'étude simulatoire démontre que le mélange de Pareto hybrides donne généralement des estimateurs de quantiles extrêmes plus précis que la méthode PoT. De plus, le mélange de Pareto hybrides procure une façon alternative d'estimer l'indice de queue d'une distribution. Enfin, l'estimateur non-paramétrique proposé est plus performant en termes de log-vraisemblance que d'autres estimateurs de densité lorsque la distribution sous-jacente est à queue lourde. Ceci ressort particulièrement sur les petits jeux de données.

Nous avons ensuite comparé le mélange de Pareto hybrides sur les données danoises d'assurance, un banc d'essai pour les valeurs extrêmes. Nous avons obtenu, pour la méthode PoT, des résultats semblables à ceux obtenus par McNeil [**34**]. Le mélange d'hybrides produit des estimateurs d'indices de queues un peu plus élevés que ceux de la méthode PoT et en conséquence, les estimateurs de quantiles sont également plus grands. Les autres estimateurs de densités fournissent des quantiles estimés plus petits. Le test binômial sur le nombre de violations d'un quantile estimé confirme que les quantiles estimés par le mélange de Pareto hybrides et la méthode PoT sont fiables. Il nous met en garde cependant contre les estimations de quantiles plus basses des mélanges de Gaussiennes et de Log-Normales.

L'objectif principal était de proposer un estimateur de densité global qui s'adapte au cas où la distribution sous-jacente peut être multi-modale, peut avoir une ou des queues lourdes et est asymétrique. Les expériences sur les jeux de données synthétiques et réelles ont démontré que le mélange de Pareto hybrides est un meilleur estimateur de densité en termes de log-vraisemblance que les autres estimateurs de densité testés. Bien que le cas où la distribution sous-jacente est multi-modale n'ait pas été testé dans le cadre de l'estimation de densité inconditionnelle, il le sera dans le prochain chapitre dans le cadre de l'estimation de densité conditionnelle.

Chapitre 4

MODÈLES PARETO HYBRIDES : DENSITÉ CONDITIONNELLE

Dans plusieurs situations de modélisation, il existe des variables X ayant un certain pouvoir prédictif sur Y, la variable d'intérêt. Il est alors avantageux de prédire Y conditionnellement à X. En finance, on s'intéresse le plus souvent à la densité des rendements futurs conditionnellement à l'information disponible. Celle-ci est généralement constituée des rendements passés de l'instrument financier considéré et de variables reliées à l'état de l'économie. Dans le milieu des assurances, afin de calculer la prime d'un client, on cherche à modéliser la distribution des réclamations étant donné le profil de ce client. Par exemple, dans le cas de l'assurance automobile, le profil du client contient de l'information sur le conducteur, sur l'auto et sur les options choisies par l'assuré. Une question qui survient fréquemment en climatologie est de pouvoir prédire la probabilité d'événements climatiques extrêmes, tels que des inondations, des tempêtes ou autres, étant donné que le niveau des gaz à effets de serre ait augmenté selon un certain scénario.

En régression, la fonction qui minimise la perte quadratique espérée est l'espérance conditionnelle $E[Y|X = x]$. Cependant, lorsque les données ont une structure complexe, l'espérance conditionnelle ne représente pas bien le processus générateur des données. C'est le cas pour une distribution multi-modale comme celle qui est illustrée à la figure 4.1, un exemple tiré du chapitre 6 de Bishop [**5**]. Nous nous intéressons donc à la modélisation de la distribution conditionnelle

$P(Y|X = x)$ complète. En effet, pour une application donnée, plusieurs quantités découlant de la distribution conditionnelle peuvent être d'intérêt, tels que les quantiles centraux ou extrêmes. Par ailleurs, comme c'est le cas dans les trois exemples mentionnés plus haut, nous voulons prendre en considération le cas où la distribution conditionnelle sous-jacente possède une ou des queues lourdes et qu'elle puisse être asymétrique ou multi-modale.

Dans ce chapitre, nous proposons donc une extension des modèles Pareto hybrides pour l'apprentissage de la densité conditionnelle. Tout comme dans le cas inconditionnel, nous allions les avantages des méthodes non-paramétriques provenant de l'apprentissage statistique et des hypothèses paramétriques rigoureuses de la théorie des valeurs extrêmes.

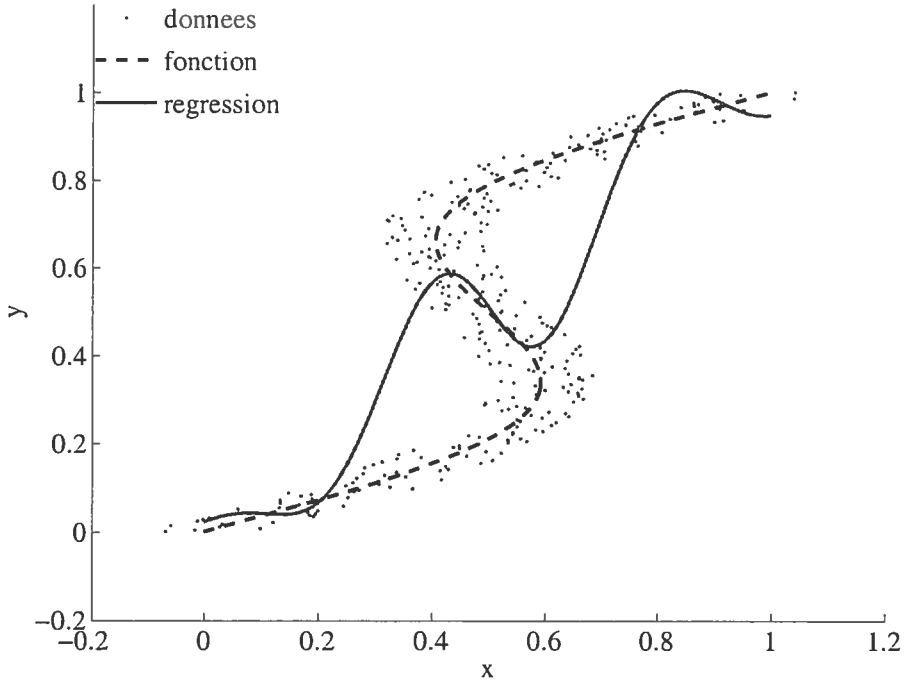

FIG. 4.1: *Exemple où l'estimation de l'espérance conditionnelle ne représente pas bien le processus générateur (voir Bishop* [**5**]*, chapitre 6). Les données sont générées de la façon suivante : $x = y + 0.3\sin(2\pi y) + \epsilon$ où ϵ est un bruit. Un réseau de neurones ayant huit unités cachées a été entraîné sur ces données pour approximer l'espérance conditionnelle.*

4.1. Estimation de densité conditionnelle

En apprentissage non-supervisé, il existe deux types de modèles pour l'apprentissage de la densité conditionnelle. Certains modèles estiment d'abord la densité conjointe de (X, Y) et l'utilisent ensuite pour construire un estimateur de la densité conditionnelle. D'autres modèles s'attaquent directement à l'estimation de la densité conditionnelle $p(Y|X = x)$. En général, il est sous-optimal d'estimer la densité conjointe lorsqu'ultimement, l'objet d'intérêt est la densité conditionnelle. En particulier, lorsque l'entrée X est de grande dimension, l'apprentissage de la densité conjointe peut s'avérer très difficile comparativement à celui de la densité conditionnelle. On peut voir l'estimation de densité conditionnelle comme l'estimation d'une fonction de x ayant certains paramètres. Cependant, les méthodes basées sur la densité conjointe sont toujours utilisées et ceci est dû en partie à l'existence de techniques simples pour l'apprentissage de la densité conjointe.

4.1.1. Modèles non-paramétriques

Les modèles d'estimation de densité conditionnelle basée sur l'estimation de la densité conjointe exploitent le fait suivant. Soit (X, Y), une paire de variables aléatoires telle que la valeur prise par Y dépende de X. Alors, si $p(x) \neq 0$, la densité conditionnelle de Y étant donné $X = x$ peut s'écrire de la manière suivante :

$$p(y|x) = \frac{p(x, y)}{p(x)}.$$

La plupart des modèles d'estimation non-paramétrique de la densité conjointe peuvent être écrits sous la forme d'une somme pondérée de densités conjointes :

$$\hat{p}(x, y) = \sum_{i=1}^{m} \pi_i p_i(x, y), \tag{4.1.1}$$

où π_i est la probabilité a priori que la composante i ait généré la paire (x, y) et $p_i(x, y)$ est la densité conjointe de la $i^{\text{ème}}$ composante. Dans le cas de l'estimateur à noyaux, $m = n$, où n est le nombre de points dans l'ensemble d'entraînement, $\pi_i = 1/n$ et $p_i(x, y)$ est un noyau gaussien multivarié centré sur (x_i, y_i). Selon

l'équation (4.1.1), l'estimateur de la densité conditionnelle est donné par :

$$\hat{p}(y|x) = \frac{\hat{p}(x,y)}{\hat{p}(x)} \tag{4.1.2}$$

$$= \frac{\sum_i \pi_i p_i(x,y)}{\sum_j \pi_j p_j(x)}, \tag{4.1.3}$$

où $\hat{p}(x) = \int \hat{p}(x,y)dy$ et $p_j(x) = \int p_j(x,y)dy$ est la densité marginale de X selon la $j^{\text{ème}}$ composante. En utilisant le fait que $p_i(x,y) = p_i(y|x)p_i(x)$ dans le numérateur de l'équation (4.1.3), on obtient :

$$\hat{p}(y|x) = \frac{\sum_i \pi_i p_i(y|x)p_i(x)}{\sum_j \pi_j p_j(x)} \tag{4.1.4}$$

$$= \sum_i \tau_i(x) p_i(y|x), \tag{4.1.5}$$

où les probabilités a posteriori sont données par $\tau_i(x) = \pi_i p_i(x) / \sum_j \pi_j p_j(x)$. L'expression de l'équation (4.1.5) met en évidence le fait que l'estimateur de densité conditionnelle construit à partir d'un estimateur non-paramétrique de la densité conjointe est lui aussi une somme pondérée de noyaux conditionnels dont les coefficients du mélange sont des fonctions de x.

Dans le cas d'un estimateur à noyaux de la densité conjointe, l'estimateur de l'équation (4.1.5) prend généralement la forme suivante (voir Rosenblatt [**42**] et Bashtannyk et Hyndman [**1**] entre autres) :

$$\hat{f}(y|x) = \frac{1}{b} \sum_{i=1}^{n} w_i(x) K\left(\frac{||y - Y_i||}{b}\right), \tag{4.1.6}$$

où $K(\cdot)$ est un noyau (c'est-à-dire une fonction réelle, non-négative, intégrable et paire qui satisfait certaines propriétés) et

$$w_i(x) = \frac{K\left(||x - X_i||/a\right)}{\sum_{j=1}^{n} K\left(||x - X_j||/a\right)}.$$

Pour l'estimateur conditionnel à noyaux de l'équation (4.1.6), les probabilités a posteriori $\tau_i(x)$ correspondent aux $w_i(x)$ et le noyau conditionnel $p_i(y|x)$ est simplement $K(||y - Y_i||/b)/b$. Les largeurs de fenêtre a et b contrôlent le voisinage d'influence du noyau dans l'espace des X et des Y respectivement. Le choix de a et b est crucial en ce qui concerne l'efficacité et la convergence de l'estimateur [**1**].

Dans la classe de modèles qui visent directement l'estimation de la densité conditionnelle, les estimateurs non-paramétriques prennent la forme d'une somme pondérée de densités conditionnelles :

$$\hat{p}(y|x) = \sum_j p(j|x)p(y|x,j), \tag{4.1.7}$$

où $p(j|x)$ est la probabilité que la composante j ait généré y étant donné que x soit observé et $p(y|x,j)$ est la densité conditionnelle de la composante j. On retrouve dans cette classe de modèles le mélange d'experts développé par Jacobs et al.[**24**] et sa version hiérarchique proposée par Jordan et Jacobs [**27**] ainsi que le mélange conditionnel de Gaussiennes de Bishop [**5**]. Dans un mélange d'experts, une fonction d'attribution appelée le "gating network", associe à chaque expert j la tâche de prédire y étant donné x selon la probabilité $p(j|x)$. Chaque expert j est une fonction de densité $p(y|x,j)$ dont les paramètres sont des fonctions de l'entrée x. Ces fonctions peuvent être modélisées par un modèle linéaire ou un réseau de neurones qui est propre à l'expert j. Chaque expert est donc spécialisé dans une région de l'espace des entrées. Si la fonction d'attribution et les fonctions associées à chaque expert sont linéaires, le mélange d'experts peut être ajusté aux données de manière efficace à l'aide de l'algorithme *EM* [**27**]. Le mélange d'experts hiérarchique [**27**] peut s'affranchir des limites d'une formulation linéaire du mélange d'experts tout en conservant l'efficacité de l'apprentissage du modèle par l'algorithme *EM*.

Le mélange conditionnel de Gaussiennes de Bishop [**5**] diffère du mélange d'experts en ceci qu'un seul réseau de neurones sert à la prédiction des probabilités $p(j|x)$ et des paramètres de chaque composante $p(y|x,j)$. Chaque composante est donc une densité gaussienne de moyenne $\mu_j(x)$ et d'écart-type $\sigma_j(x)$. Si le nombre de composantes du mélange et le nombre d'unités cachées du réseau de neurones sont choisis de manière adéquate, le mélange de Gaussiennes conditionnel sera en mesure de modéliser toute distribution conditionnelle. Les paramètres du modèle sont les poids du réseau de neurones et ceux-ci sont appris en maximisant la log-vraisemblance.

La couche cachée d'un réseau de neurones peut être vue comme une représentation de l'entrée x. Dans le mélange de Gaussiennes conditionnel, les unités cachées sont partagées par les composantes et les probabilités a priori du mélange. Ceci permet donc d'exploiter une représentation de l'entrée pour la prédiction de tous les paramètres du mélange. Nous reprenons ce type d'architecture en proposant un mélange conditionnel mais en utilisant plutôt des composantes Pareto hybrides. Tout comme pour le mélange inconditionnel, ceci permet de combiner la flexibilité d'un modèle non-paramétrique avec la capacité d'extrapolation des méthodes issues de la théorie des valeurs extrêmes.

4.1.2. Mélange conditionnel de Pareto hybrides

On peut formuler le mélange conditionnel de Pareto hybrides comme un mélange dont les paramètres dépendent de l'entrée x :

$$\phi_\theta(y|x) = \sum_{j=1}^{m} \pi_j(x) h_{\psi_j(x)}(y). \tag{4.1.8}$$

Chaque composante $h_{\psi_j(x)}(y)$ est une densité de loi Pareto hybride dont la probabilité a priori $\pi_j(x)$ et les paramètres $\psi_j(x) = (\xi_j(x), \mu_j(x), \sigma_j(x))$ sont des fonctions de la variable d'entrée x. Un réseau de neurones ayant une couche cachée sert à la prédiction des $\pi_j(x)$ et des $\psi_j(x)$, pour $j = 1, \ldots, m$. Les paramètres θ du mélange conditionnel sont donc les poids du réseau de neurones. L'architecture du mélange conditionnel de Pareto hybrides est illustrée à la figure 4.2.

Soit $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, un ensemble d'entraînement. Les paramètres θ du mélange conditionnel sont appris par maximisation de la log-vraisemblance sur $\mathcal{D}_n$:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log\left(\phi_\theta(y_i|x_i)\right). \tag{4.1.9}$$

Les réseaux de neurones ayant une couche cachée et la tangente hyperbolique comme fonction d'activation de la couche cachée (voir la section 2.5) sont une classe de fonctions appropriée pour prédire les paramètres du mélange conditionnel. En effet, il existe une méthode efficace du calcul du gradient pour ces modèles (la rétropropagation de l'erreur [**43**]) et ils possèdent la propriété d'approximation universelle (si le nombre d'unités cachées est choisi adéquatement, ces réseaux de

neurones sont en mesure d'approximer arbitrairement bien toute fonction continue). Par contre, l'optimisation des paramètres est non-convexe et l'on n'a pas de garantie de trouver le minimum global. La complexité d'un mélange conditionnel est contrôlée par le nombre d'unités cachées du réseau de neurones et le nombre de composantes du mélange.

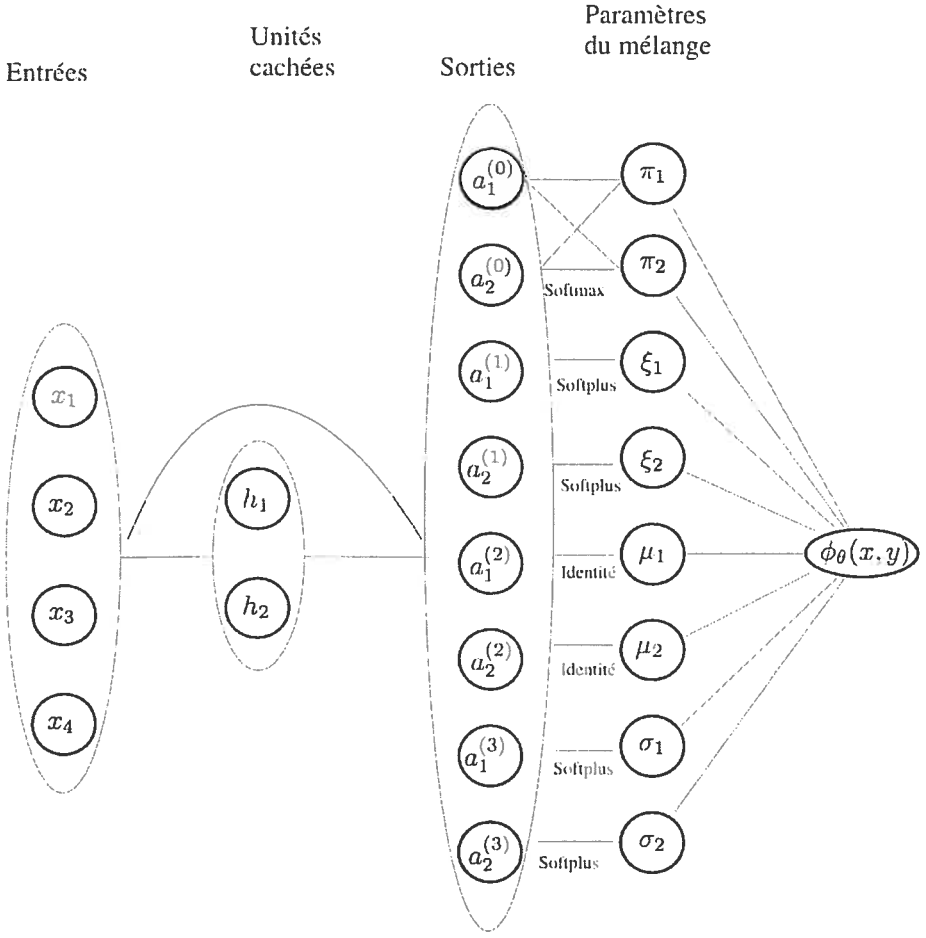

FIG. 4.2: *Mélange conditionnel avec composantes Pareto hybrides : un réseau de neurones ayant une couche cachée et la tangente hyperbolique comme fonction d'activation de la couche cachée sert à prédire les paramètres du mélange conditionnellement à l'entrée x. Des fonctions de transfert à la sortie du réseau de neurones permettent de contrôler l'intervalle de valeurs prises par les paramètres.*

Dans l'architecture de la figure 4.2, nous utilisons un réseau de neurones à une couche cachée auquel on ajoute une connexion linéaire entre l'entrée et la sortie. De cette manière, le modèle linéaire est le cas limite (aucune unité cachée) du réseau de neurones. Les sorties du réseau de neurones, les $a_j^{(i)}$ de la figure 4.2, sont

composés d'une combinaison linéaire des sorties des unités cachées $z_h(x)$ plus une combinaison linéaire des entrées x_k ainsi qu'un terme de biais b_j :

$$a_j^{(i)}(x) = b_j^{(i)} + \sum_{h=1}^{n_h} w_{jh}^{(i)} z_h(x) + \sum_{k=1}^{d} \widetilde{w}_{jk}^{(i)} x_k, \tag{4.1.10}$$

où n_h est le nombre d'unités cachées, d est la dimension de l'entrée x et

$$z_i(x) = \tanh\left(c_i + \sum_{k=1}^{d} v_{ik} x_k\right). \tag{4.1.11}$$

Les fonctions de transfert à la sortie du réseau de neurones sont choisies de sorte que les paramètres du mélange conditionnel respectent certaines contraintes sur les valeurs qu'ils peuvent prendre. Les $\pi_j(x)$ représentent la probabilité a priori que la composante j soit responsable de la génération de y étant donné x. On doit donc avoir les contraintes suivantes : $0 \leq \pi_j(x) \leq 1$ pour tout j et $\sum_{j=1}^{m} \pi_j(x) = 1$. Pour satisfaire ces contraintes, nous utilisons la fonction *softmax* :

$$\pi_i(x) = \frac{\exp(a_i^{(0)})}{\sum_{j=1}^{m} \exp(a_j^{(0)})}.$$

Parmi les paramètres de la Pareto hybride, nous voulons contraindre l'indice de queue et le paramètre d'échelle à être positifs. Nous nous assurons ainsi que la Pareto hybride ait une queue supérieure infinie. La positivité de ces paramètres est imposée à l'aide de la fonction *softplus* :

$$\text{softplus}(x) = \log(1 + e^x).$$

La *softplus* a été proposée par [**13**] ; tout comme l'exponentielle, elle prend ses valeurs dans $\mathbb{R}^+$ mais elle croît plus lentement que l'exponentielle. En effet, puisque lorsque $x > 0$, $\text{softplus}(x) = x + \log(1 + e^{-x})$, nous avons :

$$\lim_{x \to \pm\infty} \text{softplus}(x) \to \max(x, 0).$$

Grâce à cette propriété, l'optimisation numérique est plus stable. Le paramètre d'emplacement des composantes Pareto hybrides ne nécessite aucune contrainte.

4.1.3. Apprentissage

Les paramètres $\theta = (b, c, v, w, \widetilde{w})$ du mélange conditionnel de Pareto hybrides sont déterminés par la maximisation de la log-vraisemblance de l'équation (4.1.9) sur un ensemble d'entraînement $\mathcal{D}_n$. Nous utilisons un algorithme d'optimisation basé sur la descente du gradient conjugué. Nous avons donc besoin de calculer le gradient de la log-vraisemblance par rapport au vecteur de paramètres θ. Nous obtenons le gradient de la log-vraisemblance en deux étapes :

(1) Nous calculons d'abord les dérivées de $\mathcal{L}(\theta)$ par rapport aux $a_j^{(i)}$, les sorties du réseau de neurones avant les fonctions de transfert, voir la figure 4.2.

(2) Les dérivées sont ensuite rétro-propagées dans le réseau de neurones afin d'obtenir $\frac{\partial \mathcal{L}(\theta)}{\partial \theta}$. Le gradient s'obtient donc par la décomposition suivante :

$$\frac{\partial l}{\partial \theta} = \sum_i \sum_j \frac{\partial l}{\partial a_j^{(i)}} \frac{\partial a_j^{(i)}}{\partial \theta}.$$

La dérivée de l'étape (2) consiste en la rétro-propagation standard dans un réseau de neurones. Celle-ci est décrite à la section 2.5. Nous décrirons donc ici les dérivées impliquées dans l'étape (1). Soit $l = \log(\phi_\theta(x, y)) = \log(\phi_\lambda(y))$, la valeur de la log-vraisemblance pour l'exemple (x, y) où $\phi_\theta(x, y)$ est donné dans l'équation (4.1.8) et $\lambda = (\pi_1(x), \ldots, \pi_m(x), \psi_1(x), \ldots, \psi_m(x))$, c'est-à-dire l'ensemble des paramètres du mélange correspondant à l'entrée x. Les calculs des dérivées $\frac{\partial l}{\partial a_j^{(i)}}$ peuvent être séparés en deux cas :

- Si $i = 0$, $a_j^{(0)}$ est une des sorties qui gouvernent les probabilités a priori et la dérivée de la log-vraisemblance s'exprime comme suit :

$$\frac{\partial l}{\partial a_j^{(0)}} = \frac{\partial l}{\partial \phi_\lambda(y)} \sum_{k=1}^{m} \frac{\partial \phi_\lambda(y)}{\partial \pi_k} \frac{\partial \pi_k}{\partial a_j^{(0)}} \tag{4.1.12}$$

- Par ailleurs, si $1 \leq i \leq 3$, $a_j^{(i)}$, $j = 1, \ldots, m$, contrôle un paramètre d'une composante Pareto hybride et la fonction de log-vraisemblance ne dépend de $a_j^{(i)}$ que par ce paramètre :

$$\frac{\partial l}{\partial a_j^{(i)}} = \frac{\partial l}{\partial \phi_\lambda(y)} \frac{\partial \phi_\lambda(y)}{\partial \psi_{j,i}(x)} \frac{\partial \psi_{j,i}(x)}{\partial a_j^{(i)}}, \tag{4.1.13}$$

où $\psi_{j,i}(x)$ dénote le $i^{\text{ème}}$ élément de $\psi_j(x)$, le vecteur de paramètres de la $j^{\text{ème}}$ composante.

Nous développons ensuite toutes les dérivées partielles des équations (4.1.12) et (4.1.13). Dans ces deux équations, nous retrouvons :

$$\frac{\partial l}{\partial \phi_\lambda(y)} = \frac{1}{\phi_\lambda(y)}.$$

La dérivée du mélange par rapport aux probabilités a priori est donnée par :

$$\frac{\partial \phi_\lambda(y)}{\partial \pi_j} = h_{\psi_j(x)}(y).$$

En prenant la dérivée du mélange par rapport aux paramètres des composantes, $\psi_{j,i}(x)$, on obtient :

$$\frac{\partial \phi_\lambda(y)}{\partial \psi_{j,i}(x)} = \pi_j(x) \frac{\partial h_{\psi_j(x)}(y)}{\partial \psi_{j,i}(x)}. \tag{4.1.14}$$

Finalement, les dérivées des paramètres du mélange par rapport aux sorties du réseau de neurones se calculent de la manière suivante :

$$\begin{aligned}
\frac{\partial \pi_k}{\partial a_j^{(0)}} &= \begin{cases} \pi_j(1-\pi_j) & \text{si } j = k \\ -\pi_k \pi_j & \text{si } j \neq k \end{cases} \\
\frac{\partial \xi_j}{\partial a_j^{(1)}} &= 1 - \exp(-\xi_i) \quad j = 1, \ldots, m \\
\frac{\partial \mu_j}{\partial a_j^{(2)}} &= 1 \quad j = 1, \ldots, m \\
\frac{\partial \sigma_j}{\partial a_j^{(3)}} &= 1 - \exp(-\sigma_i) \quad j = 1, \ldots, m.
\end{aligned}$$

Les calculs détaillés relatifs au gradient de la fonction de densité de la Pareto hybride ($\partial h_{\psi_j(x)}(y)/\psi_{j,i}(x)$ dans l'équation (4.1.14)) ainsi que les calculs du gradient du mélange conditionnel global sont décrits dans l'annexe A.

4.1.4. Initialisation

Afin de commencer l'optimisation avec des paramètres raisonnables, nous suivons la procédure d'initialisation suivante, inspirée de celle du mélange conditionnel de Gaussiennes dans Netlab, la librairie de fonctions Matlab de Ian Nabney et Chris Bishop. Les biais $b_j^{(i)}$ du mélange conditionnel sont initialisés à l'aide de

la distribution inconditionnelle. Les étapes sont les suivantes (voir la section 3.3.1 pour l'initialisation d'un mélange de Pareto hybrides inconditionnel) :

(1) Regrouper les variables de sorties de l'ensemble d'entraînement $\{y_1, \ldots, y_n\}$ en m sous-groupes à l'aide d'un algorithme de "clustering" comme k-means ou k-medians [**38**].

(2) Estimer le vecteur de paramètres $\psi_j = (\xi_j, \mu_j, \sigma_j)$ pour $j = 1, \ldots, m$ sur chacun des m sous-groupes.

(3) Estimer les probabilités a priori du mélange inconditionnel comme les proportions d'observations dans chacun des m sous-groupes.

(4) Utiliser les fonctions de transfert inverses pour transférer les estimations des paramètres du mélange inconditionnel aux biais :

$$\begin{aligned} b_j^{(0)} &= \log(\hat{\pi}_j) & j = 1, \ldots, m-1 \\ b_j^{(1)} &= \log\left(\exp(\hat{\xi}_j) - 1\right) & j = 1, \ldots, m \\ b_j^{(2)} &= \hat{\mu}_j & j = 1, \ldots, m \\ b_j^{(3)} &= \log\left(\exp(\hat{\sigma}_j) - 1\right) & j = 1, \ldots, m \end{aligned}$$

où $\hat{\pi}$, $\hat{\xi}_j$, $\hat{\mu}_j$ et $\hat{\sigma}_j$ sont les paramètres estimés du mélange de Pareto hybrides inconditionnel.

Les autres paramètres du mélange conditionnel, c'est-à-dire c, v, w et $\widetilde{w}$ sont initialisés aléatoirement selon une distribution uniforme sur l'intervalle $[-0.9/\sqrt{k}, 0.9/\sqrt{k}]$ où $k = 1 + d$ pour les poids de la couche cachée (c et v) et $k = 1 + d + h$ pour les poids de la couche de sorties (w et $\widetilde{w}$).

4.2. ÉTUDE SIMULATOIRE

Nous utilisons à nouveau des jeux de données synthétiques pour explorer les propriétés du mélange conditionnel de Pareto hybrides. Faisant suite à l'étude simulatoire pour la densité inconditionnelle de la section 3.3, nous générons des paires d'observations (x, y) à l'aide de la loi de Fréchet conditionnelle. Étant donné X, Y suit une loi de Fréchet de paramètres $\mu(X)$, $\sigma(X)$ et $\xi(X)$. Les trois

paramètres de la Fréchet sont donc des fonctions de l'entrée X. Nous avons choisis deux types de dépendance fonctionnelle : soit linéaire ($f(X) = aX + b$), soit sinusoïdale ($f(X) = c\sin(aX + b) + d$) afin de vérifier les capacités d'apprentissage du réseau de neurones à l'intérieur d'un mélange conditionnel. La variable d'entrée X suit une loi Normale. Les paramètres de la dépendance fonctionnelle sont choisis de sorte que l'indice de queue de la Fréchet se trouve soit dans l'intervalle $[0.25, 0.5]$ dans 99% des cas soit dans l'intervalle $[0.66, 1.33]$ dans 99% des cas, ce qui correspond respectivement à une queue modéremment lourde et une queue très lourde. Les paramètres d'emplacement et de dispersion de la Fréchet sont modélisés de façon similaire. Ces deux spécifications (dépendance linéaire ou sinusoïdale et queue modérée ou lourde) donne lieu à quatre types de jeux de données qui sont illustrés aux figures 4.3 et 4.4. Nous utilisons les abbréviations suivantes pour référer à ces quatre jeux de données : Fréchet-*dep*-*queue* désigne le modèle générateur de la Fréchet conditionnelle avec dépendance linéaire des paramètres si *dep*=lin et sinusoïdale si *dep*=sin. L'indice de queue est soit modérée si *queue*=mod ou lourd si *queue*=lourd. Nous avons vu dans le cas inconditionnel que le mélange de Pareto hybrides se distingue surtout des autres estimateurs pour les petits jeux de données. Cependant, l'estimation de quantiles semble être plus stable pour les jeux de données contenant au moins 1000 observations. Nous avons donc choisi deux tailles d'ensemble d'entraînement, $n = 200$ et $n = 2000$, pour étudier le comportement des estimateurs lorsque peu de données sont disponibles. La taille de l'ensemble de test est de 10 000 observations et chaque simulation est répétée 100 fois afin de calculer la moyenne et la variance des performances pour chaque taille d'ensemble d'entraînement.

4.2.1. Entraînement et critères d'évaluation

Nous avons à nouveau deux objectifs. Tout d'abord, nous voulons comparer le mélange conditionnel de Pareto hybrides à d'autres estimateurs de densité conditionnelle en termes de log-vraisemblance et d'estimation de la queue de la distribution. Ensuite, nous voulons vérifier si l'estimation de la queue de la distribution conditionnelle provenant du mélange conditionnel de Pareto hybrides

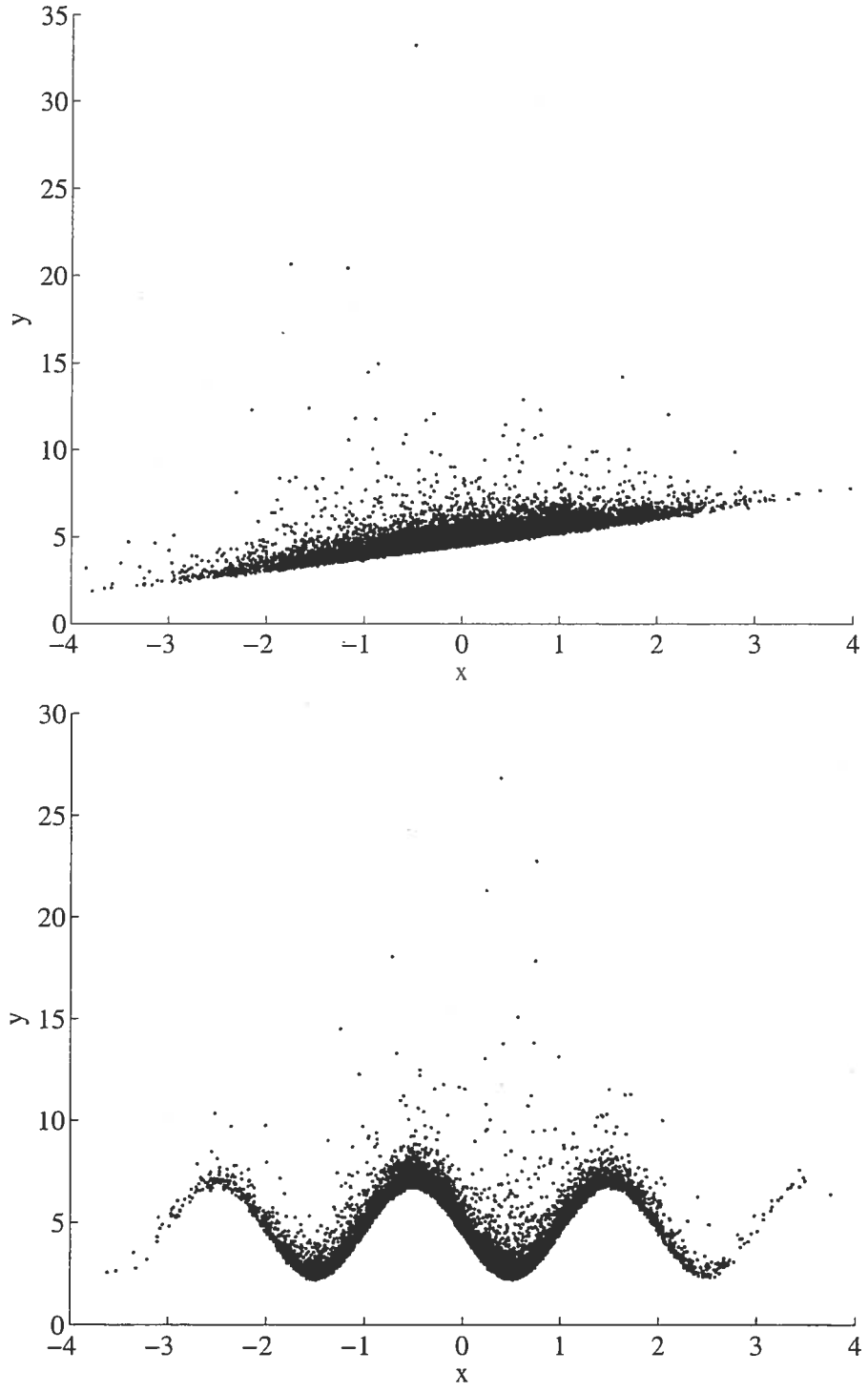

FIG. 4.3: *Jeux de données avec queue modérée $\xi \in [0.25, 0.5]$ pour la loi de Fréchet conditionnelle. Dépendance linéaire des paramètres (panneau du haut) et sinusoïdale (panneau du bas).*

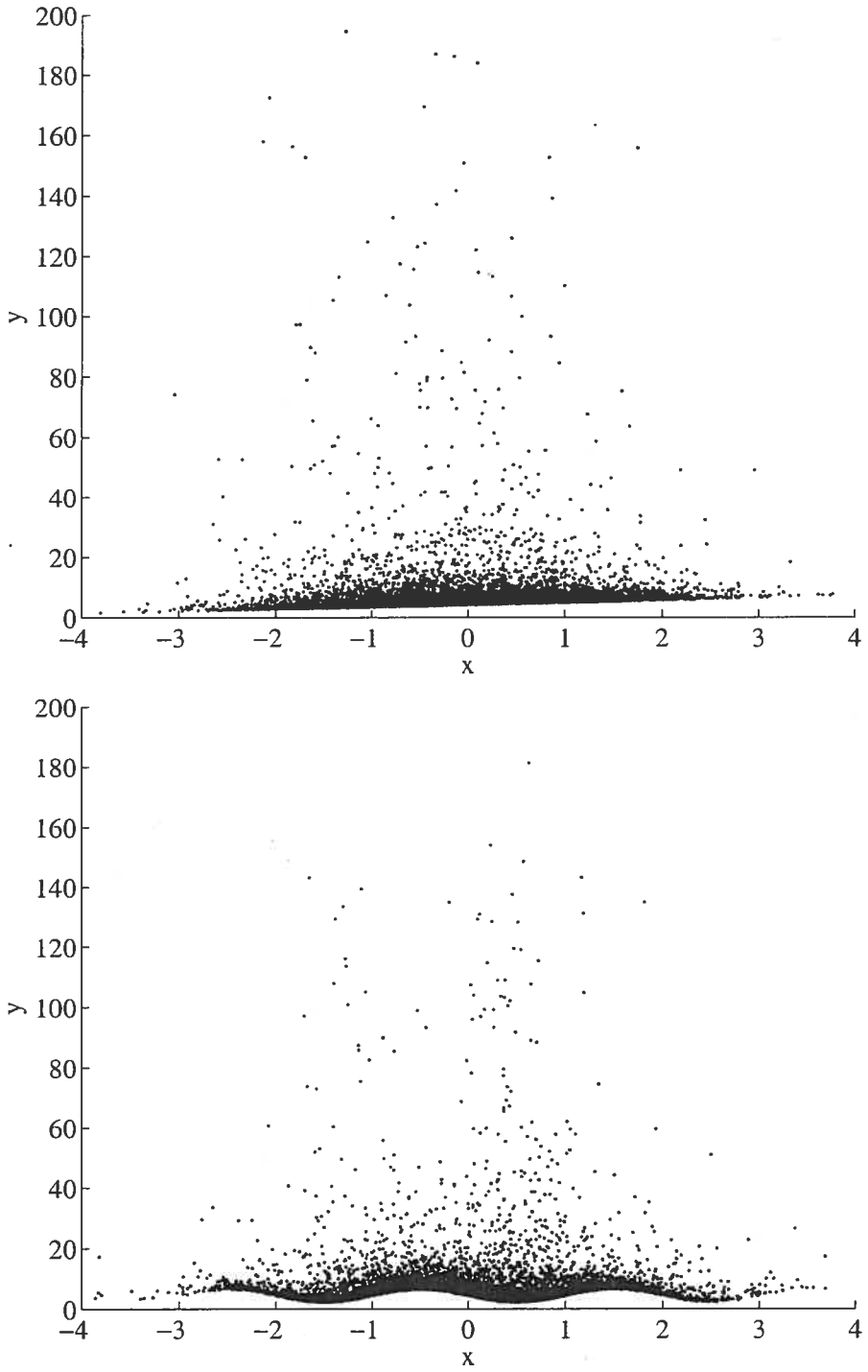

FIG. 4.4: *Jeux de données avec queue lourde* $\xi \in [0.66, 1.33]$ *pour la loi de Fréchet conditionnelle. Dépendance linéaire des paramètres (panneau du haut) et sinusoïdale (panneau du bas). L'axe des* y *est raccourci afin qu'on puisse voir la forme de la dépendance ; la valeur maximale de* y *est de l'ordre de* 10^6 .

est équivalente à la version conditionnelle de la méthode PoT issue de la théorie des valeurs extrêmes **[10, 35]** .

Dans un premier temps, nous comparons le mélange conditionnel de Pareto hybrides avec la version conditionnelle des estimateurs de densité univariée de la section 3.3. Il s'agit donc de mélanges conditionnels avec des composantes Normales ou Log-Normales et de l'estimateur de la fenêtre de Parzen pour le cas conditionnel dont la formulation est donnée à l'équation (4.1.6). À l'exception de l'estimateur de la fenêtre de Parzen conditionnel, les mélanges conditionnels pour tout type de composantes sont entraînés en maximisant la log-vraisemblance conditionnelle. Pour éviter les minima locaux, l'optimisation est relancée cinq fois avec des paramètres initiaux différents. Les paramètres ayant donné lieu à l'erreur d'entraînement la plus petite sont retenus. L'initialisation des mélanges conditionnels avec des composantes Normales ou Log-Normales est faite de manière similaire à celle du mélange conditionnel de Pareto hybrides décrite à la sous-section 4.1.4. Les hyper-paramètres, le nombre d'unités cachées et de composantes pour les mélanges et les largeurs de fenêtres pour l'estimateur de la fenêtre de Parzen, sont choisis sur une ensemble de validation qui consiste en 20% de l'ensemble d'entraînement.

Nous comparons d'abord les estimateurs de densité conditionnelle en termes de log-vraisemblance relative au modèle générateur :

$$\mathcal{R}_l(\theta) = -\frac{1}{l}\sum_{i=1}^{l} \log\left(\frac{\phi_\theta(y_i|x_i)}{p(y_i|x_i)}\right), \tag{4.2.1}$$

où la somme est sur l'ensemble de test $\mathcal{D}_l$, $p(y|x)$ est la densité du modèle générateur et $\phi_\theta(y|x)$ est la densité de l'estimateur. Puisque, à un facteur près, la log-vraisemblance relative est l'équivalent empirique du critère de Kullback-Leibler, plus ce critère est petit, plus l'estimateur est proche du modèle générateur.

Pour avoir une idée de la distribution conditionnelle apprise par les différents modèles, nous avons ensuite généré des données à partir des modèles entraînés. Soit un point x dans l'espace d'entrée, une composante j est choisie selon la distribution discrète donnée par $\pi_1(x), \ldots, \pi_m(x)$. Un point y est ensuite généré selon la distribution de la $j^{\text{ième}}$ composante de paramètres $\psi_j(x)$.

Parmi les méthodes issues de la théorie des valeurs extrêmes, Davison et Smith **[10]** ont proposé de modéliser les paramètres de la Pareto généralisée comme des fonctions d'une variable explicative. Ils proposent la modélisation suivante : $\xi(x) = \exp(\sum_{i=1}^{d} a_i x_i + b)$ et $\beta(x) = \exp(\sum_{i=1}^{d} c_i x_i + d)$. Par ailleurs, McNeil et Frey **[35]** ont dévelopé un modèle pour les séries financières. Dans un premier temps, les dépendances temporelles de la série sont capturées à l'aide d'un modèle de série chronologique (un modèle AR(1)-GARCH(1,1)). Ensuite, la méthodologie PoT est appliquée aux résidus sous l'hypothèse que ceux-ci soient indépendants et identiquement distribués. À partir de l'idée sous-jacente de ce modèle, nous proposons de faire une régression à l'aide d'un réseau de neurones et ensuite, d'appliquer la méthodologie PoT aux résidus. Soit $f(x, \theta)$, la fonction calculée par le réseau de neurones de paramètres θ pour l'entrée x. Les paramètres θ sont appris par minimisation de l'erreur quadratique moyenne. Soit u le seuil correspondant au quantile de niveau q_{PoT} sur les résidus. Le seuil pour la variable dépendante y est alors donné par $u(x) = f(x, \theta) + u$. Au-delà de ce seuil, les excès sont modélisés à l'aide de la loi de Pareto généralisée. Ceci nous permettra de comparer le mélange conditionnel de Pareto hybrides en termes de génération d'excédents conditionnels (donc d'observations dans la queue de la distribution) et d'estimation de l'indice de queue.

La PoT conditionnelle a deux hyper-paramètres : le nombre d'unités cachées du réseau de neurones qui est choisi sur l'ensemble de validation et le niveau de quantile qui détermine le seuil qui est choisi à l'aide du test d'adéquation de la Pareto généralisée de Choulakian et Stephens **[6]**.

4.2.2. Résultats des simulations

4.2.2.1. *Log-vraisemblance relative*

Le tableau 4.2 contient les résultats des simulations en termes de log-vraisemblance relative au modèle générateur pour les quatre types de jeux de données et les deux tailles d'ensemble d'entraînement. On désigne par **cmmh**, **cmmg** et **cmml**, les mélanges conditionnels à composantes Pareto hybrides, Gaussiennes et Log-Normales respectivement. On dénote l'estimateur conditionnel de la fenêtre de

parzen par **cparzen**. Il y a deux niveaux de difficulté dans ces données : la forme de la dépendance fonctionnelle des paramètres (linéaire ou sinusoïdale) et l'épaisseur de la queue de la distribution conditionnelle. Dans presque tous les cas, le mélange conditionnel de Pareto hybrides offre une performance significativement supérieure en termes de log-vraisemblance relative. Dans un seul cas, la performance du mélange d'hybrides n'est pas distinguable de celle du mélange de Log-Normales. Il s'agit du cas où l'ensemble d'entraînement contient 200 observations pour le jeu de données avec dépendance linéaire et queue lourde. Le mélange conditionnel de Log-Normales s'adapte toutefois relativement bien aux quatre types de jeux de données. Cependant, le mélange conditionnel de Gaussiennes ne parvient pas à fournir un modèle acceptable dans le cas où la queue de la distribution conditionnelle est lourde. C'est aussi le cas pour l'estimateur de la fenêtre de Parzen conditionnel. Il est intéressant de noter que, pour les mélanges conditionnels, la forme des composantes interfère avec la capacité du réseau de neurones à capter la forme de la dépendance fonctionnelle.

4.2.2.2. *Hyper-paramètres selectionnés*

Les hyper-paramètres moyens sélectionnés en validation sont fournis dans le tableau 4.3. Lorsque la dépendance est linéaire, dans la plupart des cas, les mélanges conditionnels n'ont aucune unité cachée. Lorsque la dépendance est sinusoïdale et la queue modérée, le nombre d'unités cachées augmente. Cependant, lorsque la queue de la distribution est lourde, le nombre d'unités cachées sélectionné est moindre. Ceci est probablement dû au fait qu'un grand nombre d'observations se trouvent très éloignées et perturbent la forme de la dépendance (voir la figure 4.4). Le nombre de composantes sélectionné pour les mélanges conditionnels augmente avec la taille de l'ensemble d'entraînement et l'épaisseur de la queue, tel qu'attendu. Pour les jeux de données avec queue modérée, l'estimateur de la fenêtre de Parzen sélectionne des largeurs de fenêtre qui diminuent avec la taille de l'ensemble d'entraînement. Ceci n'est pas le cas lorsque la queue de la distribution est lourde ; les largeurs de fenêtre choisies sont beaucoup plus grandes et la performance de cet estimateur se dégrade.

Le tableau 4.1 fournit les résultats des simulations pour la méthode PoT conditionnelle. Le nombre d'unités cachées sélectionné pour le réseau de neurones se comporte de manière semblable à celui des mélanges conditionnels. Il y a peu ou pas d'unités cachées sélectionnées dans les cas de dépendance linéaire ou de queue lourde alors qu'un nombre significatif d'unités cachées est sélectionné dans le cas de dépendance sinusoïdale et de queue modérée. Le niveau de quantile choisi, et donc le seuil, augmente avec la taille de l'ensemble d'entraînement. Ceci est cohérent avec le théorème de Pickands sur la convergence de l'approximation de la queue d'une distribution par la Pareto généralisée lorsque le seuil augmente.

	n	h	u	q_{PoT}	$\mathcal{R}_l$ (err. std.)
lin-mod	200	0	0.06176	0.7075	0.01128 (0.000323)
	2000	0	0.05865	0.7095	-0.01858 (9.708e-05)
sin-mod	200	2.93	0.2967	0.685	0.4563 (0.003213)
	2000	3.92	0.3303	0.747	0.2633 (0.002551)
lin-lourd	200	0.19	0.4351	0.8165	0.3962 (0.01417)
	2000	0.02	1.97	0.8865	0.35 (0.008203)
sin-lourd	200	0.46	0.529	0.751	0.5708 (0.005926)
	2000	0.02	1.718	0.893	0.6231 (0.01049)

TAB. 4.1: *Log-vraisemblance relative à la densité génératrice $\mathcal{R}_l$ (calculée sur les excès uniquement) et hyper-paramètres sélectionnés (le nombre d'unités cachées h et le niveau de quantile q_{PoT} déterminant le seuil u) pour la méthode PoT conditionnelle. La taille de l'ensemble d'entraînement est n.*

	n	cmmh	cmmg	cmml	cparzen
lin-mod	200	**0.1081 (0.01621)**	0.6401 (0.07695)	0.3059 (0.02426)	2.54 (0.3703)
	2000	**0.006835 (0.0002546)**	0.1919 (0.02726)	0.1255 (0.002318)	1.459 (0.2688)
sin-mod	200	**0.7014 (0.06141)**	2.338 (0.9337)	0.9909 (0.07663)	2.038 (0.2649)
	2000	**0.1834 (0.005721)**	0.3024 (0.02447)	0.3028 (0.005838)	1.399 (0.2735)
lin-lourd	200	**0.4176 (0.03305)**	6.583e+07 (5.612e+07)	**0.476 (0.04296)**	8.377e+04 (4.578e+04)
	2000	**0.01875 (0.0003348)**	1450 (1027)	0.1761 (0.008315)	3.929e+05 (3.841e+05)
sin-lourd	200	**0.597 (0.04247)**	8.112e+05 (5.067e+05)	2.335 (0.9769)	9.563e+05 (5.717e+05)
	2000	**0.1783 (0.006356)**	7.403e+06 (7.128e+06)	0.2628 (0.006283)	1.807e+06 (1.278e+06)

TAB. 4.2: *Log-vraisemblance (err. std) relative à la densité génératrice sur l'ensemble de test. Plus ce critère est petit, plus performant est l'estimateur. Les meilleures performances sont soulignées en caractères gras. Les données sont générées d'après une loi de Fréchet conditionnelle. La taille de l'ensemble d'entraînement est n.*

	n	$(h_{\text{cmmh}}, m_{\text{cmmh}})$	$(h_{\text{cmmg}}, m_{\text{cmmg}})$	$(h_{\text{cmml}}, m_{\text{cmml}})$	$(\lambda^x_{\text{cparzen}}, \lambda^y_{\text{cparzen}})$
lin. mod.	200	(0, 1.94)	(0.1, 2.47)	(0.01, 2.2)	(0.1106, 0.2494)
	2000	(0, 2)	(0.02, 2.54)	(0.05, 2.2)	(0.0586, 0.1495)
sin. mod.	200	(2.53, 2.34)	(2.34, 3.07)	(2.45, 2.8)	(1.006, 0.541)
	2000	(3.97, 4.36)	(4.35, 5.81)	(3.51, 4.5)	(0.00325, 0.1882)
lin. lourde	200	(0.1, 1.64)	(0, 2.11)	(0.12, 2.3)	(7.448, 24.77)
	2000	(0, 2)	(0.02, 3.75)	(0.13, 3.34)	(5.896, 41.9)
sin. lourde	200	(1.71, 2.43)	(0.78, 3.1)	(0.74, 2.83)	(16.87, 31.72)
	2000	(2.16, 6.18)	(2.73, 5.3)	(1.27, 6.98)	(35.7, 58.15)

TAB. 4.3: *Hyper-paramètres sélectionnés en validation pour les modèles d'estimation de densité conditionnelle correspondant au tableau 4.2. Pour les mélanges conditionnels, h_{cmm} est le nombre d'unités cachées et m_{cmm} est le nombre de composantes. Pour l'estimateur de Parzen conditionnel, $\lambda^x_{cparzen}$ est la largeur de fenêtre dans l'espace des entrées et $\lambda^y_{cparzen}$ la largeur de fenêtre dans l'espace des sorties.*

4.2.3. Étude des estimateurs conditionnels

4.2.3.1. *Paramètres des mélanges conditionnels : modèle générateur Fréchet-lin-mod*

Afin d'avoir une idée du mécanisme qui permet aux mélanges conditionnels de s'adapter aux différents types de données, nous avons tracé, après l'entraînement sur un jeu de données, les courbes des paramètres des mélanges en fonction de la variable d'entrée x. Dans le cas de la dépendance linéaire et de la queue de distribution modérée (modèle générateur Fréchet-lin-mod), les figures 4.5, 4.6 et 4.7 contiennent les graphiques des paramètres pour les mélanges avec composantes Pareto hybrides, Gaussiennes et Log-Normales respectivement. Si Y suit une loi Log-Normale de paramètres (m, s^2), alors $Z = \log Y$ suit une loi Normale d'espérance m et d'écat-type s. Cependant, l'espérance et l'écart-type de Y sont donnés par $\mu = \exp(m + s^2/2)$ et $\sigma = \exp(s^2 + 2m)(\exp(s^2) - 1)$. Dans le cas du mélange conditionnel de Log-Normales, plutôt que de tracer les courbes de m et s en fonction de x, nous avons tracé les courbes de μ et σ en fonction de x. Pour ces trois types de mélanges conditionnels, les paramètres μ_j représentant l'emplacement des composantes, suivent assez fidèlement le paramètre μ de la loi de Fréchet conditionnelle ayant généré les données. Pour le mélange conditionnel de Pareto hybrides, l'indice de queue de la composante dominante est décroissant en x. Lorsque l'ensemble d'entraînement contient 200 observations, l'indice de queue est grandement sur-estimé pour les valeurs négatives de x et sous-estimé pour les valeurs positives. Ceci se corrige lorsque l'ensemble d'entraînement augmente à 2000 observations.

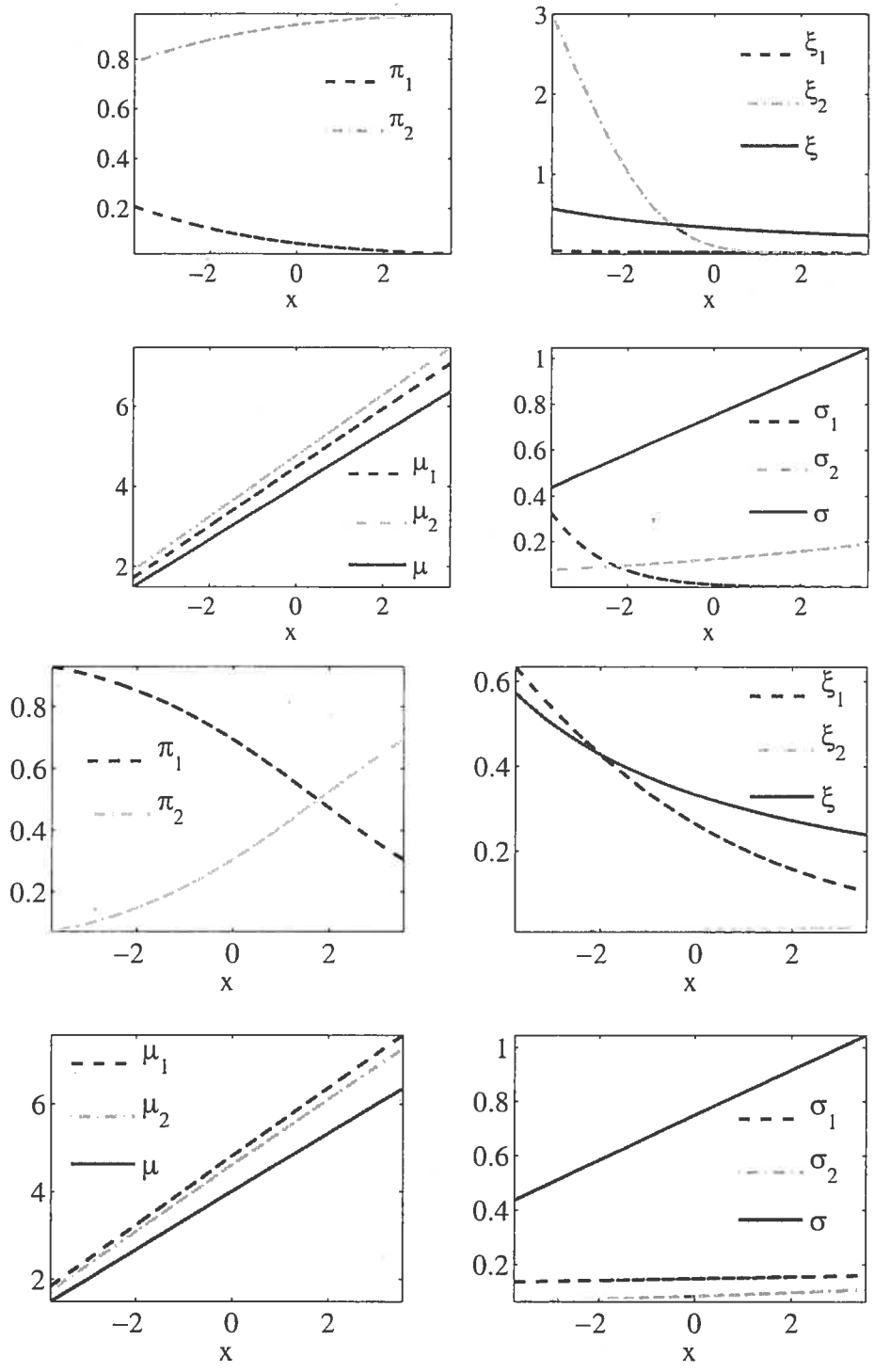

FIG. 4.5: $(\pi_j(x), \xi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMMH avec deux composantes. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-lin-mod. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

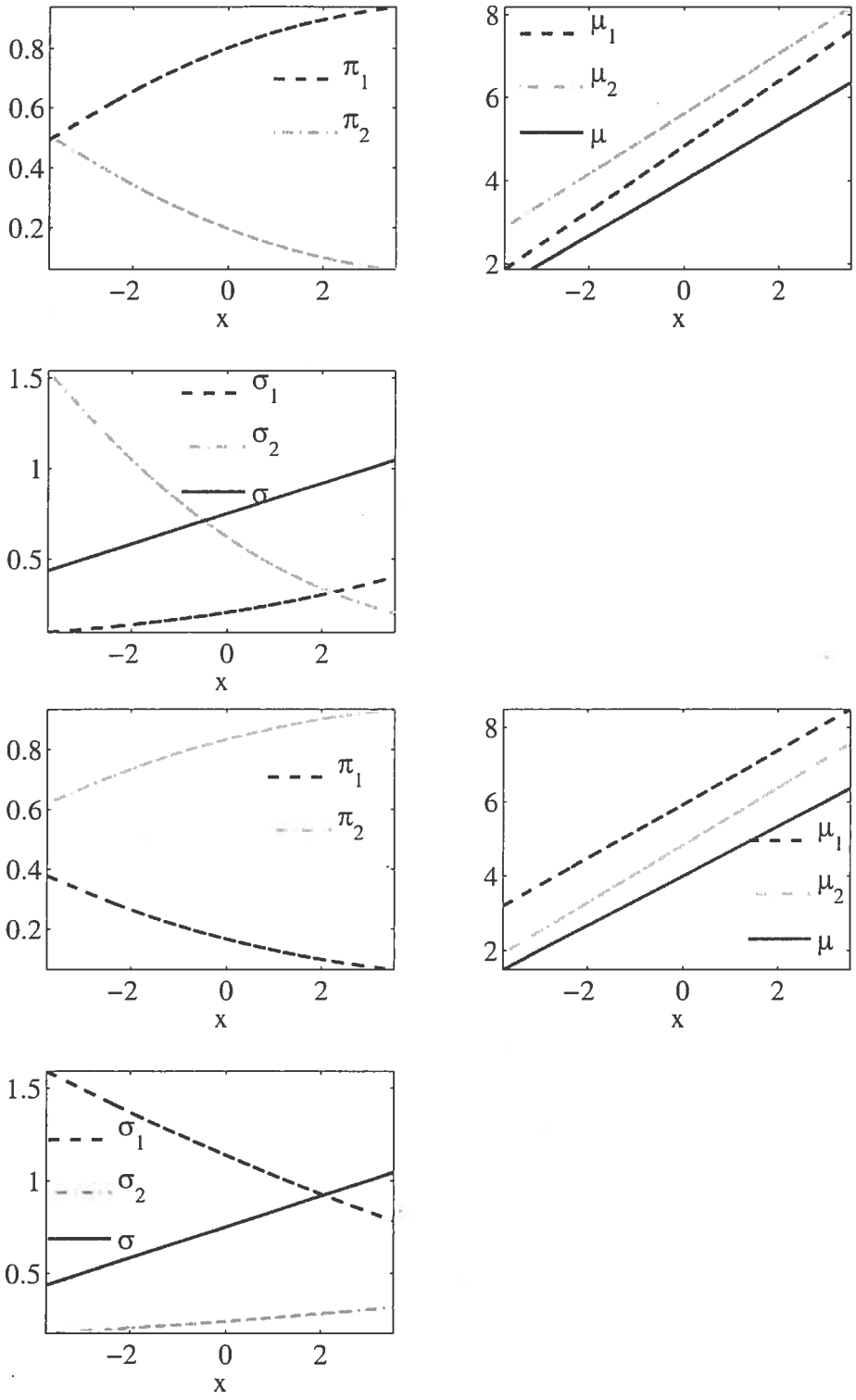

FIG. 4.6: $(\pi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMMG avec deux composantes. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-lin-mod. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

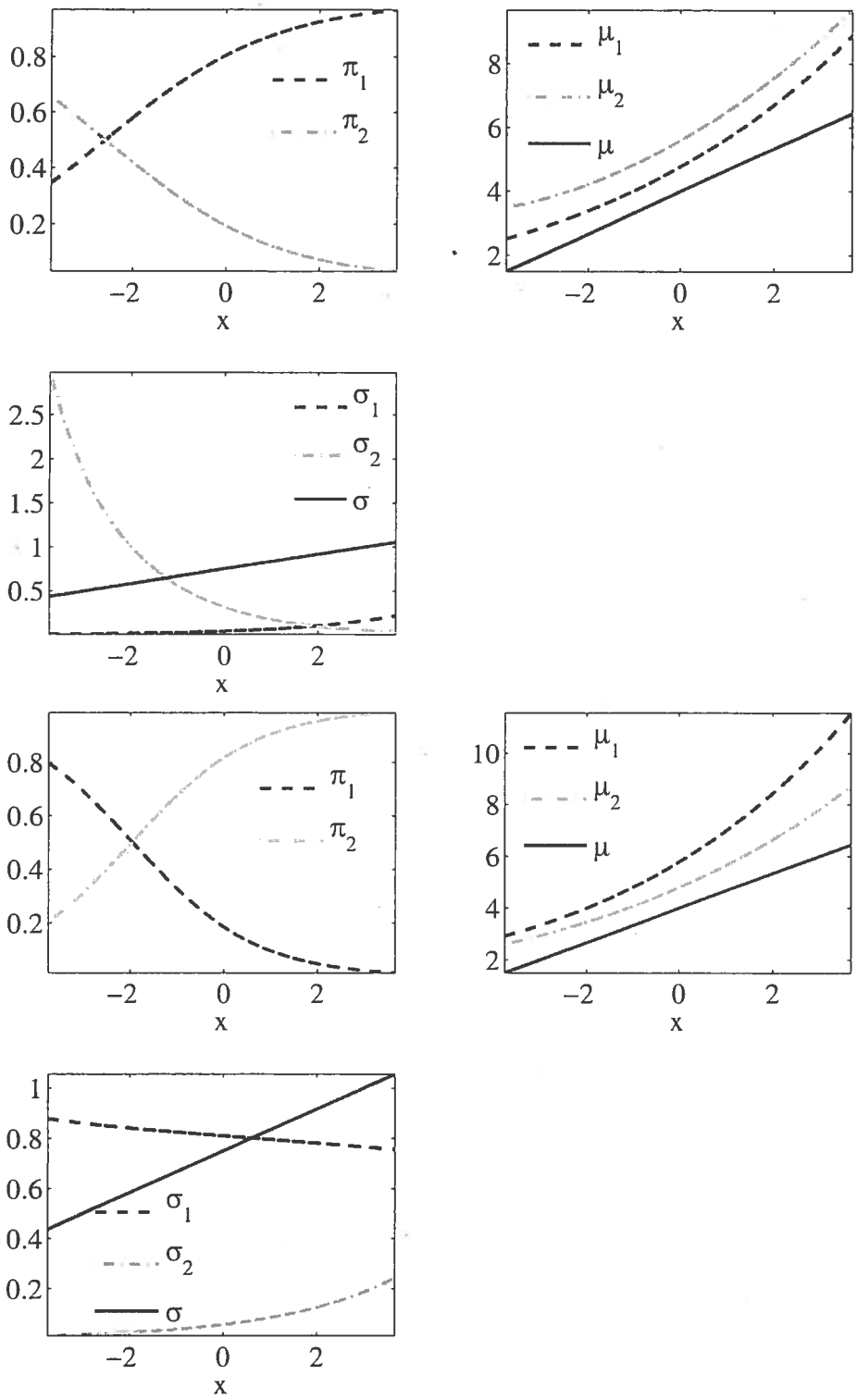

FIG. 4.7: $(\pi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMML avec deux composantes. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-lin-mod. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

4.2.3.2. *Génération de données : modèle générateur Fréchet-lin-mod*

Les figures 4.8 et 4.9 contiennent des données générées par chacun des modèles conditionnels dont les paramètres et hyper-paramètres ont été appris sur des données générées selon une loi de Fréchet conditionnelle avec dépendance linéaire des paramètres et queue modérément lourde (modèle Fréchet-lin-mod). Ces figures peuvent être comparées aux données originales des figures 4.3 et 4.4. Du fait que l'indice de queue est grandement sur-estimé lorsque l'ensemble d'entraînement contient 200 observations, le mélange conditionnel de Pareto hybrides donne lieu à de très grandes observations lorsque x est négatif. Lorsque l'ensemble d'entraînement augmente à 2 000 observations, les données générées par ce modèle ont alors le même type de disposition que les données vues en entraînement (voir la figure 4.3) à l'exception d'un point qui est très éloigné. Le mélange conditionnel avec composantes gaussiennes présente les mêmes caractérisques que le mélange inconditionnel. Les données générées sont très centrales. Lorsque l'ensemble d'entraînement atteint 2 000 observations, ce modèle génère des données légèrement plus excentrées. Cependant, on voit alors apparaître l'autre problème lié aux composantes gaussiennes : la densité est non-négligeable dans la queue inférieure de la distribution et des observations sont générées dans cette région (sous la partie linéaire dans la colonne de droite, rangée du milieu de la figure 4.8). Dans le cas du mélange conditionnel avec composantes Log-Normales, à la dernière rangée de la figure 4.8, on observe aussi peu d'extrêmes. On remarque également une légère courbe dans les données. L'estimateur de Parzen conditionnel, rangée du haut de la figure 4.9, capte assez bien la relation linéaire entre les données, particulièrement lorsque l'ensemble d'entraînement contient 2000 observations. Pour la méthode PoT conditionnelle, rangée du bas de la figure 4.9, seuls les excès au-delà du seuil sont générés. Malgré que le seuil ne dépende de la variable d'entrée x que par la fonction de régression $f(x, \theta)$ et que l'indice de queue soit le même pour tout x, il semble que le modèle générateur de cette version de la PoT conditionnelle soit semblable à celui de la Fréchet qui a servi à générer les données d'entraînement.

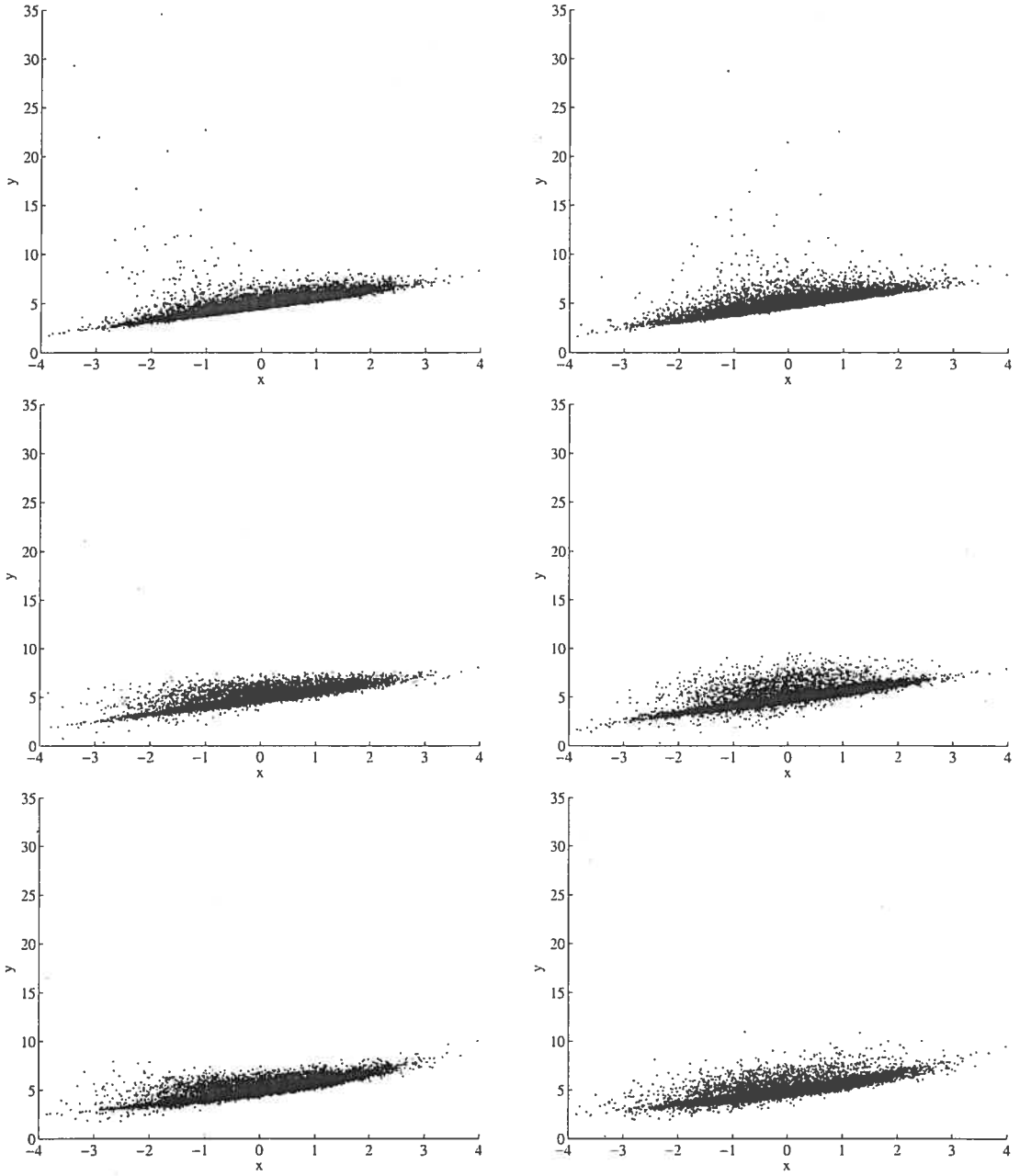

FIG. 4.8: *De haut en bas : Génération de données à partir de CMMH, CMMG et CMML. L'ensemble d'entraînement consistait en 200 (colonne de gauche) ou 2 000 observations (colonne de droite) générées à partir du modèle Fréchet-lin-mod.*

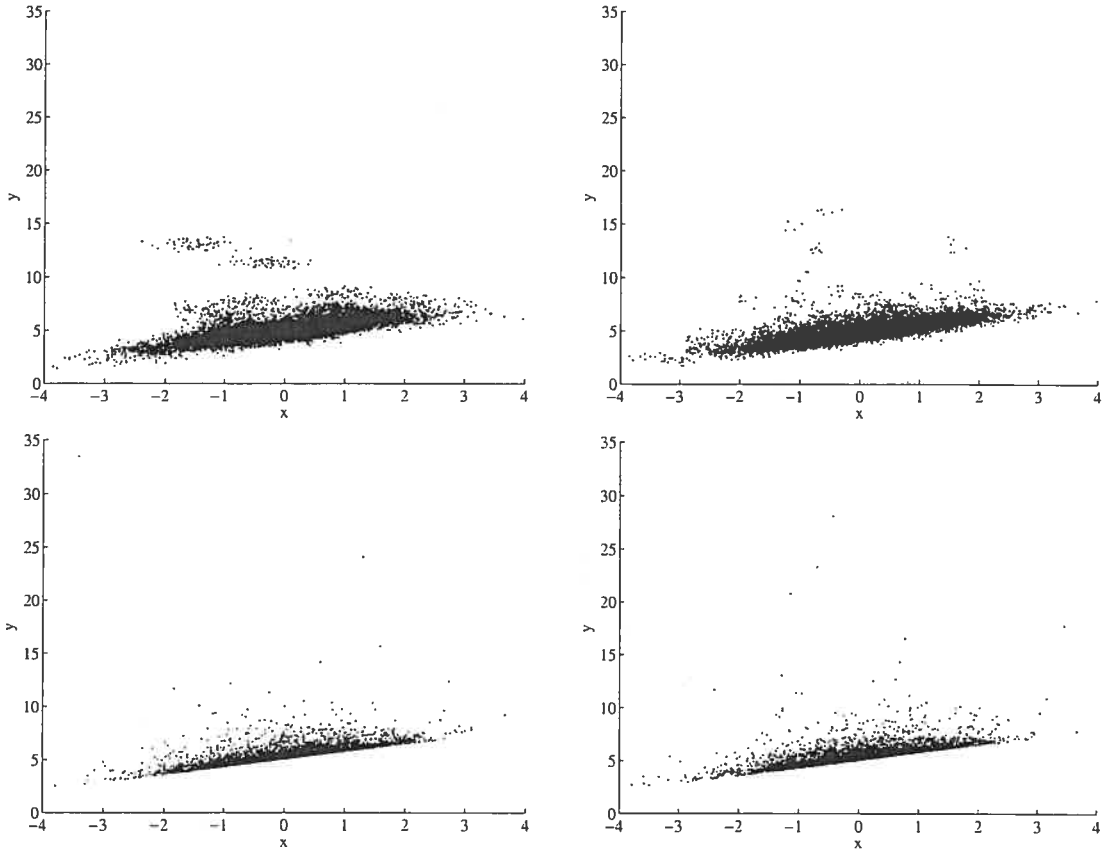

FIG. 4.9: *Génération de données à partir de CPARZEN (rangée du haut) et d'excès par CPOT (rangée du bas). L'ensemble d'entraînement consistait en 200 (colonne de gauche) ou 2 000 observations (colonne de droite) générées à partir du modèle Fréchet-lin-mod.*

4.2.3.3. *Paramètres des mélanges conditionnels : modèle générateur Fréchet-sin-mod*

Les figures 4.10, 4.11 et 4.12 représentent les paramètres des mélanges conditionnels avec composantes Pareto hybrides, Gaussiennes et Log-Normales respectivement pour les données Fréchet conditionnelle avec dépendance sinusoïdale des paramètres et queue modérée (modèle Fréchet-sin-mod). On y observe des phénomènes similaires au cas de dépendance linéaire. En particulier, la moyenne est assez bien capturée par les composantes, quel qu'en soit le type. La modélisation est plus difficile lorsque $|x|$ est grand puisqu'étant généré par une loi Normale, moins d'observations se trouvent dans cette région.

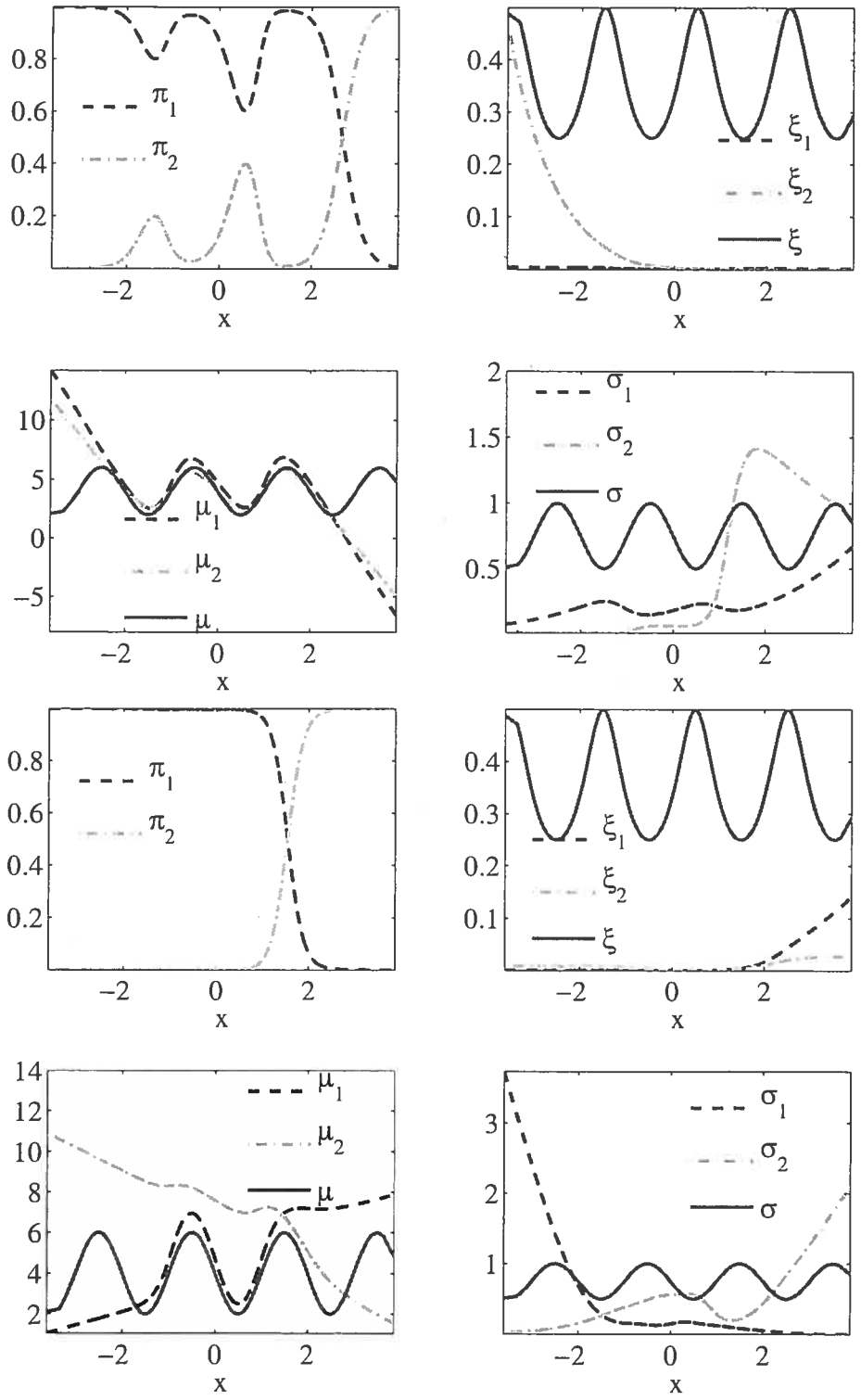

FIG. 4.10: $(\pi_j(x), \xi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMMH avec deux composantes. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-sin-mod. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

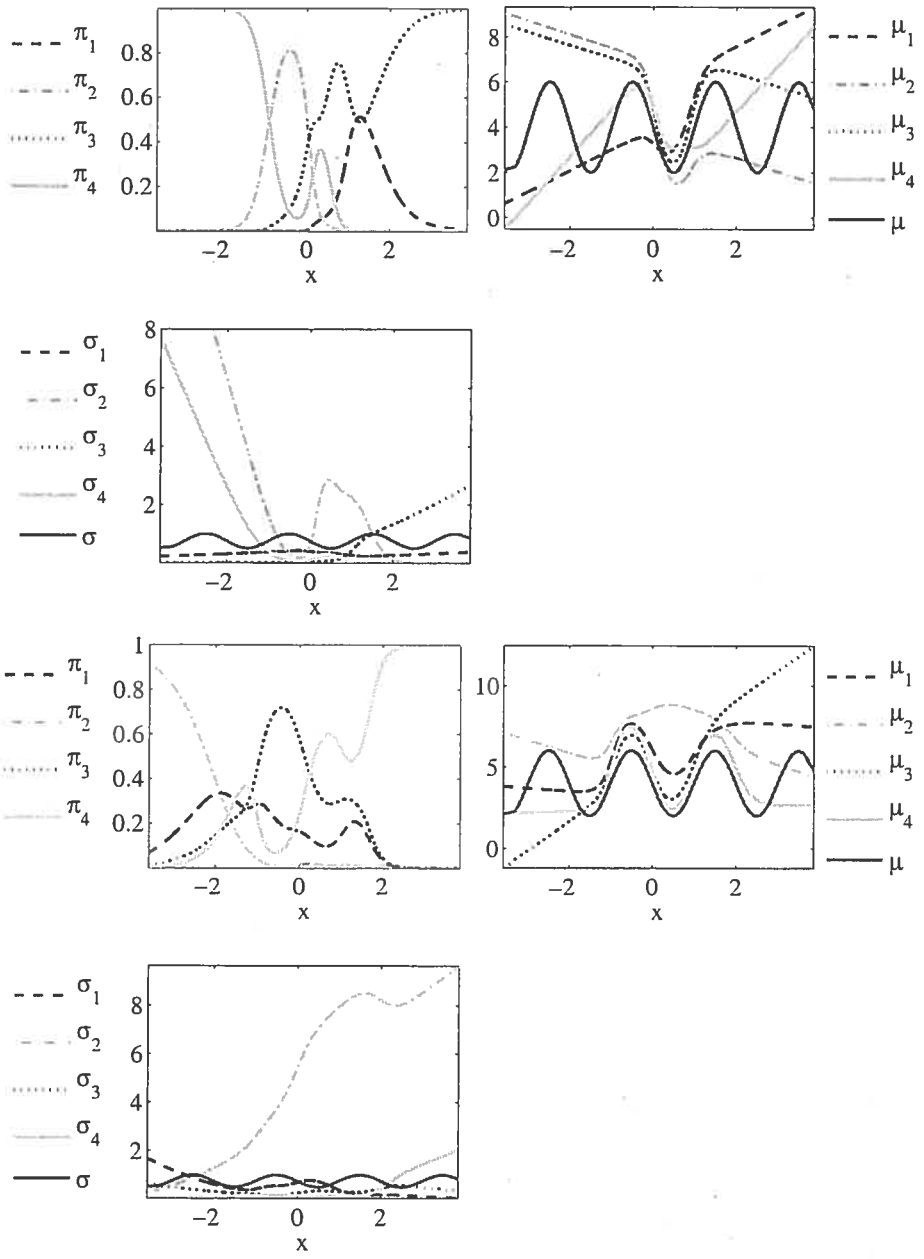

FIG. 4.11: $(\pi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMMG avec quatre composantes. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-sin-mod. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

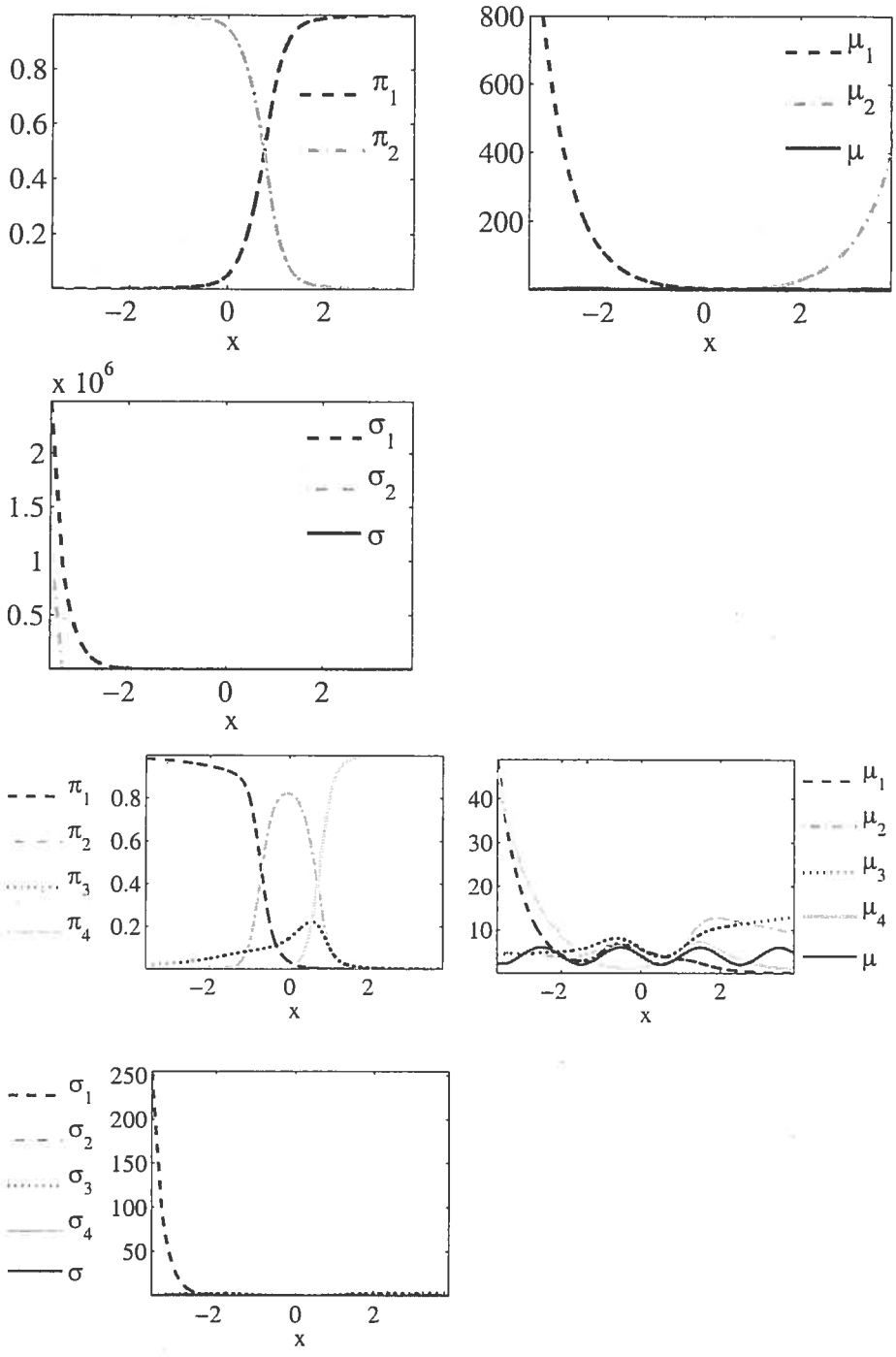

FIG. 4.12: $(\pi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMML avec quatre composantes. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-sin-mod. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

4.2.3.4. *Génération de données : modèle générateur Fréchet-sin-mod*

Les figures 4.13 et 4.14 représentent les données générées par chacun des modèles de densités conditionnels après l'entraînement pour le modèle générateur est Fréchet-sin-mod. Ces figures peuvent être comparées aux données originales des figures 4.3 et 4.4. Les mélanges conditionnels de la figure 4.13 ont bien capté la dépendance sinusoïdale dans la partie centrale mais il y a clairement plus d'incertitude quant à cette dépendance lorsque $|x|$ est grand. Le mélange conditionnel de Pareto hybrides semble avoir modélisé relativement bien le modèle générateur. On observe à nouveau pour le mélange conditionnel gaussien la présence d'observations dans la queue inférieure de la distribution (rangée du milieu de la figure 4.13). Dans ce cas-ci, le mélange conditionnel de Log-Normales produit des observations particulièrement élevées. L'estimateur de la fenêtre de Parzen (voir la rangée du haut de la figure 4.14) détecte bien la relation sinusoïdale entre les données mais il ne génère pas d'observations extrêmes. La méthode PoT conditionnelle souffre aussi des effets de bords, lorsque $|x|$ est grand.

4.2.4. Estimation de l'indice de queue et de quantiles extrêmes

La figure 4.15 illustre l'indice de queue estimé moyen pour le mélange conditionnel de Pareto hybrides et pour la méthode PoT conditionnelle. Pour cette dernière, puisque l'indice de queue est estimé sur les résidus de la régression qui sont supposés indépendants, l'indice de queue ne dépend pas de x. L'indice de queue du modèle générateur est tracé en trait plein. On remarque que l'estimation de l'indice de queue dans le cas conditionnel est beaucoup plus ardu que dans le cas inconditionnel. Le mélange de Pareto hybrides sur-estime l'indice de queue dans les extrémités du domaine et le sous-estime dans la partie centrale. Par conséquent, l'estimation de quantiles conditionnels présente des caractéristiques semblables. Pour la plus petite taille d'ensemble d'entraînement, les indices de queues estimés par le mélange conditionnel de Pareto hybrides sont particulièrement élevés pour permettre la modélisation des extrêmes. Ceci donne lieu à des quantiles extrêmes estimés très grands. Par exemple, pour une loi de Pareto généralisée centrée à zéro et de paramètres $\xi = 0.9$ et $\beta = 1.5$, les quantiles de niveaux

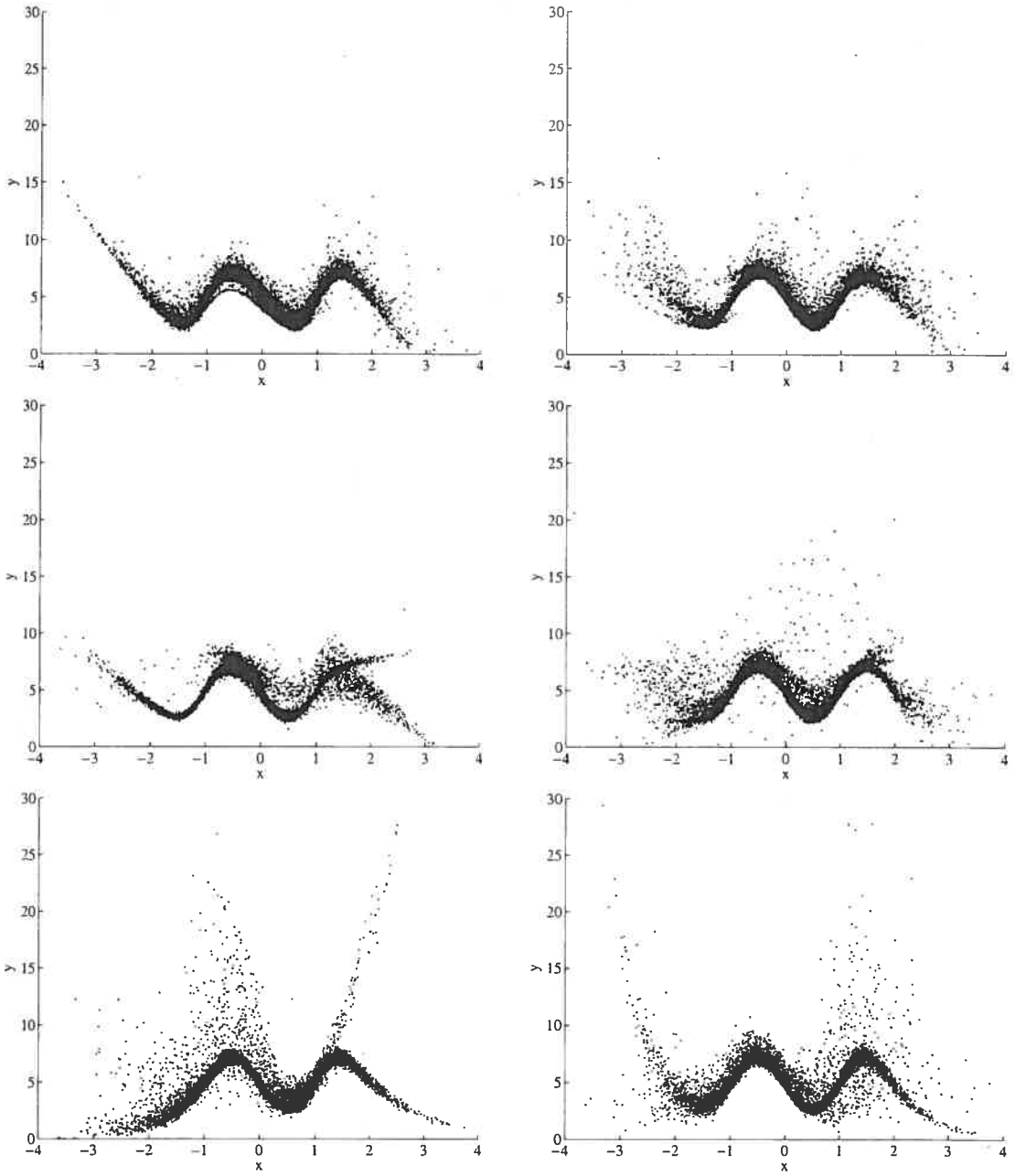

FIG. 4.13: *De haut en bas : Génération de données à partir de CMMH, CMMG et CMML. L'ensemble d'entraînement consistait en 200 (colonne de gauche) ou 2 000 observations (colonne de droite) générées à partir du modèle Fréchet-sin-mod.*

0.99, 0.999 et 0.9999 sont respectivement 103.5, 833.6 et 6 633. Nous présentons les quantiles estimés conditionnels moyens pour les jeux de données avec 2 000 observations provenant d'une loi de Fréchet avec indice de queue modéré.

Les figures 4.16, 4.17 et 4.18 représentent les quantiles conditionnels moyens de niveaux 0.99, 0.999 et 0.9999 respectivement pour le cas où les paramètres

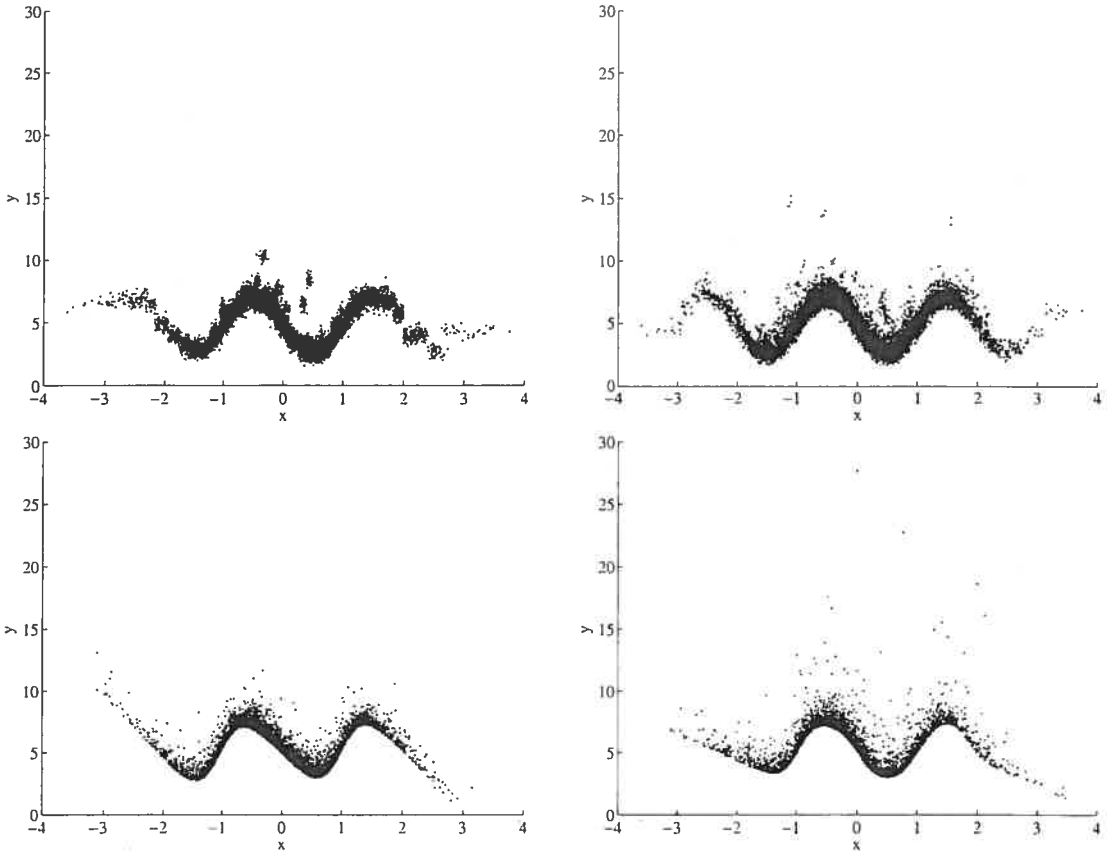

FIG. 4.14: *Génération de données à partir de CPARZEN (rangée du haut) et d'excès par CPOT (rangée du bas). L'ensemble d'entrainement consistait en 200 (colonne de gauche) ou 2 000 observations (colonne de droite) générées à partir du modèle Fréchet-sin-mod.*

dépendent linéairement de l'entrée. Les figures 4.19, 4.20 et 4.21 présentent les mêmes graphiques pour le cas où la dépendance des paramètres à l'entrée est sinusoïdale. Les quantiles conditionnels estimés sont tracés dans la région centrale de l'espace des entrées (pour la loi Normale standard, plus de 95% des observations se trouvent dans l'intervalle $[-2, 2]$). Dans le cas de la dépendance linéaire des paramètres, lorsque le niveau de quantile vaut 0.99, il y a peu de différences entre les différents estimateurs. Cependant, alors que le niveau de quantile augmente, le quantile conditionnel estimé par le mélange de Pareto hybrides se distingue des autres estimateurs. Il suit de plus près le quantile conditionnel du modèle générateur alors que le quantile conditionnel estimé par la PoT est, par construction, linéaire. Les autres estimateurs ont de plus tendance à sous-estimer le quantile

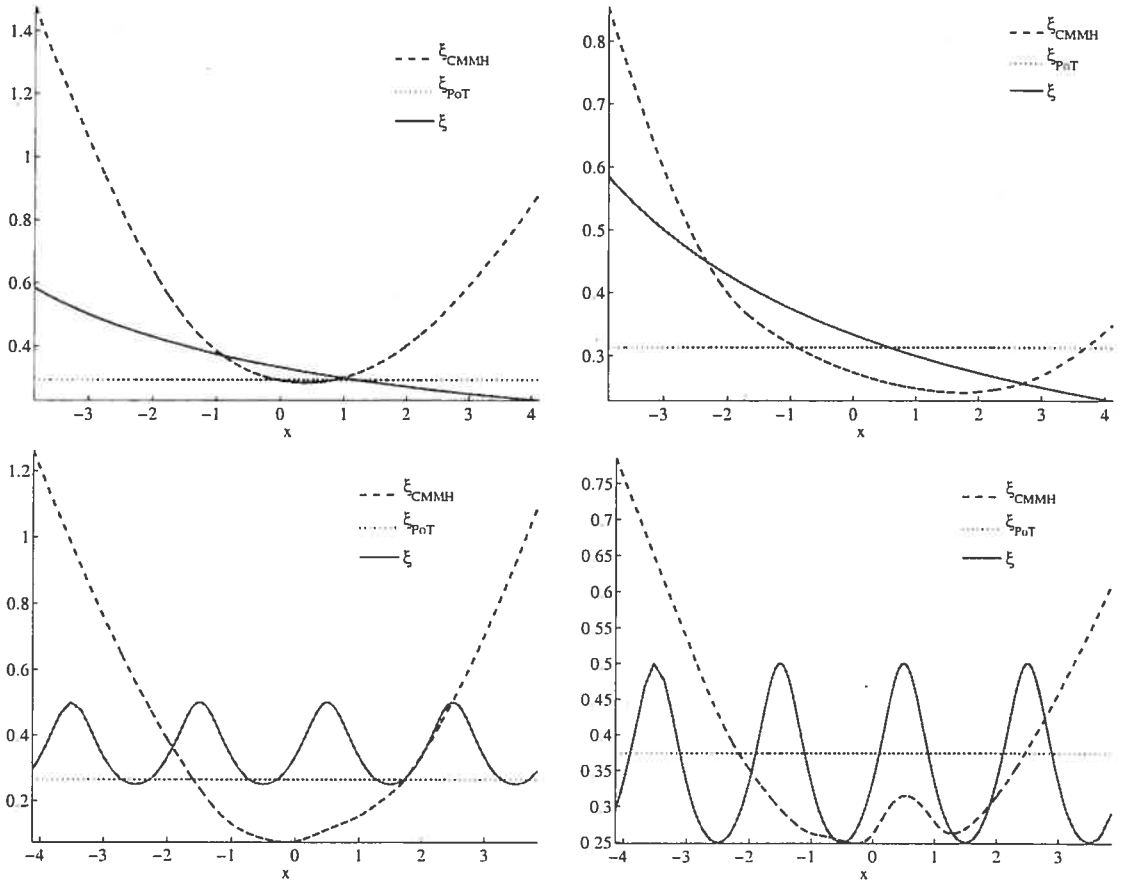

FIG. 4.15: *Indice de queue moyen tel qu'estimé par CMMH et la méthode PoT conditionnelle. Le trait plein représente l'indice de queue du modèle générateur qui est soit le Fréchet-lin-mod (rangée du haut) soit le Fréchet-sin-mod (rangée du bas). L'ensemble d'entraînement contenait 200 (colonne de gauche) ou 2 000 observations (colonne de droite).*

conditionnel. On remarque l'effet de bord pour le mélange de Pareto hybrides, lorsque $q = 0.9999$. Dans le cas de dépendance sinusoïdale, la différence de forme entre les estimateurs et la fonction de quantile conditionnel est encore plus flagrante, voir le figure 4.20. Seul le mélange conditionnel de Pareto hybrides suit de près la fonction de quantile conditionnel. Dans le cas du niveau de quantile de 0.9999, les quantiles estimés par le mélange de Pareto hybrides sont tellement grands qu'ils sortent du cadre du graphique. Dans ce cas, aucun des estimateurs n'est satisfaisant.

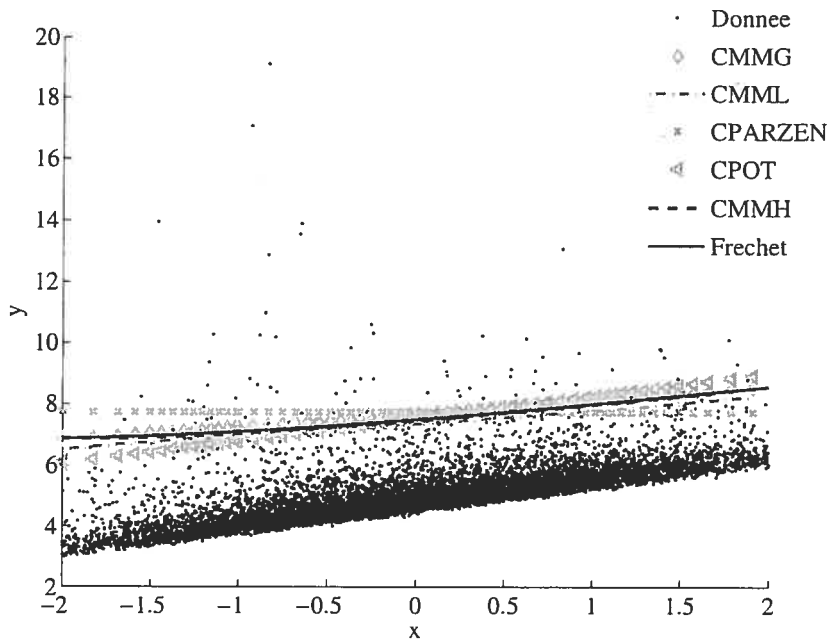

FIG. 4.16: *Quantiles conditionnels moyens de niveau $q = 0.99$ des modèles de densité et de la méthode PoT conditionnelle. L'ensemble d'entraînement contient 2 000 points générés par le modèle Fréchet-lin-mod. Les données de test sont aussi illustrées.*

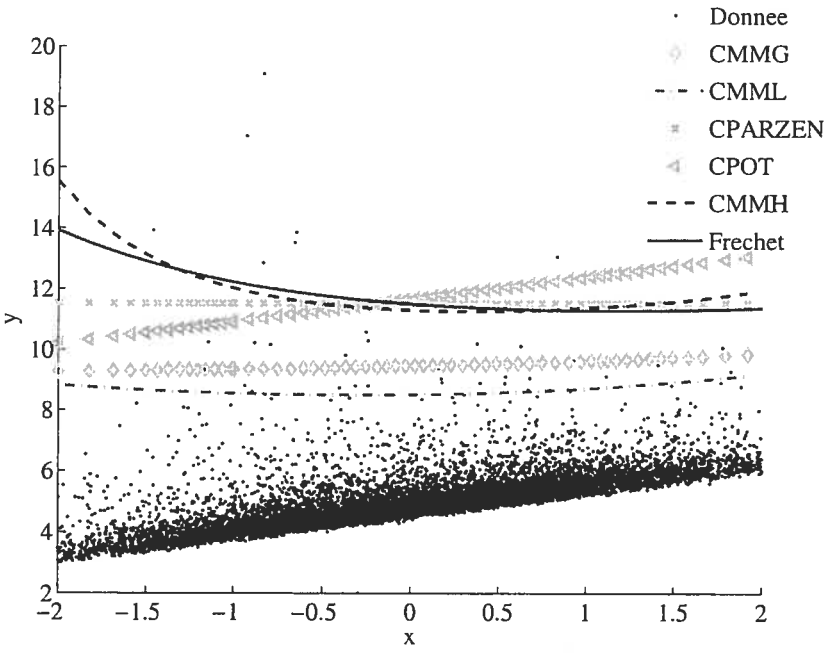

FIG. 4.17: *Quantiles conditionnels moyens de niveau $q = 0.999$ des modèles de densité et de la méthode PoT conditionnelle. L'ensemble d'entraînement contient 2 000 points générés par le modèle Fréchet-lin-mod. Les données de test sont aussi illustrées.*

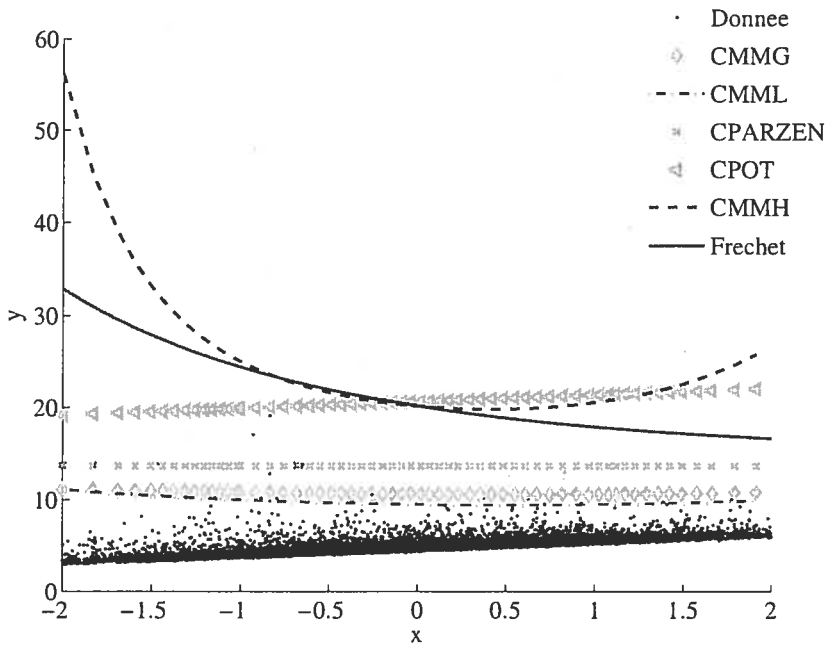

FIG. 4.18: *Quantiles conditionnels moyens de niveau $q = 0.9999$ des modèles de densité et de la méthode PoT conditionnelle. L'ensemble d'entraînement contient 2 000 points générés par le modèle Fréchet-lin-mod. Les données de test sont aussi illustrées.*

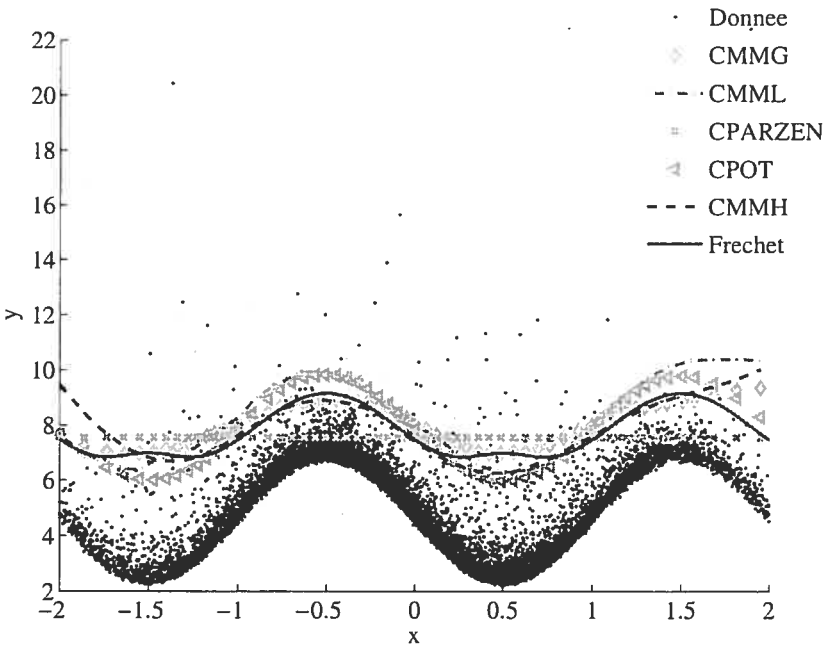

FIG. 4.19: *Quantiles conditionnels moyens de niveau $q = 0.99$ des modèles de densité et de la méthode PoT conditionnelle. L'ensemble d'entraînement contient 2 000 points générés par le modèle Fréchet-sin-mod. Les données de test sont aussi illustrées.*

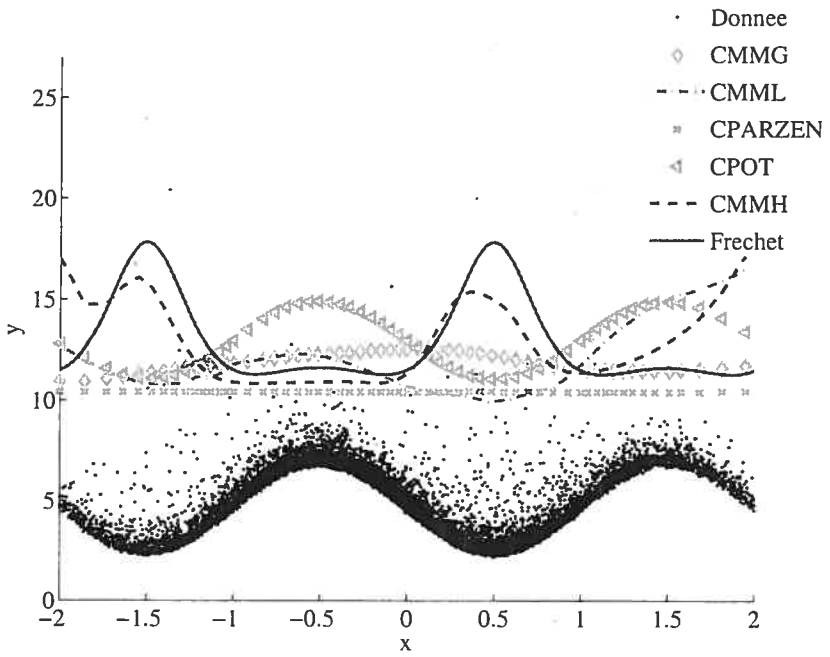

FIG. 4.20: *Quantiles conditionnels moyens de niveau $q = 0.999$ des modèles de densité et de la méthode PoT conditionnelle. L'ensemble d'entraînement contient 2 000 points générés par le modèle Fréchet-sin-mod. Les données de test sont aussi illustrées.*

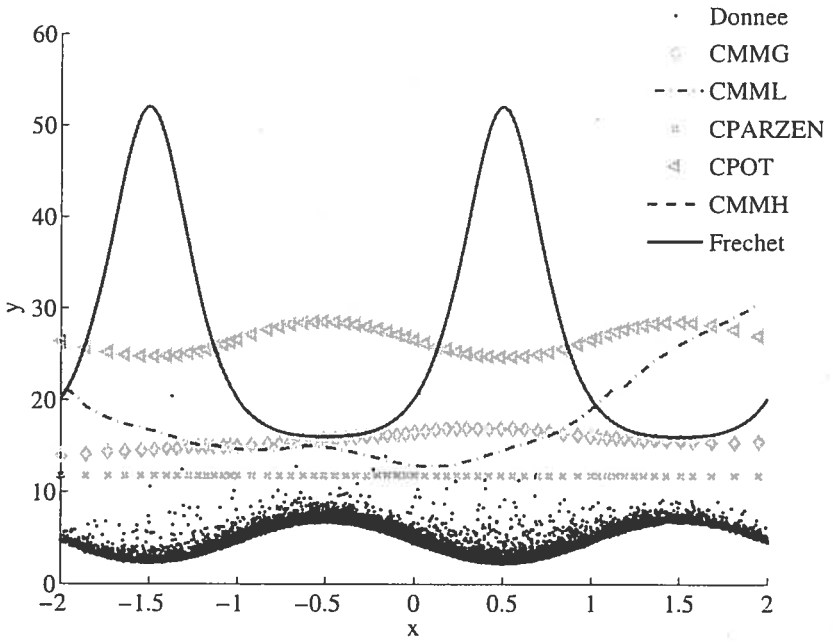

FIG. 4.21: *Quantiles conditionnels moyens de niveau $q = 0.9999$ des modèles de densité et de la méthode PoT conditionnelle. L'ensemble d'entraînement contient 2 000 points générés par le modèle Fréchet-sin-mod. Les données de test sont aussi illustrées.*

4.3. Données d'assurance

Le premier jeu de données réelles sur lequel nous évaluons le modèle de densité conditionnelle proposé provient du domaine des assurances. Ces données nous ont été fournies par une compagnie d'assurance sous couvert de l'anonymat. La variable dépendante Y est le montant de la réclamation d'un client divisé par la durée de la police d'assurance. La variable explicative X est un vecteur de dimension 140 composé essentiellement de variables binaires qui décrivent le profil du client. La distribution complète des réclamations inclut un point de masse à zéro, c'est-à-dire qu'il y a une probabilité positive qu'un client ne fasse aucune réclamation. On peut aborder ce problème de la façon suivante. Dans un premier temps, il s'agit d'un problème de classification : étant donné le profil du client X, quelle est la probabilité que celui-ci fasse une réclamation. Formellement, on cherche à estimer $P(Y = 0|X)$, la probabilité que le client ne fasse pas de réclamation, et $P(Y > 0|X)$, la probabilité qu'il en fasse une. Dans ce dernier cas, il faut ensuite estimer la distribution des réclamations positives, $P(Y|X, Y > 0)$. C'est cette partie du problème qui nous intéresse. Nous disposons des données d'assurance pour une année ; en ne considérant que les clients qui ont fait une réclamation, nous avons un total de 54 119 observations. La figure 4.22 illustre la distribution des réclamations positives inférieures à 5 000$ au moyen d'un histogramme. On y voit clairement la nature multimodale de la distribution. Aussi, il est apparent que la queue supérieure de la distribution est lourde. La plus grande réclamation est de l'ordre de 10^6 alors que la moyenne des réclamations est de l'ordre de 10^3. La présence de valeurs extrêmes influence la moyenne de sorte que 75% des réclamations sont inférieures à la moyenne.

4.3.1. Entraînement et critère d'évaluation

Nous avons utilisé l'analyse en composantes principales [**26**] pour réduire la taille de l'entrée. Nous retenons un nombre de composantes suffisant pour permettre d'expliquer au moins 90% de la variance. Selon la taille de l'ensemble d'entraînement, ce nombre varie de 49 à 69. Nous comparons le mélange conditionnel de Pareto hybrides en tant qu'estimateur de densité conditionnel aux mélanges

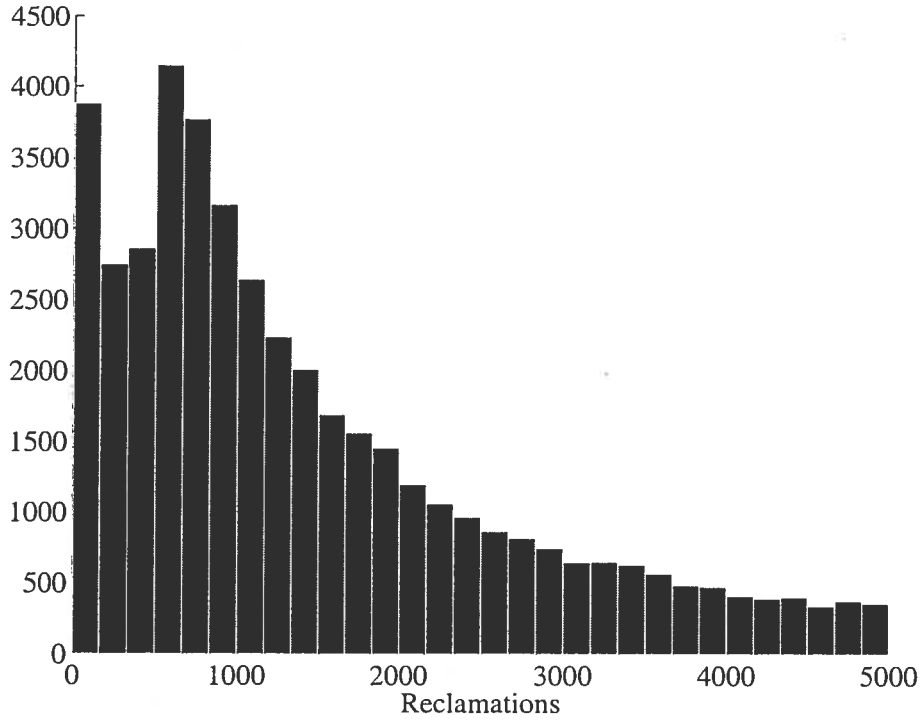

FIG. 4.22: *Histogramme des réclamations positives d'assurance dont le montant est inférieur à 5 000$.*

conditionnels avec composantes Gaussiennes ou Log-Normales ainsi qu'à l'estimateur conditionnel de la fenêtre de Parzen. Les trois types de mélanges sont entraînés par maximisation de la vraisemblance conditionnelle. Pour éviter que l'optimisation ne s'enlise dans un minimum local, l'optimisation est relancée cinq fois avec des paramètres initiaux différents. Les paramètres ayant fait atteindre la plus petite erreur d'entraînement sont retenus. L'initialisation des mélanges conditionnels se fait selon le schéma décrit à la sous-section 4.1.4. Les hyperparamètres des mélanges conditionnels (le nombre d'unités cachées du réseau de neurones et le nombre de composantes du mélange) et de l'estimateur conditionnel de la fenêtre de Parzen (les largeurs de fenêtre dans l'espace d'entrée et de sortie) sont choisis sur l'ensemble de validation qui constitue 20 % de l'ensemble d'entraînement. Pour cette expérience, nous avons choisi trois tailles d'ensemble d'entraînement soit 200, 2 000 et 20 000. Comme la queue de la distribution est très lourde, l'utilisation d'un test statistique pour déterminer un gain significatif en performance entre deux modèles mène au rejet de tous les modèles. En effet, même si la performance est meilleure en moyenne, la variance est très grande.

La sélection de modèle se fait donc uniquement en mesurant une augmentation de la performance moyenne. La généralisation des modèles sélectionnés ainsi est supérieure à celle des modèles sélectionnés par un test statistique (en fait, ce test retient le modèle le plus simple). Cependant, la sélection des hyper-paramètres est moins robuste.

Comme il s'agit de données réelles et que la distribution génératrice est inconnue, la performance est mesurée en termes de log-vraisemblance relative au mélange conditionnel de Pareto hybrides :

$$\mathcal{R}(x, y) = \log(\phi_\theta^{\text{cmmh}}(x, y)) - \log(\phi_\theta^{\text{alt}}(x, y)),$$

où $\phi_\theta^{\text{cmmh}}$ est l'estimateur du mélange conditionnel de Pareto hybrides et ϕ_θ^{alt} est un estimateur alternatif.

Nous utilisons la version conditionnelle de la méthode PoT présentée dans l'étude simulatoire, section 4.2, pour comparer la performance du mélange conditionnel de Pareto hybrides dans la queue de la distribution. Le nombre d'unités cachées de la PoT conditionnelle est choisi en validation alors que le niveau de quantile déterminant le seuil est sélectionné à l'aide du test d'adéquation de la Pareto généralisée [6].

4.3.2. Résultats des expériences

Les tableaux 4.4 et 4.5 contiennent respectivement la log-vraisemblance relative au mélange conditionnel de Pareto hybrides sur l'ensemble de test avec l'erreur standard entre parenthèses et les hyper-paramètres sélectionnés pour les trois tailles d'ensembles d'entraînement. On note cmmh, cmmg et cmml les mélanges conditionnels avec composantes Pareto hybrides, gaussiennes ou Log-Normales respectivement. L'estimateur de Parzen conditionnel est noté cparzen. On remarque que le mélange conditionnel de Pareto hybrides fournit une meilleure performance pour toutes les tailles d'ensemble d'entraînement. La différence de performance est particulièrement frappante pour le plus petit ensemble d'entraînement. Pour l'ensemble d'entraînement contenant 20 000 observations, la différence de performance avec le mélange conditionnel de gaussiennes n'est pas significative. On verra cependant que les courbes de densité conditionnelles générées par

ces deux modèles sont très différentes. Pour les mélanges conditionnels, il y a deux façons de contrôler la complexité du modèle, soit par le nombre d'unités cachées, soit par le nombre de composantes. Dans le tableau 4.5, on remarque qu'un nombre positif d'unités cachées est sélectionné lorsque l'ensemble d'entraînement est de petite taille. Lorsque l'ensemble d'entraînement est grand, une dépendance linéaire suffit à modéliser les données. Ce résultat contre-intuitif (un modèle plus complexe pour un ensemble de données plus petit) s'explique par la grande variabilité de la sélection de modèles lorsque l'ensemble d'entraînement est petit.

n	cmmg	cmml	cparzen
200	2.774 (0.1312)	0.1983 (0.01412)	78.2 (56.23)
2 000	0.3562 (0.06264)	0.7903 (0.02752)	68.45 (55.18)
20 000	*0.0687 (0.03117)*	0.03902 (0.01603)	54.55 (50.44)

TAB. 4.4: *Log-vraisemblance moyenne relative au mélange conditionnel de Pareto hybrides (cmmh) sur l'ensemble de test pour les données d'assurances. La taille de l'ensemble d'entraînement est n. Une valeur positive de log-vraisemblance relative signifie que cmmh donne de meilleurs résultats que le modèle alternatif envisagé. Le résultat en italique dénote une performance qui n'est pas significativement positive.*

n	$(h_{\text{cmmh}}, m_{\text{cmmh}})$	$(h_{\text{cmmg}}, m_{\text{cmmg}})$	$(h_{\text{cmml}}, m_{\text{cmml}})$	$(\lambda^x_{\text{cparzen}}, \lambda^y_{\text{cparzen}})$
200	(5, 2)	(15, 2)	(15, 16)	(1e+05, 1e+06)
2 000	(0, 2)	(0, 4)	(5, 1)	(1e+05, 1e+06)
20 000	(0, 8)	(0, 16)	(0, 2)	(1, 1e+06)

TAB. 4.5: *Hyper-paramètres sélectionnés en validation pour les modèles d'estimation de densité conditionnelle correspondant au tableau 4.4. Pour les mélanges conditionnels, h_{cmm} est le nombre d'unités cachées et m_{cmm} est le nombre de composantes. Pour l'estimateur de Parzen conditionnel, $\lambda^x_{cparzen}$ est la largeur de fenêtre dans l'espace des entrées et $\lambda^y_{cparzen}$ la lageur de fenêtre dans l'espace des sorties.*

Les figures 4.23, 4.24, 4.25 et 4.26 donnent un aperçu des densités conditionnelles générées par les mélanges avec composantes Pareto hybrides, gaussiennes et Log-Normales et l'estimateur conditionnel de la fenêtre de Parzen respectivement.

Pour chaque taille d'ensemble d'entraînement, 20 points $\{x_{j_1}, \ldots, x_{j_{20}}\} \subset \mathcal{D}_{\text{test}}$ ont été choisis aléatoirement parmi l'ensemble de test. Les courbes de densité conditionnelles $p(Y|X = x_{j_i})$, $i = 1, \ldots, 20$ sont ensuite tracées pour chacun des modèles. D'une manière générale, ces graphiques montrent, à l'exception de quelques cas, une grande variété dans les courbes de densité conditionnelle. On peut donc conclure qu'il existe bien une relation de dépendance entre l'entrée X et la sortie Y qui est utile à la modélisation de la densité conditionnelle. Pour les mélanges conditionnels, la multimodalité de la partie centrale de la distribution peut être captée soit par le réseau de neurones, qui permet de changer la forme et l'emplacement de la distribution, soit par le nombre de composantes. Pour le mélange de Pareto hybrides, à la colonne de gauche de la figure 4.23, la densité conditionnelle est unimodale et varie selon l'entrée pour les tailles d'ensemble d'entraînement de 200 et 2 000. À 20 000, on observe plusieurs modes dans la densité conditionnelle. Ceci est cohérent avec les hyper-paramètres sélectionnés, voir le tableau 4.5. La densité du mélange conditionnel de gaussiennes, à la colonne de gauche de la figure 4.24, affiche clairement deux types de densités conditionnelles unimodales dont l'emplacement et l'écart-type varient selon l'entrée, ce qui produit une densité bi-modale. Ce phénomène est moins présent pour le plus grand ensemble d'entraînement. La densité conditionnelle du mélange conditionnel avec composantes Log-Normales a de nombreux artéfacts lorsque l'ensemble d'entraînement est petit, voir la colonne de gauche de la figure 4.25. Ceci est causé par le grand nombre de composantes sélectionné. Pour les deux autres tailles d'ensemble d'entraînement, la densité conditionnelle est uni-modale avec la hauteur et l'emplacement du mode qui changent selon l'entrée. Finalement, l'estimateur conditionnel de la fenêtre de Parzen, figure 4.26, ne semble pas capter d'aspect conditionnel dans la densité pour les tailles d'ensemble d'entraînement de 200 et 2 000. À 20 000, on observe au contraire des densités très différentes d'une entrée à l'autres, certaines uni-modales et d'autres bi-modales. En ce qui a trait à la queue supérieure de la distribution, la colonne de droite des graphiques, le mélange conditionnel de Pareto hybrides produit des queues parfois légères parfois lourdes. Comme attendu, le mélange à composantes gaussiennes génère des

queues de distributions décroissant rapidement vers zéro. La queue supérieure du mélange à composantes Log-Normales est généralement lourde. L'estimateur conditionnel de la fenêtre de Parzen extrapole difficilement au-delà des données vues en entraînement et la queue supérieure de cet estimateur décroît donc très rapidement. Ce phénomène s'atténue lorsque la taille de l'ensemble d'entraînement atteint 20 000.

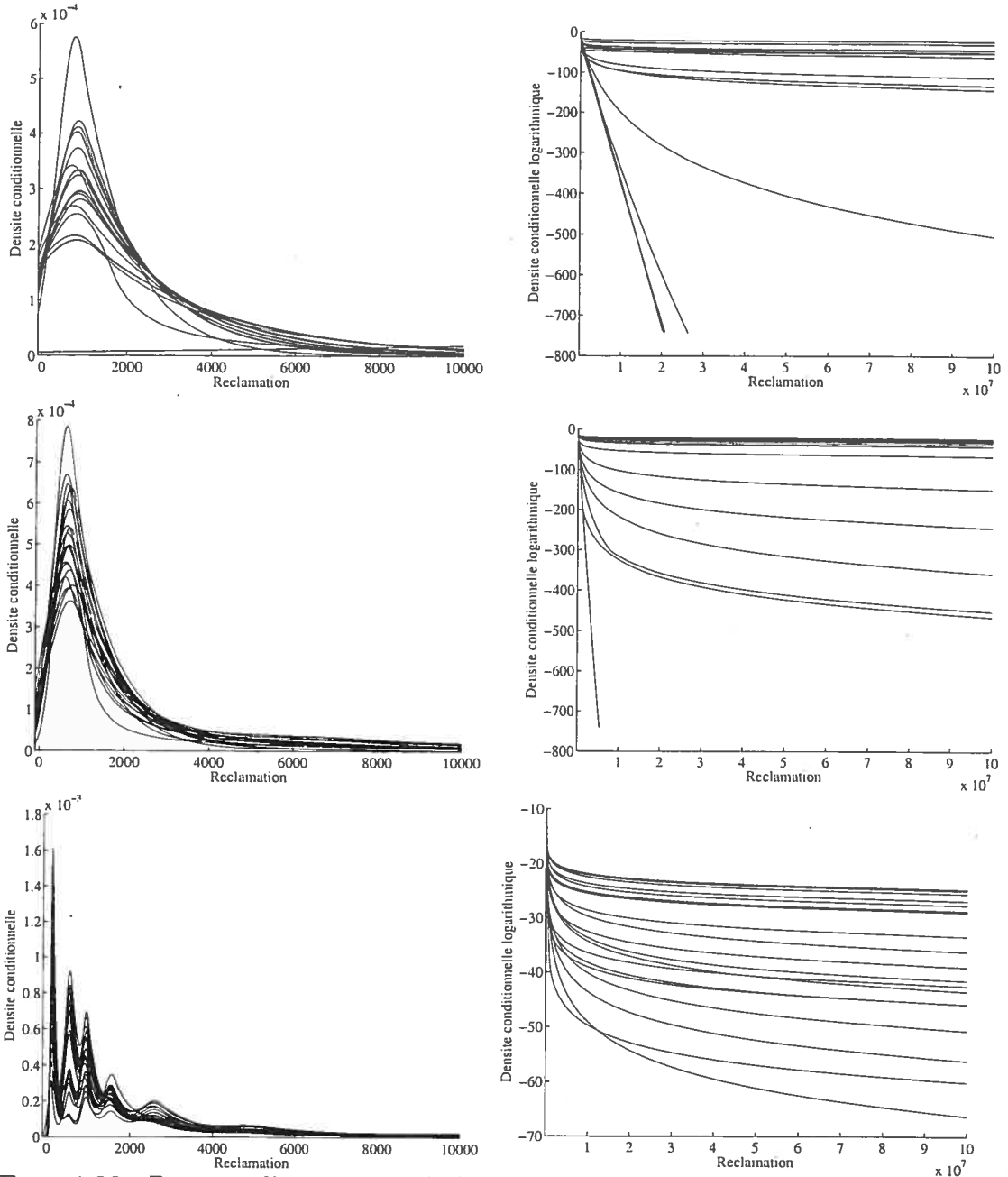

FIG. 4.23: *Données d'assurance : de haut en bas, densité conditionnelle du modèle CMMH pour 20 points de l'ensemble de test choisis aléatoirement pour les tailles d'ensemble d'entraînement de 200, 2 000 et 20 000 respectivement. La colonne de gauche illustre la partie centrale de la densité et celle de droite la queue supérieure en échelle logarithmique.*

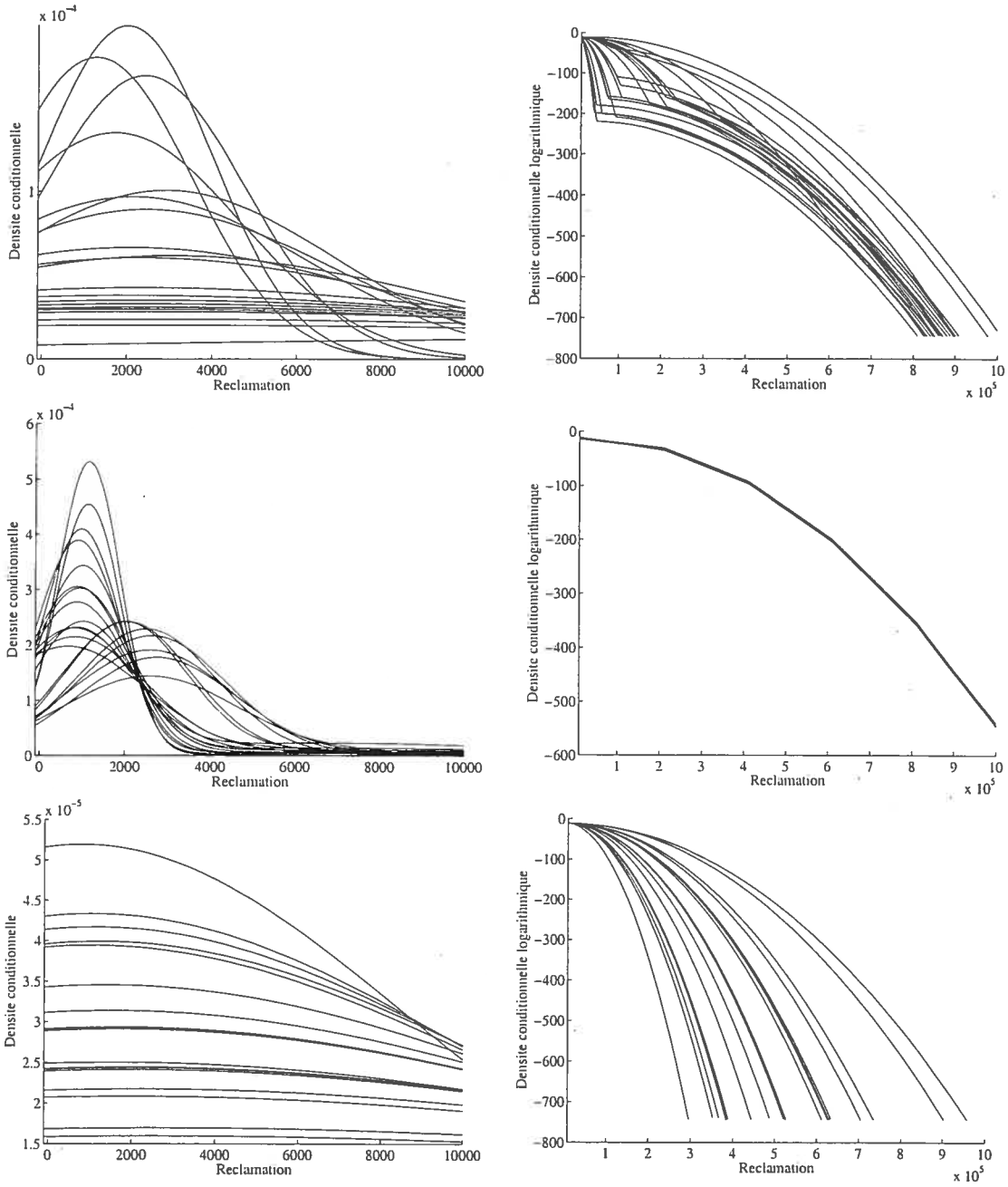

FIG. 4.24: *Données d'assurance : de haut en bas, densité conditionnelle du modèle CMMG pour 20 points de l'ensemble de test choisis aléatoirement pour les tailles d'ensemble d'entraînement de 200, 2 000 et 20 000 respectivement. La colonne de gauche illustre la partie centrale de la densité et celle de droite la queue supérieure en échelle logarithmique.*

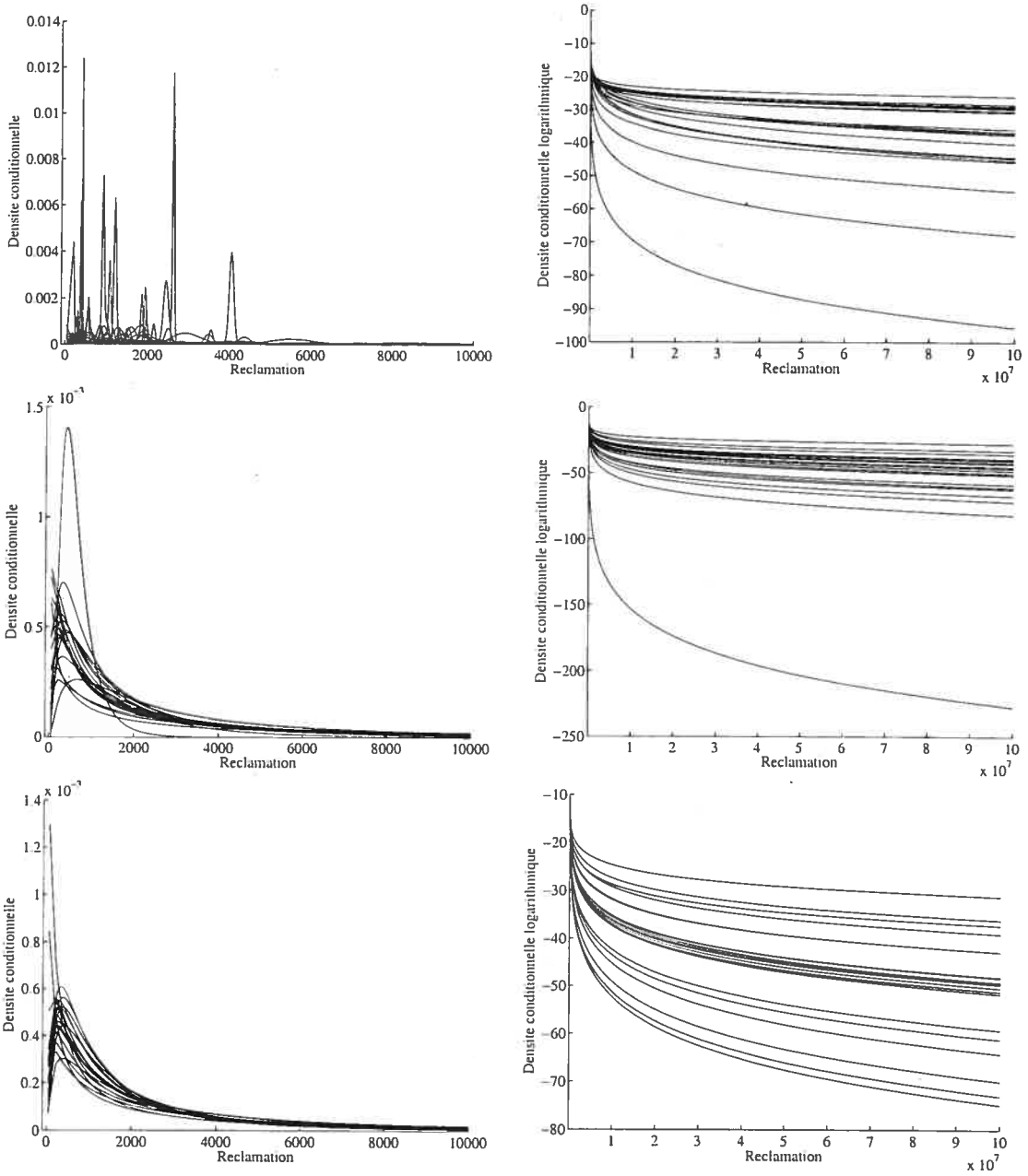

FIG. 4.25: *Données d'assurance : de haut en bas, densité conditionnelle du modèle CMML pour 20 points de l'ensemble de test choisis aléatoirement pour les tailles d'ensemble d'entraînement de 200, 2 000 et 20 000 respectivement. La colonne de gauche illustre la partie centrale de la densité et celle de droite la queue supérieure en échelle logarithmique.*

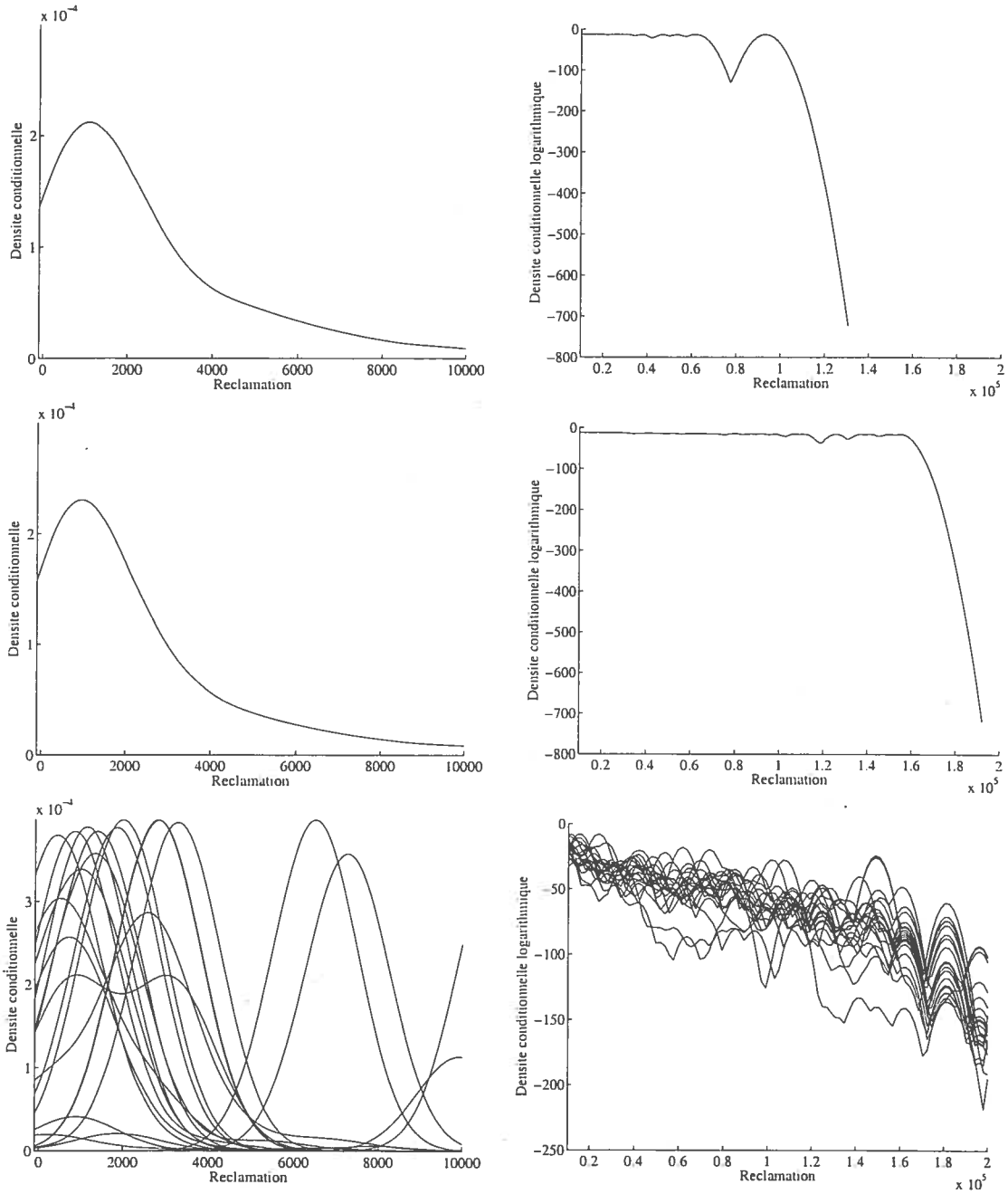

FIG. 4.26: *Données d'assurance : de haut en bas, densité conditionnelle du modèle CPARZEN pour 20 points de l'ensemble de test choisis aléatoirement pour les tailles d'ensemble d'entraînement de 200, 2 000 et 20 000 respectivement. La colonne de gauche illustre la partie centrale de la densité et celle de droite la queue supérieure en échelle logarithmique.*

Le tableau 4.6 fournit les résultats pour la méthode PoT conditionnelle sur les données d'assurance. Le nombre d'unités cachées ainsi que le niveau de seuil augmentent avec la taille de l'ensemble d'entraînement alors que l'indice de queue estimé diminue. La figure 4.27 illustre la densité conditionnelle provenant de la PoT conditionnelle en 20 points aléatoires de l'ensemble de test. Les courbes de densités sont tracées dans la région de la queue supérieure en échelle logarithmique. Le seuil de la PoT conditionnelle est donné par $u(x) = u + f(x, \theta)$. Les excès au-delà de ce seuil sont considérés comme indépendants et sont modélisés par une Pareto généralisée. Il y a donc un seul indice de queue. Les courbes de densité sont superposées car, malgré que le seuil varie en fonction de x, la région du graphe en est si éloignée qu'il n'y a plus de différence entre les densités. La queue de la densité conditionnelle du mélange conditionnel de Pareto hybrides présente beaucoup de variabilité. Pour certaines valeurs de l'entrée, la queue conditionnelle de ce modèle décroît rapidement vers zéro (décroissance exponentielle) alors que pour d'autres valeurs, cette décroissance est très lente (décroissance polynômiale). Puisque la PoT n'utilise qu'un seul indice de queue, il faut sans doute qu'elle fasse un compromis entre ces différents types de décroissance. La variabilité de l'indice de queue tel qu'estimé par le mélange conditionnel de Pareto hybrides est illustrée à la figure 4.28 au moyen d'histogrammes pour les trois tailles d'ensemble d'entraînement. Dans les trois cas, environ 60 % des indices de queue estimés sont inférieurs à 1. Pour l'ensemble d'entraînement contenant 2 000 points, les indices de queue prennent des valeurs particulièrement grandes.

n	h	u	q_{PoT}	$\hat{\xi}$
200	0	336.2	0.1	0.7343
2 000	5	5133	0.9	0.4817
20 000	5	9579	0.9	0.5097

TAB. 4.6: *Données d'assurance : hyper-paramètres sélectionnés (le nombre d'unités cachées h et le niveau de quantile q_{PoT} déterminant le seuil u) pour la méthode PoT conditionnelle et indice de queue ξ estimé. La taille de l'ensemble d'entraînement est n.*

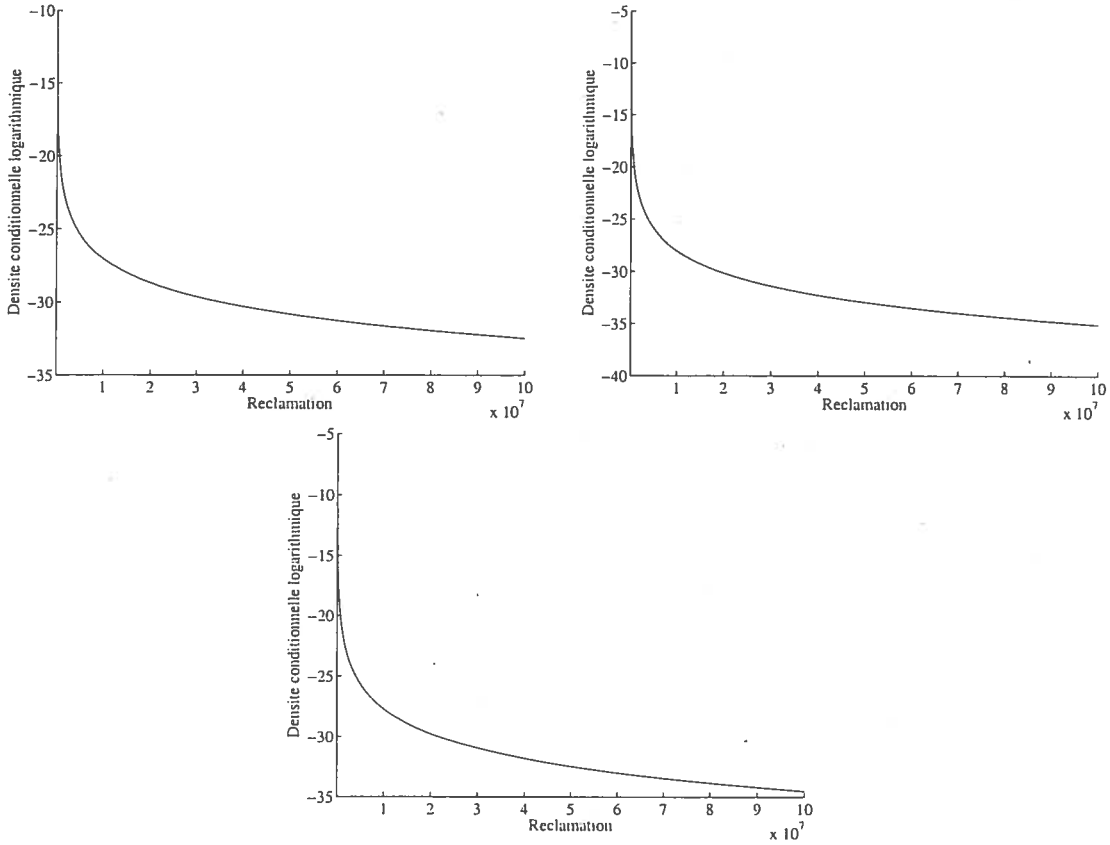

FIG. 4.27: *Données d'assurance : dans le sens des aiguilles d'une montre, densité conditionnelle du modèle CPOT pour 20 points de l'ensemble de test choisis aléatoirement pour l'ensemble d'entraînement de taille 200, 2 000 et 20 000 respectivement. La densité de la queue supérieure est tracée en échelle logarithmique.*

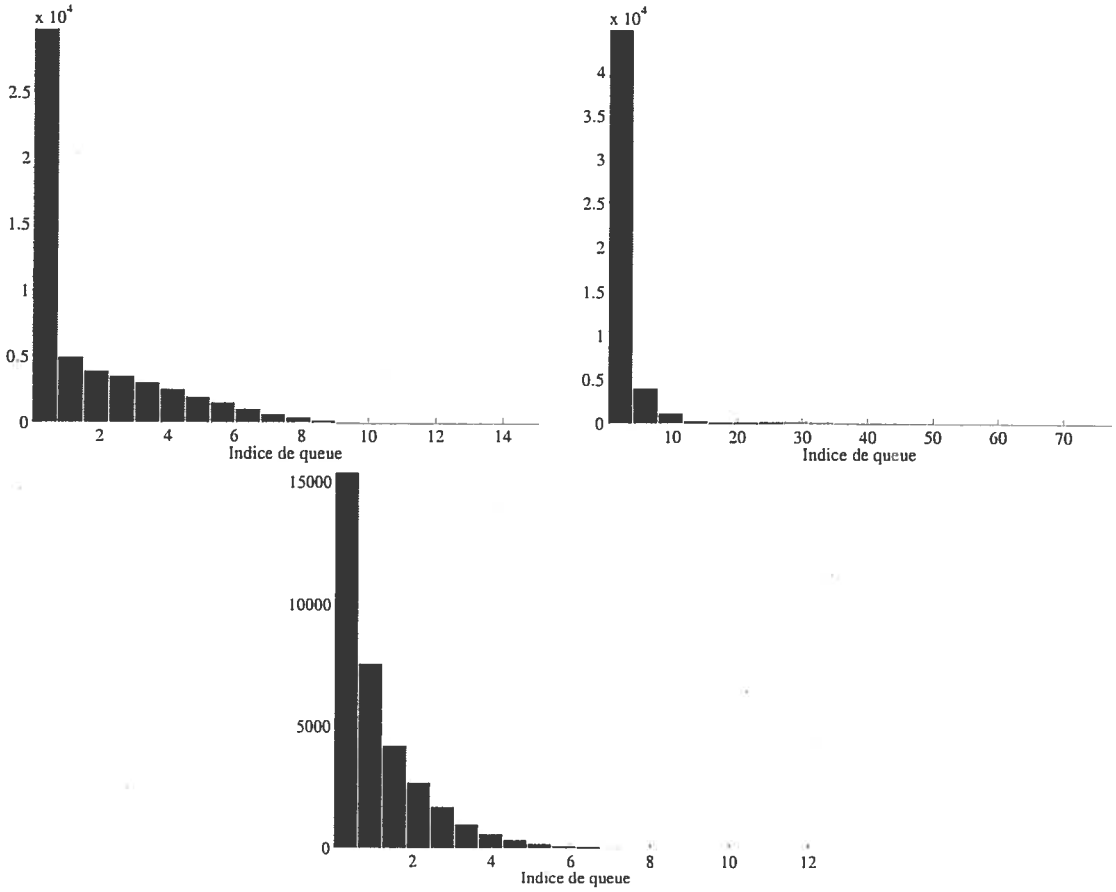

FIG. 4.28: *Données d'assurance : dans le sens des aiguilles d'une montre, indice de queue du modèle CMMH estimé sur l'ensemble de test pour l'ensemble d'entraînement de taille 200, 2 000 et 20 000 respectivement.*

4.4. Données KDD cup 98

Comme deuxième jeu de données réelles, nous utilisons les données fournies par KDD Cup 98 (http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html). Il s'agit de données relatives à une campagne de promotion afin de maximiser les profits provenant de dons. La variable dépendante Y est le montant donné à une organisation nationale de vétérans. La variable explicative X a 479 entrées qui décrivent le profil du donateur. Une variable binaire indique si la personne a répondu ou non à la campagne de promotion : on observe un don positif seulement si cette variable est active (si elle vaut 1). Tout comme pour les données d'assurance (voir la section 4.3), on pourrait d'abord évaluer la probabilité, étant donné le profil X, qu'une personne fasse un don positif. Cependant, nous nous intéressons uniquement à l'estimation de la distribution des dons positifs étant donné le profil, c'est-à-dire de $P(Y|X, Y > 0)$. C'est pourquoi nous n'avons conservé que les profils ayant résulté en un don positif. Le jeu de données que nous utilisons contient 9 716 observations. Un histogramme des dons inférieurs à 50 $ est illustré à la figure 4.29. La variable Y prend souvent, mais pas seulement, des valeurs entières. On note que les montants qui sont des multiples de 5$ sont particulièrement fréquents. Aussi, 75 % des dons sont inférieurs à 20$ alors que le don maximal est de 500$.

4.4.1. Entraînement et critère d'évaluation

Pour diminuer la dimension de la variable explicative, nous avons utilisé la sélection de variables de Georges et Milley [**20**]. Cinq variables servent à décrire le profil. Deux variables, AVGGIFT et LASTGIFT, proviennent directement des données originales. Les trois autres variables sont des transformations de variables d'origine : le don moyen pour les promotions de 94NK à 96NK, le ratio entre NGIFTALL et NUMPROM et la somme de RAMNT_8, 9, 12 and 14. Ces variables explicatives ont de plus été normalisées. Nous reprenons le schéma d'expérience utilisé avec les données d'assurance de la section précédente. Nous comparons le mélange conditionnel de Pareto hybrides avec les mélanges conditionnels de Gaussiennes et de Log-Normales ainsi qu'à l'estimateur conditionnel

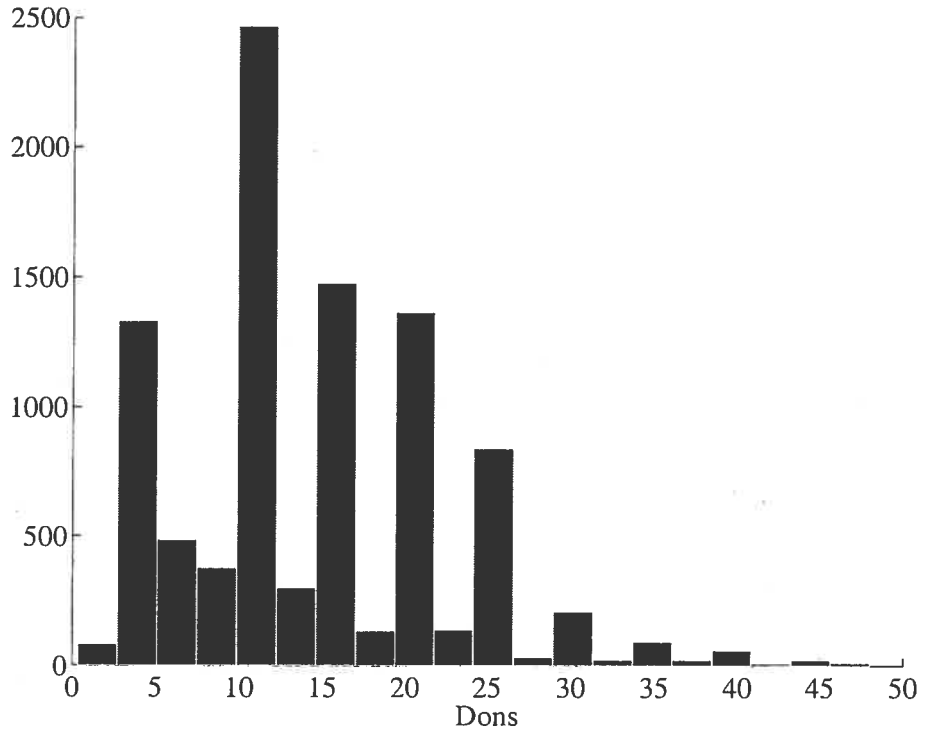

FIG. 4.29: *Histogramme des dons positifs pour KDD cup 98 dont le montant est inférieur à 50$.*

de la fenêtre de Parzen. Les mélanges conditionnels, quel que soit le type de composante utilisée, sont entraînés par maximisation de la log-vraisemblance. Pour éviter d'être pris dans un optimum local, l'optimisation est relancée cinq fois avec des valeurs initiales différentes pour les paramètres du modèle. Les paramètres ayant donné lieu à la plus petite erreur d'entraînement sont retenus. L'initialisation des mélanges se fait selon la méthode décrite à la sous-section 4.1.4. Nous avons recours à nouveau à la méthode PoT conditionnelle qui consiste à ajuster un réseau de neurones aux données pour modéliser la relation entre l'entrée et la sortie et ensuite à appliquer la méthode PoT non-conditionnelle aux résidus du modèle. Le nombre d'unités cachées et de composantes pour les mélanges conditionnels, les largeurs de fenêtres pour Parzen conditionnel et le nombre d'unités cachées pour la PoT conditionnelle sont choisis sur un ensemble de validation qui représente 20% de l'ensemble d'entraînement. Un modèle est sélectionné si sa performance moyenne est meilleure que celle des autres modèles. Le niveau de quantile q_{PoT} servant à déterminer le seuil pour les résidus de la méthode PoT

conditionnelle est choisi à l'aide du test d'adéquation de Choulakian et Stephens [6].

Pour ce jeu de données, nous avons choisi des ensembles d'entraînement de taille 200 et 2 000. Nous comparons les modèles de densité conditionnelle par rapport au mélange conditionnel de Pareto hybrides en termes de log-vraisemblance relative :

$$\mathcal{R}(x,y) = \log(\phi_\theta^{\text{cmmh}}(x,y)) - \log(\phi_\theta^{\text{alt}}(x,y)),$$

où $\phi_\theta^{\text{cmmh}}$ est l'estimateur du mélange conditionnel de Pareto hybrides et ϕ_θ^{alt} est un estimateur alternatif. Nous examinons également les courbes de densité conditionnelle des modèles entraînés pour comparer ces modèles sur la partie centrale et sur la queue supérieure de la distribution.

4.4.2. Résultats des expériences

La moyenne de la log-vraisemblance relative au mélange conditionnel de Pareto hybrides sur l'ensemble de test ainsi que l'erreur standard sont présentées dans le tableau 4.7 pour les deux tailles d'ensemble d'entraînement. Le tableau 4.8 contient les hyper-paramètres sélectionnés pour chaque modèle et pour chaque taille d'ensemble d'entraînement. Dans tous les cas, sauf pour l'estimateur de Parzen conditionnel, la log-vraisemblance relative est significativement positive ce qui signifie que le mélange conditionnel de Pareto hybrides est un meilleur estimateur en termes de log-vraisemblance. L'estimateur de Parzen conditionnel met la majeure partie de la densité autour des exemples d'entraînement et peine à extrapoler au-delà de ces exemples. Ceci explique la grande variance de la log-vraisemblance relative pour cet estimateur. Pour ce jeu de données, le niveau de complexité des modèles semble être plus élevé puisque le nombre d'unités cachées et de composantes sélectionnées est grand.

Les figures 4.30, 4.31, 4.32 et 4.33 contiennent un échantillon de la densité conditionnelle générées par les mélanges avec composantes Pareto hybrides, gaussiennes et Log-Normales et l'estimateur conditionnel de la fenêtre de Parzen respectivement. Pour chaque taille d'ensemble d'entraînement, 20 points $\{x_{j_1}, \ldots, x_{j_{20}}\} \subset$

n	cmmg	cmml	cparzen
200	0.5818 (0.04452)	2.045 (0.02841)	108.3 (47.93)
2 000	1.03 (0.0332)	2.001 (0.02726)	431.2 (405.3)

TAB. 4.7: *Log-vraisemblance moyenne relative au mélange conditionnel de Pareto hybrides (cmmh) sur l'ensemble de test pour les données KDD cup 98. La taille de l'ensemble d'entraînement est n. Une valeur positive de log-vraisemblance relative signifie que cmmh donne de meilleurs résultats que le modèle alternatif envisagé.*

n	$(h_{\text{cmmh}}, m_{\text{cmmh}})$	$(h_{\text{cmmg}}, m_{\text{cmmg}})$	$(h_{\text{cmml}}, m_{\text{cmml}})$	$(\lambda^x_{\text{cparzen}}, \lambda^y_{\text{cparzen}})$
200	(20, 20)	(5, 40)	(10, 8)	(10, 0.1)
2 000	(20, 50)	(10, 20)	(15, 30)	(10, 0.01)

TAB. 4.8: *Hyper-paramètres sélectionnés en validation pour les modèles d'estimation de densité conditionnelle correspondant au tableau 4.7. Pour les mélanges conditionnels, h_{cmm} est le nombre d'unités cachées et m_{cmm} est le nombre de composantes. Pour l'estimateur de Parzen conditionnel, $\lambda^x_{cparzen}$ est la largeur de fenêtre dans l'espace des entrées et $\lambda^y_{cparzen}$ la lageur de fenêtre dans l'espace des sorties.*

$\mathcal{D}_{\text{test}}$ ont été choisis aléatoirement parmi l'ensemble de test. Les courbes de densité conditionnelles $p(Y|X = x_{j_i})$, $i = 1, \ldots, 20$ sont ensuite tracées pour chacun des modèles. On remarque que les modèles ont généralement bien capté l'aspect multi-modal des données. Lorsque l'ensemble d'entraînement augmente, la multimodalité est encore plus présente puisque plus d'exemples variés ont été vus à l'entraînement. Pour les mélanges conditionnels, l'emplacement et la hauteur des modes varient d'une courbe à l'autre, ce qui signifie que d'après le modèle, pour un profil donné, certaines valeurs de dons sont plus probables que d'autres. Par contre, les courbes produites par l'estimateur de la fenêtre de Parzen conditionnel ne reflètent pas l'effet de l'information conditionnante. Plusieurs modes sont apparents qui représentent les valeurs de dons vues en entraînement. On observe les phénomènes habituels au niveau des queues de distribution des modèles : les queues des densités provenant du mélange conditionnel de gaussiennes décroissent rapidement alors que celles des mélanges conditionnels de Pareto hybrides ou de

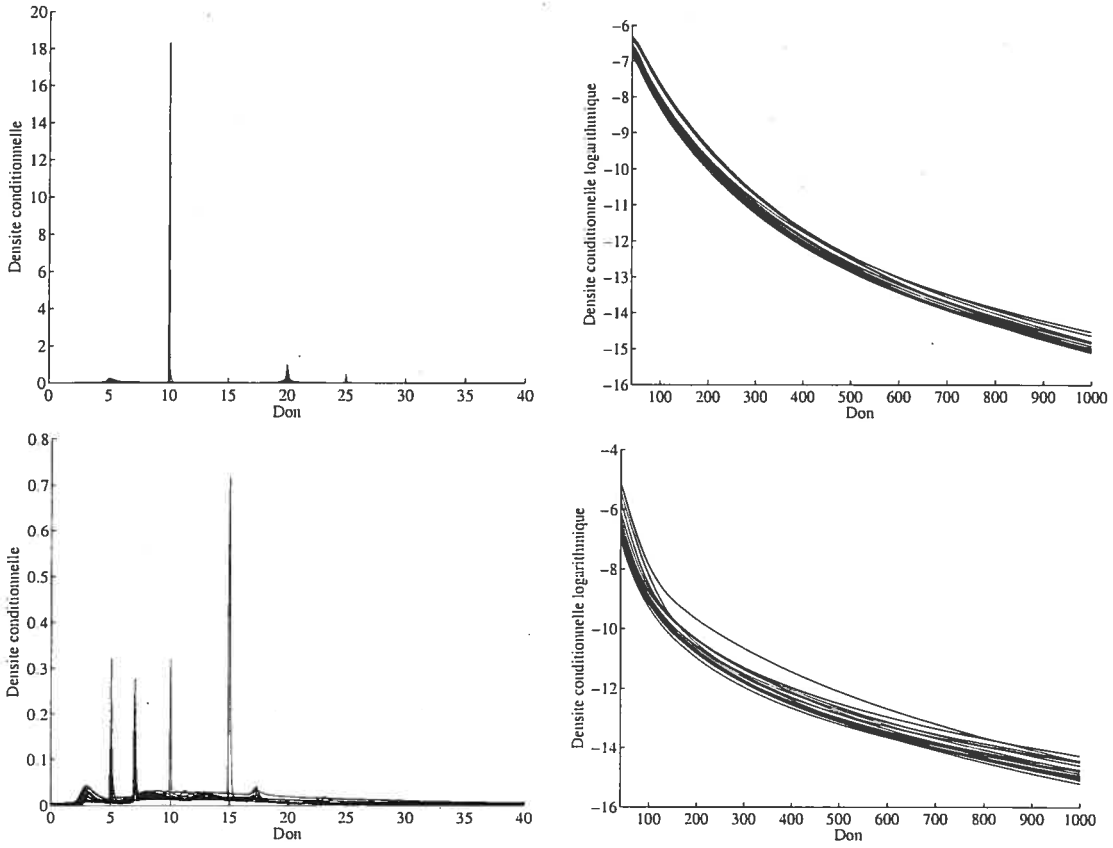

FIG. 4.30: *Données KDD cup 98 : de haut en bas, densité conditionnelle du modèle CMMH pour 20 points de l'ensemble de test choisis aléatoirement pour les tailles d'ensemble d'entraînement de 200 et 2 000 respectivement. La colonne de gauche illustre la partie centrale de la densité et celle de droite la queue supérieure en échelle logarithmique.*

Log-Normales se comportent de façon similaire et décroissent beaucoup plus lentement. L'estimateur conditionnel de la fenêtre de Parzen ne met pratiquement aucune densité dans la queue de la distribution.

Le tableau 4.9 contient les résultats de la sélection de modèles pour la méthode PoT conditionnelle. Pour ces données, le réseau de neurones réussit à apprendre assez bien les données d'entraînement mais il généralise très mal aux données de validation (phénomène du sur-apprentissage). Pour cette raison, le modèle le plus simple est sélectionné, c'est-à-dire que le réseau de neurones est en fait une fonction linéaire de l'entrée. Le seuil de la PoT conditionnelle $u(x) = u + f(x, \theta)$ dépend de x par la fonction calculée par le réseau de neurones. La figure 4.34 illustre les courbes de densité conditionnelle de la PoT conditionnelle en échelle

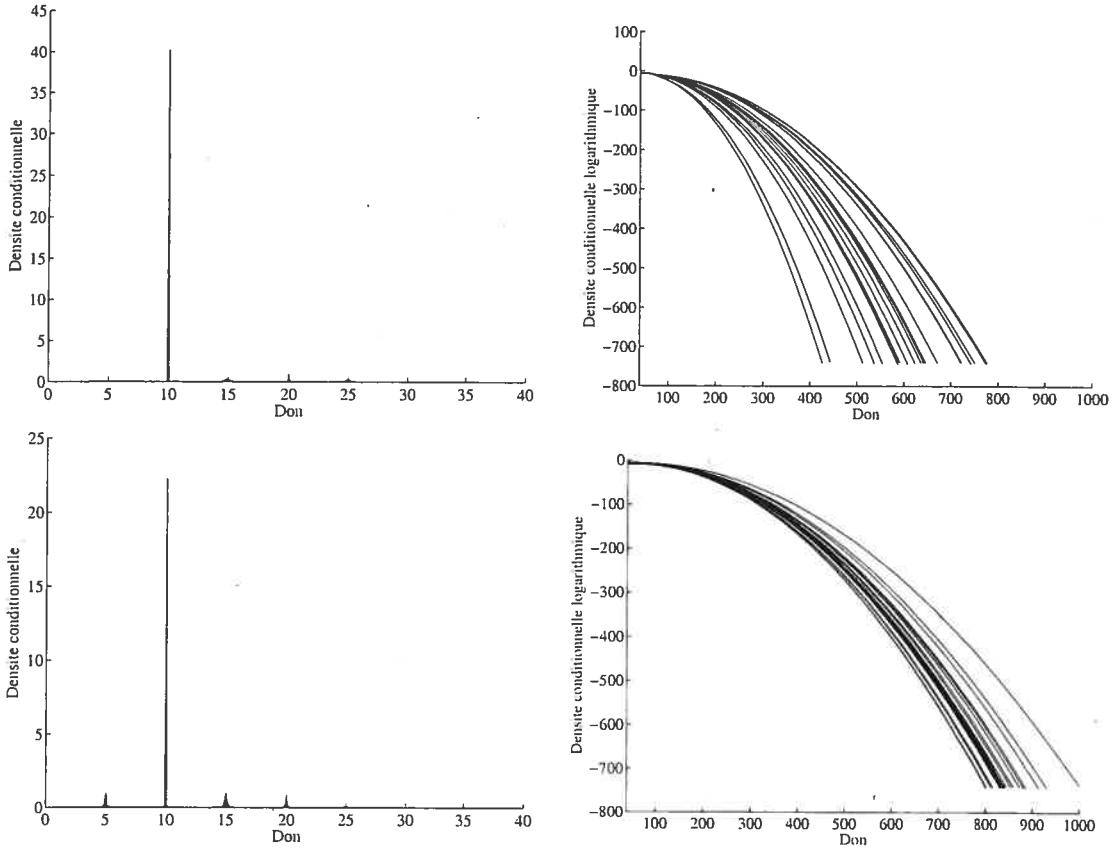

FIG. 4.31: *Données KDD cup 98 : de haut en bas, densité conditionnelle du modèle CMMG pour 20 points de l'ensemble de test choisis aléatoirement pour les tailles d'ensemble d'entraînement de 200 et 2 000 respectivement. La colonne de gauche illustre la partie centrale de la densité et celle de droite la queue supérieure en échelle logarithmique.*

logarithmique dans la queue de la distribution. On observe un peu de variabilité dans les courbes au début de celles-ci. L'épaisseur de la queue de la distribution estimée par la PoT conditionnelle est semblable à celle des mélanges conditionnels avec composantes Pareto hybrides ou Log-Normales bien que ceux-ci introduisent plus de variabilité entre les queues de distribution selon la valeur de l'entrée.

La figure 4.35 illustre à l'aide d'histogramme les indices de queue tels qu'estimés par le mélange conditionnel de Pareto hybrides. La très grande majorité des indices de queue (plus de 90 %) sont inférieurs à un (donc l'espérance de la loi conditionnelle existe). Cependant, pour certaines valeurs de la variable d'entrée, l'indice de queue estimé peut être très grand.

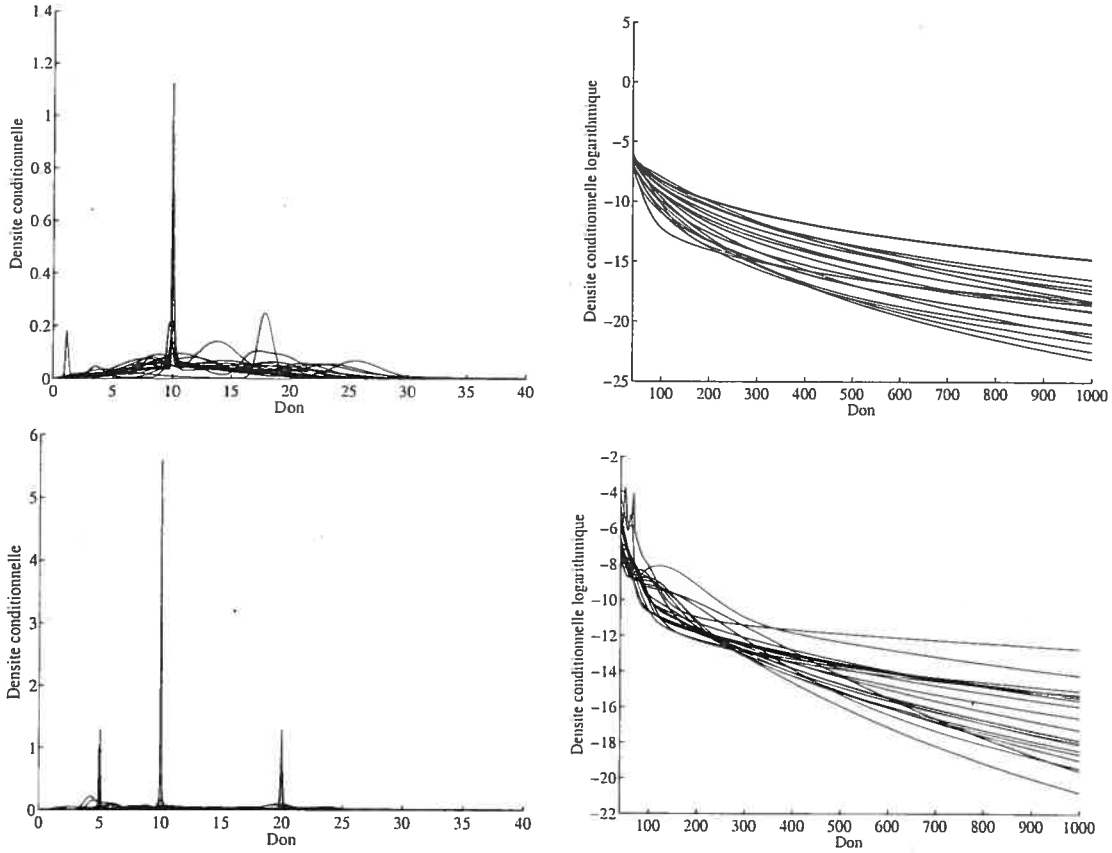

FIG. 4.32: *Données KDD cup 98 : de haut en bas, densité conditionnelle du modèle CMML pour 20 points de l'ensemble de test choisis aléatoirement pour les tailles d'ensemble d'entraînement de 200 et 2 000 respectivement. La colonne de gauche illustre la partie centrale de la densité et celle de droite la queue supérieure en échelle logarithmique.*

n	h	u	q_{PoT}	$\hat{\xi}$
200	0	-0.3973	0.60	0.4318
2 000	5	13.5677	0.95	0.3720

TAB. 4.9: *Données KDD cup 98 : hyper-paramètres sélectionnés (le nombre d'unités cachées h et le niveau de quantile q_{PoT} déterminant le seuil u) pour la méthode PoT conditionnelle et indice de queue ξ estimé. La taille de l'ensemble d'entraînement est n.*

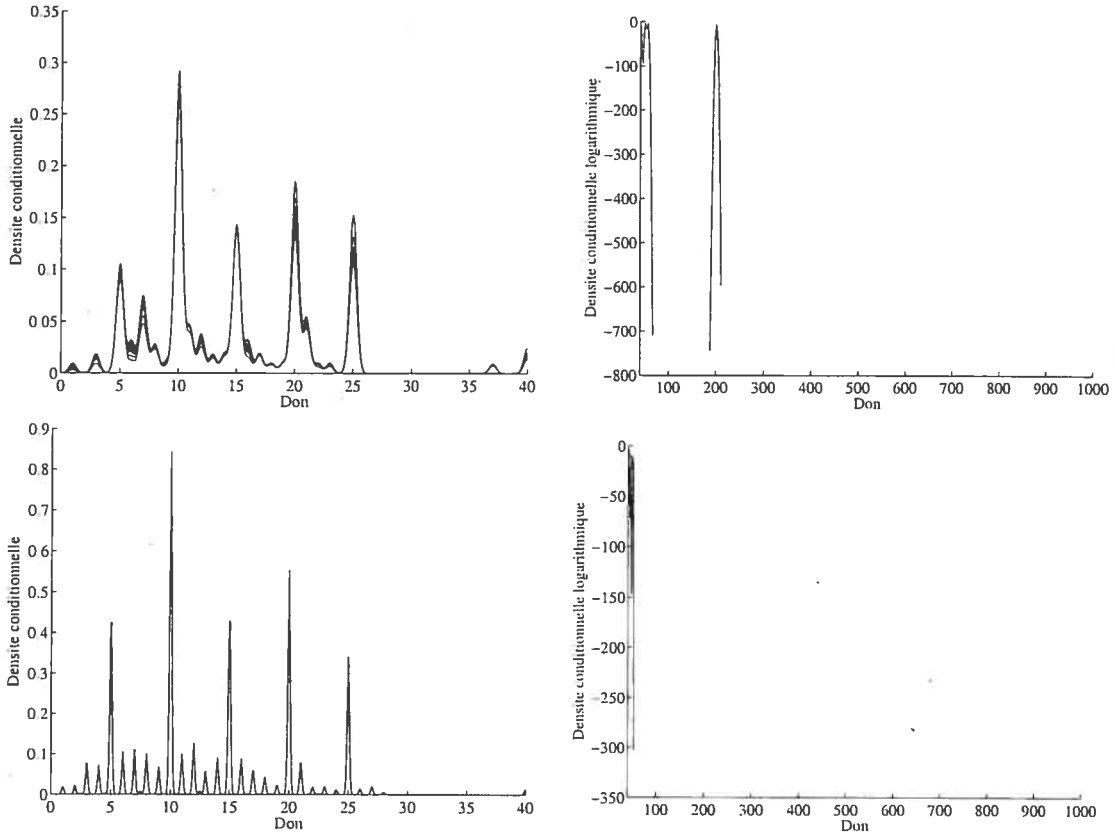

FIG. 4.33: *Données KDD cup 98 : de haut en bas, densité conditionnelle du modèle CPARZEN pour 20 points de l'ensemble de test choisis aléatoirement pour les tailles d'ensemble d'entraînement de 200 et 2 000 respectivement. La colonne de gauche illustre la partie centrale de la densité et celle de droite la queue supérieure en échelle logarithmique.*

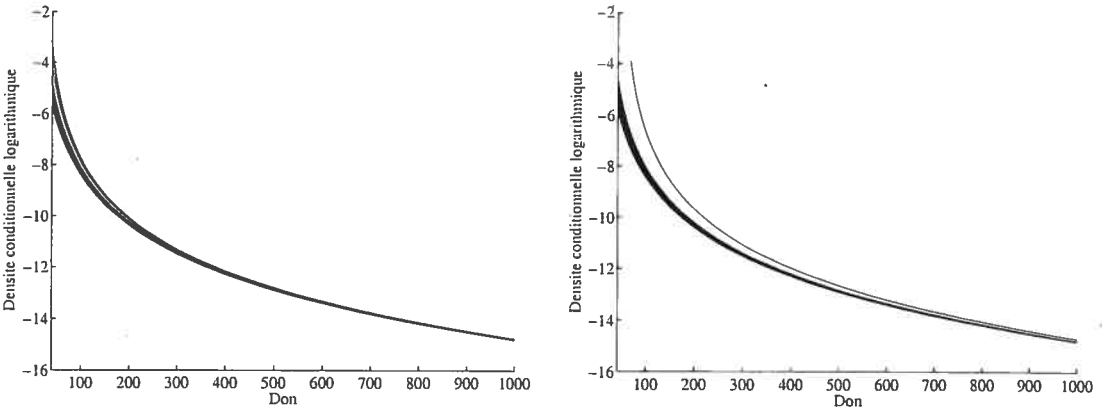

FIG. 4.34: *Données KDD cup 98 : de gauche à droite, densité conditionnelle du modèle CPOT pour 20 points de l'ensemble de test choisis aléatoirement pour l'ensemble d'entraînement de taille 200 et 2 000 respectivement. La densité de la queue supérieure est tracée en échelle logarithmique.*

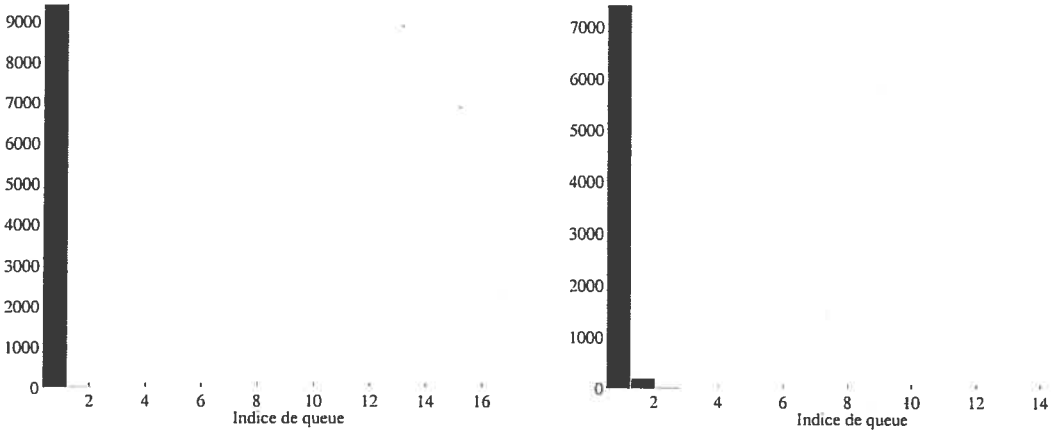

FIG. 4.35: *Données KDD cup 98 : de gauche à droite, indice de queue du modèle CMMH estimé sur l'ensemble de test pour l'ensemble d'entraînement de taille 200 et 2 000 respectivement.*

4.5. Prédiction de rendements financiers

McNeil et Frey [**35**] ont développé un modèle pour les séries financières. Un modèle AR(1)-GARCH(1,1) sert à modéliser les log-rendements Y_t d'un instrument financier :

$$Y_t = \mu_t + \sigma_t \eta_t, \tag{4.5.1}$$

où $\mu_t = \phi Y_{t-1}$ est la partie AR(1) du modèle, $\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2$ avec $\epsilon_t = Y_t - \mu_t$ est la partie GARCH(1,1) et η_t , le résidu est supposé indépendent et identiquement distribué (i.i.d.). Le modèle standard fait l'hypothèse que les résidus suivent une loi Normale. Cependant, la densité conditionnelle des log-rendements est en général sous-estimée par ce modèle. Afin de tenir compte de l'épaisseur de la distribution conditionnelle, une autre approche proposée suppose que les résidus suivent une loi t de Student. Le nombre de degré de liberté de la Student ν est l'inverse multiplicatif de l'indice de queue de la Pareto généralisée (autrement dit $\nu = 1/\xi$). L'inconvénient de cette approche est que cette distribution est symétrique alors que la distribution conditionnelle des log-rendements est le plus souvent asymétrique. McNeil et Frey proposent plutôt d'utiliser la méthode PoT classique (donc inconditionnelle) sur les résidus η_t. Les queues supérieures et inférieures de la distribution des η_t peuvent être modélisées de manière indépendante, ce qui tiendrait compte de la nature asymétrique de la distribution. Ce modèle permet d'obtenir des estimations de quantiles dans la queue inférieure de la distribution (ce qui permet d'évaluer la VaR) supérieures à celles des autres méthodes. Notre objectif est de comparer le mélange conditionnel de Pareto hybrides avec le modèle proposé par McNeil et Frey [**35**]. Nous comparons également le mélange conditionnel de Pareto hybrides avec le mélange conditionnel de gaussiennes et l'estimateur conditionnel de la fenêtre de Parzen en termes de log-vraisemblance relative et d'estimation de quantiles. Les composantes Log-Normales ne peuvent pas être utilisées puisqu'il y a des log-rendements positifs et négatifs.

4.5.1. Entraînement et évaluation de la performance

L'apprentissage du modèle AR(1)-GARCH(1,1) est basé sur la méthode présenté dans McCullough et Renfro [**33**]. Ce type de modèle requiert des valeurs initiales pour Y_0, $\hat{\epsilon}_0$ et $\hat{\sigma}_0$. Des valeurs initiales différentes produiront des estimations différentes des paramètres du modèle. Nous utilisons l'initialisation suivante : $Y_0 = (1/T)\sum_{t=1}^{T} Y_t$, où T est le nombre total d'observations, et $\hat{\epsilon}_0 = \hat{\sigma}_0 = Y_1 - \phi Y_0$. Les paramètres $\zeta = (\phi, \omega, \alpha, \beta)$ sont estimés en maximisant la log-vraisemblance conditionnelle :

$$L(\zeta) = -\sum_{t=1}^{T} \left(\log \sigma_t + \frac{\epsilon_t^2}{2\sigma_t^2} \right).$$

L'optimisation est effectuée à l'aide d'une méthode de descente de gradients conjugués. La méthode PoT classique est ensuite appliquée aux résidus η_t. McNeil et Frey choisissent le nombre d'observations considérées comme excédentaires, ce qui correspond à choisir le niveau de seuil, à l'aide d'une étude simulatoire basée sur la loi t de Student. Nous sélectionnons plutôt le seuil par la même méthode considérée auparavant qui consiste à choisir un niveau de quantile correspondant au seuil à l'aide du test d'adéquation de la Pareto généralisée de Choulakian et Stephens [**6**].

Afin de modéliser les queues supérieure et inférieure de la distribution conditionnelle des rendements logarithmiques, nous utilisons dans le mélange conditionnel de Pareto hybrides deux types de composantes, la Pareto hybride et la Pareto hybride inversée. Pour simplifier la recherche d'hyper-paramètres, le mélange contient le même nombre de composantes de chaque type. Les hyper-paramètres retenus pour chaque modèle sont ceux qui donnent la meilleure performance moyenne sur un ensemble de validation. L'ensemble de validation constitue 20% de l'ensemble d'entraînement.

Nous utilisons la série des rendements logarithmiques de l'indice boursier S&P 500 disponible sur le site http://www.ma.hw.ac.uk/~mcneil/data.html. Comme variable prédictive X_{t+1}, nous considérons les log-rendements en t et en $t-1$ et des moyennes et écart-types mobiles ayant une fenêtre de cinq jours (une semaine), 10

jours (deux semaines), 40 jours (deux mois) et 120 jours (six mois). L'entrée est donc de dimension 10. Pour ces deux modèles, une fenêtre de longueur $n = 1000$, correspondant aux observations de $t - n + 1$ à t, est utilisée pour apprendre les paramètres. Pour le modèle de McNeil et Frey, ces paramètres servent à prédire la distribution conditionnelle au temps $t + 1$. Comme le mélange conditionnel de Pareto hybrides est beaucoup plus long à entraîner, le modèle entraîné sur les données $t - n + 1$ à t sert à prédire la distribution conditionnelle aux temps $t + 1$ à $t + 5$. Cette fenêtre est ensuite déplacée dans le temps pour apprendre un autre modèle et prédire la distribution conditionnelle à la période suivante. La même stratégie pour l'apprentissage et la prédiction est utilisée pour le mélange conditionnel de gaussiennes et l'estimateur conditionnel de la fenêtre de Parzen. Il y a un total de 8 414 observations pour la série des rendements boursiers du S&P 500. Si on exclut les données nécessaires aux calculs des moyennes et écart-types mobiles ainsi que la longueur de fenêtre utilisée pour l'entraînement du premier modèle, il reste donc 7 294 données de test.

4.5.2. Résultats

Le tableau 4.10 donne la log-vraisemblance moyenne relative au mélange conditionnel de Pareto hybrides en test et les hyper-paramètres moyens sélectionnés en validation pour chacun des estimateurs de densité conditionnelle. On remarque que, bien que la log-vraisemblance relative soit positive, l'erreur standard entre parenthèses est élevée. Ceci signifie que la performance du mélange conditionnel de Pareto hybrides n'est pas significativement meilleure dans ce cas-ci.

La figure 4.36 fournit un aperçu du modèle AR(1)-GARCH(1,1) couplé avec la méthode PoT classique. Il s'agit d'un extrait de 750 observations qui contient la période du crash boursier d'octobre 1987. Le panneau du haut représente la série des rendements logarithmiques du S&P 500, le panneau du milieu illustre l'écart-type conditionnel du modèle GARCH(1,1) et le panneau du bas donne l'indice de queue conditionnel tel qu'estimé par la méthode PoT. Une figure semblable est incluse dans l'article de McNeil et Frey [**35**] qui contient les panneaux du

	cmmh	cmmg	cparzen
$\mathcal{R}_l$ *(err.std)*		0.03072 (0.03456)	0.119 (0.1609)
	$(h_{\text{cmmh}}, m_{\text{cmmh}})$	$(h_{\text{cmmg}}, m_{\text{cmmg}})$	$(\lambda^x_{\text{cparzen}}, \lambda^y_{\text{cparzen}})$
hyper-paramètres	(1.441, 3.5)	(1.637, 3.82)	(2238, 0.517)

TAB. 4.10: *Log-vraisemblance moyenne relative au mélange conditionnel de Pareto hybrides $\mathcal{R}_l$ sur les données de test S&P500 et hyper-paramètres moyens sélectionnés. Pour le mélange conditionnel de Pareto hybrides, m_{cmmh} inclut les composantes Pareto hybride standard et inversée. L'ensemble d'entraînement est une fenêtre contenant 1000 observations que l'on fait rouler dans le temps.*

haut et du milieu. Notre implantation du modèle AR(1)-GARCH(1,1) est donc raisonnablement proche de la leur. Il est intéressant de noter que l'écart-type conditionnel σ_t augmente ponctuellement suite au crash d'octobre 1987 alors que pour l'indice de queue conditionnel, le niveau global augmente pour tous les jours subséquents.

La figure 4.37 est la réciproque de la figure 4.36 pour le mélange conditionnel de Pareto hybrides. Il s'agit donc des mêmes 750 observations de test incluant le crash boursier d'octobre 1987. Le panneau du haut contient également les log-rendements du S&P 500, les panneaux du milieu et du bas illustrent les variations des indices de queue pour les queues supérieure et inférieure respectivement tel qu'estimé par le mélange conditionnel de Pareto hybrides. Les estimateurs des indices de queue ont un comportement semblable à l'écart-type conditionnel du modèle GARCH(1,1) (voir le panneau du milieu de la figure 4.36). Il y a une augmentation très marquée des indices de queue supérieure et inférieure suite au crash d'octobre 87. Les estimateurs de l'indice de queue conditionnel du mélange conditionnel de Pareto hybrides prennent des valeurs dans un intervalle beaucoup plus grand que ceux résultants de la méthode PoT couplée au modèle AR(1)-GARCH(1,1).

À partir des modèles de densité conditionnelle, nous calculons des quantiles conditionnels qui se trouvent dans la queue inférieure de la distribution conditionnelle. McNeil et Frey introduise un test d'hypothèse basé sur la loi Binômiale que nous avons déjà utilisé sur les données danoises de la section 3.4. La version

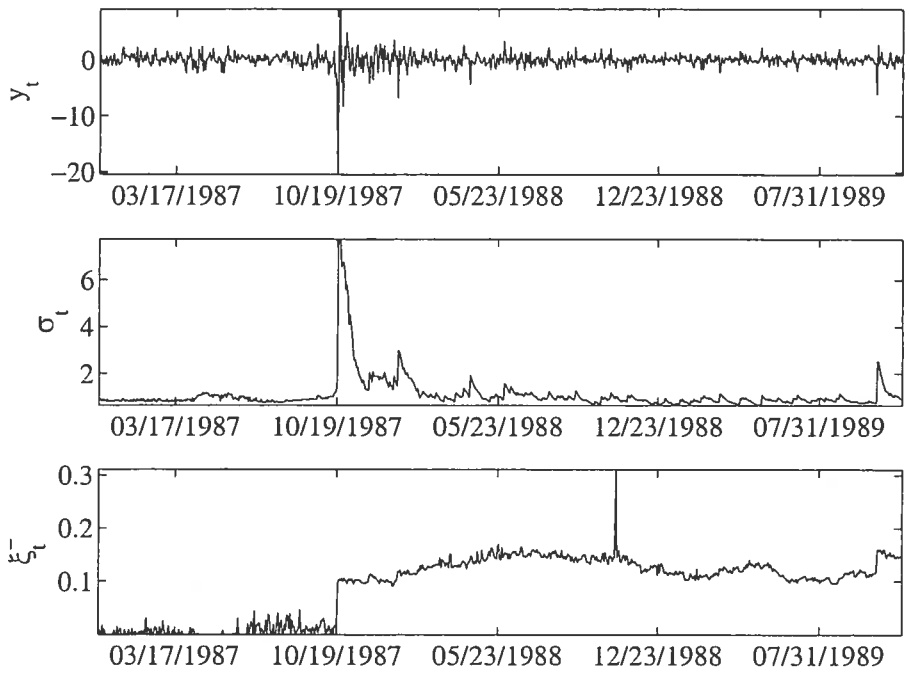

FIG. 4.36: *S&P 500 : Le panneau du haut contient les log-rendements du S&P 500 pour une période de 750 jours incluant le crash d'octobre 1987. Le panneau du milieu fournit l'écart-type conditionnel estimé par le modèle GARCH(1,1) et le panneau du bas donne l'indice de queue inférieure estimé par la méthode PoT classique appliquée aux résidus du modèle AR(1)-GARCH(1,1).*

conditionnelle de ce test est semblable à la version inconditionnelle. On dit qu'il y a violation d'un quantile conditionnel de niveau q lorsque $y_t \leq y_t^q$ (puisque dans ce cas-ci, on s'intéresse aux quantiles de la queue inférieure). Sous l'hypothèse nulle que le modèle considéré soit approprié pour les données, le nombre de violations d'un quantile conditionnel de niveau q sur un ensemble de test de longueur T suit une loi Binômiale $B(T, q)$. Le tableau 4.11 contient les résultats du test binômial pour les estimateurs de quantiles de niveau $\{0.05, 0.001, 0.005\}$ sur les données de test pour le mélange conditionnel de Pareto hybrides (cmmh), le mélange conditionnel de gaussiennes (cmmg), l'estimateur conditionnel de la fenêtre de Parzen (cparzen) et le modèle AR(1)-GARCH(1,1)-PoT de McNeil et Frey. La proportion de violations espérée est donnée par q dans la première colonne du tableau. Pour chaque niveau de quantile q, le tableau fournit pour chaque modèle la valeur P associée à la proportion observée de violations ainsi qu'un intervalle de confiance de niveau 95 % entre parenthèses. Plus la valeur P

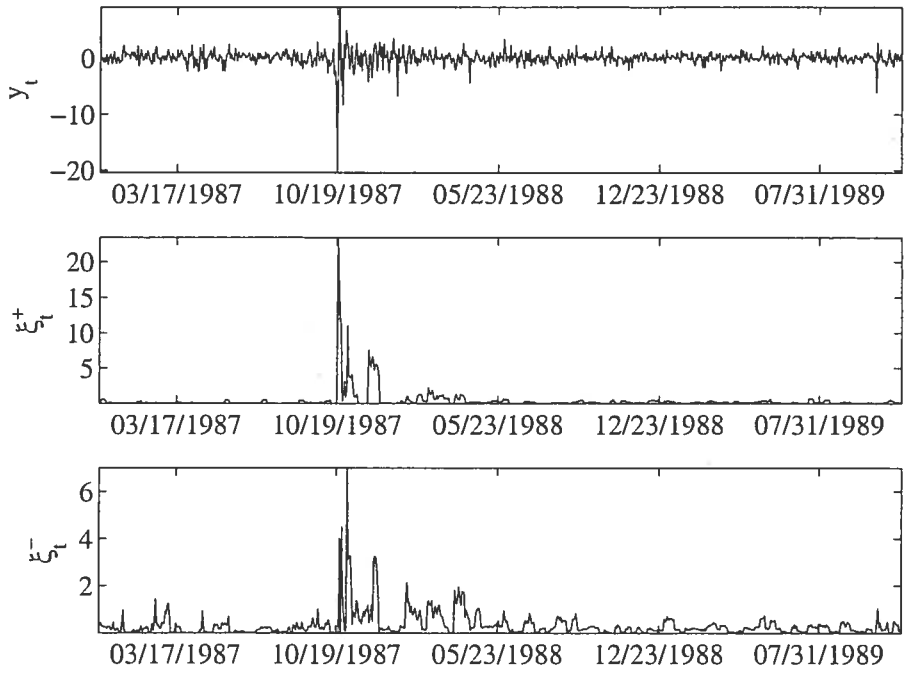

FIG. 4.37: *S&P 500 : Le panneau du haut contient les log-rendements du S&P 500 pour une période de 750 jours incluant le crash d'octobre 1987. Les panneaux du milieu et du bas fournissent l'indice de queue conditionnel pour les queues supérieure et inférieure respectivement tel qu'estimés par le modèle CMMH.*

est grande, plus la proportion observée de violations est plausible. Si l'hypothèse nulle est juste, on s'attend à ce que la proportion espérée de violations se trouve dans l'intervalle de confiance. Les valeurs P pour le modèle AR(1)-GARCH(1,1)-PoT sont généralement plus petites que celles obtenues par McNeil et Frey mais elles sont suffisamment grandes pour ne pas rejeter l'hypothèse nulle avec une bonne marge. L'ensemble de test dont nous disposons est un peu plus petit que le leur (120 observations de moins, qui sont utilisées dans le calcul des moyennes et écart-types mobiles). Ceci change donc les valeurs initiales du processus et affecte aussi les paramètres trouvés par maximum de vraisemblance. Les valeurs P pour le mélange conditionnel de Pareto hybrides sont également suffisamment grandes pour ne pas entraîner le rejet de l'hypothèse nulle. La performance du mélange de Pareto hybrides en termes d'estimation de quantiles extrêmes est donc semblable à celle du modèle proposé par McNeil et Frey. Les quantiles estimés par le mélange conditionnel de gaussiennes sont rejetés par le test binômial pour les trois niveaux de quantiles. Les quantiles estimés par l'estimateur conditionnel de

la fenêtre de Parzen ne sont rejetés qu'en une occasion, lorsque $q = 0.05$ et que la proportion de violations est sous-estimée par le modèle.

Les quantiles conditionnels estimés par ces quatre modèles sont illustrés aux figures 4.38, 4.39 et 4.40 pour les niveaux de quantiles $q = 0.05$, $q = 0.01$ et $q = 0.005$ respectivement. Le quantile conditionnel du mélange conditionnel de Pareto hybrides est en général plus grand et plus variable que ceux des autres méthodes.

q	cmmh	cmmg
0.05	**0.1624 (0.04855 0.05902)**	4.381e-05 (0.05536 0.06646)
0.01	**1 (0.007731 0.01242)**	7.945e-12 (0.01592 0.02231)
0.005	**0.5069 (0.003003 0.006188)**	3.113e-15 (0.01030 0.01560)
q	cparzen	AR(1)-GARCH(1,1)-PoT
0.05	2.428e-05 (0.03513 0.04421)	**0.6479 (0.04619 0.05644)**
0.01	**0.5177 (0.007126 0.01165)**	**0.5177 (0.007126 0.01165)**
0.005	**0.5603 (0.003921 0.007460)**	**0.6188 (0.003116 0.006348)**

TAB. 4.11: *Test binômial pour l'estimation de quantiles sur les données S&P500. La proportion espérée de violations est q. Pour chaque estimateur, on donne la valeur P pour la proportion espérée suivie, entre parenthèse, d'un intervalle de confiance de niveau 95%. Plus la valeur P est élevée, plus la proportion observée de violations est vraisemblable. Si l'hypothèse nulle est juste, l'intervalle de confiance devrait contenir la proportion espérée de violations. Les résultats en caractères gras indiquent le succès du test.*

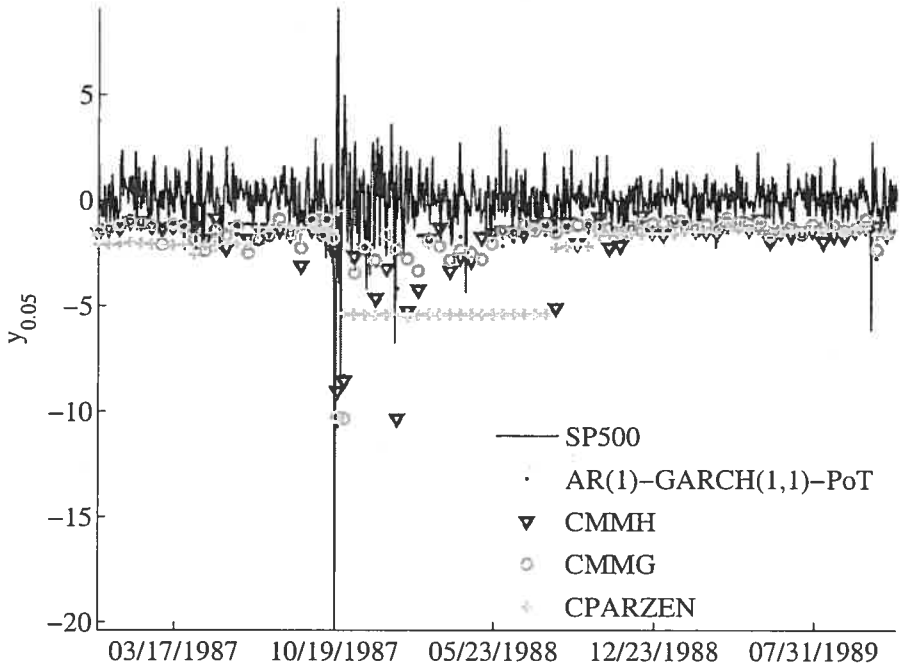

FIG. 4.38: *S&P 500 : Estimation de quantile conditionnel de niveau 0.05 pour une période de 750 jours incluant le crash d'octobre 1987.*

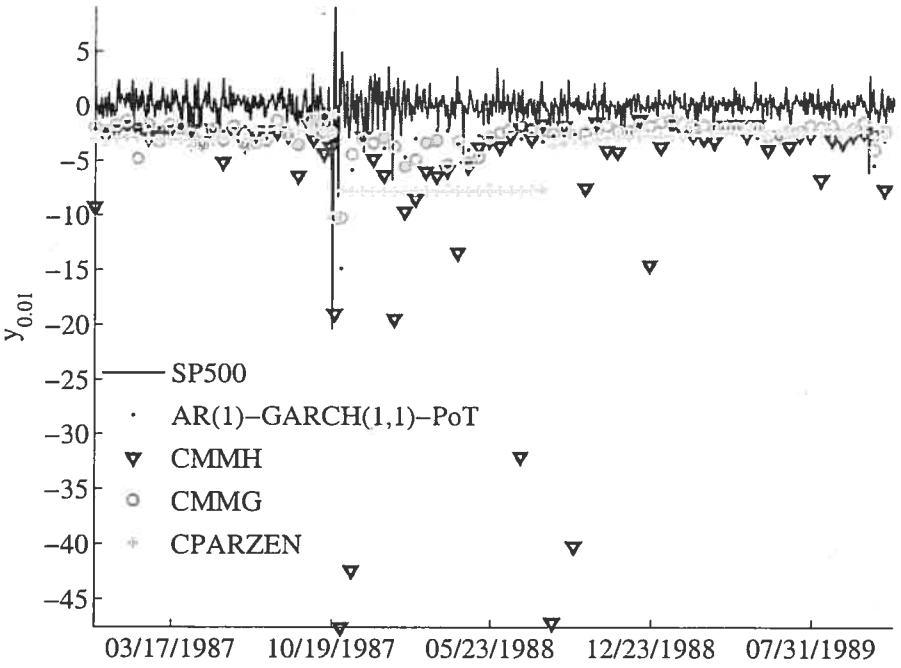

FIG. 4.39: *S&P 500 : Estimation de quantile conditionnel de niveau 0.01 pour une période de 750 jours incluant le crash d'octobre 1987.*

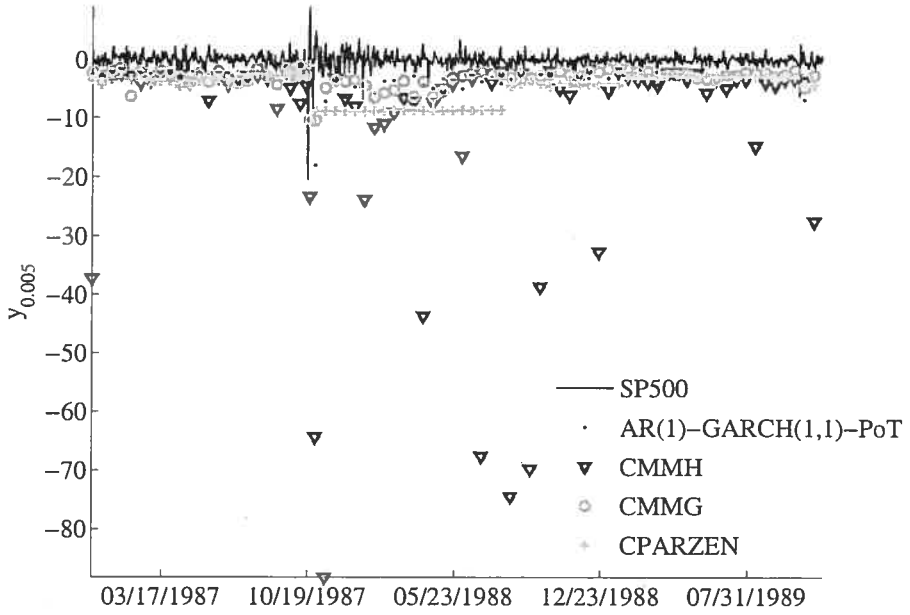

FIG. 4.40: *S&P 500 : Estimation de quantile conditionnel de niveau 0.005 pour une période de 750 jours incluant le crash d'octobre 1987.*

4.6. Conclusion

Il existe, dans plusieurs domaines d'application, des variables explicatives qui fournissent de l'information pertinente sur la variable d'intérêt. Les compagnies d'assurance disposent de renseignements sur le profil des clients qui influencent la distribution des réclamations. En finance, les arbitragistes se basent sur les rendements des derniers jours, les tendances du marché à plus long termes (captées par de moyennes ou écart-types mobiles) et des informations sur l'état de l'économie pour prédire les mouvements du marché. Les climatologues élaborent des scénarios probables de l'augmentation des gaz à effet de serre et cherchent à prédire la probabilité d'événements climatiques extrêmes pour chacun de ces scénarios.

La régression consiste en l'estimation de l'espérance conditionnelle $E[Y|X = x]$. Lorsque les données présentent une structure complexe, par exemple lorsqu'il y a multi-modalité, l'espérance conditionnelle ne représente pas bien le processus générateur des données. C'est pourquoi nous nous intéressons à l'estimation de la densité conditionnelle $P(Y|X = x)$ qui permet ensuite le calcul de quantités d'intérêt propres à une application donnée.

Dans ce chapitre, nous avons présenté une extension du mélange de Pareto hybrides qui permet d'estimer la densité conditionnelle en présence de multi-modalité, d'asymétrie et de queues lourdes. Cet estimateur combine l'idée du mélange conditionnel avec composantes gaussiennes de Bishop [5] avec les méthodes provenant de la théorie des valeurs extrêmes. La distribution Pareto hybride, développée à la section 3.1, permet d'intégrer la Pareto généralisée dans un mélange de distribution tout en en conservant les propriétés d'approximation des queues de distribution. L'estimateur proposé est un mélange de Pareto hybrides dont les paramètres sont prédits par un réseau de neurones.

L'étude simulatoire permet la comparaison de l'estimateur proposé avec d'autres estimateurs de densité sur des jeux de données dont le modèle générateur est connu. Dans le cas conditionnel, il y a deux niveaux de difficulté pour les estimateurs de densité : capter la relation entre l'entrée et les paramètres de la Fréchet et modéliser l'épaisseur de la queue de la distribution. Pour les quatre jeux de

données étudiés et pour les deux tailles d'ensemble d'entraînement utilisées, le mélange conditionnel de Pareto hybride offre une performance supérieure aux autres estimateurs considérés. La différence de performance est particulièrement frappante lorsque l'ensemble d'entraînement est plus petit, que la relation entre l'entrée et les paramètres du modèle générateur est non-linéaire et que la queue de la distribution est très lourde. Lorsqu'utilisé pour générer des données, le mélange conditionnel de Pareto hybrides produit des nuages de points dont la forme est très semblable à celle du modèle utilisé pour générer les données d'entraînement. Par contre, l'estimation de l'indice de queue conditionnel par le mélange conditionnel de Pareto hybrides n'est pas satisfaisante. Ultimement, ce que nous tenons à vérifier est que l'estimateur proposé estime adéquatement la queue de la distribution conditionnelle. Le mélange de Pareto hybrides réussit relativement bien à capter la forme de la dépendance des quantiles extrêmes conditionnels. Dans certains cas, par exemple lorsque le niveau de quantile est de 0.9999 ou que peu de données sont disponibles, l'estimation est cependant plus difficile.

Nous avons ensuite évalué le mélange conditionnel de Pareto hybrides sur des données d'assurance. À nouveau, la performance de l'estimateur proposé est supérieure à celles des autres estimateurs considérés. Ceci est particulièrement marquant pour les plus petits ensembles d'entraînement. L'examen des courbes de densité conditionnelle associée au mélange conditionnel de Pareto hybrides fait ressortir les faits suivants. Cet estimateur arrive à capter la nature multi-modale de la distribution conditionnelle, soit par le biais de la fonction de dépendance des paramètres du mélange (modélisée par le réseau de neurones), soit par l'utilisation de plusieurs composantes. Aussi, l'estimateur est en mesure de modéliser différents types de queues de la distribution conditionnelle, de la décroissance exponentielle à la décroissance polynômiale. Finalement, nous avons comparé l'estimation de quantiles extrêmes du mélange conditionnel de Pareto hybrides avec l'estimateur proposé par McNeil et Frey [**35**] sur la série des rendements logarithmique de l'indice boursier du S&P500. Le test binômial utilisé par ces auteurs confirment que les quantiles estimés par le mélange conditionnel de Pareto hybrides sont comparables aux estimations produites par leur modèle.

Chapitre 5

CONCLUSION

Les événements extrêmes provoquent généralement des perturbations majeures lorsqu'ils surviennent. Que ce soit un crash boursier qui génère de fortes répercussions dans les marchés financiers, un ouragan qui laisse des milliers de personnes sans abri ou encore un tremblement de terre extrême qui détruit des villes. La théorie des valeurs extrêmes est née du besoin de modéliser de tels événements pour tenter de se prémunir contre eux. Elle repose sur le fait qu'il est possible de caractériser le comportement asympotique de la queue d'une distribution univariée avec une condition très souple. En particulier, la loi de Pareto généralisée permet d'approximer arbitrairement bien la queue de toute distribution connue. La méthode PoT consiste en l'ajustement d'une loi de Pareto généralisée aux observations qui excèdent un seuil donné. Ce seuil spécifie quelles observations sont considérées comme appartenant à la queue de la distribution. La principale difficulté dans l'application de la Pareto généralisée est la détermination de ce seuil. Un seuil trop bas entraîne un biais élevé de l'approximation de la queue de la distribution par la Pareto généralisée. D'un autre côté, un seuil trop élevé laisse peu d'observations pour l'estimation des paramètres ce qui augmente la variance des estimateurs.

Les méthodes non-paramétriques ont été developpées pour répondre au besoin de modéliser des données sans faire d'hypothèses a priori quant à la forme de la relation entre ces données. En estimation de densité, des modèles comme le mélange de gaussiennes permet d'approximer arbitrairement bien toute distribution si le nombre de composantes augmente de façon appropriée avec la taille de

l'ensemble d'entraînement. Cependant, plus la queue de la distribution est lourde, plus il faut un grand ensemble d'entraînement pour que le mélange de gaussiennes donne des résultats satisfaisants. Nous proposons une nouvelle distribution, la Pareto hybride, qui permet de transférer les propriétés d'approximation de la Pareto généralisée aux méthodes non-paramétriques. La Pareto hybride consiste en la juxtaposition d'une loi Normale tronquée avec une loi de Pareto généralisée. Au point de jonction, où la Normale est tronquée, des conditions de continuité sont imposées à la densité et à sa dérivée. Comme la Pareto hybride est continue sur l'axe des réels, elle peut facilement être utilisée dans un mélange de distributions. La densité de la Pareto généralisée est zéro sous le seuil et elle est discontinue en ce point. Pour l'utiliser dans un mélange, il faut déterminer quel est le seuil approprié puisque celui-ci ne peut être appris lors de l'entraînement du mélange. La Pareto hybride subvient à la question de la sélection du seuil puisque celui-ci est une fonction des paramètres de l'hybride. Il s'adapte donc aux données lors de l'entraînement du mélange.

La queue de la Pareto hybride est, à un facteur près, la même que celle de la Pareto généralisée. Elle permet entre autres de modéliser des distributions à queues lourdes comme la loi t de Student, à queue modérée comme la loi Log-Normale et à queue légère telle que la loi Normale. Dans un mélange de Pareto hybrides, toutes les données participent à l'estimation de la queue de la distribution alors que dans l'utilisation classique de la Pareto généralisée, seules les données au-delà du seuil servent à l'estimation du modèle. Les données centrales influencent l'estimation des paramètres du mélange mais il ressort des simulations sur des jeux de données synthétiques que l'emploi de plusieurs composantes permette au mélange d'estimer la queue de la distribution de façon comparable à la méthode PoT. Les expériences sur des jeux de données synthétiques et réelles ont démontré que la performance du mélange de Pareto hybrides en termes de log-vraisemblance est supérieure à celle des autres estimateurs considérés. Ceci est particulièrement frappant pour les petits ensembles d'entraînement.

La recherche sur les événements extrêmes porte essentiellement sur la modélisation de la densité inconditionnelle. Cependant, dans de nombreuses applications, il est nécessaire de considérer de l'information de laquelle dépend la variable d'intérêt. Il est alors naturel de chercher à modéliser la densité conditionnelle. L'information dont on dispose prend la forme d'une variable prédictive qui peut être de haute dimension. Les techniques existantes pour représenter la densité conditionnelle ne sont pas toujours appropriées lorsque la distribution conditionnelle sous-jacente présente des queues lourdes, est multi-modale et asymétrique. Nous proposons comme estimateur de densité conditionnelle, un mélange conditionnel de Pareto hybrides. Il s'agit d'un mélange de Pareto hybrides dont les paramètres sont des fonctions de la variable prédictive. Ces fonctions peuvent être représentées par un modèle paramétrique ou non-paramétrique. Il suffit que la classe de fonctions choisie puisse être entraînée à l'aide du gradient par rapport aux paramètres de cette classe car l'apprentissage des paramètres du mélange conditionnel se fait par la maximisation de la log-vraisemblance conditionnelle. Nous avons utilisé comme classe de fonctions les réseaux de neurones artificiels. Les réseaux de neurones ont déjà été utilisés pour des tâches semblables. Ils possèdent par ailleurs la capacité d'approximer arbitrairement bien toute fonction continue si la complexité du réseau, c'est-à-dire le nombre d'unités cachées, est bien choisie. Le mélange conditionnel de Pareto hybrides se compare favorablement aux autres estimateurs testés en termes de log-vraisemblance sur des jeux de données synthétiques et réelles.

Les défis actuels en théorie des valeurs extrêmes concernent la modélisation des extrêmes multivariés [**36**]. La difficulté principale dans le cas multivarié est d'estimer la fonction de dépendance extrêmale entre les variables aléatoires. Les solutions actuelles s'adressent à des problèmes de basse dimension (au plus trois) et supposent le plus souvent que toutes les variables aient le même schéma de dépendance extrêmale (soit elles sont toutes dépendantes ou toutes indépendantes entre elles). L'apprentissage statistique devrait être en mesure de proposer une alternative non-paramétrique à ce problème qui serait applicable à toute dimension.

BIBLIOGRAPHIE

[1] D. M. Bashtannyk and R. J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3) :279–298, May 2001.

[2] J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, 2(2) :177–200, June 1999.

[3] J. Beirlant, E. Joossens, and J. Segers. Unbiased tail estimation by an extension of the generalized pareto distribution. Technical report, Tilburg University, 2005.

[4] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics, 1995.

[5] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford, 1995.

[6] V. Choulakian and M. A. Stephens. Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, 43(4) :478–484, November 2001.

[7] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, 2001.

[8] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambert w function. *Advances in Computational Mathematics*, 5 :329–359, 1996.

[9] J. Danielsson and C. G. de Vries. Beyond the sample : Extreme quantile and probability estimation. Technical Report dp298, Financial Markets Group, July 1998. available at http ://ideas.repec.org/p/fmg/fmgdps/dp298.html.

[10] A. C. Davison and R. L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3) :393–442, 1990.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39(1) :1–38, 1977.

[12] L. Devroye and L. Gyorfi. *Nonparametric Density Estimation : The L_1 View.* New York : John Wiley, 1985.

[13] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. A universal approximator of convex functions applied to option pricing. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

[14] D. J. Dupuis. Exceedances over high thresholds : A guide to threshold selection. *Extremes*, 1(3) :251–261, 1998.

[15] P. Embrechts, C. Kluppelberg, and T. Mikosch. *Modelling Extremal Events.* Applications of Mathematics, Stochastic Modelling and Applied Probability. Springer, 1997.

[16] E. F. Fama. The behavior of stock market prices. *Journal of Business*, 38 :34–105, 1965.

[17] R. A. Fisher. Theory of statistical estimation. In *Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725, 1925.

[18] A. Frigessi, O. Haug, and H. Rue. A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5 :219–235, 2002.

[19] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1) :1–58, 1992.

[20] J. Georges and A. H. Milley. Kdd'99 competition : Knowledge discovery contest. *SIGKDD Explorations*, 1(2), January 2000.

[21] T. Heskes. Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6) :1425–1433, 1998.

[22] K. M. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251–257, 1991.

[23] J. R. M. Hosking and J. R. Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29 :339–349, 1987.

[24] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixture of local experts. *Neural Computation*, 3(1) :79–87, 1991.

[25] G. M. James. Variance and bias for general loss functions. *Machine Learning*, 51 :115–135, 2003.

[26] I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

[27] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. In *IJCNN'93-Nagoya*, volume 2, pages 1339–1344. International joint conference on neural networks, 1993.

[28] S. Kang and R. F. Serfozo. Extreme values of phase-type and mixed random variables with parallel-processing examples. *Journal of Applied Probability*, 36 :194–210, 1999.

[29] A. N. Kolmogorov. On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition. *Doklady Akademija Nauk SSSR*, 114(5), 1957.

[30] S. Kullback and R. A. Leibler. On the information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 1951.

[31] C. Liu and D. B. Rubin. Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*, 5 :19–39, 1995.

[32] B. Mandelbrot. The variation of certain speculative prices. *Journal of Business*, 36 :394–419, 1963.

[33] B. D. McCullough and C. G Renfro. Benchmarks and software standards : A case study of garch procedures. *Journal of Economic and Social Measurement*, 25 :59–71, 1998.

[34] A. J. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *Astin Bulletin*, 27 :117–137, 1997.

[35] A. J. McNeil and R. Frey. Estimation of tail-related risk measures for heteroscedatic financial time series : an extreme value approach. *Journal of Empirical Finance*, 7 :271–300, 2000.

[36] T. Mikosch. How to model multivariate extremes if one must ? Research Report 21, The Danish National Research Foundation, 2004.

[37] J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3 :119–131, 1975.

[38] D. Pollard. Strong consistency of the k-means clustering. *The Annals of Statistics*, 9(1) :135–140, January 1981.

[39] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical recipes in FORTRAN : the art of scientific computing.* Cambridge University Press, 2nd edition, 1992.

[40] C. E. Priebe. Adaptive mixtures. *Journal of the American Statistical Association*, 89 :796–806, 1994.

[41] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26 :1443–1471, 2002.

[42] M. Rosenblatt. Conditional probability density and regression estimators. In P. R. Krishnaiah, editor, *Multivariate Analysis II*, pages 25–31. Academic Press, New York, 1969.

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing : Explorations in the Macrostructure of Cognition*, volume 1, pages 318–362. Bradford Books, Cambridge, MA, 1986.

[44] R. L. Smith. Estimating tails of probability distributions. *The Annals of Statistics*, 15(3) :1174–1207, September 1987.

[45] A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4) :595–601, 1949.

[46] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1) :1–26, 1982.

Annexe A

ALGORITHMES

A.1. Loi Pareto hybride

Dans cette section se trouvent les algorithmes qui servent au calcul de la densité (Algorithme 1) et du gradient de la log-vraisemblance (Algorithme 3) de la loi Pareto hybride. Nous donnons également l'algorithme qui sert à calculer la fonction W de Lambert (Algorithme 2).

Algorithme 1 Densité de la loi Pareto hybride : $h_\psi(x)$

Entrées: x, $\psi = (\xi, \mu, \sigma)$ avec $\sigma > 0$

Sorties: $y \leftarrow h_\psi(x)$

{Requiert les fonctions $W(\cdot)$ (Algorithme 2) et `erf` [39]}

$z \leftarrow W\left((1+\xi)^2/(2\pi)\right)$

$\gamma \leftarrow 3/2 + 1/2\ \texttt{erf}(\text{sign}(1+\xi)\sqrt{z/2})$

$\beta \leftarrow \sigma\sqrt{(1+\xi)^2/z}$

$\alpha \leftarrow \text{sign}(1+\xi)\sigma\sqrt{z} + \mu$

si $x \leq \alpha$ **alors**

  $y \leftarrow \exp(-(x-\mu)^2/(2\sigma^2))/(\sqrt{2\pi}\sigma)$

sinon

  si $\xi \neq 0$ **alors**

    $y \leftarrow \frac{1}{\beta}(1 + \xi(x-\alpha)/\beta)^{-1/\xi - 1}$

  sinon

    $y \leftarrow \frac{1}{\beta}\exp(-(x-\alpha)/\beta)$

Algorithme 2 Fonction W de Lambert : $W(z)$

ENTRÉES: z, w_0, tol_{rel} et $tol_{abs} \in \mathbb{R}$

SORTIES: w tel que $w\exp(w) = z$

{Valeurs par défaut}

si w_0 manquant **alors**

$w_0 \leftarrow 0.5$

sinon si tol_{rel} manquant **alors**

$tol_{rel} \leftarrow 10^{-6}$

sinon si tol_{abs} manquant **alors**

$tol_{abs} \leftarrow 10^{-6}$

$err_{abs} \leftarrow |z - w_0\exp(w_0)|$

$err_{rel} \leftarrow err_{abs}/z$

tantque $err_{abs} > tol_{abs}$ and $err_{rel} > tol_{rel}$ **faire**

$e_0 \leftarrow \exp(w_0)$

$w \leftarrow w_0 - (w_0e_0 - z)/\left(e_0(w_0+1) - (w_0+2)(w_0e_0 - z)/(2w_0+2)\right)$

$err_{abs} \leftarrow |z - w\exp(w)|$

$err_{rel} \leftarrow err_{abs}/z$

$w_0 \leftarrow w$

Algorithme 3 Gradient de la log-vraisemblance négative pour la Pareto hybride

ENTRÉES: x, $\psi = (\xi, \mu, \sigma)$ et $\sigma > 0$

SORTIES: $y \leftarrow \nabla l = \left(\frac{\partial l}{\partial \xi}, \frac{\partial l}{\partial \mu}, \frac{\partial l}{\partial \sigma}\right)$, où $l = -\log h_\psi(x)$

{Requiert les fonctions $W(\cdot)$ (Algorithme 2) et `erf` [39]}

$z \leftarrow W\left((1+\xi)^2/(2\pi)\right)$

$\gamma \leftarrow 3/2 + 1/2\ \texttt{erf}(\text{sign}(1+\xi)\sqrt{z/2})$

$\beta \leftarrow \sigma\sqrt{(1+\xi)^2/z}$

$\alpha \leftarrow \text{sign}(1+\xi)\sigma\sqrt{z} + \mu$

{Dérivées des paramètres dépendants, γ, β et α, p/r à ξ, μ et σ}

$\frac{\partial \gamma}{\partial \xi} \leftarrow (\exp(-z/2)\sqrt{z}) / \left(\sqrt{2\pi(1+\xi)^2}(1+z)\right)$

$\frac{\partial \beta}{\partial \xi} \leftarrow \text{sign}(1+\xi)\sigma\sqrt{z}/(1+z)$

$\frac{\partial \beta}{\partial \sigma} \leftarrow \sqrt{(1+\xi)^2/z}$

$\frac{\partial \alpha}{\partial \xi} \leftarrow \sigma\sqrt{z} / \left(\sqrt{(1+\xi)^2}(1+z)\right)$

$\frac{\partial \alpha}{\partial \sigma} \leftarrow \text{sign}(1+\xi)\sqrt{z}$

$\frac{\partial \gamma}{\partial \mu} \leftarrow 0, \frac{\partial \gamma}{\partial \sigma} \leftarrow 0, \frac{\partial \beta}{\partial \mu} \leftarrow 0, \frac{\partial \alpha}{\partial \mu} \leftarrow 1$

{Dérivée de $\omega = -\log g_{\xi,\beta}(x-\alpha)$ p/r à ξ, β et α}

si $\xi \neq 0$ **alors**

$\frac{\partial \omega}{\partial \xi} \leftarrow (1/\xi + 1)(x-\alpha)/\left(\beta + \xi(x-\alpha)\right) - 1/\xi^2 \log(1 + \xi(x-\alpha)/\beta)$

$\frac{\partial \omega}{\partial \beta} \leftarrow 1/\beta(1 - (1+\xi)(x-\alpha)/(\beta + \xi(x-\alpha)))$

$\frac{\partial \omega}{\partial \alpha} \leftarrow -(1+\xi)/\left(\beta + \xi(x-\alpha)\right)$

sinon

$\frac{\partial \omega}{\partial \xi} \leftarrow 0, \frac{\partial \omega}{\partial \alpha} \leftarrow -1/\beta$

$\frac{\partial \omega}{\partial \beta} \leftarrow 1/\beta(1 - (x-\alpha)/\beta)$

{Dérivée de $l = -\log h_\psi(x)$ p/r à ξ, μ et σ}

si $x \leq \alpha$ **alors**

$\frac{\partial l}{\partial \xi} \leftarrow \frac{1}{\gamma}\frac{\partial \gamma}{\partial \xi}$

$\frac{\partial l}{\partial \mu} \leftarrow -(x-\mu)/\sigma^2$

$\frac{\partial l}{\partial \sigma} \leftarrow 1 - (x-\mu).^2/\sigma^2$

sinon

$\frac{\partial l}{\partial \xi} \leftarrow \frac{1}{\gamma}\frac{\partial \gamma}{\partial \xi} + \frac{\partial \omega}{\partial \xi} + \frac{\partial \omega}{\partial \beta}\frac{\partial \beta}{\partial \xi} + \frac{\partial \omega}{\partial \alpha}\frac{\partial \alpha}{\partial \xi}$

$\frac{\partial l}{\partial \mu} \leftarrow \frac{\partial \omega}{\partial \alpha}\frac{\partial \alpha}{\partial \mu}$

$\frac{\partial l}{\partial \sigma} \leftarrow \frac{\partial \omega}{\partial \alpha}\frac{\partial \alpha}{\partial \sigma} + \frac{\partial \omega}{\partial \beta}\frac{\partial \beta}{\partial \sigma}$

$y \leftarrow \{\frac{\partial l}{\partial \xi}, \frac{\partial l}{\partial \mu}, \frac{\partial l}{\partial \sigma}\}$

A.2. Mélange conditionnel de Pareto hybrides

Cette section présente les algorithmes nécessaires à l'apprentissage d'un mélange conditionnel de Pareto hybrides. La log-vraisemblance négative est calculée par l'algorithme 4 et son gradient par l'algorithme 5. Ce dernier algorithme requiert le calcul du gradient de la densité de la Pareto hybride qui fourni à l'algorithme 6.

Algorithme 4 Log-vraisemblance négative pour CMMH : $\mathcal{L}(\theta)$

Entrées: $y \in \mathbb{R}$, $x \in \mathbb{R}^d$, $n_h \in \mathbb{N}$, $m \in \mathbb{N}$, $\theta = (b, c, v, w, \tilde{w})$

Sorties: $l \leftarrow -\log\left(\phi_\theta(x, y)\right)$, $\lambda \leftarrow (\pi_1(x), \ldots, \pi_m(x), \psi_1(x), \ldots, \psi_m(x))$ et $z_i \leftarrow \tanh\left(c_i + \sum_{k=1}^{d} v_{ik} x_k\right)$

{1. Calcul des activations des unités cachées}

$z_i \leftarrow \tanh\left(c_i + \sum_{k=1}^{d} v_{ik} x_k\right)$ pour $i = 1 \ldots n_h$

{2. Calcul des sorties du réseau de neurones}

$a_j^{(i)} \leftarrow b_j^{(i)} + \sum_{h=1}^{n_h} w_{jh}^{(i)} z_h + \sum_{k=1}^{d} \tilde{w}_{jk}^{(i)} x_k$

{3. Transformer les sorties du réseau de neurones}

$\pi_i(x) \leftarrow \exp(a_i^{(0)}) / \sum_{j=1}^{m} \exp(a_j^{(0)})$ pour $i = 1, \ldots, m$

$\xi_i(x) \leftarrow \text{softplus}(a_i^{(1)})$, pour $i = 1, \ldots, m$

$\mu_i(x) \leftarrow a_i^{(2)}$, pour $i = 1, \ldots, m$

$\sigma_i(x) \leftarrow \text{softplus}(a_i^{(3)})$, pour $i = 1, \ldots, m$

$\psi_i(x) \leftarrow (\xi_i(x), \mu_i(x), \sigma_i(x))$

$\lambda \leftarrow (\pi_1(x), \ldots, \pi_m(x), \psi_1(x), \ldots, \psi_m(x))$

{4. Calcul de la log-vraisemblance négative}

$l \leftarrow -\log\left(\sum_{i=1}^{m} \pi_i(x) h_{\psi_i(x)}(y)\right)$

Algorithme 5 Gradient de la log-vraisemblance négative pour CMMH

ENTRÉES: $y \in \mathbb{R}$, $x \in \mathbb{R}^d$, $n_h \in \mathbb{N}$, $m \in \mathbb{N}$, $\theta = (b, c, v, w, \tilde{w})$

SORTIES: $g \leftarrow \frac{\partial l}{\partial \theta}$ où $l = -\log\left(\phi_\theta(x, y)\right) = -\log\left(\phi_\lambda(y)\right)$

{Exécuter l'algorithme 4 pour obtenir z_i et $\lambda = (\pi_1, \ldots, \pi_m, \psi_1, \ldots, \psi_m)$}

{Requiert la fonction $\nabla h_\psi(y) = \left(\frac{\partial h_\psi(y)}{\partial \xi}, \frac{\partial h_\psi(y)}{\partial \mu}, \frac{\partial h_\psi(y)}{\partial \sigma}\right)$ (algorithme 6).}

$\frac{\partial l}{\partial \phi_\lambda(y)} \leftarrow -1/\phi_\lambda(y)$

$\frac{\partial \phi_\lambda(y)}{\partial \pi_j} \leftarrow h_{\psi_j(x)}(y)$

$\frac{\partial \phi_\lambda(y)}{\partial \psi_{j,i}} \leftarrow \pi_j \frac{\partial h_{\psi_j}(y)}{\partial \psi_{j,i}}$ où $j = 1, \ldots, m$, $\psi_{j,1} = \xi_j$, $\psi_{j,2} = \mu_j$ et $\psi_{j,3} = \sigma$

$\frac{\partial l}{\partial \psi_{j,i}} \leftarrow \frac{\partial l}{\partial \phi_\lambda(y)} \frac{\partial \phi_\lambda(y)}{\partial \psi_{j,i}}$

{Dérivée de l p/r aux sorties $a_j^{(i)}$}

$\frac{\partial l}{\partial a_j^{(0)}} \leftarrow \frac{\partial l}{\partial \phi_\lambda(y)} \pi_j h_{\psi_j}(y) + \pi_j$ pour $j = 1 \ldots m$

$\frac{\partial l}{\partial a_j^{(1)}} \leftarrow \frac{\partial l}{\partial \psi_{j,1}}(1 - \exp(-\xi_j))$

$\frac{\partial l}{\partial a_j^{(2)}} \leftarrow \frac{\partial l}{\partial \psi_{j,2}}$

$\frac{\partial l}{\partial a_j^{(3)}} \leftarrow \frac{\partial l}{\partial \psi_{j,3}}(1 - \exp(-\sigma_j))$

{Dérivée de l p/r à θ}

$\frac{\partial l}{\partial b_j^{(i)}} \leftarrow \frac{\partial l}{\partial a_j^{(i)}}$

$\frac{\partial l}{\partial w_{jh}^{(i)}} \leftarrow \frac{\partial l}{\partial a_j^{(i)}} z_h$

$\frac{\partial l}{\partial \tilde{w}_{jk}^{(i)}} \leftarrow \frac{\partial l}{\partial a_j^{(i)}} x_k$

$\frac{\partial l}{\partial z_h} \leftarrow \sum_{j=1}^{m} \sum_{i=0}^{3} \frac{\partial l}{\partial a_j^{(i)}} w_{jh}^{(i)}$

$\frac{\partial l}{\partial c_i} \leftarrow \frac{\partial l}{\partial z_i}(1 - z_i^2)$

$\frac{\partial l}{\partial v_{ik}} \leftarrow \frac{\partial l}{\partial z_i}(1 - z_i^2) x_k$

Algorithme 6 Gradient de la densité de la Pareto hybride : $\nabla h_\psi(y)$

ENTRÉES: y, $\psi = (\xi, \mu, \sigma)$ et $\sigma > 0$

SORTIES: $g \leftarrow \nabla h_\psi(y)$

{Requiert les fonctions $W(\cdot)$ (Algorithme 2) et `erf` [39]}

$z \leftarrow W\left((1+\xi)^2/(2\pi)\right)$

$\gamma \leftarrow 3/2 + \texttt{erf}(\text{sign}(1+\xi)\sqrt{z/2})/2$

$\beta \leftarrow \sigma\sqrt{(1+\xi)^2/z}$

$\alpha \leftarrow \text{sign}(1+\xi)\sigma\sqrt{z} + \mu$

$\frac{\partial h_\psi}{\partial \xi} \leftarrow -\sqrt{z}\text{sign}(1+\xi)/(\gamma\sqrt{2\pi}\exp(z/2)(1+z)(1+\xi))$

si $y \leq \alpha$ **alors**

$\frac{\partial h_\psi}{\partial \mu} \leftarrow (y-\mu)/\sigma^2$

$\frac{\partial h_\psi}{\partial \sigma} \leftarrow (y-\mu)^2/\sigma^3 - 1/\sigma$

sinon

si $\xi \neq 0$ **alors**

$\frac{\partial g_{\xi;\beta}}{\partial \xi} \leftarrow (1+\xi)(\alpha - y)/(\xi(\beta + \xi(y-\alpha))) + \log\left(1 + \xi(y-\alpha)/\beta\right)/\xi^2$

$\frac{\partial g_{\xi;\beta}}{\partial \beta} \leftarrow (y - \beta - \alpha)/(\beta(\beta + \xi(y-\alpha)))$

$\frac{\partial g_{\xi;\beta}}{\partial \alpha} \leftarrow (1+\xi)/(\beta + \xi(y-\alpha))$

sinon

$\frac{\partial g_{\xi;\beta}}{\partial \xi} \leftarrow 0$

$\frac{\partial g_{\xi;\beta}}{\partial \beta} \leftarrow (y - \beta - \alpha)/\beta^2$

$\frac{\partial g_{\xi;\beta}}{\partial \alpha} \leftarrow 1/\beta$

$\frac{\partial h_\psi}{\partial \xi} \leftarrow \frac{\partial h_\psi}{\partial \xi} + \frac{\partial g_{\xi;\beta}}{\partial \beta}(\text{sign}(1+\xi)\sigma/\sqrt{z} - \sigma|1+\xi|/(\sqrt{z}(1+z)(1+\xi))) + \ldots$

$\frac{\partial g_{\xi;\beta}}{\partial \alpha}\sigma\sqrt{z}/((1+z)|1+\xi|)$

$\frac{\partial h_\psi}{\partial \mu} \leftarrow \frac{\partial g_{\xi;\beta}}{\partial \alpha}$

$\frac{\partial h_\psi}{\partial \sigma} \leftarrow \frac{\partial g_{\xi;\beta}}{\partial \alpha}\text{sign}(1+\xi)\sqrt{z} + \frac{\partial g_{\xi;\beta}}{\partial \beta}|1+\xi|/\sqrt{z}$

$g \leftarrow \left(\frac{\partial h_\psi}{\partial \xi}, \frac{\partial h_\psi}{\partial \mu}, \frac{\partial h_\psi}{\partial \sigma}\right)$

Annexe B

ESTIMATION DES PARAMÈTRES DE LA PARETO HYBRIDE

Ce chapitre présente des résultats complémentaires (voir la sous-section 3.1.4) sur la convergence des estimateurs de maximum de vraisemblance des paramètres de la Pareto hybride. Des paires d'ensembles d'entraînement et de test $(\mathcal{D}_n, \mathcal{D}_l)$ sont générés selon une loi Pareto hybride de paramètres $\psi = (\xi, \mu, \sigma)$ où n est choisi de plus en plus grand et l est fixé à 10 000. Soit $\hat{\psi}_n$, l'estimateur par maximum de vraisemblance de ψ. La performance de $\hat{\psi}_n$ est mesurée en termes de log-vraisemblance relative :

$$\mathcal{R}_l(\widehat{\psi}_n; \psi) = \frac{1}{l} \sum_{i=1}^{l} (\log h_{\psi}(Z_i) - \log h_{\widehat{\psi}_n}(Z_i)).$$

Plus $\mathcal{R}_l(\widehat{\psi}_n; \psi)$ est petit, plus la densité estimée est près du modèle générateur. Pour chaque valeur de n, 100 ensembles d'entraînement sont générés et l'estimateur de maximum de vraisemblance $\widehat{\psi}_n^i$ est calculé pour chacun de ces ensembles. Ceci nous permet de rapporter la log-vraisemblance relative moyenne en test $(1/100 \sum_{i=1}^{100} \mathcal{R}_l(\widehat{\psi}_n^i; \psi))$, de calculer des intervalles de confiance autour de cette moyenne et d'estimer le biais carré et la variance des paramètres estimés.

La log-vraisemblance moyenne des estimateurs relative à la densité génératrice ainsi qu'un intervalle de confiance de niveau 5 % sont illustrés aux figures B.1, B.2 et B.3 lorsque le vecteur de paramètres de la Pareto hybride est $\psi = (0.4, 0, 1)$, $\psi = (0, 0, 1)$ et $\psi = (-0.25, 0, 1)$ respectivement et que la taille de l'ensemble

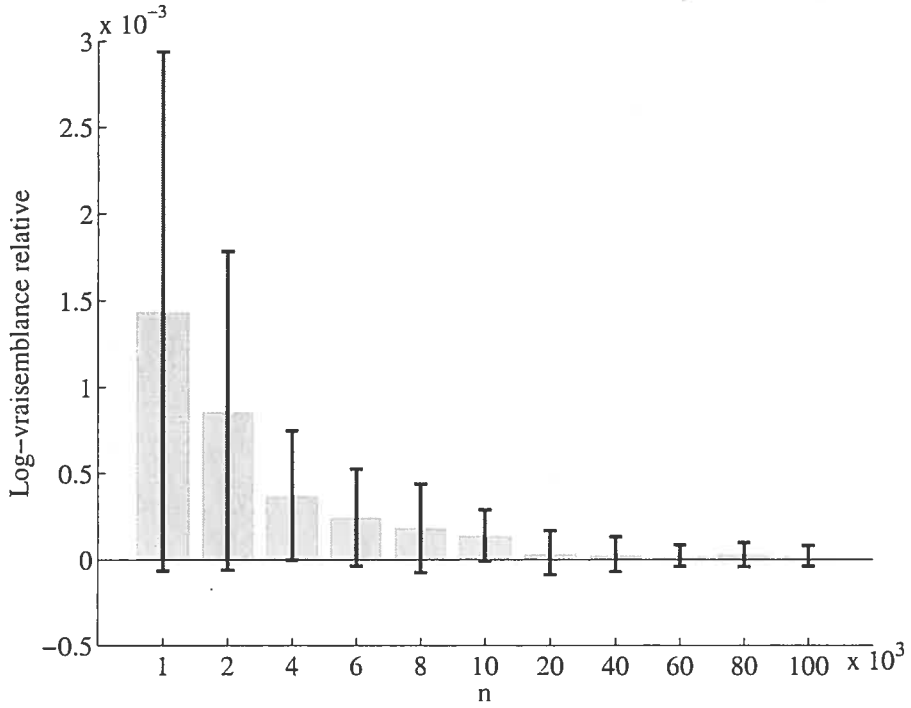

FIG. B.1: *Estimation du vecteur de paramètres $\psi = (0.4, 0, 1)$ de la Pareto hybride : log-vraisemblance moyenne relative sur l'ensemble de test et intervalle de confiance de niveau 5 % lorsque la taille n de l'ensemble d'entraînement augmente.*

d'entraînement croît. Le biais carré et la variance des estimateurs de maximum de vraisemblance correspondants sont fournis aux figures B.4, B.5 et B.6.

La log-vraisemblance relative décroît et l'intervalle de confiance se rétrécit lorsque la taille de l'ensemble d'entraînement augmente, ce qui indique que les estimateurs de maximum de vraisemblance sont convergents. Bien que le biais carré fluctue un peu pour les plus petits ensembles d'entraînement, il est au plus de l'ordre de 10^{-4} et diminue éventuellement avec n. La variance est décroissante en n dans tous les cas.

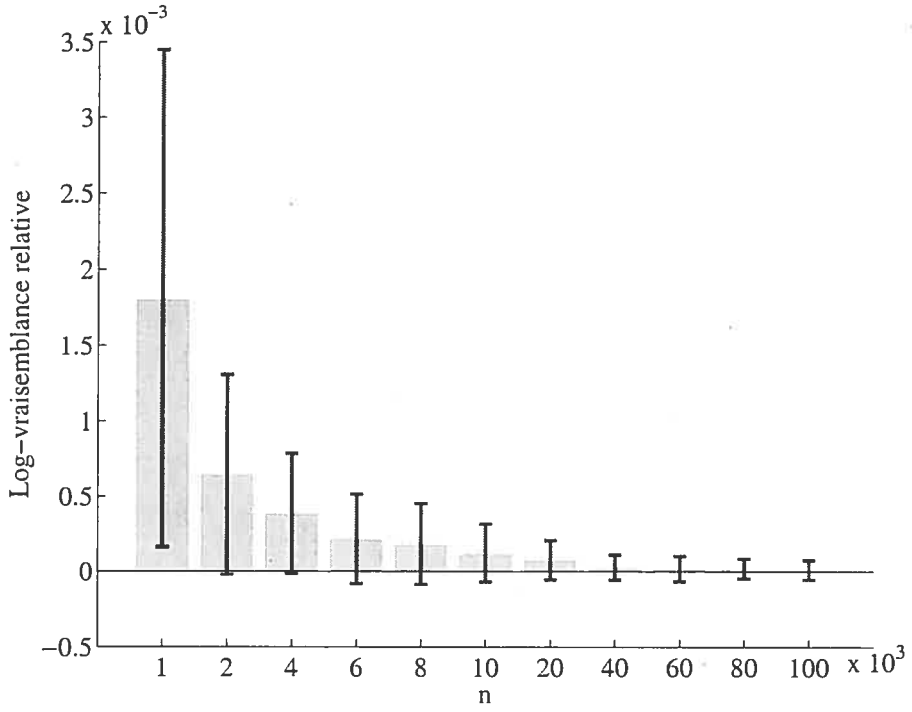

FIG. B.2: *Estimation du vecteur de paramètres $\psi = (0, 0, 1)$ de la Pareto hybride : log-vraisemblance moyenne relative sur l'ensemble de test et intervalle de confiance de niveau 5 % lorsque la taille n de l'ensemble d'entraînement augmente.*

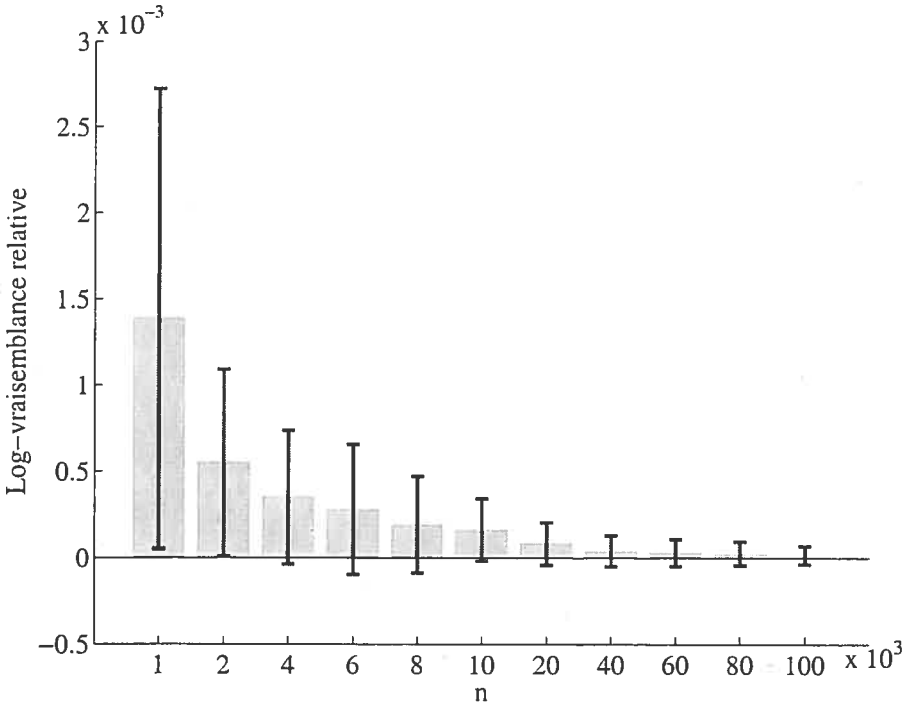

FIG. B.3: *Estimation du vecteur de paramètres $\psi = (-0.25, 0, 1)$ de la Pareto hybride : log-vraisemblance moyenne relative sur l'ensemble de test et intervalle de confiance de niveau 5 % lorsque la taille n de l'ensemble d'entraînement augmente.*

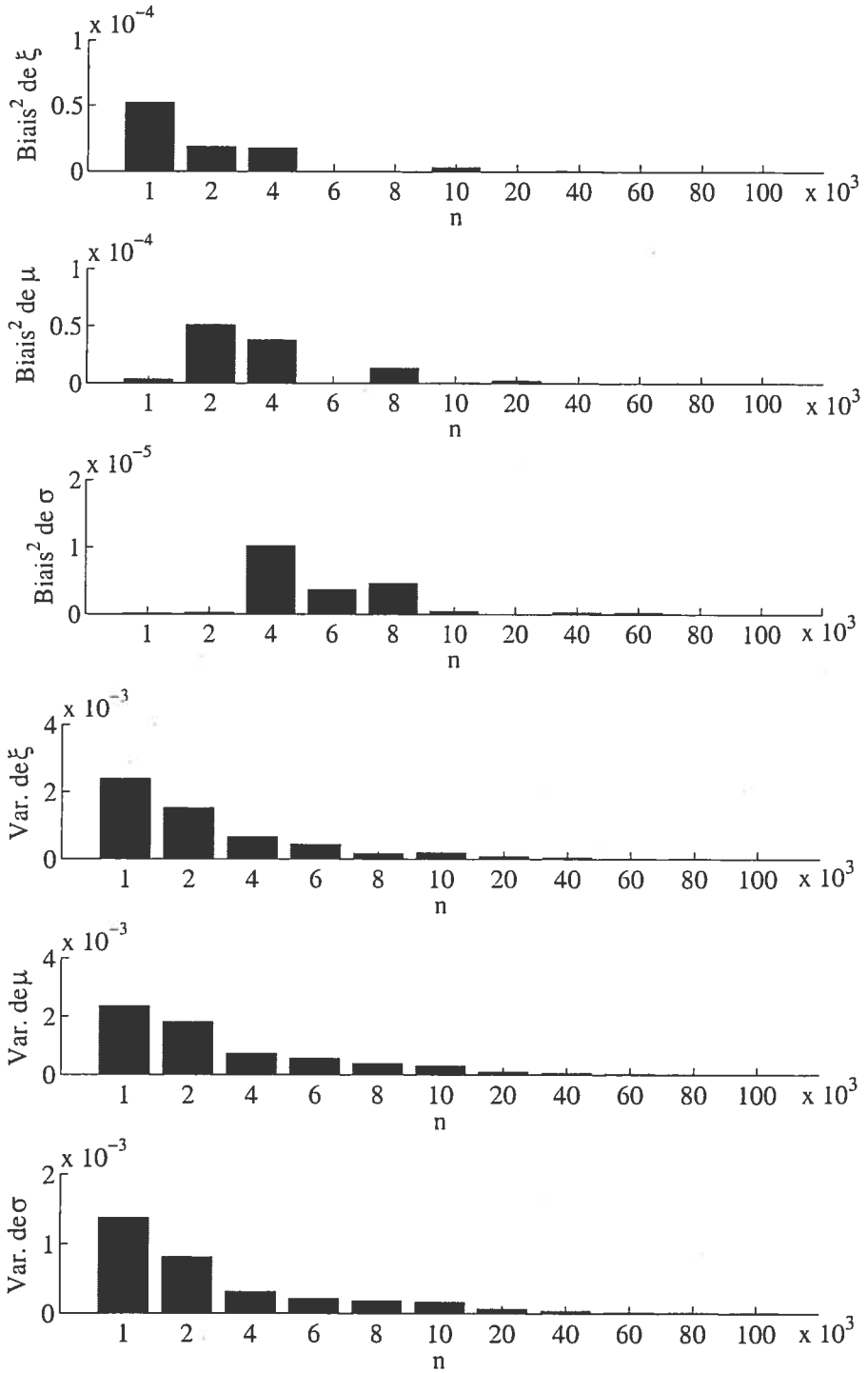

FIG. B.4: *Biais carré (panneau du haut) et variance (panneau du bas) estimés pour les estimateurs de maximum de vraisemblance des paramètres de la Pareto hybride avec* $\psi = (0.4, 0, 1)$ *lorsque la taille* n *de l'ensemble d'entraînement augmente.*

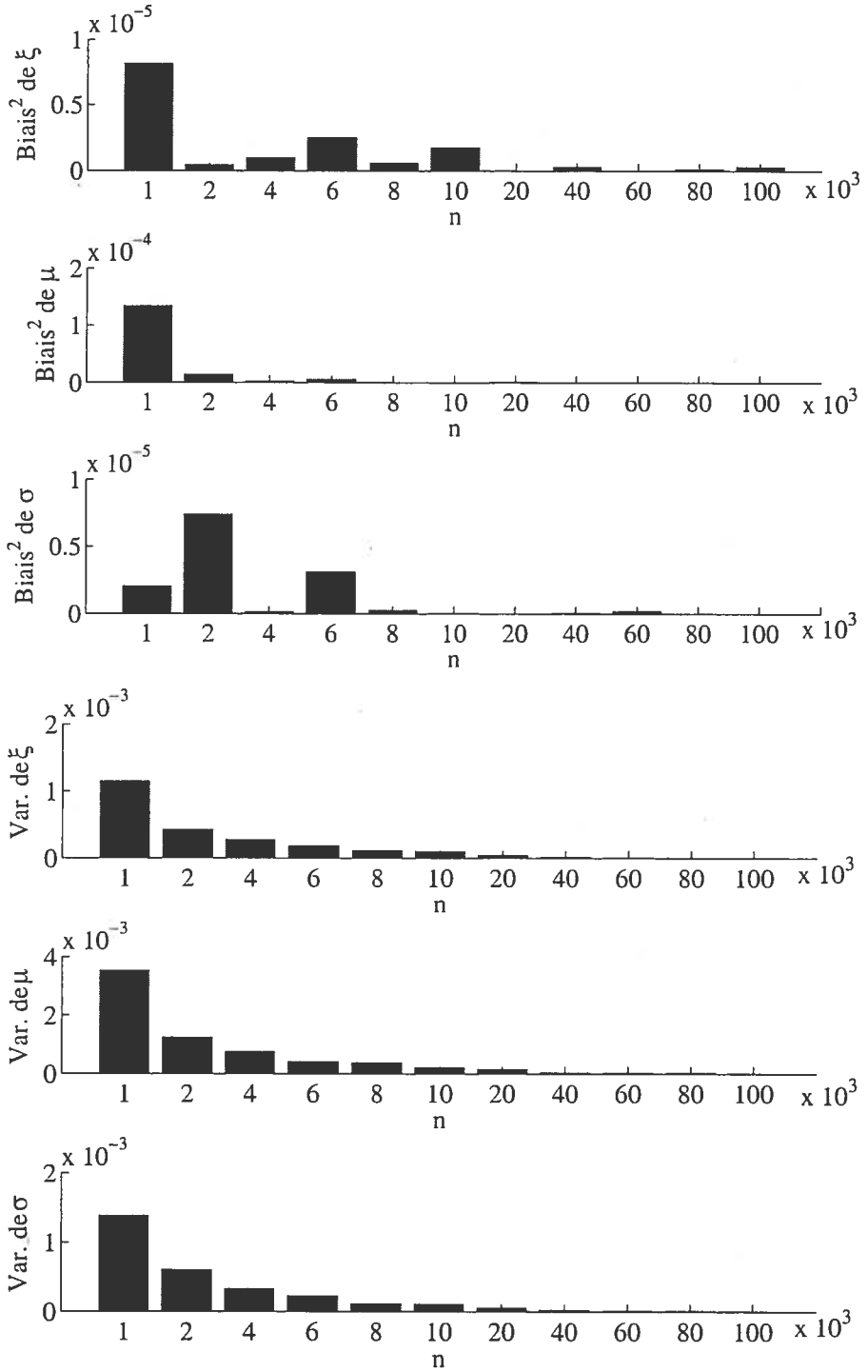

FIG. B.5: *Biais carré (panneau du haut) et variance (panneau du bas) estimés pour les estimateurs de maximum de vraisemblance des paramètres de la Pareto hybride avec* $\psi = (0, 0, 1)$ *lorsque la taille* n *de l'ensemble d'entraînement augmente.*

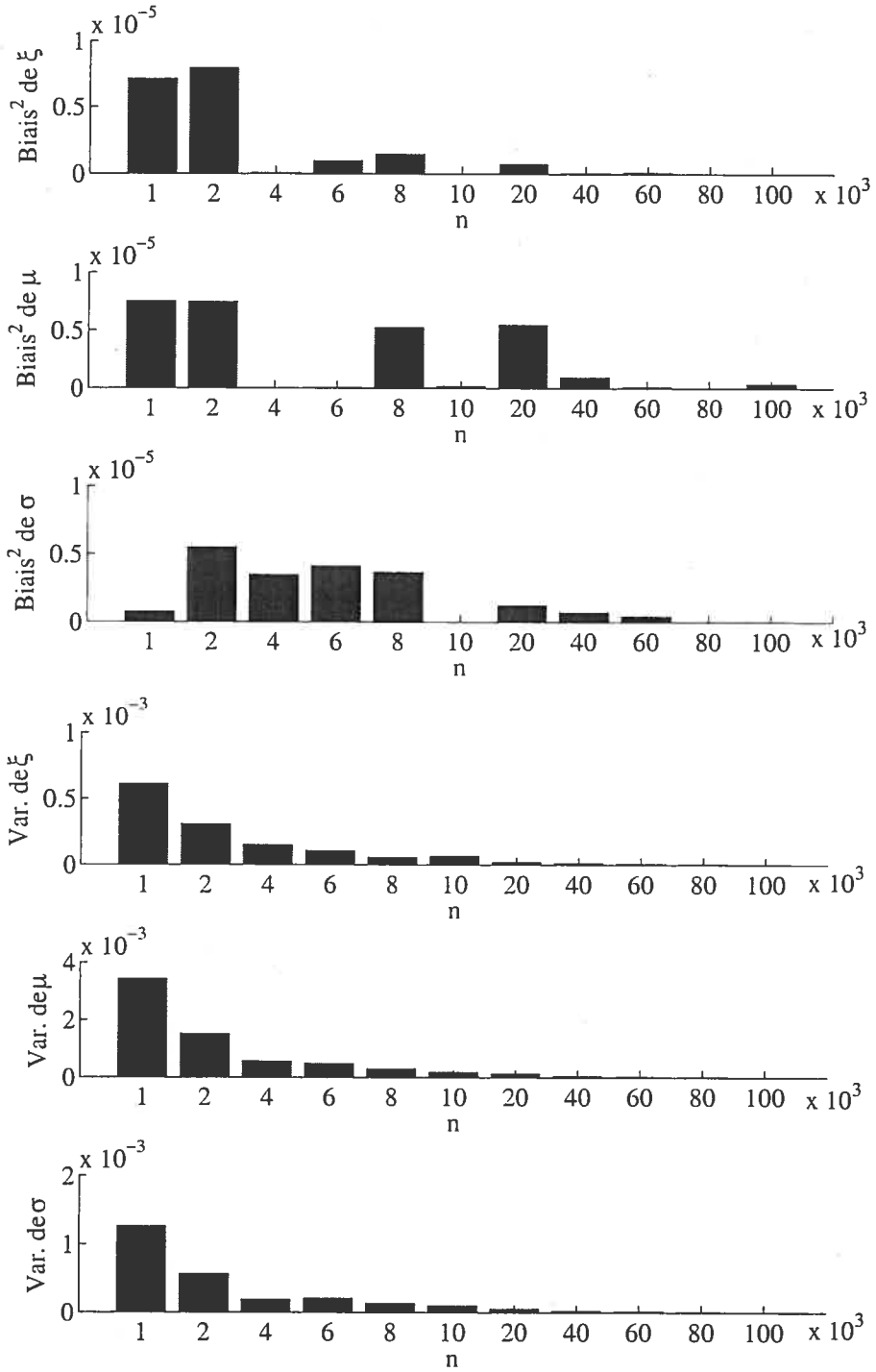

FIG. B.6: *Biais carré (panneau du haut) et variance (panneau du bas) estimés pour les estimateurs de maximum de vraisemblance des paramètres de la Pareto hybride avec* $\psi = (-0.25, 0, 1)$ *lorsque la taille* n *de l'ensemble d'entraînement augmente.*

Annexe C

ÉTUDE DES ESTIMATEURS CONDITIONNELS : DONNÉES FRÉCHET CONDITIONNELLE À QUEUE LOURDE

Ce chapitre donne des résultats complémentaires pour les expériences sur les données générées par la loi de Fréchet conditionnelle avec queue lourde (voir 4.2.3). Les paramètres de la Fréchet ont une dépendance soit linéaire ($f(X) = aX + b$) soit sinusoïdale de l'entrée ($f(X) = c\sin(aX + b) + d$). La variable d'entrée X est générée selon une loi Normale standard. Les fonctions de dépendance (linéaire ou sinusoïdale) sont choisies de sorte que l'indice de queue de la Fréchet se trouve dans l'intervalle $[0.66, 1.33]$ dans 99% des cas, ce qui correspond à une queue lourde. Pour le cas où la fonction de dépendance est linéaire, les figures C.1, C.2 et C.3 montrent les paramètres des mélanges conditionnels après l'apprentissage avec composantes Pareto hybrides, Gaussiennes et Log-Normales respectivement en fonction de l'entrée X. Pour la loi Log-Normale, nous avons tracé l'espérance et l'écart-type de chaque composante plutôt que ses paramètres d'emplacement et de dispersion. Les figures C.4 et C.5 contiennent des observations générées par chacun des modèles après l'entraînement.

Pour le cas où la fonction de dépendance est sinusoïdale, les paramètres des composantes des mélanges conditionnels après l'apprentissage avec composantes Pareto hybrides, Gaussiennes et Log-Normales sont tracés aux figures C.6, C.7 et C.8 respectivement. Pour la loi Log-Normale, il s'agit en fait de l'espérance

et de l'écart-type de chaque composante. Les figures C.9 et C.10 contiennent des observations générées par chacun des modèles après l'entraînement.

Puisque l'indice de queue conditionnel de la Fréchet est plus lourd, l'apprentissage de la densité en est plus difficile. On observe tout de même des phénomènes semblabes aux résultats fournis à la sous-section 4.2.3. Dans le cas linéaire, l'indice de queue de la composante dominante du mélange conditionnel de Pareto hybride est assez fidèle à l'indice de queue du modèle générateur lorsque l'ensemble d'entraînement contient 2000 observations. Les données générées par le mélange conditionnel de Gaussiennes (figure C.2, rangée du milieu) mettent en évidence le problème de la queue de la distribution inférieure. En effet, la densité est nulle dans cette région pour le modèle générateur alors que le mélange conditionnel de Gaussiennes y génére de nombreuses observations. Le cas où la dépendance des paramètres du modèle générateur est sinusoïdale est encore plus difficile pour la modélisation. Les mélanges conditionnels de Pareto hybrides et de Log-Normales (figure C.9, rangée du haut et du bas respectivement) parviennent à capter en partie la forme de la dépendance. Pour la méthode PoT conditionnelle (figure C.10, rangée du bas), il semble que le fait que la dépendance des paramètres et la distribution des excès soient modélisées en deux temps nuise à la modélisation. En effet, la présence de très grandes observations extrêmes cache la nature de la dépendance. L'estimateur de la fenêtre de Parzen (figure C.10, rangée du haut) ne parvient pas non plus, dans ce cas, à détecter la forme de la dépendance entre les données.

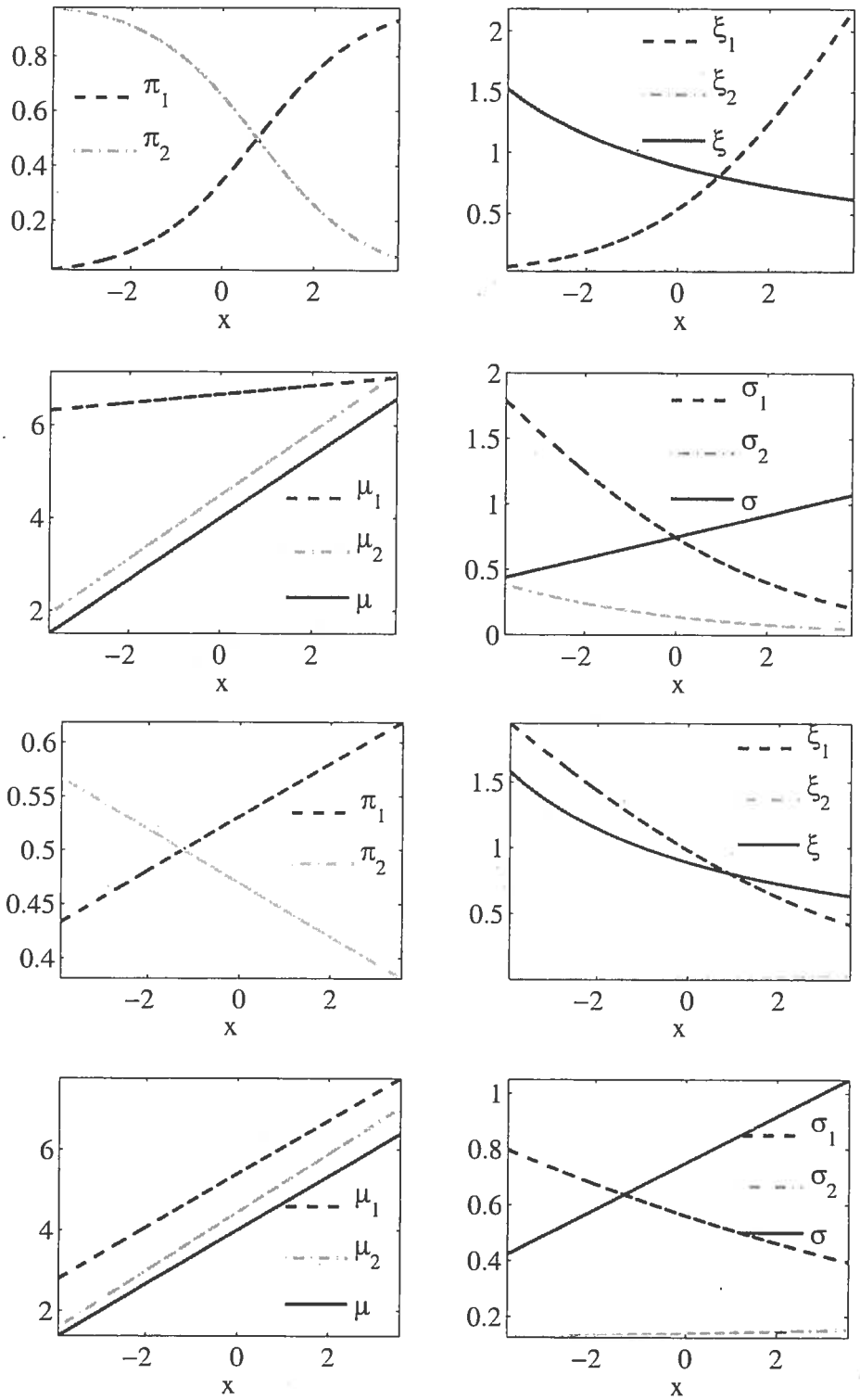

FIG. C.1: $(\pi_j(x), \xi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMMH. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-lin-lourde. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

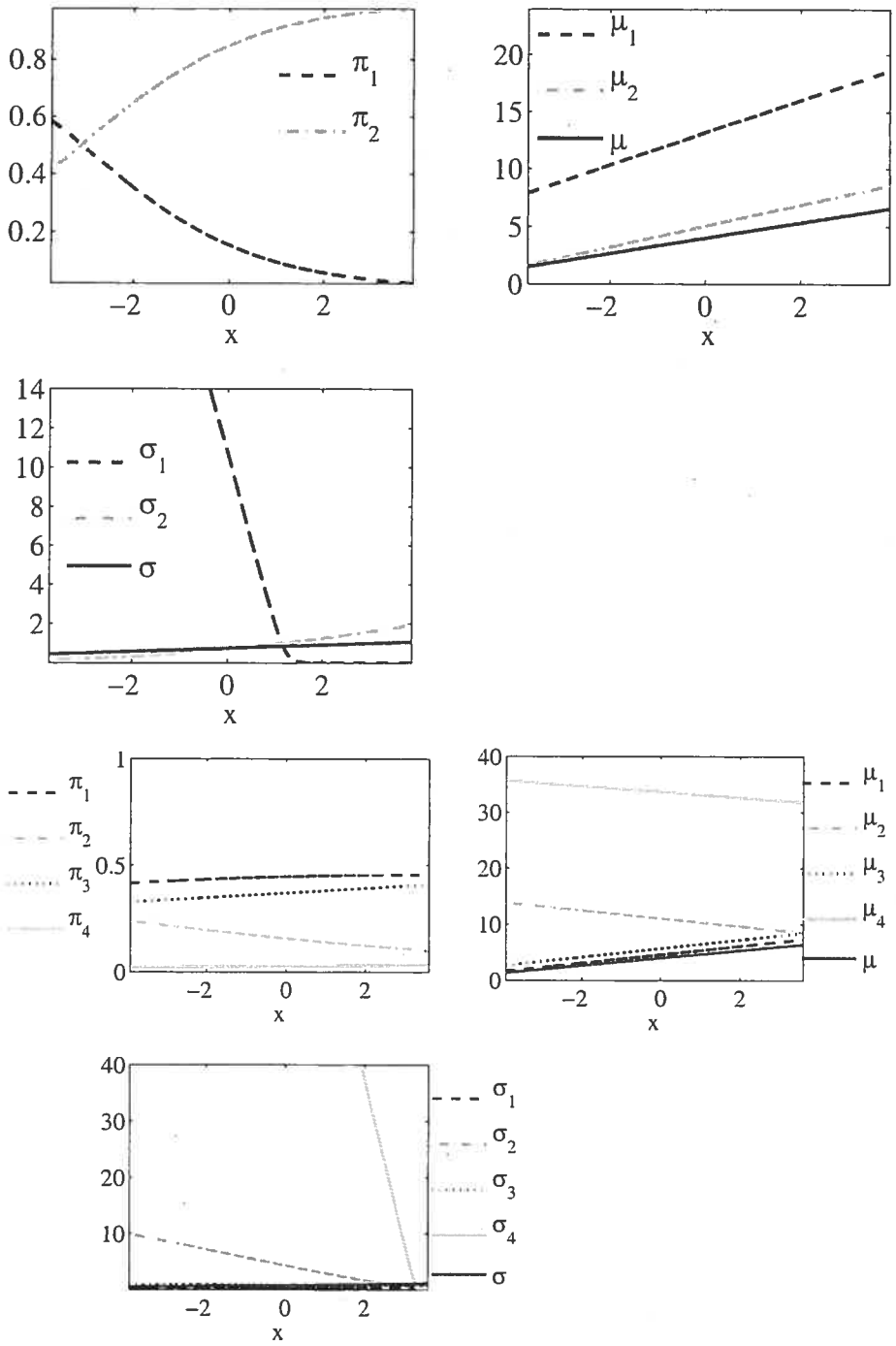

FIG. C.2: $(\pi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMMG. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-lin-lourde. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

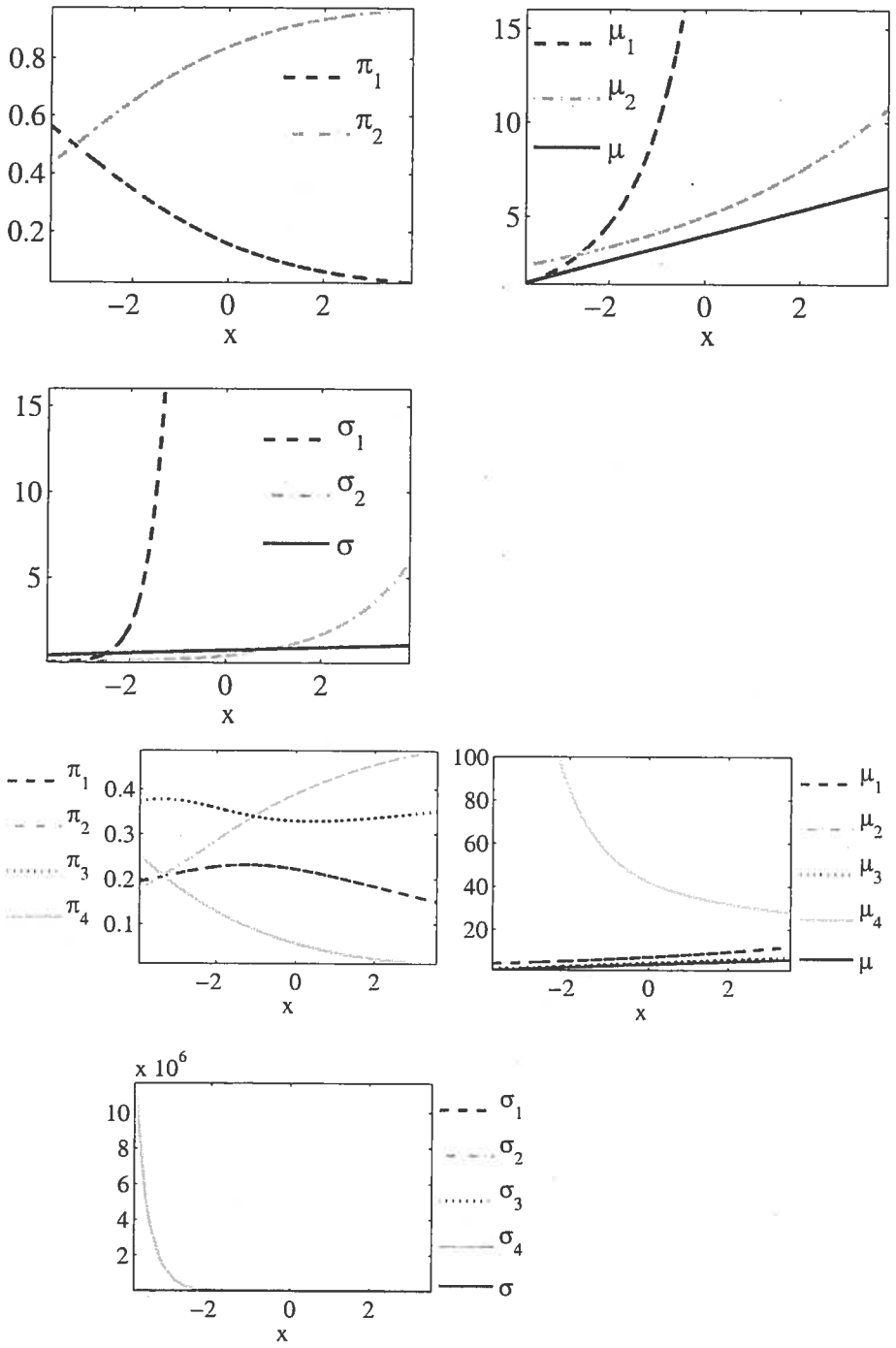

FIG. C.3: $(\pi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMML. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-lin-lourde. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

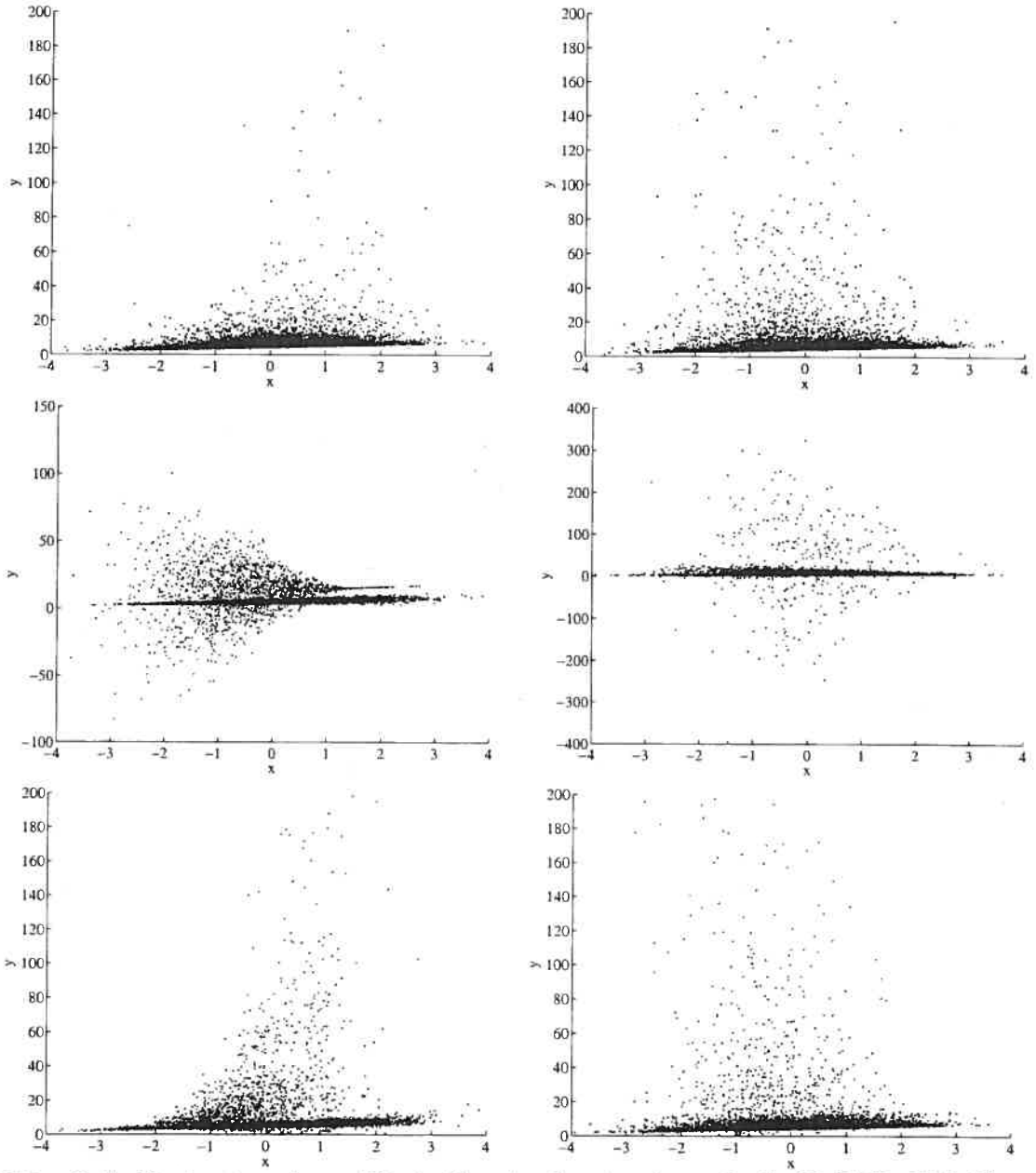

FIG. C.4: *De haut en bas : Génération de données à partir de CMMH, CMMG et CMML. L'ensemble d'entraînement consistait en 200 (colonne de gauche) ou 2 000 observations (colonne de droite) générées à partir du modèle Fréchet-lin-lourde.*

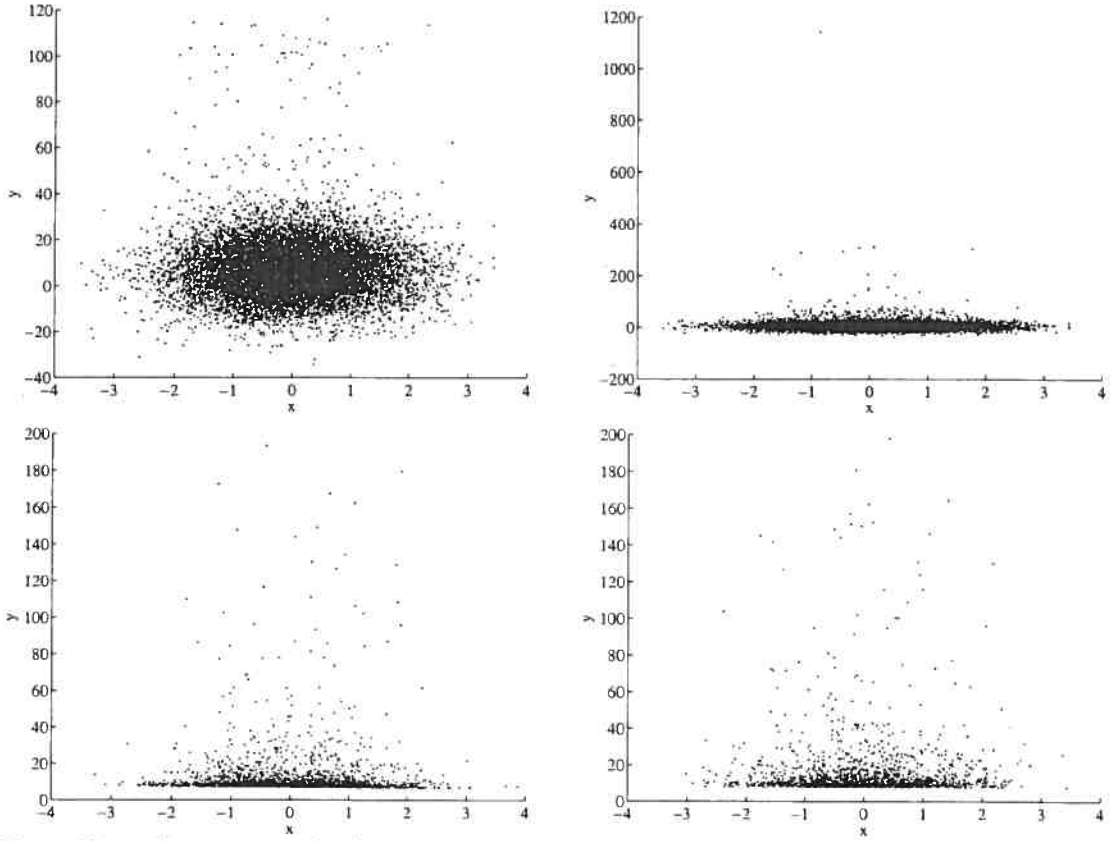

FIG. C.5: *Génération de données à partir de CPARZEN (rangée du haut) et d'excès par CPOT (rangée du bas). L'ensemble d'entraînement consistait en 200 (colonne de gauche) ou 2 000 observations (colonne de droite) générées à partir du modèle Fréchet-lin-lourde.*

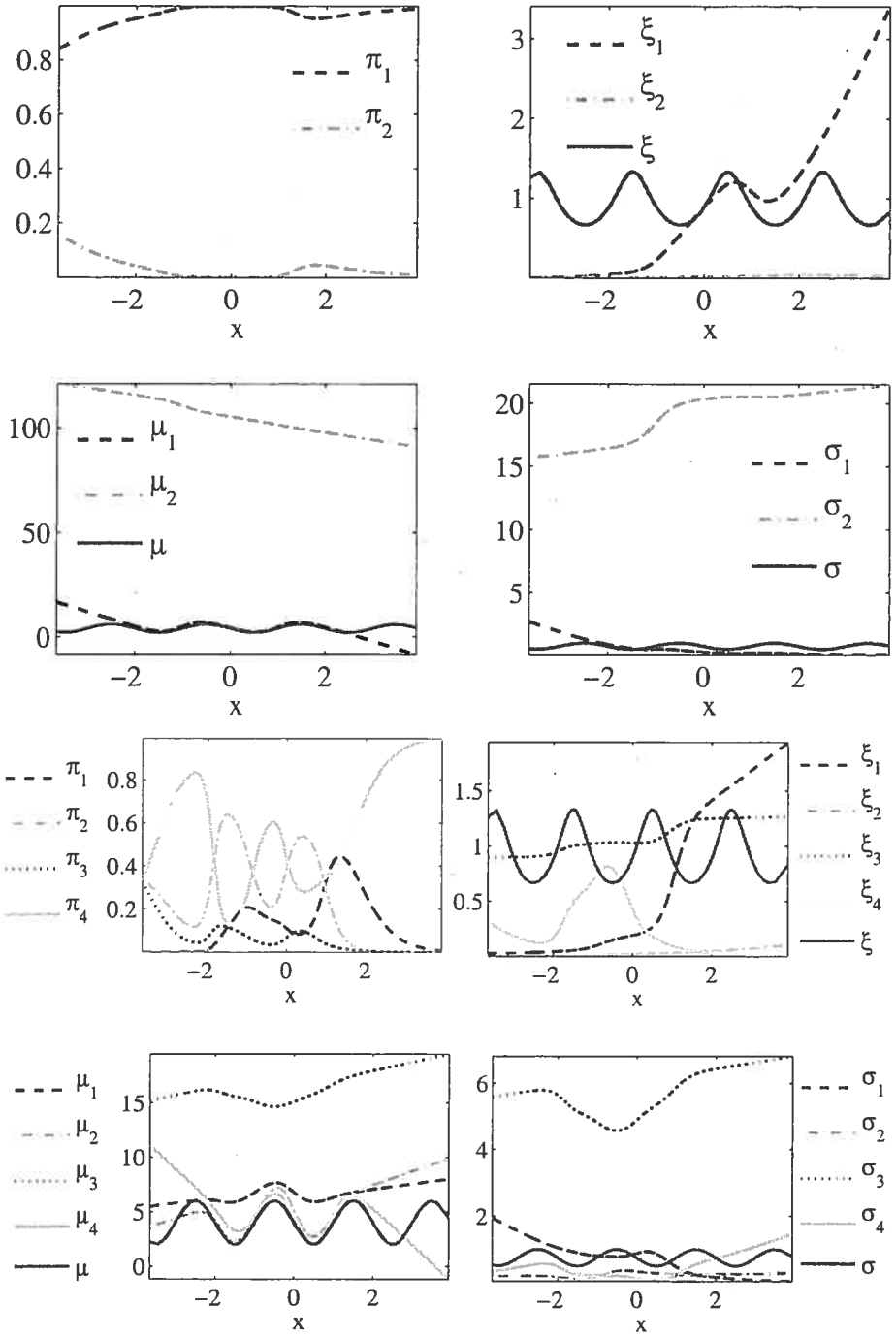

FIG. C.6: $(\pi_j(x), \xi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMMH. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-sin-lourde. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

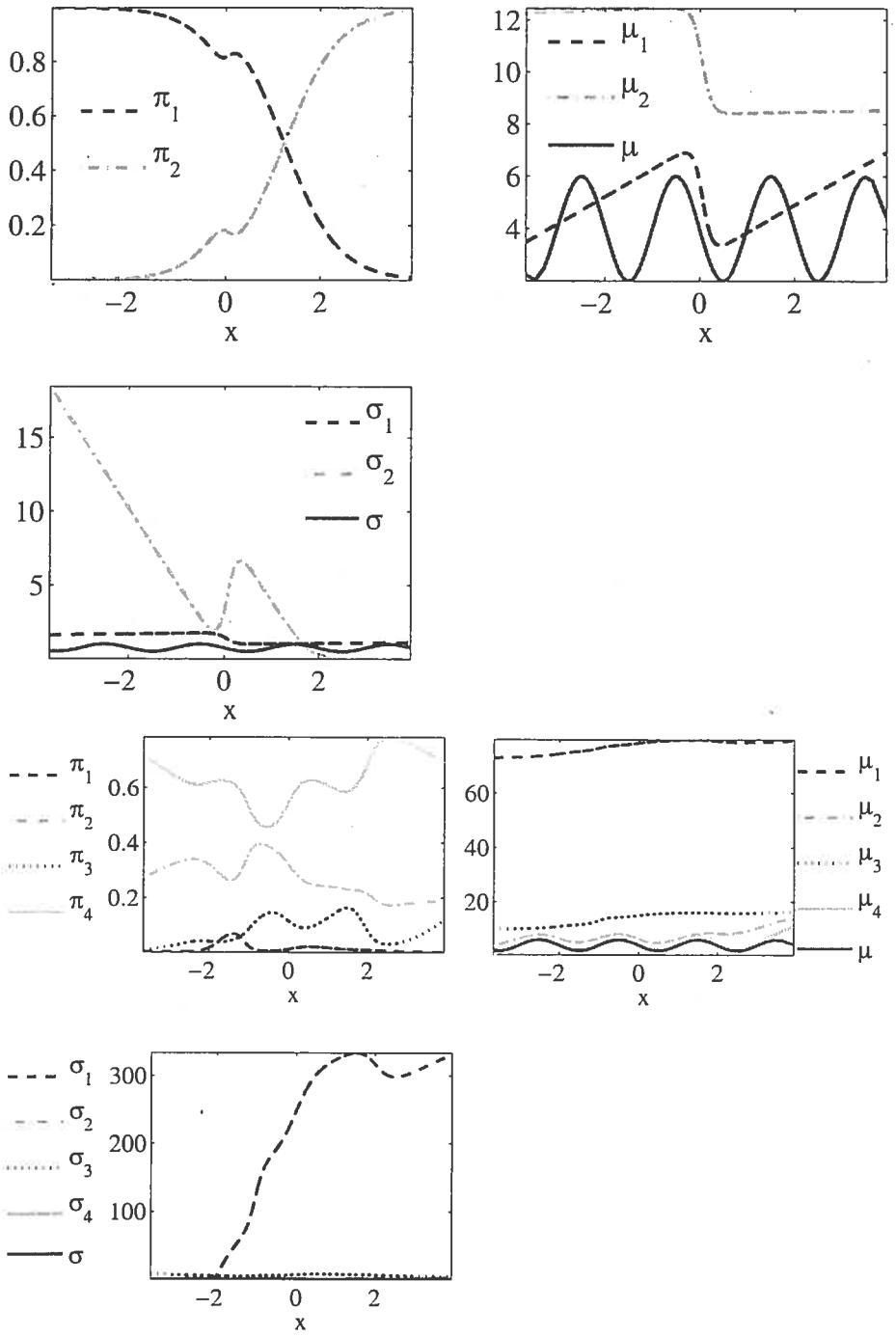

FIG. C.7: $(\pi_j(x), \mu_j(x), \sigma_j(x))$ *pour CMMG. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-sin-lourde. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

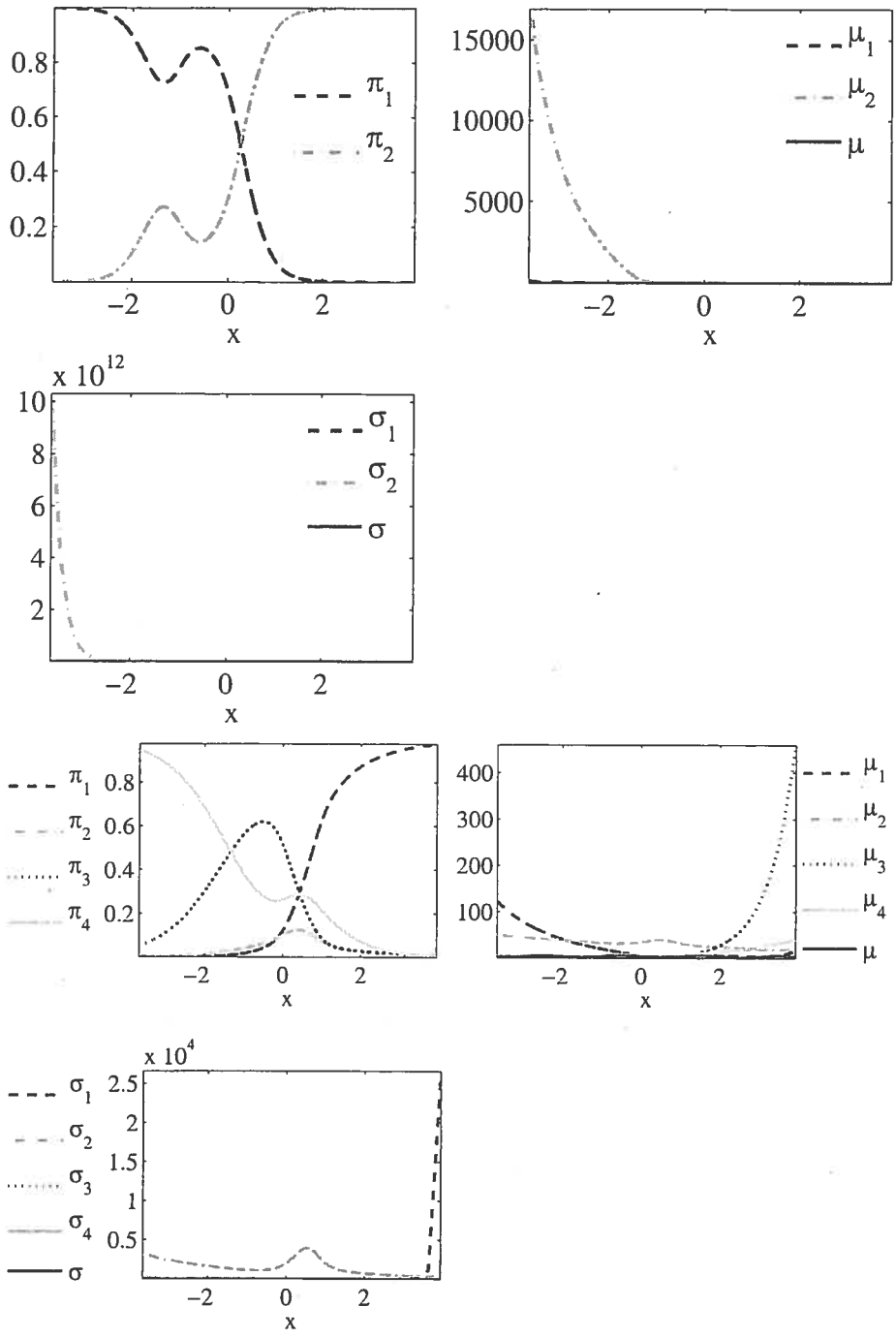

FIG. C.8: *$(\pi_j(x), \mu_j(x), \sigma_j(x))$ pour CMML. L'ensemble d'entraînement contenait 200 observations pour le panneau supérieur et 2 000 pour le panneau inférieur. Le modèle générateur est le Fréchet-sin-lourde. Les paramètres du modèle générateur sont dessinés en trait plein gris foncé.*

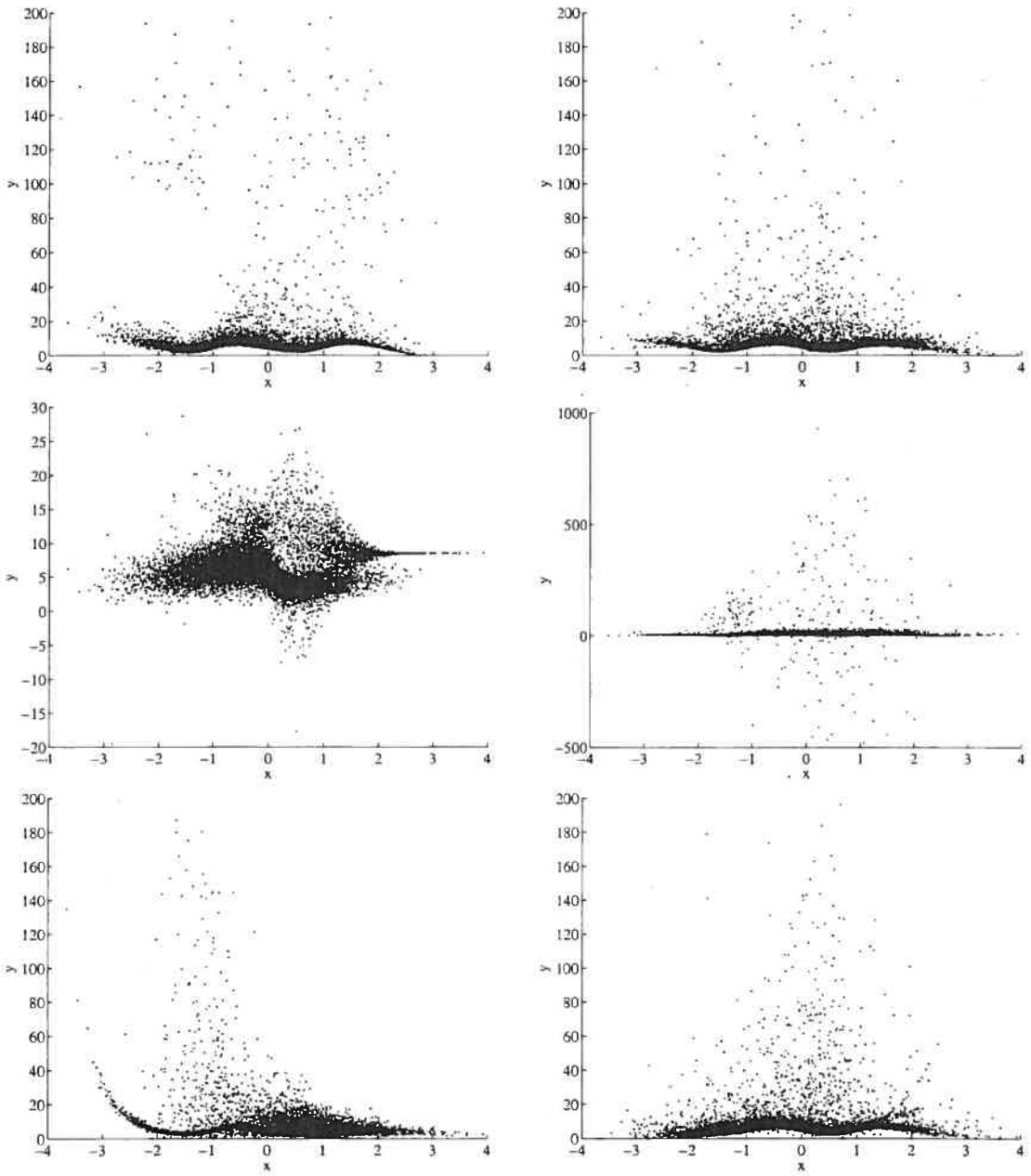

FIG. C.9: *De haut en bas : Génération de données à partir de CMMH, CMMG et CMML. L'ensemble d'entraînement consistait en 200 (colonne de gauche) ou 2 000 observations (colonne de droite) générées à partir du modèle Fréchet-sin-lourde.*

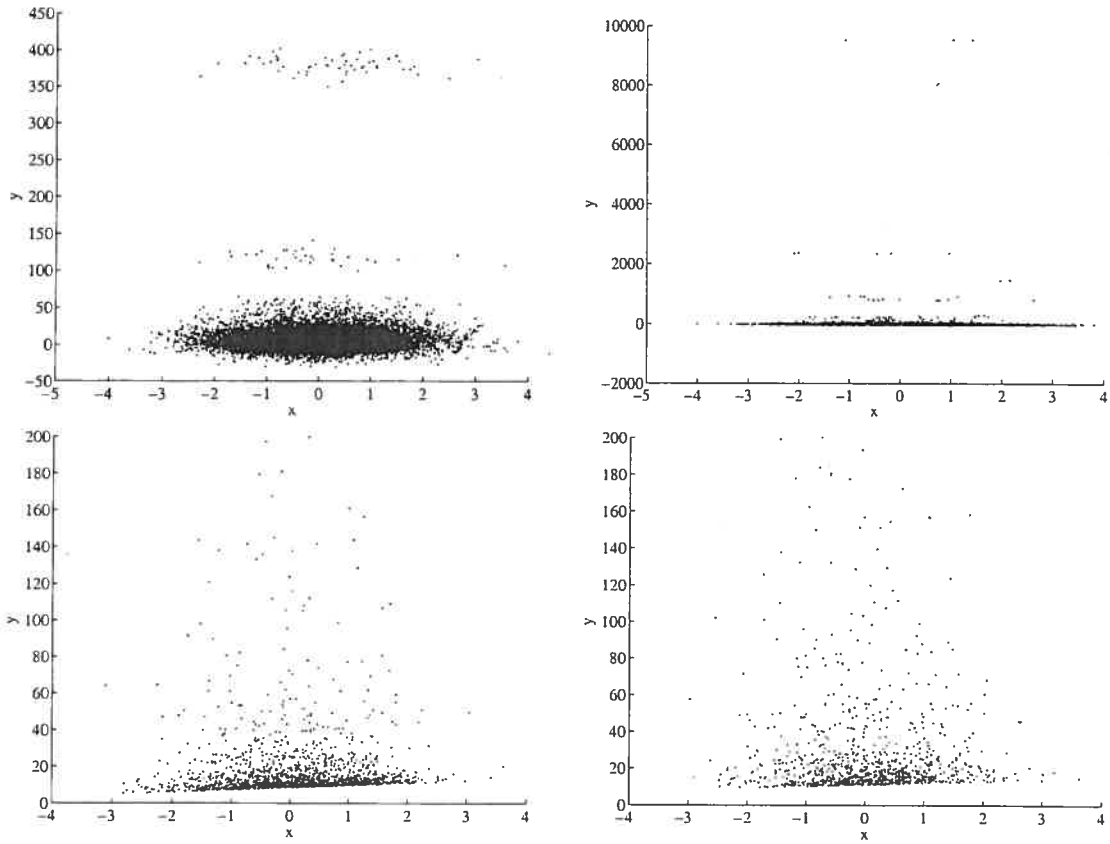

FIG. C.10: *Génération de données à partir de CPARZEN (rangée du haut) et d'excès par CPOT (rangée du bas). L'ensemble d'entraînement consistait en 200 (colonne de gauche) ou 2 000 observations (colonne de droite) générées à partir du modèle Fréchet-sin-lourde.*